KDDM1 - Statistical Data Science

Roman Kern <rkern@tugraz.at>
Version 2.1.0

> **Motivation**:
Statistics provide us with powerful tools and a immense body of knowledge.
> **Goal**:
Understand which tools and approaches are suitable, and also the importance of assumptions and given limitations.

Outline

Statistical Basics

Key techniques and methods

Background

Data science vs. (inferential) statistics

- Data mining
  - The data is the **complete representation** of the world and of the phenomena we are studying
- Statistics
  - The data is obtained from an **underlying generative process**, that is what we really care about

> There is a long discussion if data science is just statistics, or something different.

> Thus, the statement on this slide only portraits a single perspective.

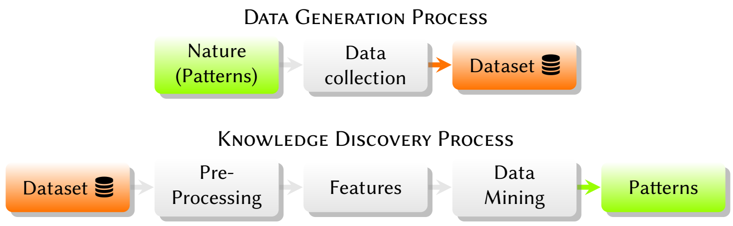
Background

Example: Information (data) from two online communities C_1 and C_2, regarding whether each post is in a given topic T.

- **Data mining**
  - What fraction of posts in C_1 are related to T?
  - What fraction of posts in C_2 are related to T?
- **Statistics**
  - What is the probability that a post from C_1 is related to T?
  - What is the probability that a post from C_2 is related to T?

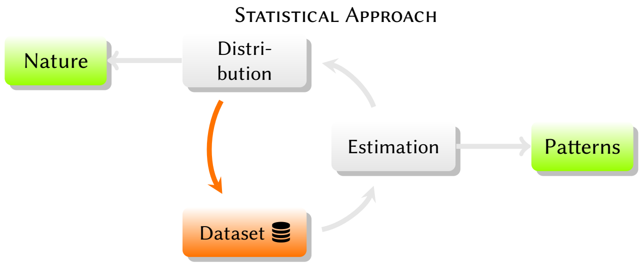
> Important here: One cannot **compute** the probability from finite data (dataset), but only **estimate** it.
> The **law of large numbers** informs us that, under favourable settings, the estimate will converge to the true value given more data.

Knowledge Discovery Process



> Recall the knowledge discovery process from the introduction lecture.

Example of a Statistical Approach



> First, making assumptions about nature (e.g., follows a certain distribution) allows to rigorously derive insights.
> Please note, the arrow does not point from Nature to Distribution to highlight that the assumption about the distribution does not "naturally" follow from Nature.

> Crucially, only if the assumptions are correct, the insights are expected to hold.

Background I

Data Mining - Grand Checklist

1. Linearity: scatter plot, common sense, and knowing your problem, transform including interactions if useful
2. t-statistics: are the coefficients significantly different from zero? Look at width of confidence intervals
3. F-tests for subsets, equality of coefficients
4. R^2: is it reasonably high in the context?

Background II

5. Influential observations, outliers in predictor space, dependent variable space
6. Normality: plot histogram of the residuals
7. Studentized residuals
8. Heteroscedasticity: plot residuals with each x variable, transform if necessary, Box-Cox transformations
9. Autocorrelation: ”time series plot”

Background III

10. Multicollinearity: compute correlations of the x variables, do signs of coefficients agree with intuition?
11. Principal Components
12. Missing Values

http://ocw.mit.edu/courses/sloan-school-of-management/15-062-data-mining-spring-2003/lecture-notes/

Background

Types of data science projects

- Hypothesis-driven
  - E.g. Is there a quality impairment, if parameter X is changed?
  - E.g. Can the quality be approximated by process measurements
- Data-driven
  - What insights can be generated from the data?
  - Do the data contain critical changes?
- Simulation-driven
  - Can Machine Learning being utilised to simulate (and then optimise) a process?

Hypothesis-driven == postivism || Data-driven == constructivism

Correlation

Relationship between variables

Correlation and Dependency

- Statistical dependence
  - X and Y (random variables) are positively dependent if the conditional probability, $P(X|Y)$, of X given Y is greater than the probability, $P(X)$
  - $P(X, Y) > P(X)P(Y)$
  - They are negatively dependent if the inequalities are reversed

Correlation and Dependency

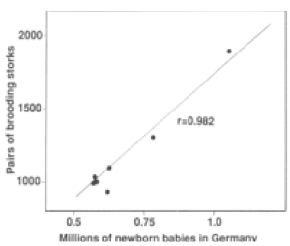
- Correlation
  - σ_X^2, σ_Y^2 being the variances of X and Y, ρ is the correlation coefficient
  - $\sigma_{Y|X}^2 = E[Y - E[Y|X = x]]^2$
  - Correlation ratio: $\eta_{Y|X}^2 = 1 - \frac{\sigma_{Y|X}^2}{\sigma_Y^2}$
  - $\eta_{Y|X}^2 = 0$, if X and Y are independent
  - $\eta_{Y|X}^2 = \rho^2$, if X and Y are linearly dependent

Properties

- Correlation does not imply causation![1]
- Correlation analysis can only be the first step
  - Followed by tests and model building
- In big data sets there will be many pairs of variables
  - Some pairs will correlation just by chance

[1] Exceptions in special cases, e.g. time series (but watch out for `post hoc ergo propter hoc`)

Spurious correlations



H. Sies, Nature 332:495 (1988); adapted from R. Wilson, http://users.physics.harvard.edu/~wilson/soundscience/ALF_Science.html
Spurious correlations, http://www.tylervigen.com/spurious-correlations

Common Properties of Correlation Measures

- Most correlations provide values $[-1, +1]$ (some $[0, +1]$)
  - ... but cannot be compared between correlation measures
  - (might be skewed, i.e. 0.5 does not imply "half" correlated)
- Correlation and dependency are different (but related) concepts

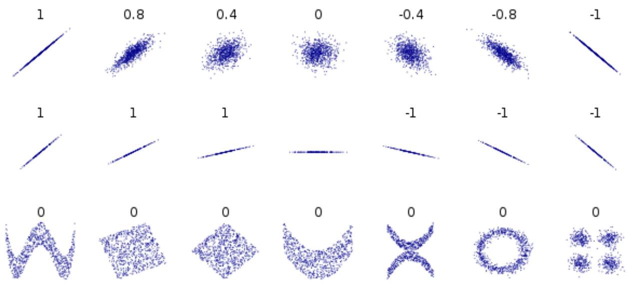
Pearson's Correlation

Linear correlation - Pearson's product moment correlation coefficient (PC)

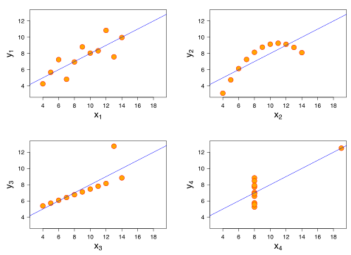
- Intuition: $Y = \beta_0 + \beta_1 X + \epsilon$ (Predict Y by observing X)
- $r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$
- $PC = 0 \not\Rightarrow X \perp Y$
- $X \perp Y \Rightarrow PC = 0$
  - If X and Y are independent, PC will be zero
- Sensitive to outliers

> Linear regression serves as intuition.
> PC can also been seen as normalised version of the covariance.
$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$

Examples of PC



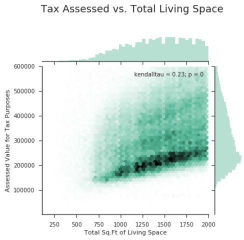
Non-Linear Relationship and PC



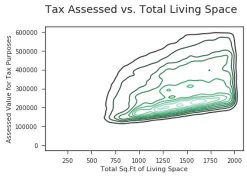
Same correlation of 0.816

> Obviously, PC is not a suitable choice, if the dependencies b/w the variables are not linear, i.e., the underlying assumption of PC (linear) does not hold.

Visual Tools for Non-Linear Correlation



Hexabin Plot



Contour Plot

> Taken from: https://github.com/arvindbetrabet/Practical_Statistics_for_Data_Scientists/blob/master/Chapter_1/PSFDS_Notebook_Explore_Hexabin_Countour_Plots.ipynb

Types of non-linear correlations

- Rank correlation
  - Detect the presence of a monotonic relationship between two random variables
- Transformation based
- Information theoretic

Types of non-linear correlations

- Spearman ρ
- Kendal τ (Kendall rank correlation coefficient)

Intuition

Order all values of each variable according to their rank
$\rightarrow$ compare the rankings, i.e. do the rankings agree
e.g. is the first entry in X also the first entry in Y?

Types of non-linear correlations

Maximal Correlation

- Definition: $mCor(X, Y) = \max_{f,g:\mathbb{R}\rightarrow\mathbb{R}} Cor(f(X), g(Y))$
- Theoretical properties:
  - Value of mCor is the correlation coefficient of transformed inputs
  - Provides results [0, 1]
  - $mCor(X, Y) = 0 \Leftrightarrow X \perp Y$
- Estimation problem: What if $f(x_i) = y_i, g(y_i) = y_i$?

> Here, *Cor* is the linear correlation - the intuition is to map the non-linear relationship into a linear one.
> The functions $f(\cdot), g(\cdot)$ could have any form.

Maximal Correlation

Solution: Alternating Conditional Expectations (ACE) algorithm

Iteratively solve $\min_{f,g} \mathbb{E}(f(X) - g(Y))$:

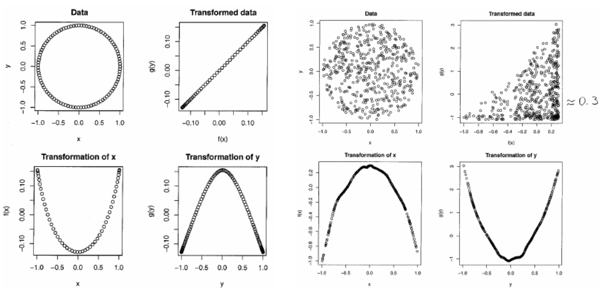
- repeat the alternating loop until convergence:
  - center and scale f(x)
  - set f to conditional expectation of f(x) given g(y)
  - do last two steps for g(y)

Maximal Correlation

- Caveats:
  - Estimation of conditional expectation (e.g. with splines) is hard
  - Overestimation of correlation value in certain cases

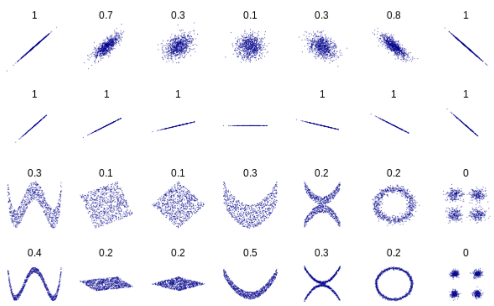
Maximal Correlation

Circle   Ball



> See also [1].

Distance Correlation



> Values for distance correlation (on top) for various scenarios, where the distance correlation appears to capture the dependencies well.

Distance Correlation

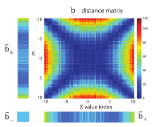
- Provides results $[0, 1]$
- $dCor(X, Y) = 0 \Rightarrow X \perp Y$
- The value itself cannot be directly be interpreted

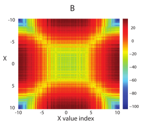
Confidence value

- Combination of distance correlation and permutation test
- ... allows to estimate $X \perp Y$ (with a confidence value)

Distance Correlation

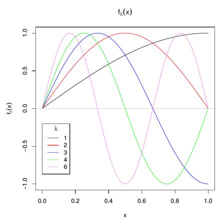
- Compute pair-wise distances for both variables (X, Y)
  - $\Rightarrow$ two symmetric, square matrices (D_X, D_Y)
- Transform both matrices via doubly centering
  - $\Rightarrow$ two symmetric, square matrices $(D_X^{centre}, D_Y^{centre})$
- Compute the pairwise distances of the two matrices

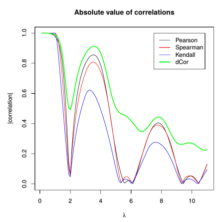




Original distance matrix (D) Centred distance matrix (D^{centre})

Insight on Correlations



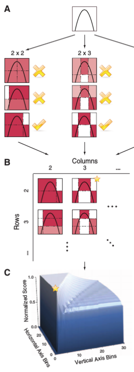


Non monotone function (for $\lambda > 1$): $f_\lambda(x) = sin(\lambda \frac{\pi}{2} x)$
Key take away: correlation $\neq$ dependency

Maximal Information Coefficient (MIC)

- Idea: Compute mutual information I between X and Y via bins
- Definition: $MIC(X, Y) = \max_{X, Y_{total} < B} \frac{I(X;Y)}{log_2(min(X,Y))}$,
  X, Y_{total} number of bins, B hyperparameter
- Theoretical properties:
  - Provides results $[0, 1]$
  - MIC values tend to 0 in case of statistical independence
  - MIC values tend to 1 for many noiseless functional relations
  - MIC is symmetric: MIC(X, Y) = MIC(Y, X)
  - MIC is not an estimate of mutual information!

Computing MIC



> Taken from: [2]

Comparisons of Correlation

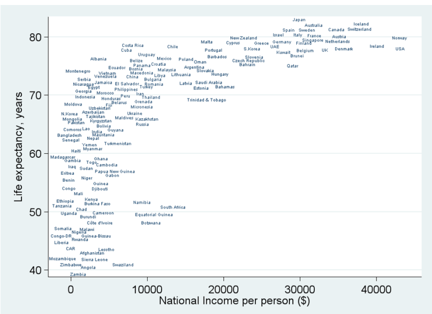
Relationship Type	MIC	Pearson	Spearman	Mutual Information (KDE)	Mutual Information (Kraskov)	CorGC (Principal Curve-Based)	Maximal Correlation
Random	0.18	-0.02	-0.02	0.01	0.03	0.19	0.01
Linear	1.00	1.00	1.00	5.03	3.89	1.00	1.00
Cubic	1.00	0.61	0.69	3.09	3.12	0.98	1.00
Exponential	1.00	0.70	1.00	2.09	3.62	0.94	1.00
Sinusoidal (Fourier frequency)	1.00	-0.09	-0.09	0.01	-0.11	0.36	0.64
Categorical	1.00	0.53	0.49	2.22	1.65	1.00	1.00
Periodic/Linear	1.00	0.33	0.31	0.69	0.45	0.49	0.91
Parabolic	1.00	-0.01	-0.01	3.33	3.15	1.00	1.00
Sinusoidal (non-Fourier frequency)	1.00	0.00	0.00	0.01	0.20	0.40	0.80
Sinusoidal (varying frequency)	1.00	-0.11	-0.11	0.02	0.06	0.38	0.76

Comparison with distance correlation (dCor): MIC requires large sample sizes, dCor more robust in small sample sizes

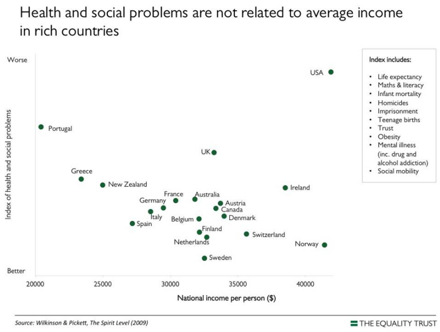
Related methods

- Non-pairwise correlations
  - Partial correlation, confounders
- Similarity measures, distance measures
- Statistical significance test
- Symbolic regression
- Copulas

Appears to be Correlated

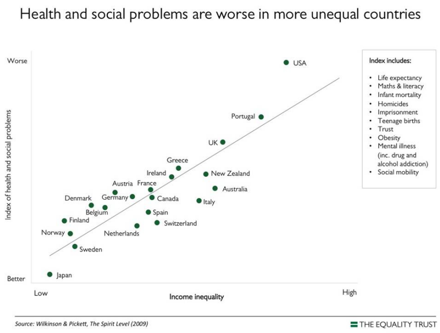


No Correlation



> Key takeaway: There might be correlation on a global scale, but also smaller correlation structures in parts of the dataset (e.g., only for specific ranges of certain variables).

Found Correlation



Hypothesis Testing

Is it statistical significant?

Hypothesis Testing

Let $x_S = f(\mathcal{D})$ the value of the *test statistic* for our dataset $\mathcal{D}$.

Let X_S be the random variable describing the value of the test statistic **under the null hypothesis** H_0 (i.e., when H_0 is true)

p-value: $p = \mathbb{P}(X_S$ more extreme than $x_S : H_0$ is true)

"X_S more extreme than x_S": depends on the test, may be $X_S \geq x_S$ or $X_S \leq x_S$ or something else...

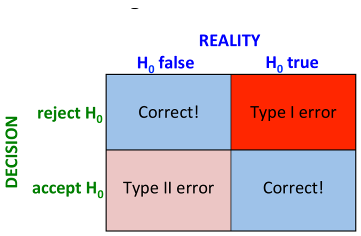
Rejection rule: Given a statistical level α *in* $(0, 1)$: reject H_0 iff $p \leq \alpha \Rightarrow \mathcal{S}$ is significant!

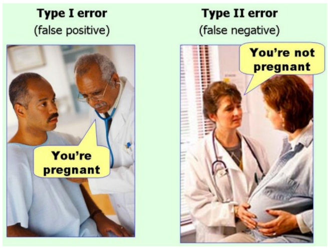
Hypothesis Testing

There are two types of errors we can make

- Type I error: reject H_0 when H_0 is true $\Rightarrow$ flag S as significant when it is not (false discovery)
- Type II error: do not reject H_0 when H_0 is false $\Rightarrow$ do not flag S as significant when it is

Hypothesis Testing





> The matrix (on the left) is a confusion matrix.

Fisher's exact test

- 2x2 contingency table, e.g., 2 groups and 1 binary feature/attribute
- $\rightarrow$ do the 2 groups statistically significant differ w.r.t the feature
- Directly provides a p-value

	Group 1	**Group 2**
True	a	b
False	c	d

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

Hypothesis Testing

Permutation test

- Fisher's permutation test
- Recall hypothesis test
  - What is the chance to observe a certain behaviour if randomly sampled?
- Idea: Just do that, e.g., random permutation of the data set
  - Count how often a certain condition has been met
- Downsides: computationally expensive, formally imprecise

Testing multiple hypothesis

- Sequentially testing multiple hypothesis at same α level will yield spurious results
- Family-Wise Error Rate (FWER)
  - Guarantees on the (expected) number of false discoveries
- Corrections
  - Bonferroni correction, Bonferroni-Holm procedure
  - LAMP
- Use validation dataset
  - ... need statistical significant result on multiple splits

Hypothesis Testing

Bonferroni correction

- Problem: random rejection of null hypothesis due to multiple tests
  - Avoid accumulation of α error
- Idea: Adaptation of significance level for n tests

$$p^* < \frac{\alpha}{n}$$

  - Alternative: Adapt p-value of individual tests
- Caveat: the Bonferroni method is conservative

Hypothesis Testing

p-Hacking

- Tricks to get the p-value below the significance threshold
  - Motivation: getting “good” scientific results (publication bias)
  - Motivation: commercial interests
- Approaches
  - Repeat experiment until - by chance - a statistical significant result has been achieved
  - Increase the amount of hypothesis
    - Try out all combination of variables, until - by chance - a statistical significant relationship has been found
    - HARKing - Hypothesizing After the Results are Known

Hypothesis Testing

Data mining without expert knowledge

If the data requires an interpretation, then results may dramatically be different, e.g., “ Twenty teams (69%) found a statistically significant positive effect and nine teams (31%) observed a non-significant relationship. Overall 29 different analyses used 21 unique combinations of covariates.” [3]

Causality

How to detect causal relationships?

Causality

Basics

- Study the impact of
  - hypothetical actions, or
  - interventions
- E.g., would there be fewer strokes if we would eat fewer Wienerschnitzel?

Causality

Preferred solution

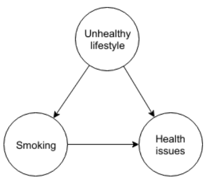
- Conduct a randomised control study
  1. Build two groups (randomly assigned)
  2. Apply the intervention on the treatment group, no treatment for the control group
  3. Observe the difference

Note: Often such a study cannot be conducted, e.g., not ethical.

Causality

Causal graphs

- Each variable is represented as a node
- Connections indicate causal relationship
  - With the arrow points from the parent (cause) to the effect
- Allow for an intuitive understanding of
  - Indirect causes (path relationships)
  - Forks (one cause, multiple effects)
  - Mediators (variables in the causal path)
  - Collider (multiple causes, one effect)



Causality

- Confounders
  - Causes that create the impression of (causal) relationships between observed variables

Causality

- Simpson's paradox
  - Given an event Y, and two variables X, Z
    - $\mathbb{P}(Y|X) < \mathbb{P}(Y|\neg X)$
    - $\mathbb{P}(Y|X, Z = z) > \mathbb{P}(Y|\neg X, Z = z)$, for all values of Z
  - Goes against the intuition, if a trend, which is true for all subpopulations, should also hold for the total population
  - Often caused by external factors (i.e., other than X, Z)

Causality

- Explaining away (collider bias)
  - Two independent variables by appear to be (negatively) correlated
    - If both are the cause for an observed variable
    - Result of the sample strategy

Limitations

Known limitations on data processing

Motivation

Goals of Knowledge Discovery

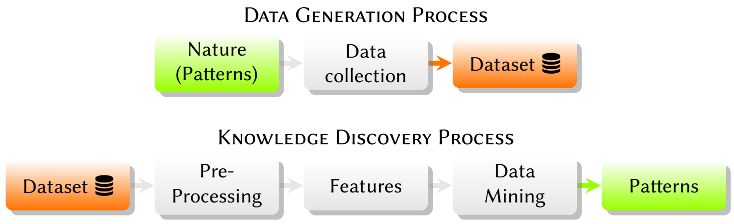
- Given data
  - ... find patterns
- Thus, our goal is
  - ... learn how data and patterns are related
- Let's start with the other direction
  - ... how is data generated

Data Processing Inequality

... you cannot invent data/information!

> https://www2.isye.gatech.edu/~yxie77/ece587/Lecture4.pdf

Data Processing



Limits of Data Processing

How hard is this problem?

- Can we re-construct the (original) patterns from the dataset?
  - Only possible, if the information (of the patterns) is still available
- For example
  - If the sensors do not record the pattern
  - ... we will not be able to recover it later

Limits of Data Processing

Model the data processing pipeline

- Model the pipeline as Markov chain
  - $X \to Y \to Z$
  - e.g., pre-processing, feature extraction, clustering
- $P(X, Y, Z) = P(Z|Y)P(Y|X)P(X)$
  - Joint probability can be factored out as conditional probabilities
  - Also, X, Z conditionally independent given Y
- Assuming time flow
  - Past and future are conditionally independent given the present

> Information flows through Y, knowing Y will explain all shared information between X and Z

Limits of Data Processing

Information Theoretical View

- X, Z conditionally independent given Y
  - $I(X; Z|Y) = 0$
- And also
  - $I(X; Y) \geq I(X; Z)$
  - $I(Y; Z) \geq I(X; Z)$
- In other words
  - Along the processing pipeline
  - ... we can only loose information!

> Once we lost critical information, there is no way to recover.
> Also applies to multi-layer neural network.

Limits of Data Processing

Practical considerations (Yes, but)

- We may want to use multiple datasets (sources of evidence)
  - Data fusion
  - ... part of feature engineering
- Even if we loose information in processing
  - Some algorithms will perform “better”
  - ... as they can better exploit the existing information
    - e.g., when extracting/parsing text we loose information, but feeding a plain text to algorithms will not work

> Once we lost critical information, there is no way to recover.

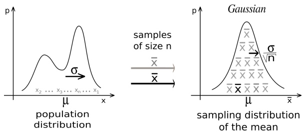
Central Limit Theorem

Relation to non-Gaussian distributions

Limitations
Central Limit Theorem

Wikipedia

[...] in many situations, when independent random variables are **summed up**, their properly normalized sum tends toward a normal distribution (informally a bell curve) even if the original variables themselves are not normally distributed.



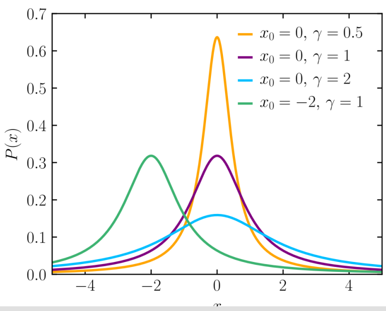
e.g., sample mean will be normally distributed

> https://en.wikipedia.org/wiki/Central_limit_theorem
> Additive assumption.

Limitations
Central Limit Theorem

- Sum of multiple random variables
  - ... normalised
  - ... with finite variance
- For example
  - X, Y ... normally distributed, independent
  - $\rightarrow Z = X + Y$
- What about
  - $Z = X/Y$
    - e.g., we want to use the ratio as feature
  - $\rightarrow$ Cauchy distribution

Limitations
Central Limit Theorem



> Taken from: https://en.wikipedia.org/wiki/File:Cauchy_pdf.svg

Limitations
Central Limit Theorem

- Cauchy breaks the CLT
  - As Cauchy does not have a finite variance
  - Sample mean of i.i.d. Cauchy is again Cauchy
- We cannot estimate the mean
  - ... as there is none
- But, we can estimate the median

> Still, Cauchy is stable.

> This also applies to the **law of large numbers**.

Robust Statistics

- Sub-field of statistics
  - Study methods not (or less) affected by
  - e.g., outliers
  - e.g., (small) violation of modelling assumptions
- Best known examples: median, interquartile range
- e.g., asymptotic breakdown point
  - Number (or fraction) of outliers a statistic is not affected by outliers

Central Limit Theorem

Practical considerations

- Please, do not consider everything is normal
- Check for outliers
  - Visually, and
  - Algorithmically
- Consider techniques like winsorising, robust statistics
  - $x' = min(LargeNumber, max(-LargeNumber, x))$

Thank You!

... for your attention

References I

[1] L. E. O. Breiman and J. H. Friedman, **Estimating optimal transformations for multiple regression and correlation,**, pp. 580–598, September 1985.

[2] D. N. Reshef, Y. A. Reshef, H. K. Finucane, et al., **Detecting novel associations in large data sets**, science, vol. 334, no. 6062, pp. 1518–1524, 2011.

[3] R. Silberzahn, E. Uhlmann, D. Martin, et al., **Crowdsourcing data analysis: Do soccer referees give more red cards to dark ski n toned players**, Center for Open Science, https://osf. io/j5v8f, 2015.