

Analysis of clinical data

jarkovsky@iba.muni.cz

Lecture 1

Organisational information - teaching materials

- This presentation in IS.MUNI + presentation and examples of Statistica software + other summary documents
- www.matematickabiologie.cz/res/file/ucebnice/pavlik-biostatistika.pdf
- portal.matematickabiologie.cz/index.php?pg=applied-analysis-of-clinical-and-biological-data--biostatistics-for-mathematical-biology
- Tables of statistical distributions
- Any basic statistics textbook - e.g.
 - https://www.amazon.com/Biostatistical-Analysis-5th-Jerrold-Zar/dp/0131008463/ref=sr_1_1?ie=UTF8&qid=1505890489&sr=8-1&keywords=zar+biostatistical+analysis
 - https://www.amazon.com/Medical-Statistics-Glance-Aviva-Petrie/dp/140518051X/ref=sr_1_sc_1?s=books&ie=UTF8&qid=1505890508&sr=1-1-spell&keywords=avive+petria
 - https://www.amazon.com/Statistics-Veterinary-Animal-Science-Petrie/dp/0470670754/ref=sr_1_sc_3?s=books&ie=UTF8&qid=1505890522&sr=1-3-spell&keywords=avive+petria

Organisational information - software

- Software
 - University licenses in inet.muni.cz (same login and passwd as for is.muni.cz)
 - Statistica - www.statsoft.com, www.statsoft.cz
 - SPSS - www.ibm.com/analytics/us/en/technology/spss/
 - [R](http://www.r-project.org) - www.r-project.org, www.rstudio.com
 - Python - <https://www.python.org/>
 - Stata - www.stata.com

Statistics in scientific practice

The position of statistical analysis in science and clinical practice

The importance of statistical outputs


Annotation

- Statistical analysis of biological data is one of the tools we use to try to find answers to our questions about understanding living nature.
- Like any tool, statistical analysis must be used correctly on the one hand and not overestimated on the other.
- A key fact in statistical data analysis is viewing reality through a sample and accepting that the results of our analysis are only as good as our sample.
- The representativeness, independence and randomness of the sample, together with its size, are important factors affecting the plausibility of our conclusions.

Life is beautiful with data analysis



What does statistical data analysis mean for a biologist/physician?

- **Mathematical statistics** is a scientific discipline on the borderline between descriptive statistics and applied mathematics. It deals with the theoretical analysis and design of methods for obtaining and analysing empirical data containing an element of randomness, i.e. the theory of planning experiments, sampling, statistical estimation, hypothesis testing and statistical models.
- **Statistics** is the science and practice of developing human knowledge using empirical data. It is based on mathematical statistics, which is a branch of applied mathematics.
- **Biostatistics** = application of statistical data analysis in biological and clinical research
 - A tool for grasping our research data
 - Necessary to understand the principles and limitations
 - Detailed mathematical knowledge is not required
- **Easy to understand**  **hard to master**



Research, reality, statistics

- Research is our way of understanding reality
- But how accurate and true is our understanding?

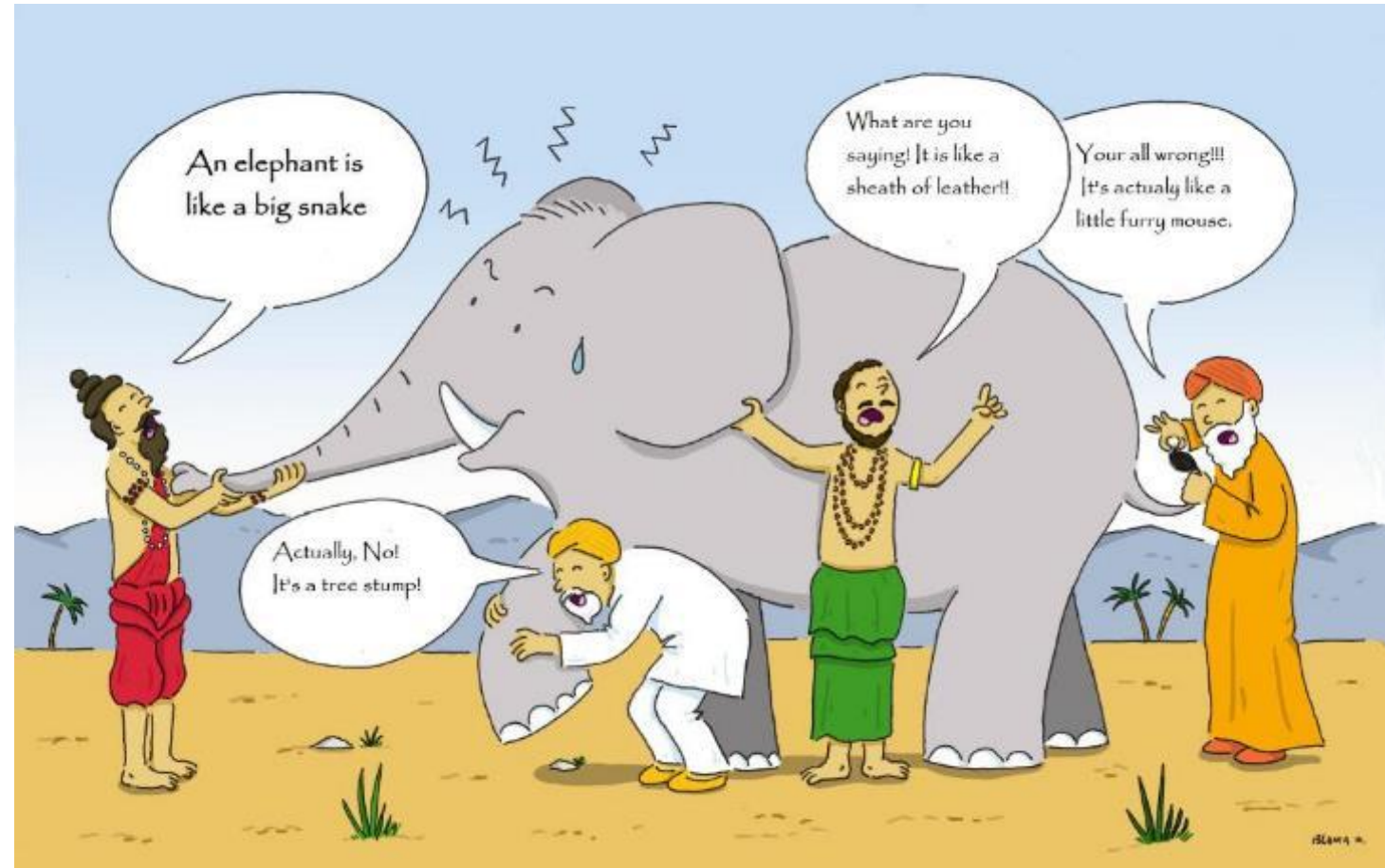


- **Statistics** is one of the tools for describing and communicating research results.
- But it's only the tool, what really matters is the **data**.



Reality and data

- A key question in the research and subsequent statistical analysis is how well our data describes the reality
- Without good data, there is no good statistics or good research.
- Any mistake made in the initial phase of the research will be multiplied in later phases and probably cannot be eliminated



Variability as a basic concept in statistics

- Our reality is variable and statistics is the science of variability
- Correct analysis of variability and its understanding provides useful information about our reality
- In a deterministic world, statistical analysis would not be necessary

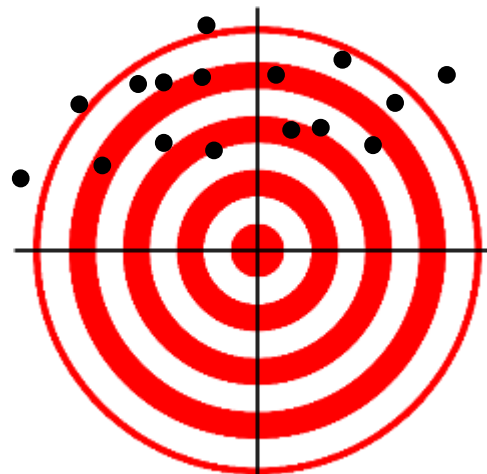


?

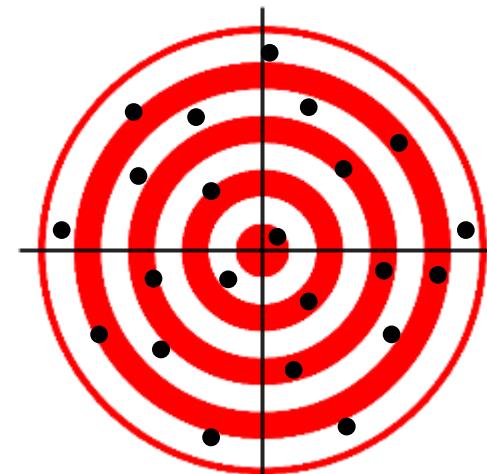


Reliability and measurement accuracy

- Data quality is crucial for any statistical evaluation
- Without reliable and accurate data, it is not possible to obtain reliable and accurate statistical evaluation results
- In the statistical analysis of data, we need to take into account both the mean of measurement and variability and think about the accuracy of the description of reality



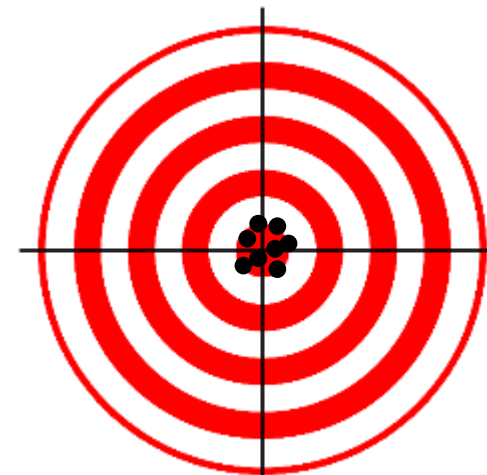
Unreliable, inaccurate



Unreliable, accurate



Reliable, inaccurate



Reliable, accurate

Variability and mean value

- Standard = 5 grams of salt per 1 kg of rice

You don't mix



0g salt / 1 kg rice



10g salt / 1 kg rice



Average: 5g salt / 1 kg rice
All OK !!!

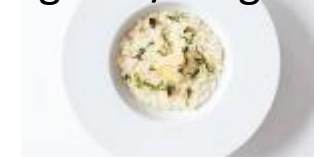
You mix



5g salt / 1 kg rice



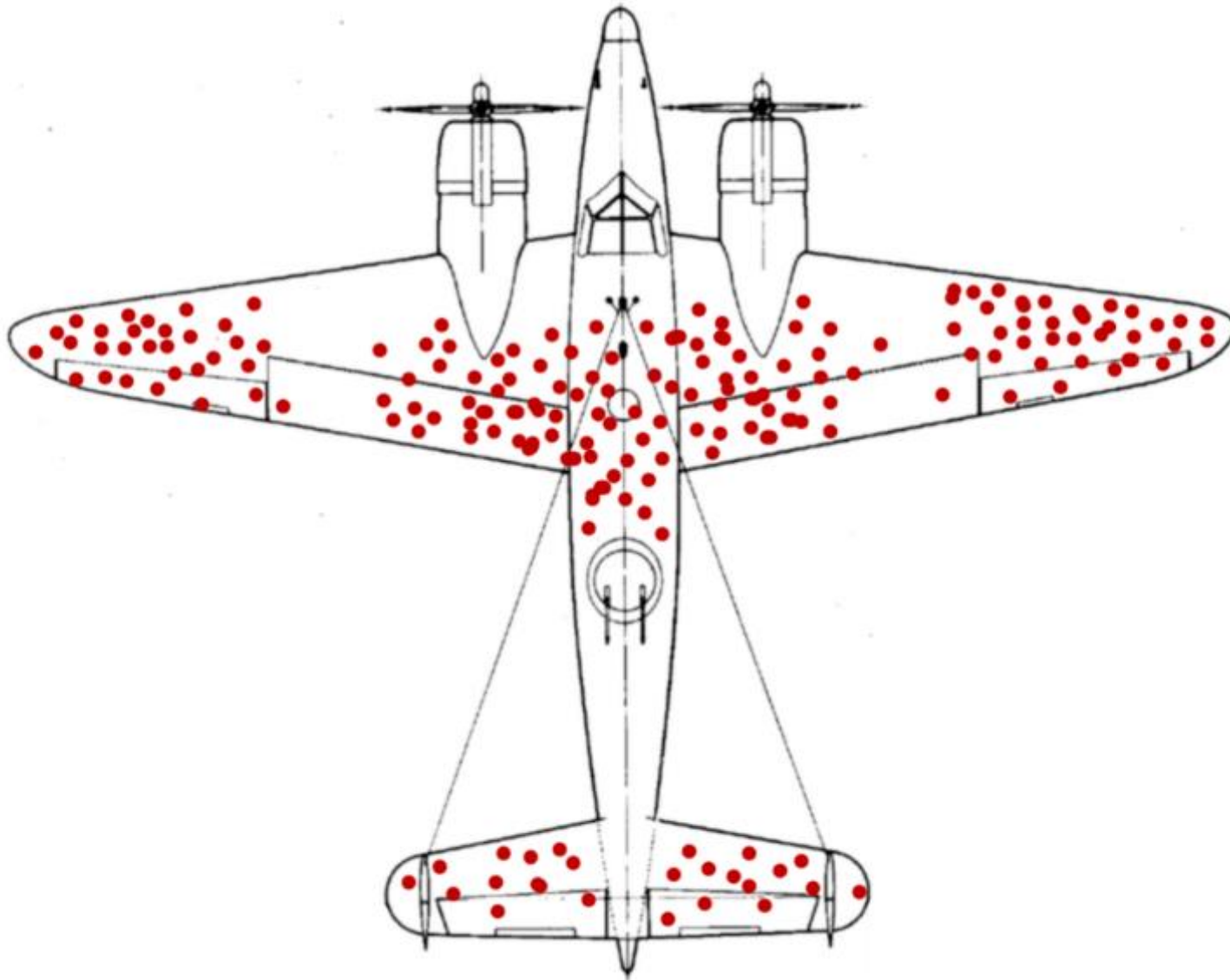
5g salt / 1 kg rice



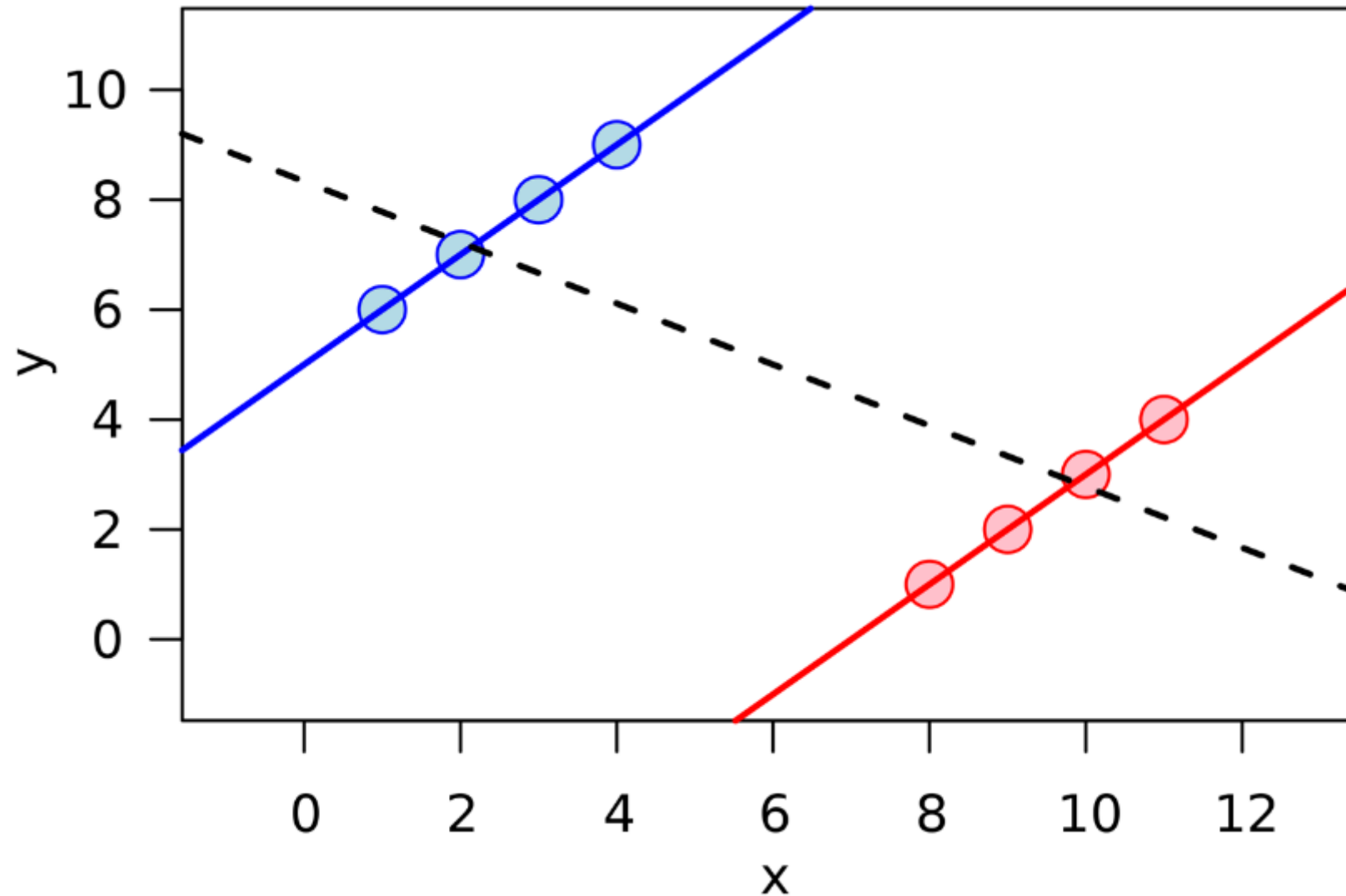
Average: 5g salt / 1 kg rice
All OK !!!

**Average is not
everything, it is essential
to take into account
variability**

Is reality really reality? Survivor bias

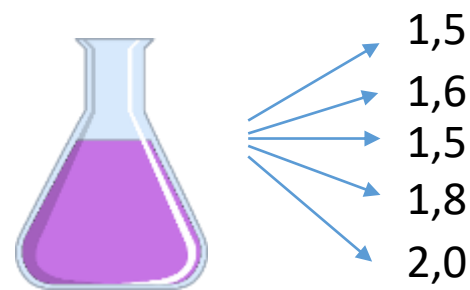


Are we interpreting the results correctly? Simpson paradox

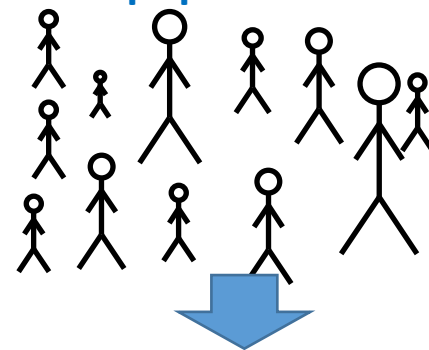


Different levels of variability

Variability of repeated measurements

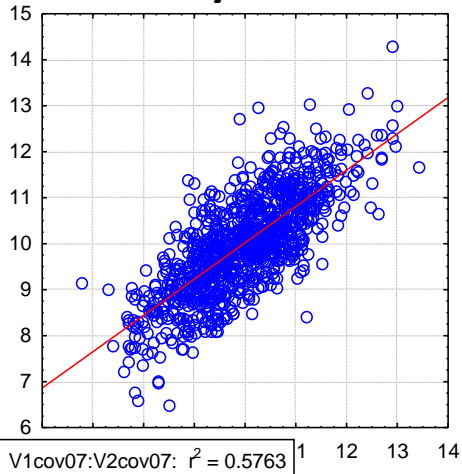


Data variability in the population

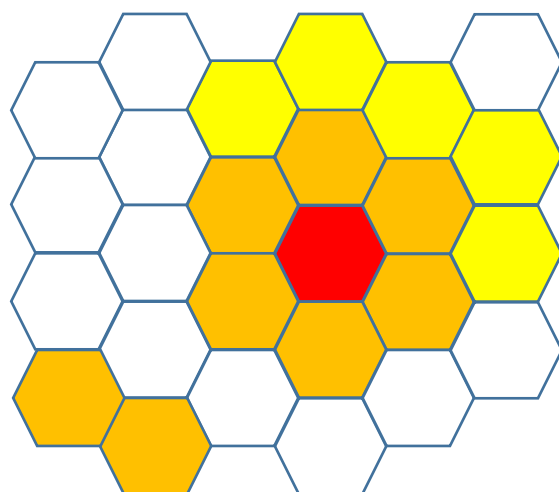


Main topic of the course

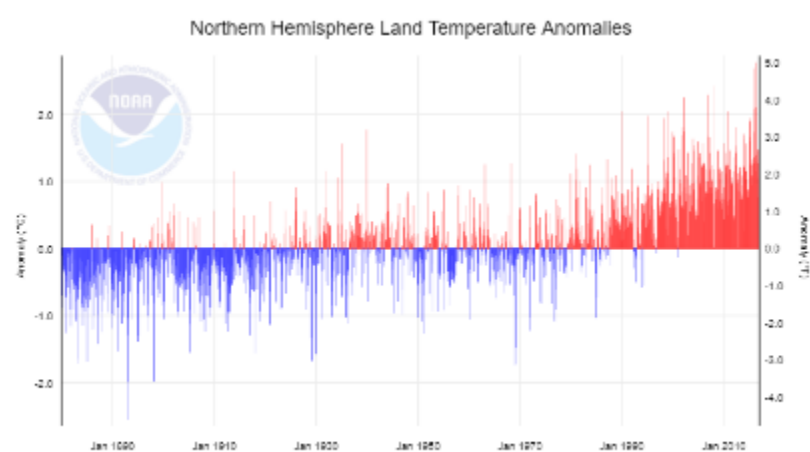
Variability in models



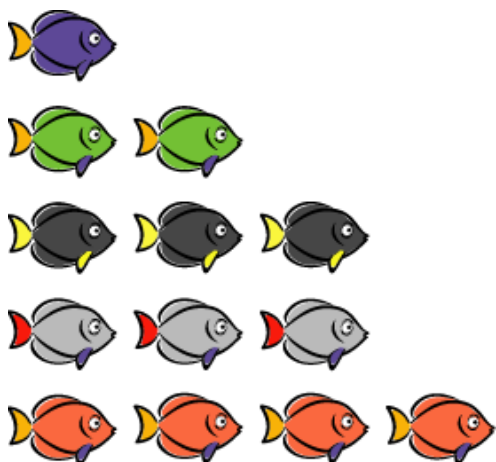
Geographical variability



Time series variability

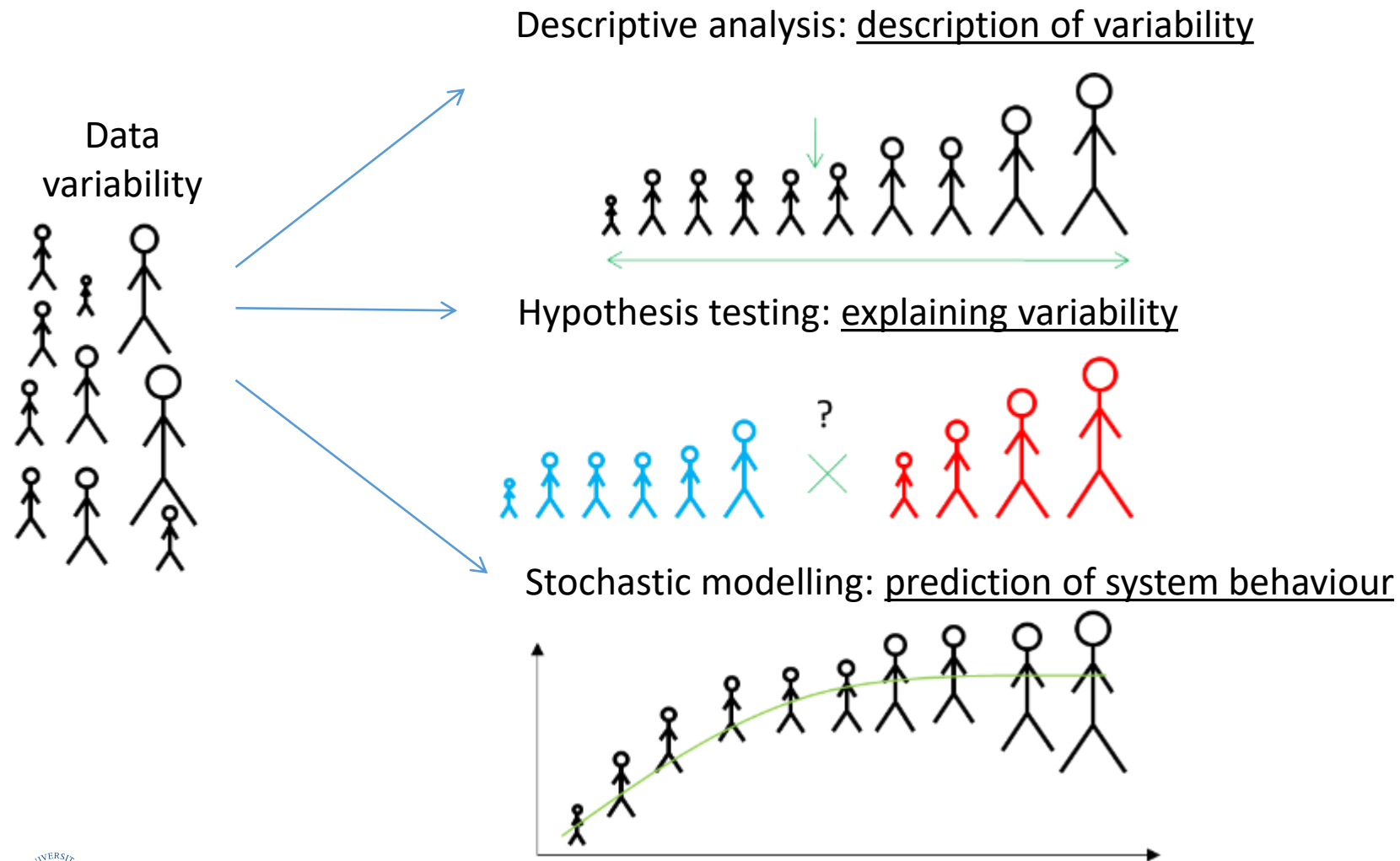


Biodiversity



Working with variability in data analysis

- There are three main approaches to dealing with variability in data analysis



Statistics - definition

WWW.WIKIPEDIA.ORG:

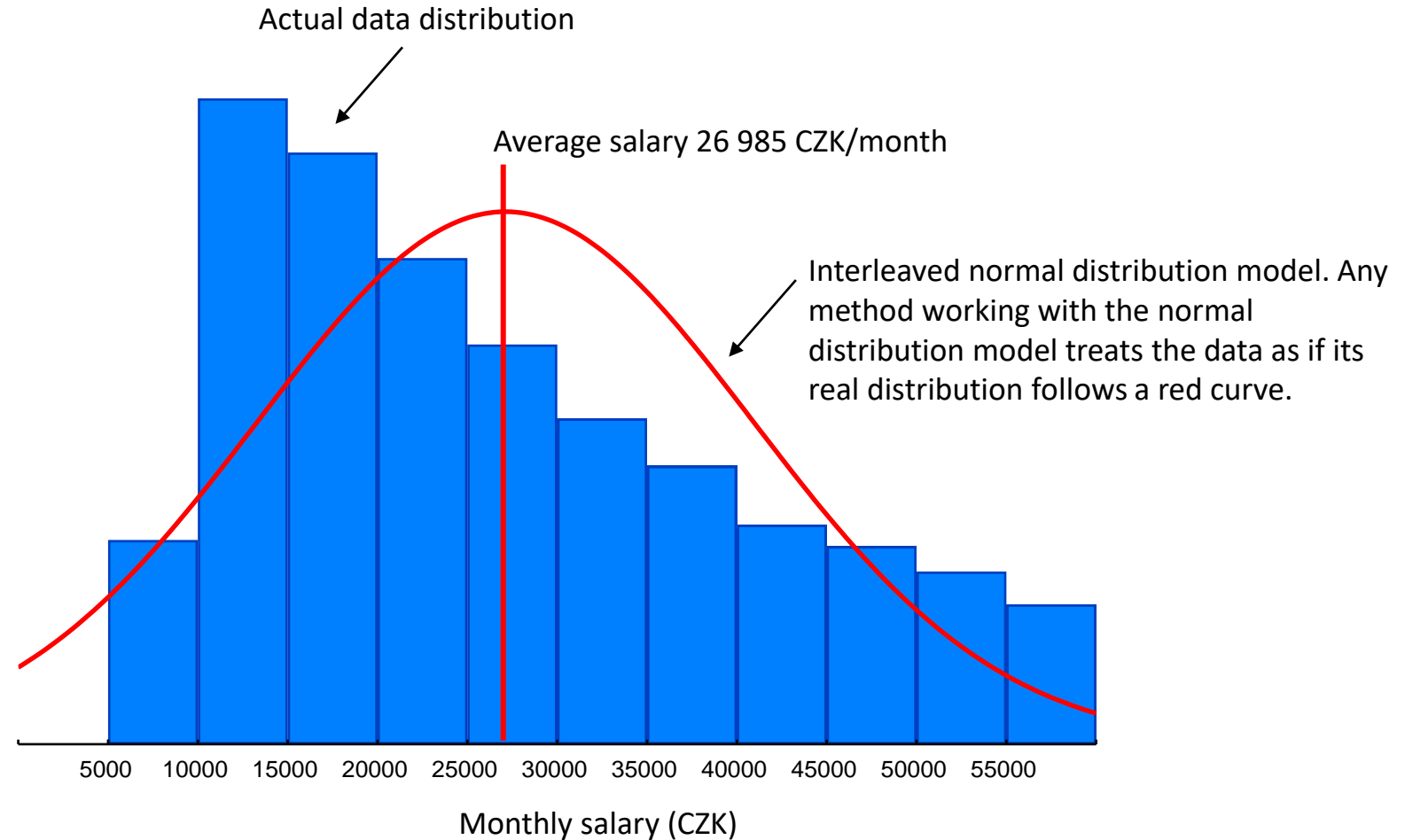
Statistics is the mathematical science concerned with the collection, analysis, interpretation, explanation and presentation of data. It can be applied in a wide range of scientific disciplines from the natural to the social sciences. Statistics is also used as a basis for decision-making, but it can be misused intentionally or unintentionally.



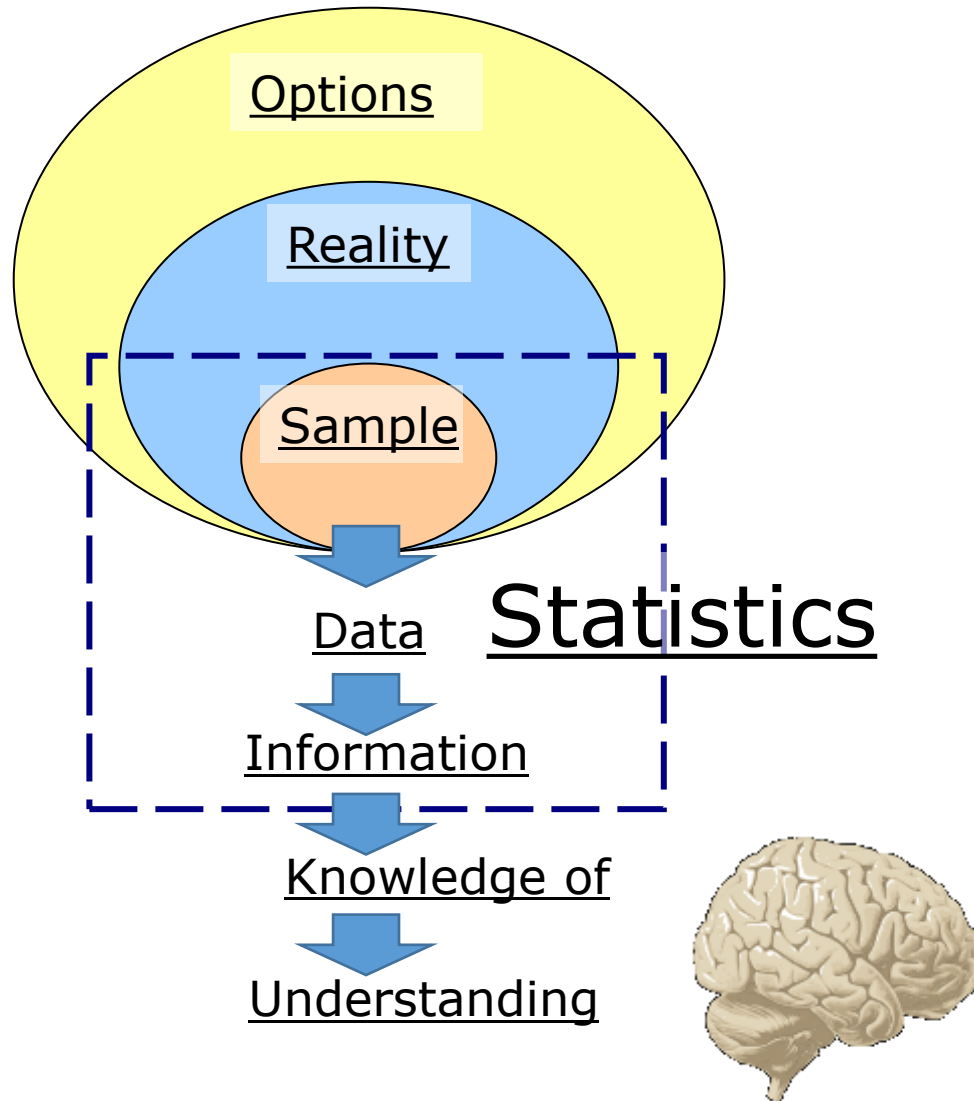
Statistics uses mathematical models of reality to generalize the results of experiments and sampling. Statistics works correctly only if the assumptions of its methods and models are met.

Incorrect model application -> biased conclusions

- Different descriptive statistics and tests are associated with different model distributions
- For a correct interpretation it is necessary to verify the fit of the real data with the model
- Some statistics can always be calculated, but their interpretation is limited if the assumptions are not met



What can statistics say about our reality?



Statistics is unable to draw conclusions about phenomena not included in our sample.

Statistics is deployed in the process of extracting information from the sampled data and is a support in gaining our knowledge and understanding of the problem.

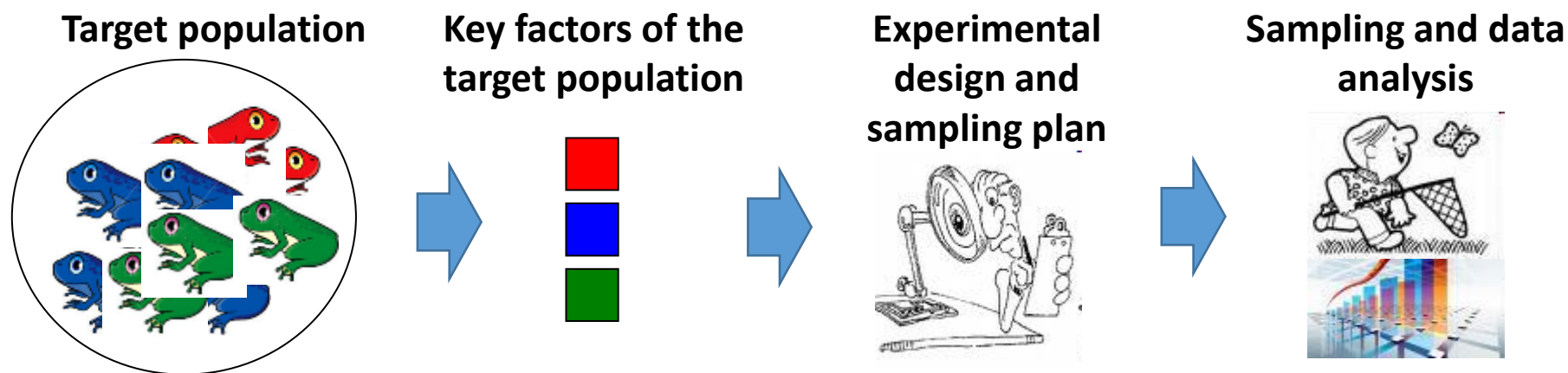
Statistics are no substitute for our intelligence !!!

What do we need to know before starting a study or experiment?

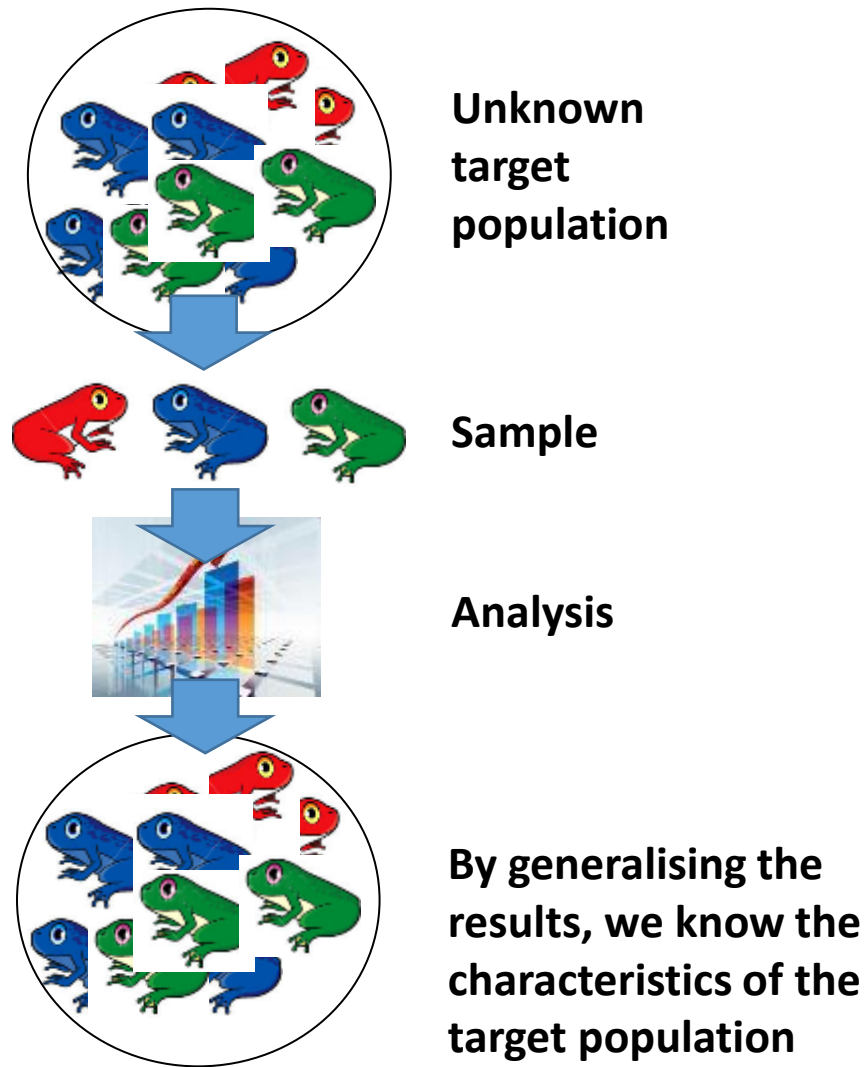
- Target population
 - The group of subjects (patients, sites, etc.) on which the study is focused
- Primary hypotheses
 - The main question asked in the study - sample size estimation and study design is developed in relation to the primary hypothesis (in many cases, formal power analysis cannot be developed in real research, but reflection on sample size is always necessary)
- Secondary hypotheses
 - Secondary questions that the study should answer
- Selection of an adequate methodology
 - Hypotheses are answered through specific variables (endpoints) - their type (binary, categorical, continuous variables, biodiversity, survival, mortality, etc.) determines the choice of statistical treatment

Target population

- Target population - a key concept in statistical processing
 - Group of objects we want to learn about (e.g. locations in a given catchment area, laboratory organisms in given conditions, patients with a given diagnosis, all people over 60, haemoglobin measurements in a given laboratory)
 - Must be defined before data collection begins
 - Data sampling is performed on the target population, which must characterize the target population well (representative)



Statistics and generalisation of results



- The aim of the analysis is not simply to describe and analyse the sample, but to generalise the results from the sample to its target population
- If the sample is not representative of the target population, generalisation leads to erroneous conclusions

Sampling and its importance in statistics

- Statistics speak of reality through sample!!!
- Statistical assumptions for correct sampling

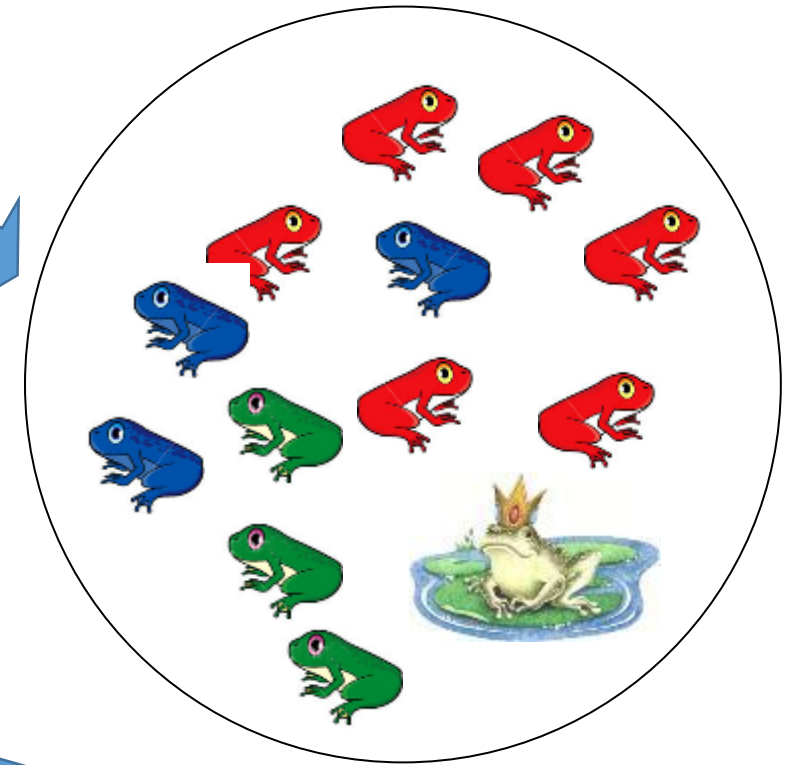
- **Representativeness:** the sample design must reflect reality as much as possible



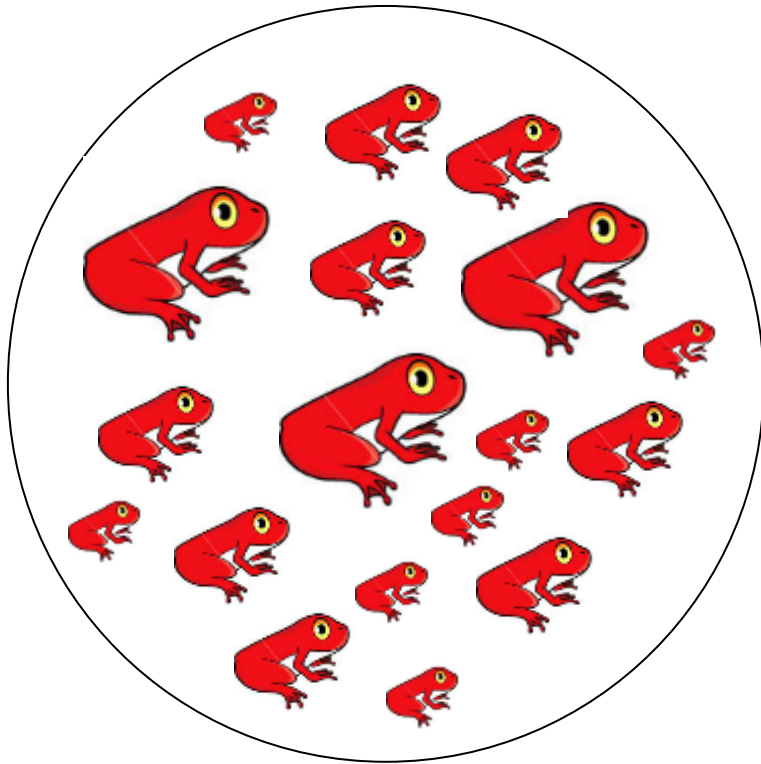
- **Independence:** multiple sampling of the same object provides no new information from a statistical point of view



- **Randomness:** ensures random influence of confounding factors



Sample size and reliability of statistical outputs



- There is a true distribution and a true mean of the measured variable
- We're not gonna get anything from one measurement



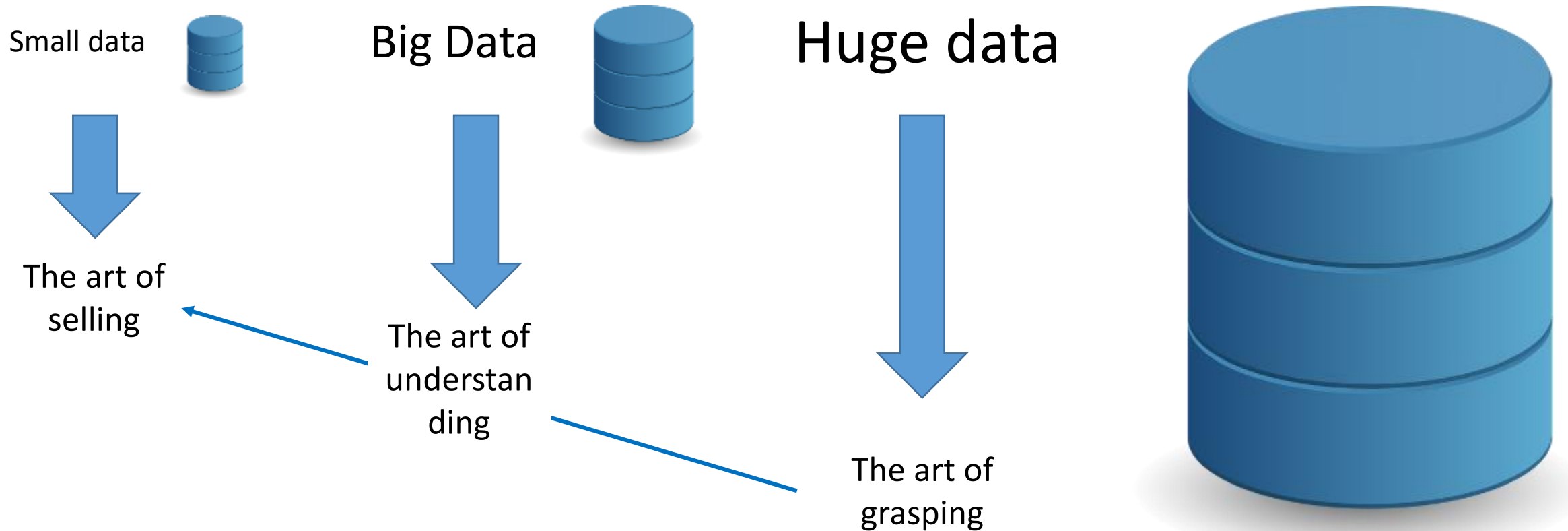
- A sample of a certain size provides an estimate of the fair value with defined reliability



- Sampling all existing objects will provide the true value of a given descriptive statistic, but this approach is unrealistic in most cases.

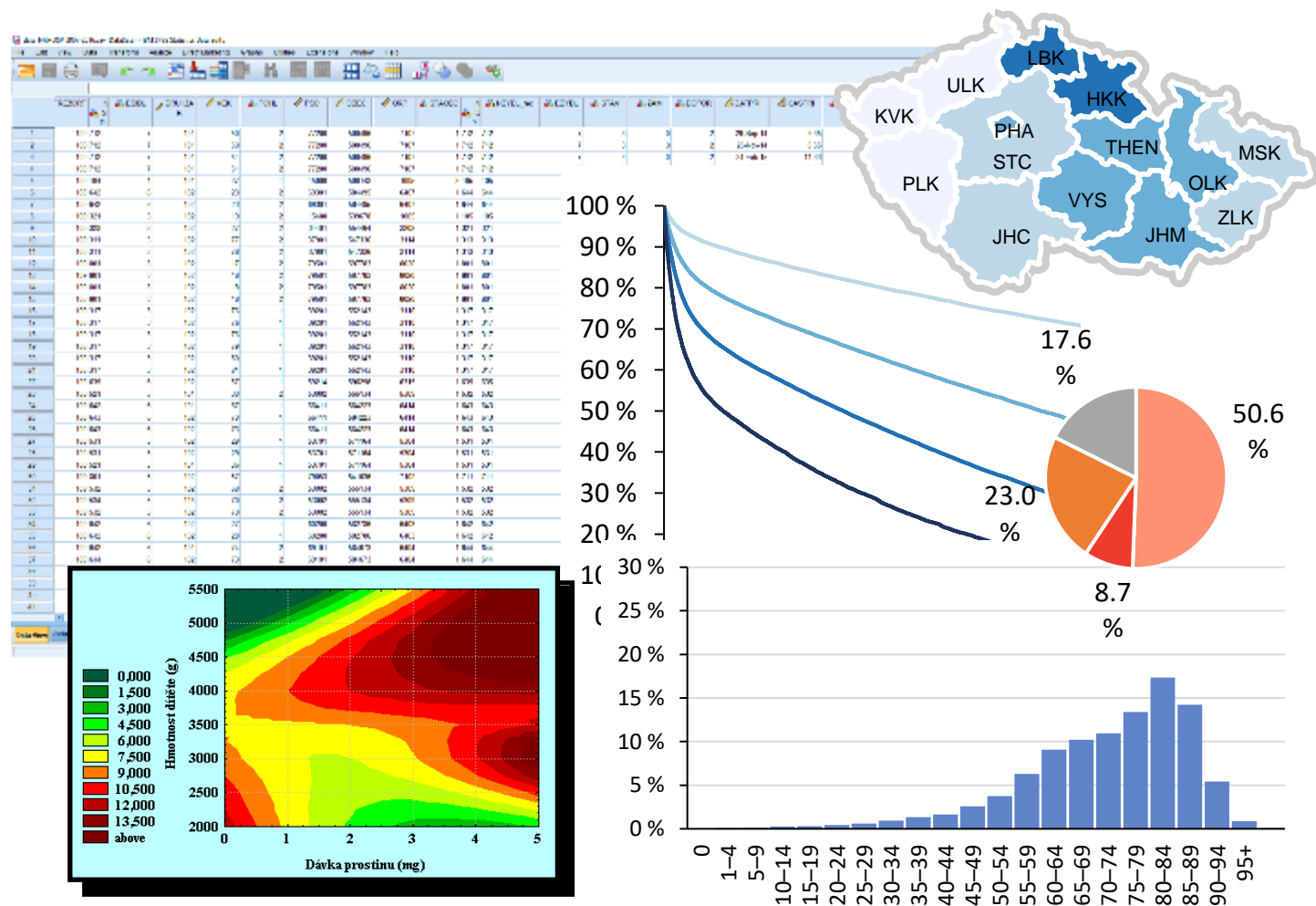
Different sample sizes - different data analysis tasks

- The complexity of data analysis increases with the volume of data
- Even with the biggest data, the ability to sell the data = meaningful interpretation and presentation is still key



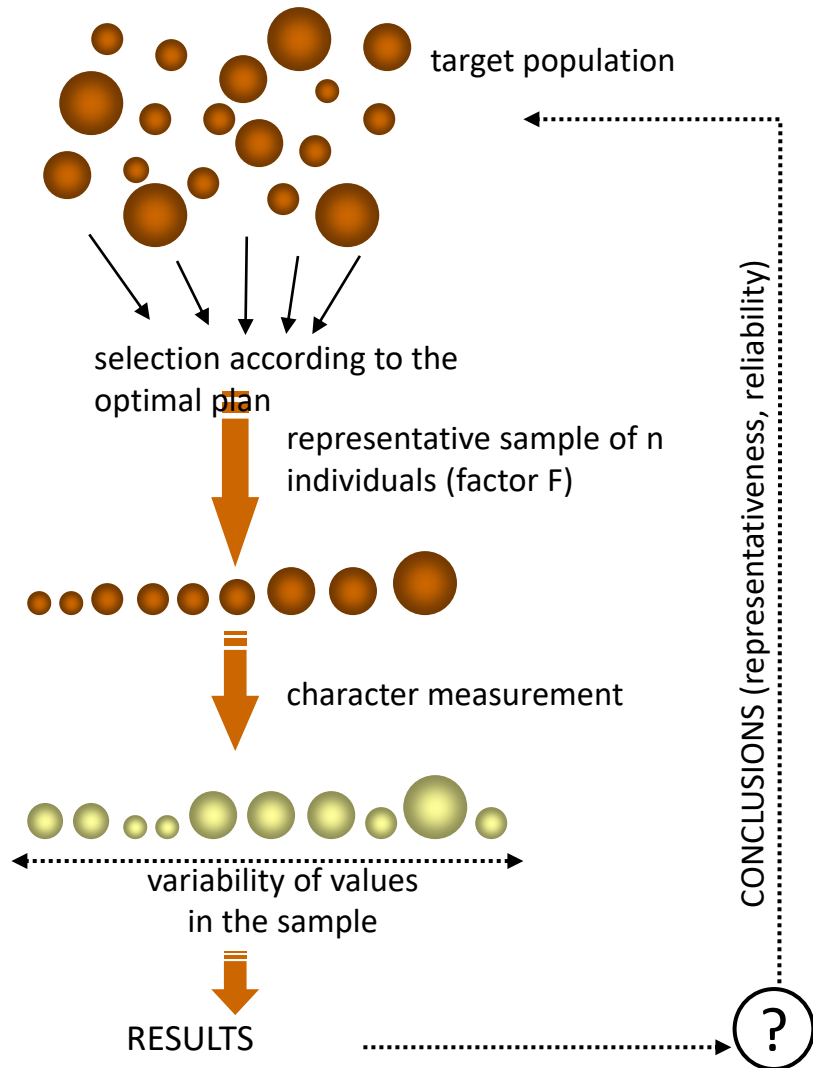
Biostatistics approach

- Ability to: see data - communicate - interpret - sell



Experimental design: essential equipment for the biologist

Purpose of analysis: descriptive

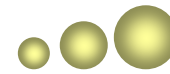


?

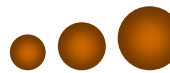
Representativeness

Reliability

Accuracy

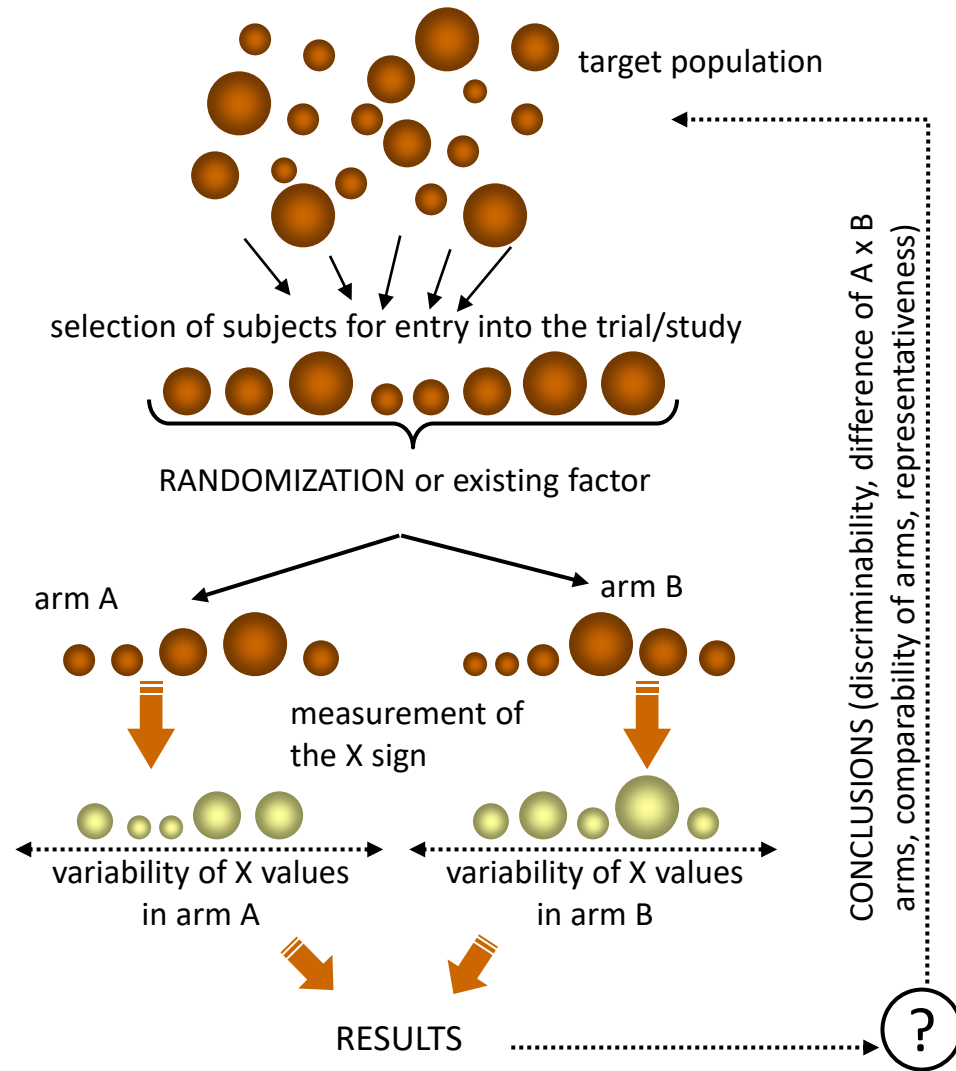


... analyzed trait of the target population (X)



... other significant factor characterising the target population (F)

Experimental design: essential equipment for the biologist



Purpose of analysis: comparative (2 groups)

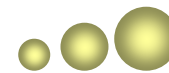
?

Representativeness

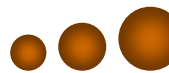
Comparability

Reliability

Accuracy

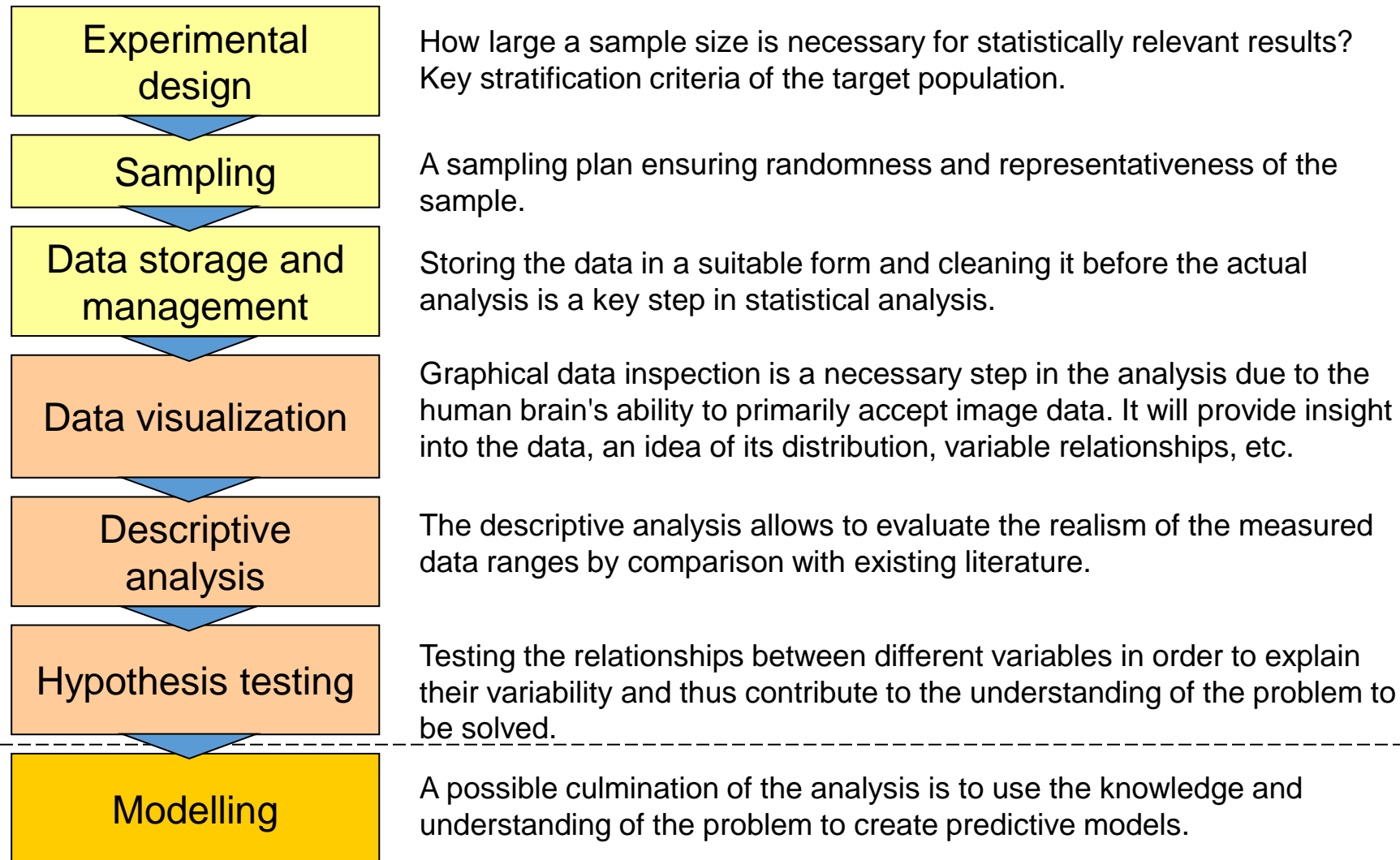


... analyzed trait of the target population (X)



... other significant factor characterising the target population (F)

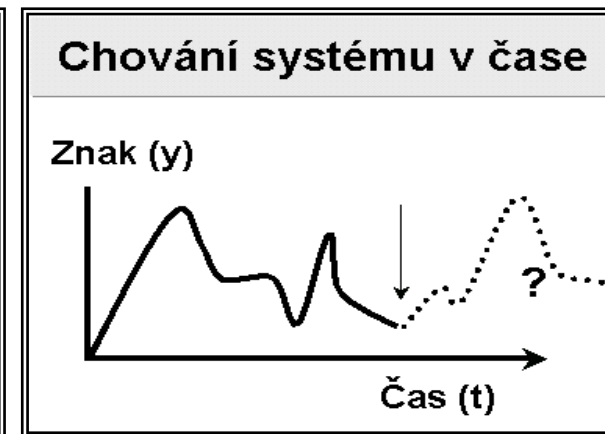
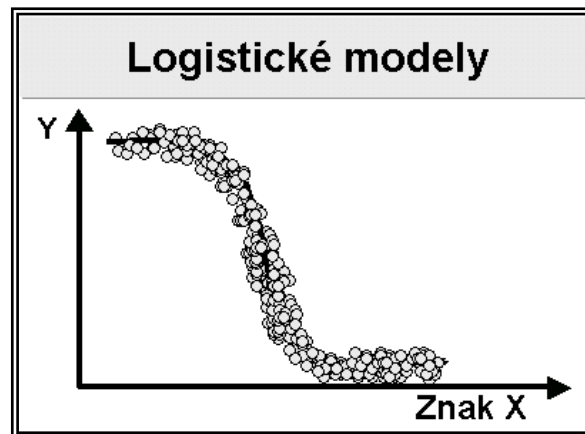
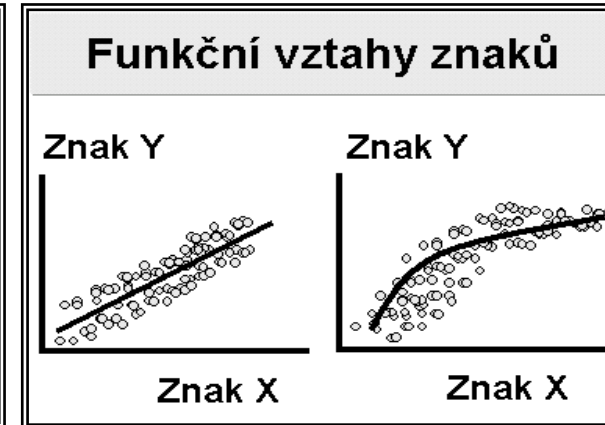
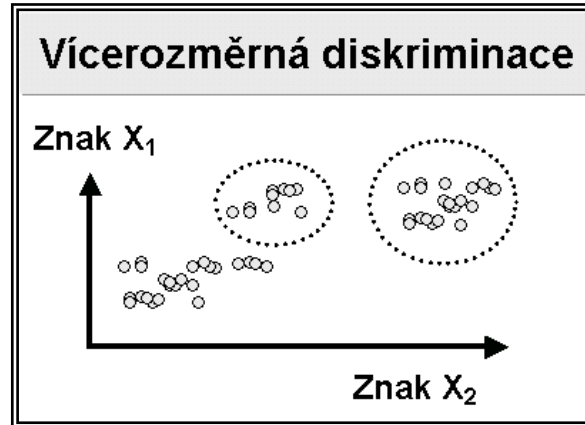
General scheme of using statistical analysis



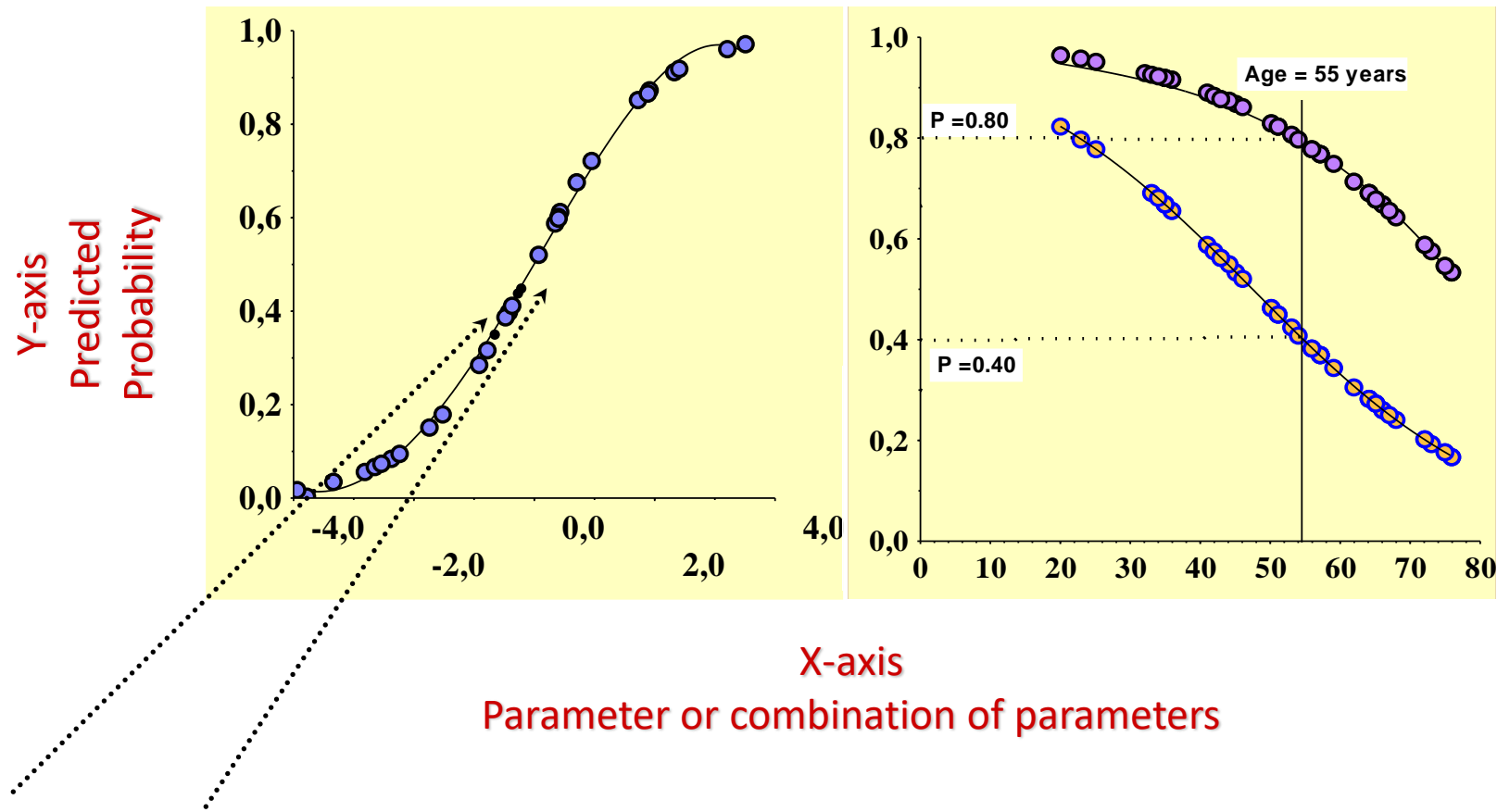
Stochastic modelling: prediction of uncertain phenomena

- Prospectively - model - it affects the behaviour of phenomena while respecting variability

Pravděpodobnostní vztahy					
Anamnéza x Výsledek vyšetření pacienta					
	Karcinom	Benigní léze	Benigní riziková	Zdravá	
Pozitivní anamnéza	2,22	34,44	0,00	63,33	100%
Negativní anamnéza	1,06	28,23	0,96	69,75	100%
p < 0.05					



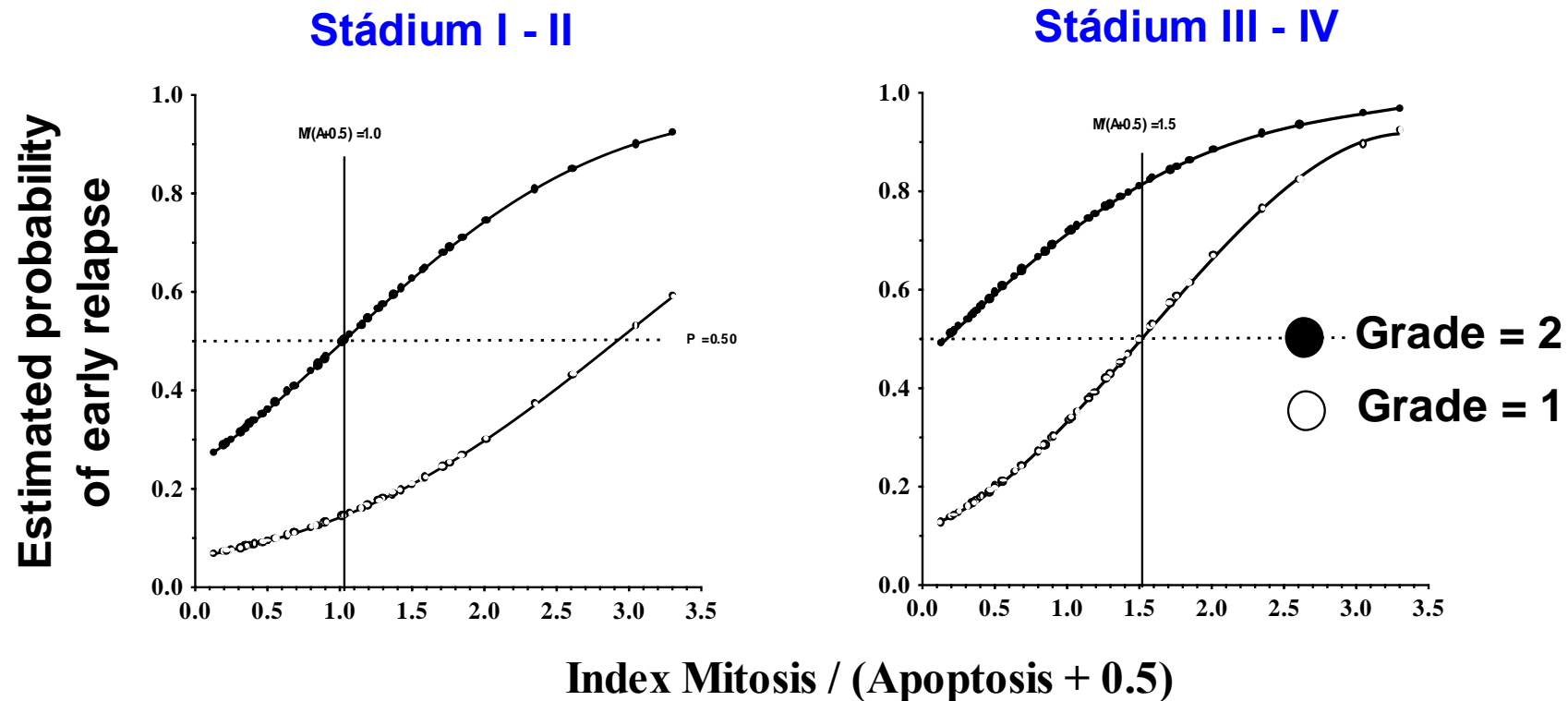
Stochastic modelling: prediction of uncertain phenomena



Object-specific data for direct assessment

Stochastic modelling: prediction of uncertain phenomena

- Ability to: create actionable tools



Lecture 2

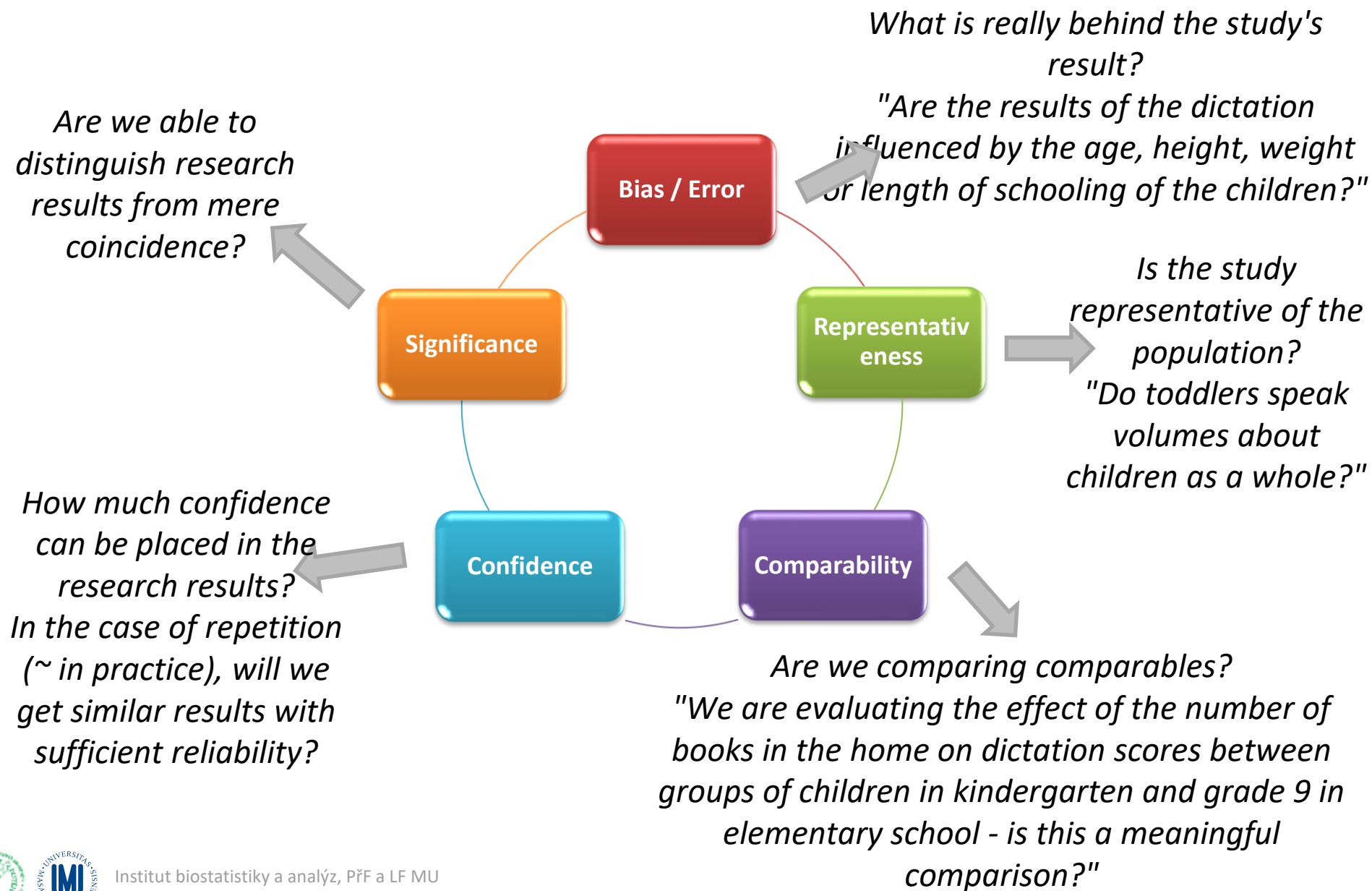
Key principles of biostatistics

Bias, representativeness, comparability, reliability significance

Annotation

- In the statistical analysis of biological and clinical data, we must always think about the research and its results in the context of the 5 key principles of biostatistics.
- Distortion - do we really see what we think we see?
- Representativeness - does our analysis tell us about the group of objects we are interested in?
- Comparability - what are we actually comparing in the analysis?
- Reliability - how reliable are our results, can they be repeated?
- Significance - how likely is it that we are observing the results of mere chance?
- Neglecting these principles can lead to misinterpretation of results.

Key principles of biostatistics

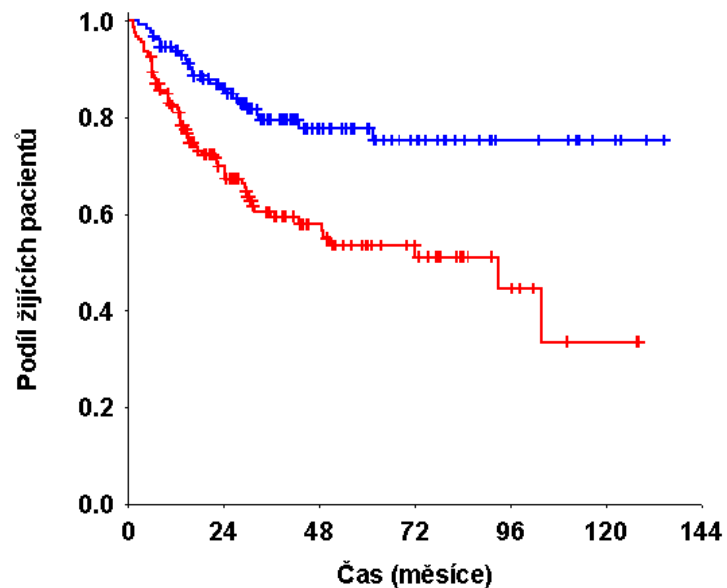


Key Principles – Bias / error

- In any evaluation, we try to avoid biased results - that is, biasing the results by factors other than those that are the objectives of the research.
- Statistical comparisons are never 100% reliable, there is randomness and therefore the probability of a wrong judgement - this cannot be controlled.
- We want to use adequate methods to remove influences that would bias the results and are not random (e.g., gender distribution, altitude).

Key Principles – Bias / error

- What causes the difference in saprobic pollution of a watercourse?
- What causes the difference in the measured biochemical parameters?
- What could account for the observed difference in 10-year survival?



Treatment?

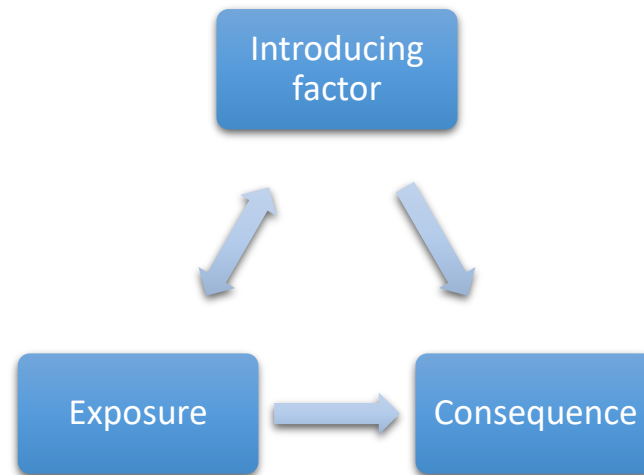
Any prognostic factor?

The stage of the disease?

Age?

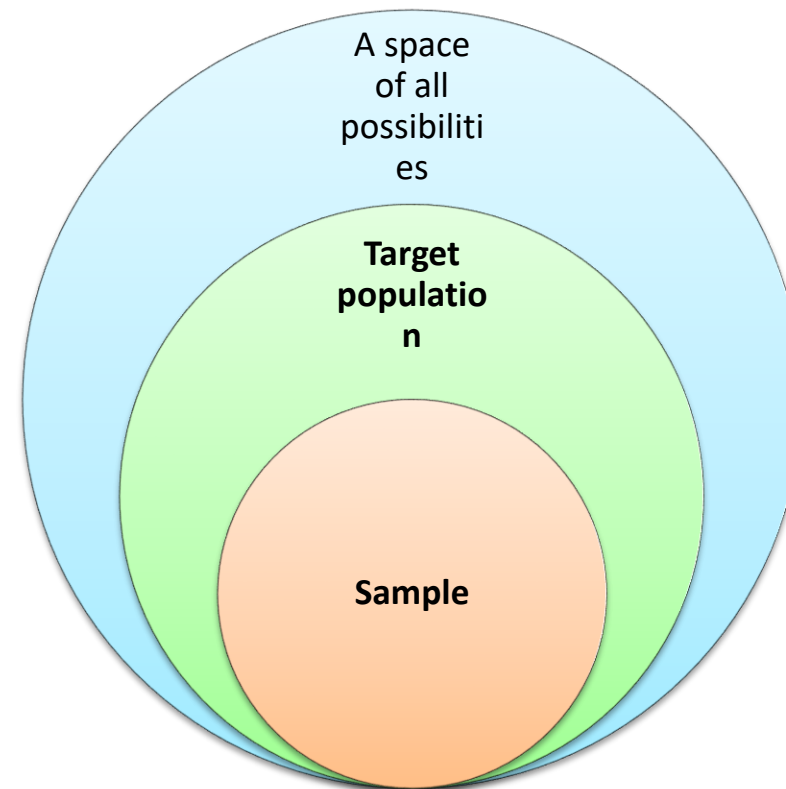
Key Principles – Bias / error

- The concept of confounding factor
- For the confounding factor, it is also true that
 - directly or indirectly influences the outcome of interest,
 - is related to the exposure being studied,
 - is not an intermediate step between exposure and consequence.

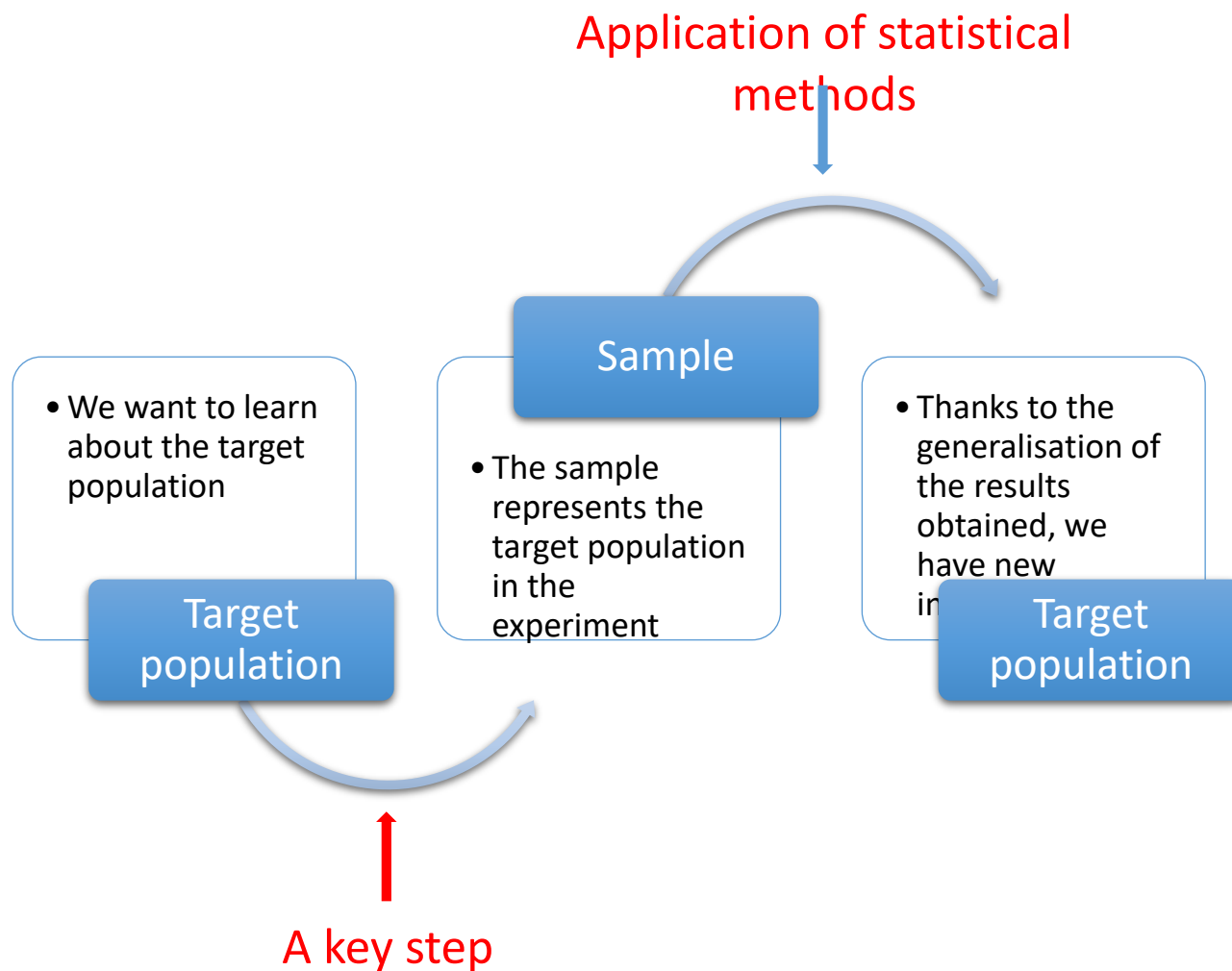


Key Principles - Representativeness

- Target population - a group of subjects about whom we want to find out some information.
- The concept of an experimental sample - a subset of the target population that is "available to us".
 - They must match the characteristics of the target population.
 - We want to generalize the results to the whole target population.
 - Relation to random selection.



Key Principles - Representativeness



Key principles - comparability

- Correct results in comparative analyses can only be obtained when comparing the comparable.
- In strictly controlled trials, comparability is ensured by randomization.
- In studies without randomization, the topic of comparability of groups must be addressed.
- Methods of adjustment, matching, propensity scores.



Key Principles - Confidence

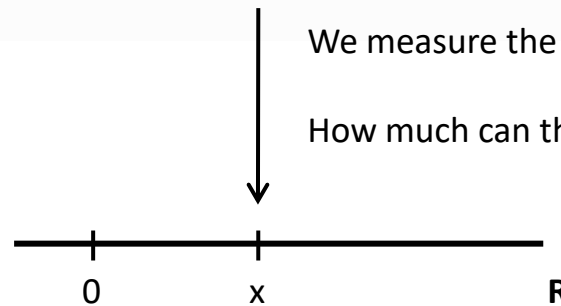
- In most studies, we are interested in quantifying the effect or characteristic of interest, generally a random variable, in the form of a single number, a point estimate.
- However, the point estimate itself is insufficient.
- It must be supplemented with an interval estimate that corresponds to the probabilistic behaviour of the observed variable, i.e. corresponds to a certain reliability of the result.

Key Principles - Confidence

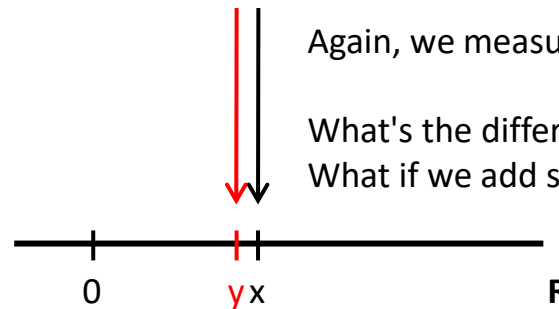


We measure the quantity of interest and then calculate an estimate.

How much can this point estimate be generalized to the target population?



Key Principles - Confidence

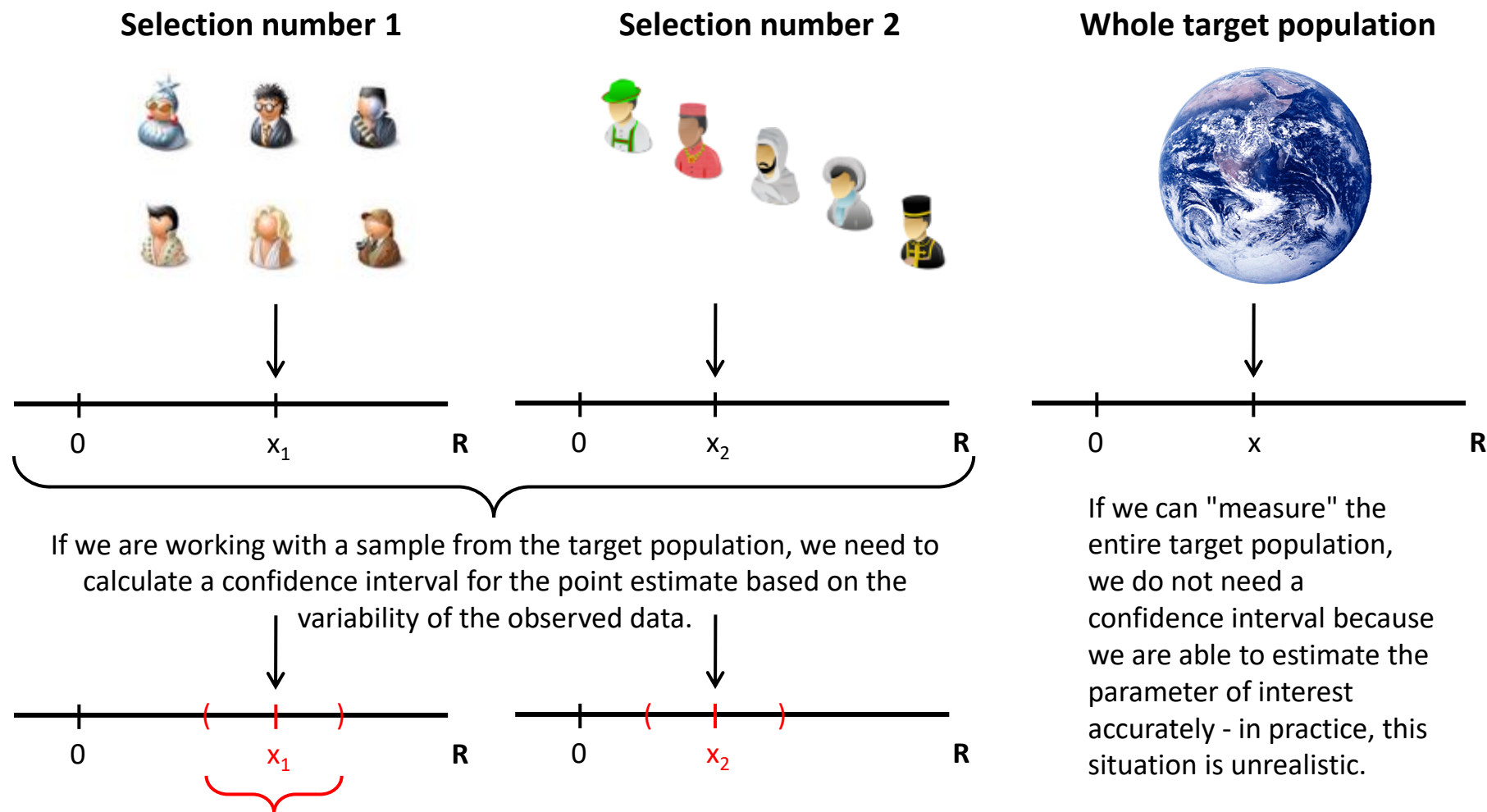


Again, we measure the quantity being monitored.

What's the difference?

What if we add someone else?

Key Principles - Confidence



Confidence interval based on selection number 1.

If we can "measure" the entire target population, we do not need a confidence interval because we are able to estimate the parameter of interest accurately - in practice, this situation is unrealistic.

Key Principles - Significance

- The analytical results of the study may not correspond to reality and reality. Statistical significance may simply not mean causality!
- Statistical significance only indicates that the observed difference is not random (in the sense of the stated hypothesis).
- Equally important is practical relevance, i.e. relevance from the point of view of a physician or biologist.
- Statistical significance can be affected by sample size.

Key Principles - Significance

		Practical significance	
		YES	NO
Statistical significance	YES	OK, practical and statistical significance are in agreement.	A significant result is a statistical artifact, practically unusable.
	NO	The result may be a mere coincidence, an inconclusive result.	OK, practical and statistical significance are in agreement.



A statistically insignificant result does not mean that the observed difference does not actually exist! It may be due to insufficient information in the observed data!

Data preparation

The key importance of correct storage of acquired data

Data storage rules

Data cleaning before analysis

Annotation

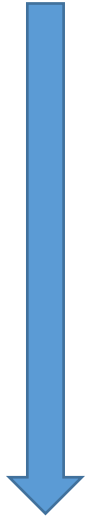
- Current statistical analysis cannot do without data processing using statistical software.
- A prerequisite for success is the correct storage of data in the form of a "database" table that allows its processing in any application.
- It is equally important to pay attention to data cleaning prior to the actual analysis.
- Any error that is made or not found in the data preparation phase will affect all subsequent steps and may cause invalid results and the need to repeat the analysis.

DATA - sample data file layout

Parameters, signs, characteristics, variables



Records



Pacient	Clovek	aLeu	aTy%	aSe%	aNeu%	aLy%	aTy	aSe	aNeu	aLy	aHtc	aCLsk	aCLNeus	aCLOZ	aCLNeuO
		cell.10 ⁶ /	%	%	%	%	cell.10 ⁶ /	cell.10 ⁶ /	cell.10 ⁶ /	cell.10 ⁶ /	%	mV.s.10 ³	mV.s.10 ³	mV.s.10 ³	mV.s.10 ³
3	1	4									33	72		32	
4	2	7,6	8	58	66	24	0,6	4,4	5,0	1,8	33	95	19	48	10
8	3	4	3	52	55	40	0,1	2,1	2,2	1,6	22	77	35	33	15
11	4	6,1	5	59	64	35	0,3	3,6	3,9	2,1	33	103	26	49	13
12	5	6,9	3	85	88	9	0,2	5,9	6,1	0,6	37	81	13	45	7
14	6	5,9	15	55	70	19	0,9	3,3	4,1	1,1	32	137	33	61	15
16	7	8	18	75	93	7	1,4	6,0	7,4	0,6	34	151	20	59	8
20	8	9,6	3	72	75	23	0,3	6,9	7,2	2,2	40	77	11	38	5
21	9	6	10	67	77	19	0,6	4,0	4,6	1,1	32	120	26	52	11
22	10	3,3	4	55	59	39	0,1	1,8	2,0	1,3	28	81	42	24	12
37	11	3,8	10	60	70	30	0,4	2,3	2,7	1,1	32	111	42	29	11
38	12	6,4	2	76	78	17	0,1	4,9	5,0	1,1	25	366	73	115	23
39	13	6,8	1	57	58	39	0,1	3,9	3,9	2,7	20	234	59	71	18
49	14	8,5	7	67	74	26	0,6	5,7	6,3	2,2	30	156	25	108	17
51	15	9,3	7	57	64	35	0,7	5,3	6,0	3,3	35	129	21	23	4
52	16	2,2	10	56	66	34	0,2	1,2	1,5	0,7	33	46	30	12	8
55	17	9,9	3	78	81	10	0,3	7,7	8,0	0,1	30	189	24	140	18
56	18	5	2	80	82	13	0,1	4,0	4,1	0,7	26	101	25	54	13
6	1	8,8	11	72	83	12	1,0	6,3	7,3	1,1	44	268	36,6	145	19,9
9	2	9,2	2	66	68	28	0,2	6,1	6,3	2,6	42	168	26,9	76	12,2
13	3	10,0	7	83	90	8	0,7	8,3	9,0	0,8	54	181	20,1	81	9
15	4	9,6	1	75	76	23	0,1	7,2	7,3	2,2	45	343	47	124	16,9
17	5	6,0									45	40		21	

Data table and its possible problems

Unique ID necessary for identification and possible linking to documentation.

The column must not contain a combination of text and numbers.

Incorrect date.

Typo in category name, it behaves like a new category when processing data.

Unrealistic outliers, probably swapped age and height.

The 0 given is probably instead of the missing value, the cell should be left blank.

Combination of two possible categorizations (0/1 or N/A), you need to choose one of them.

It should be listed in separate columns for diastolic and systolic pressure.

ID	Pohlaví	Věk	Výška	Zařazen	Alergie	TKD/TKS
9	M	53	177	13.9.2001	N	80/120
14	M	41	167	10.9.2001	N	75/119
19	M	52	182	14.90.2001	N	91/145
22	M	26	193	17.9.2001	A	78/130
23	MM	53	neznámo	17.9.2001	N	80/120
29	M	23	197	4.10.2001	0	75/119
30	M	58	158	4.10.2001	N	91/145
32	Z	198	45	5.10.2001	N	78/130
33	Z	51	191	5.10.2001	1	80/120
34	M	44	169	5.10.2001	1	75/119
35	Z	22	0	5.10.2001	N	91/145
38	M	42	163	5.10.2001	A	78/130

Data storage policy

- Correct and clear data storage is the basis for later analysis
- It is advisable to think in advance how the data will be stored
- For computer data processing it is necessary to store data in tabular form
- The most suitable way is to store the data in the form of a database table
 - Each column contains only a single data type, identified by the column header
 - Each row contains a minimum unit of data (e.g. patient, one patient visit, etc.)
 - It is not allowed to combine numeric and text values in one column
 - Comments are stored in separate columns
 - For text data, it is necessary to check for typos in category names
 - A specific type of data is dates where it is necessary to check whether the dates are stored in the correct format
- Data arranged in this way can be converted to any output table in spreadsheet or database programs
- For basic storage and cleaning of smaller data it is possible to use MS Office applications

Data visualization

Types of graphic visualisation

Risks of desinterpretation of graphical data display

Annotation

- The first step in data analysis is data visualization.
- Different types of data allow us to get an idea of the distribution of data, the representation of categories and the relationships of variables to each other.
- Through visualization, we gain insight into the data and begin to form hypotheses about patterns among the variables in the data set being evaluated.

What to create charts in

- Different software - different options
 - MS Office - basic charts, easy editing, can be inventively modified, easy replicability by data exchange
 - R - various libraries (e.g. ggplot) - higher input investment, various types of graphs, automation
 - SPSS, Statistica - fast creation of a large number of graphs, many types of graphs
- Criteria
 - Selection of different chart types
 - Ease of editing and appearance modification
 - Easy replication/automation/quick charting

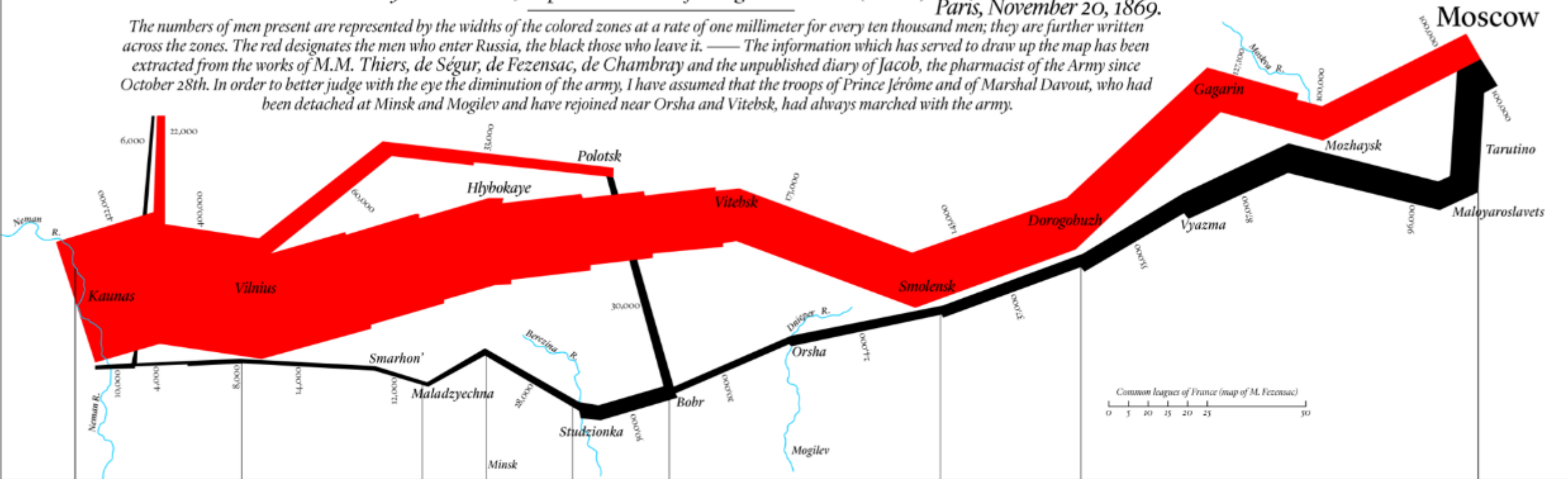
Charles Joseph Minard - Napoleon's campaign in Russia

Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812 ~ 1813

Drawn by M. Minard, Inspector General of Bridges and Roads (retired).

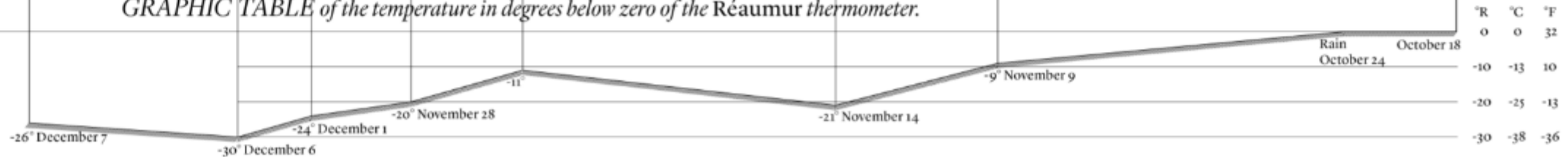
Paris, November 20, 1869.

The numbers of men present are represented by the widths of the colored zones at a rate of one millimeter for every ten thousand men; they are further written across the zones. The red designates the men who enter Russia, the black those who leave it. — The information which has served to draw up the map has been extracted from the works of M.M. Thiers, de Ségur, de Fezensac, de Chambray and the unpublished diary of Jacob, the pharmacist of the Army since October 28th. In order to better judge with the eye the diminution of the army, I have assumed that the troops of Prince Jérôme and of Marshal Davout, who had been detached at Minsk and Mogilev and have rejoined near Orsha and Vitebsk, had always marched with the army.



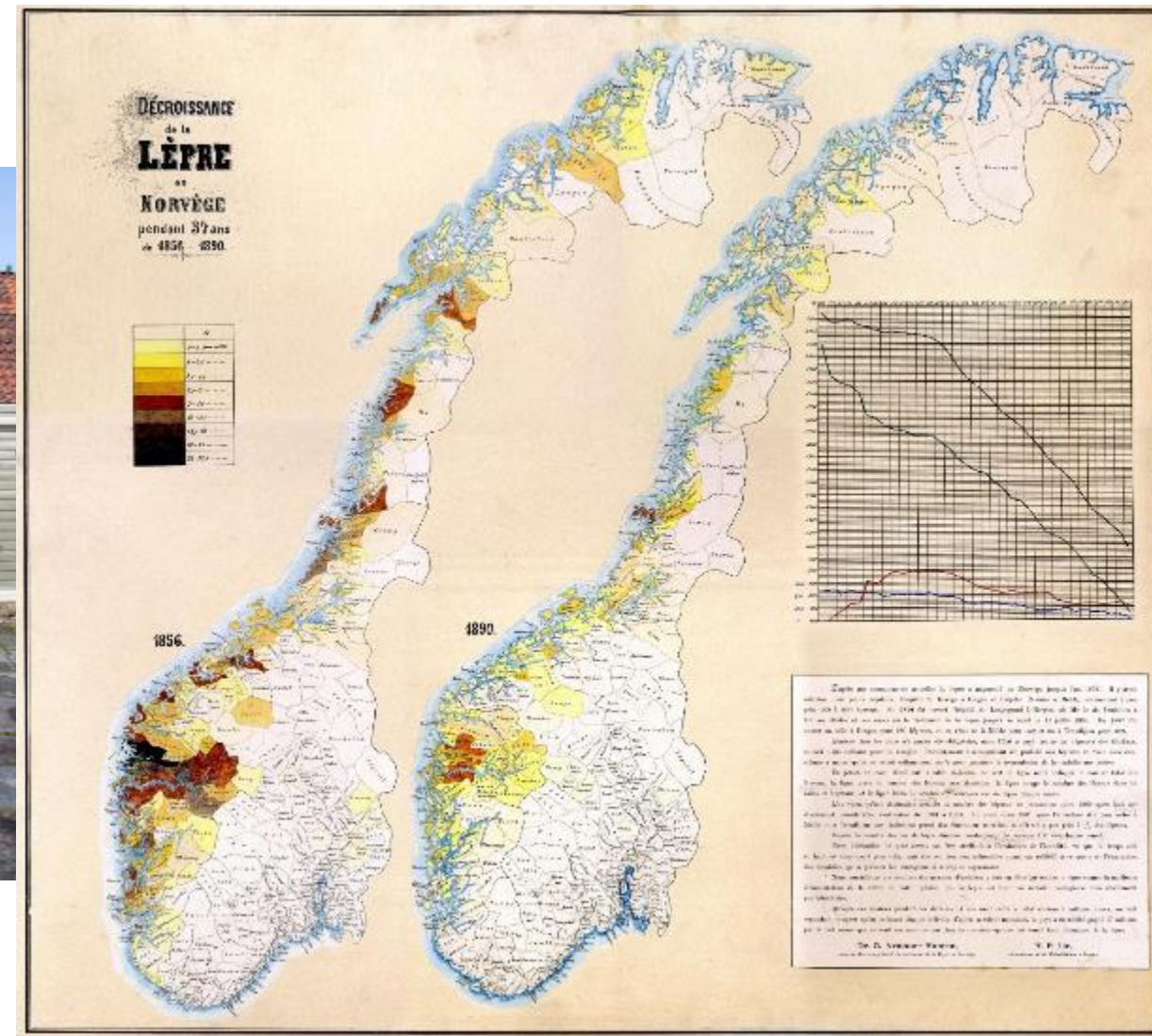
GRAPHIC TABLE of the temperature in degrees below zero of the Réaumur thermometer.

The Cossacks pass the frozen Neman at a gallop.



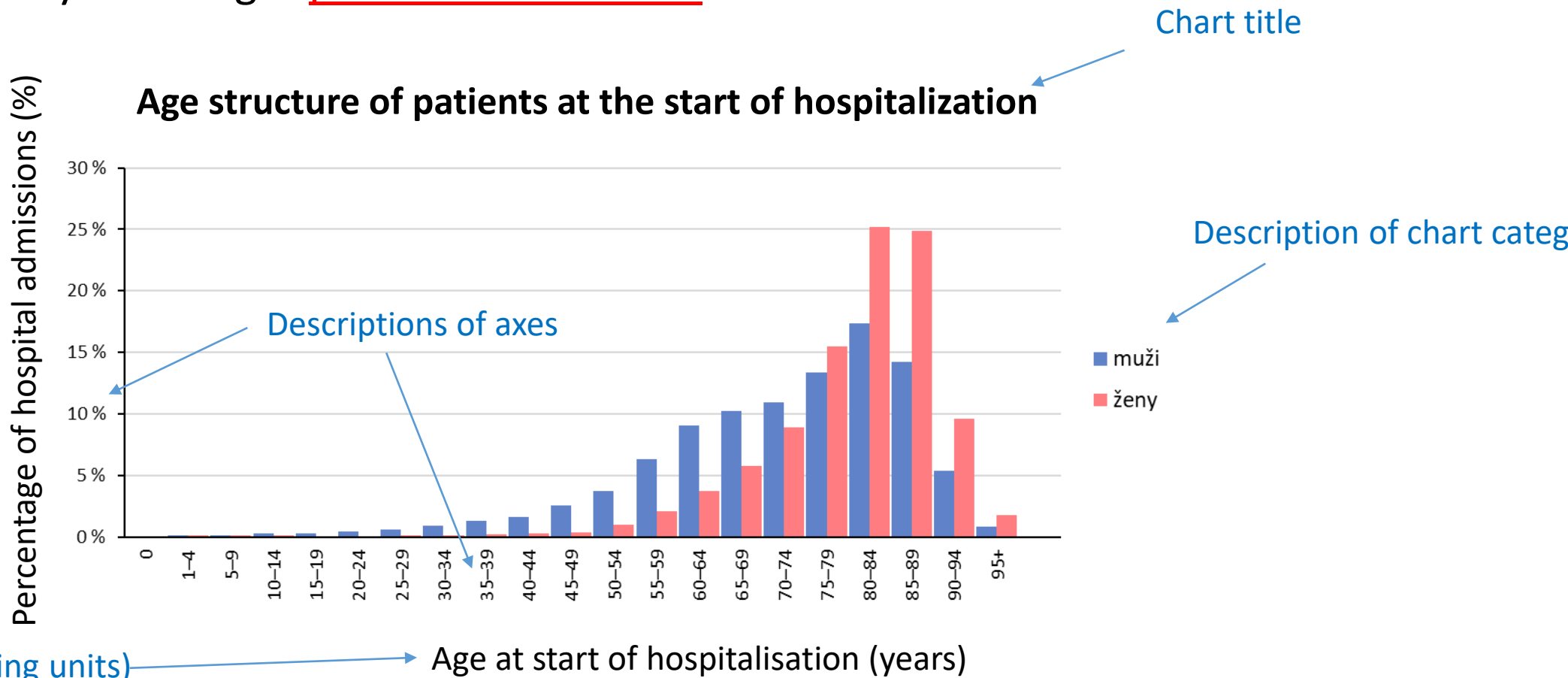
Famous charts: leprosy eradication in Norway

- 1856 - national leprosy registry in Norway established in Bergen -> analysis of the data -> measures to eradicate leprosy in Norway
- Gerhard Armauer Hansen



What not to miss on the chart

- Each graph must be uniquely described - self explained
- A graph that says nothing is pointless to draw !!!

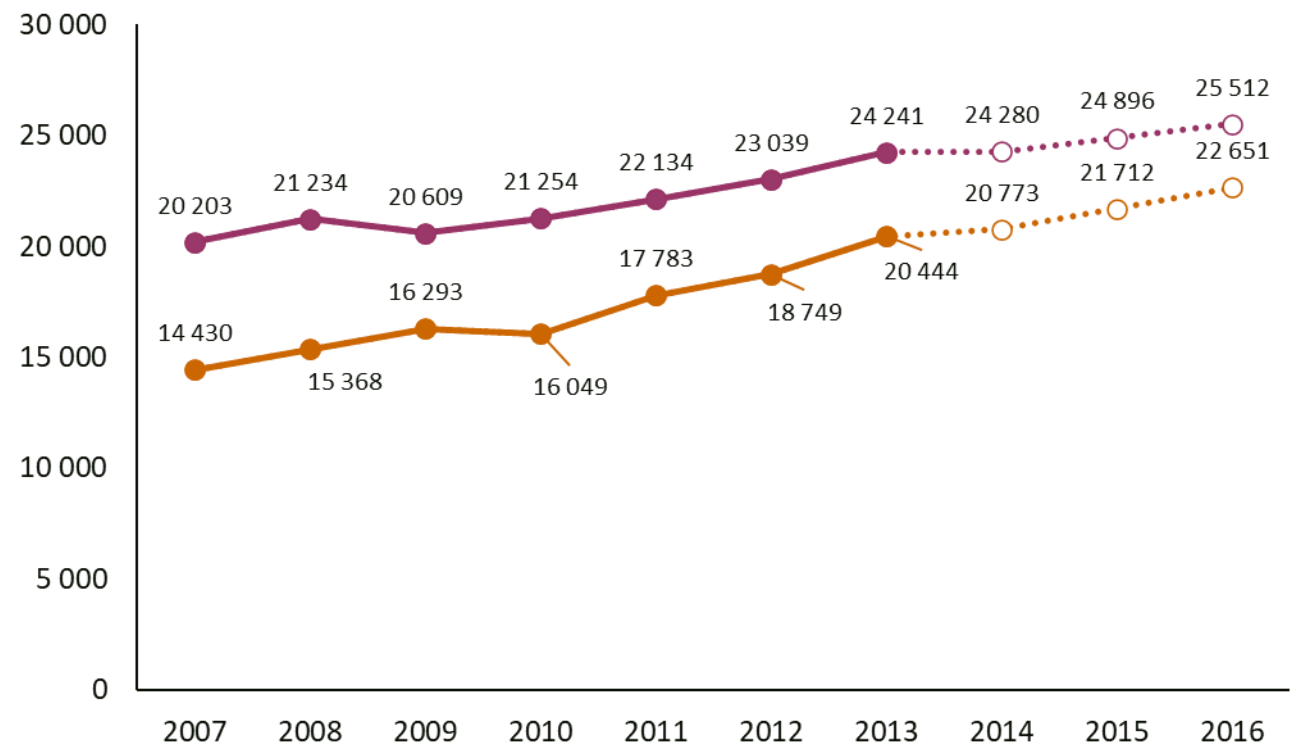
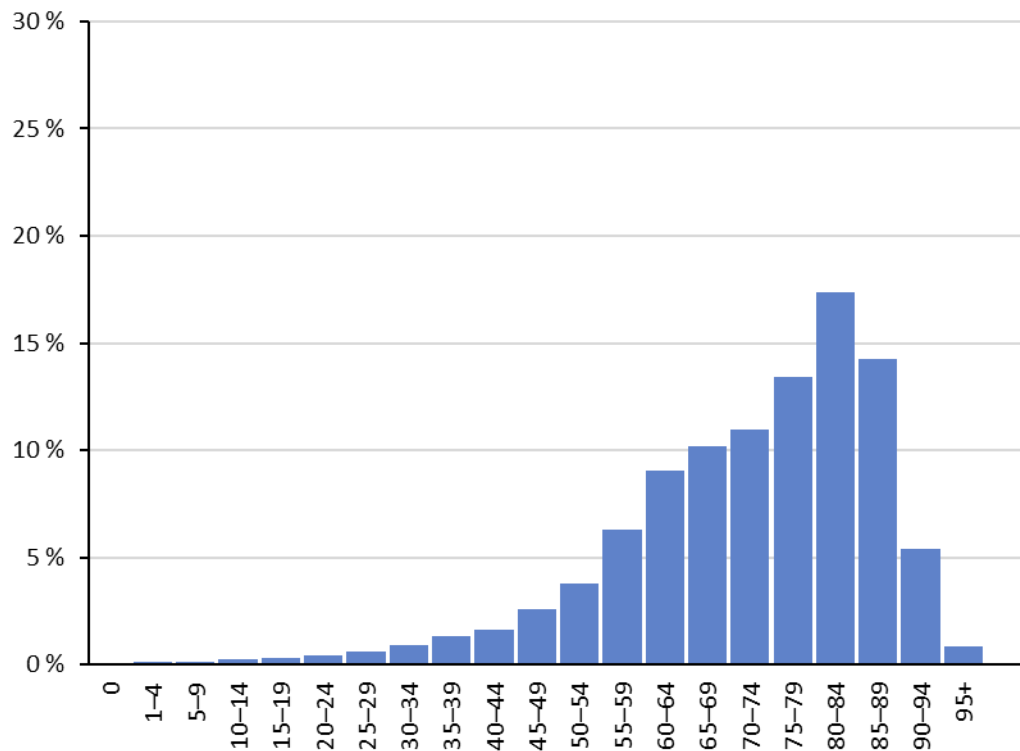


Axis headings (including units)

Age at start of hospitalisation (years)

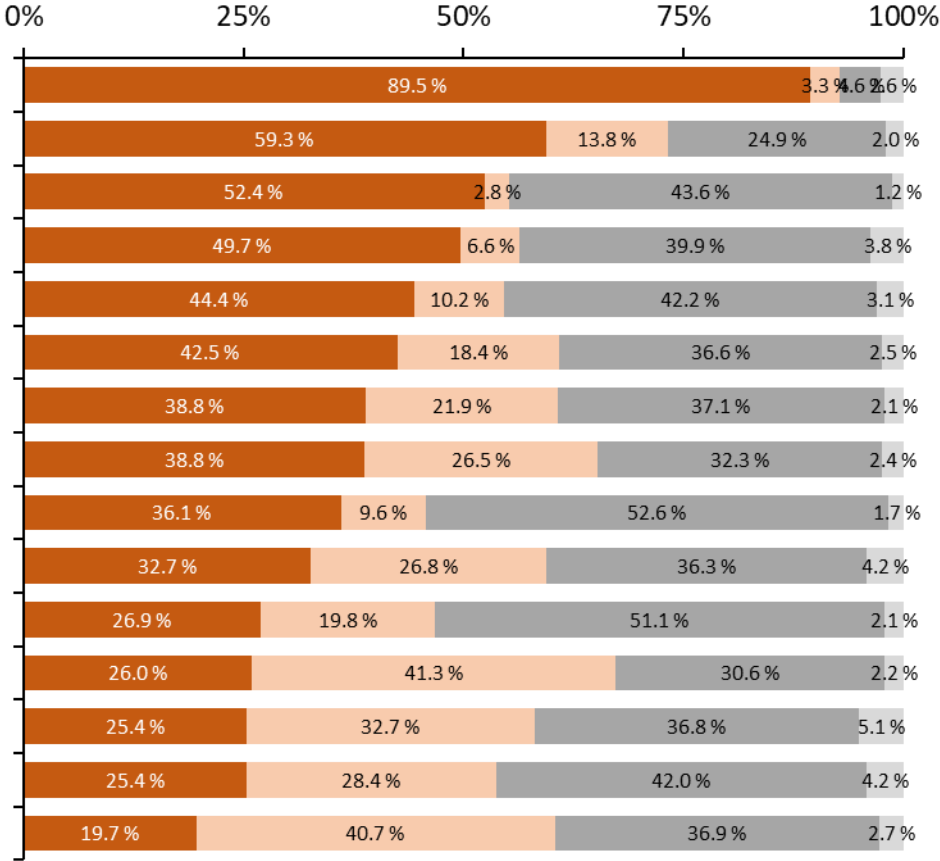
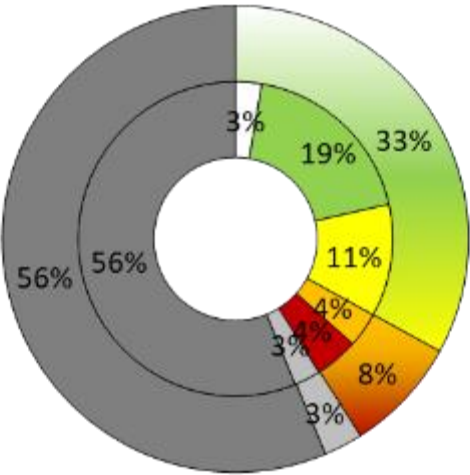
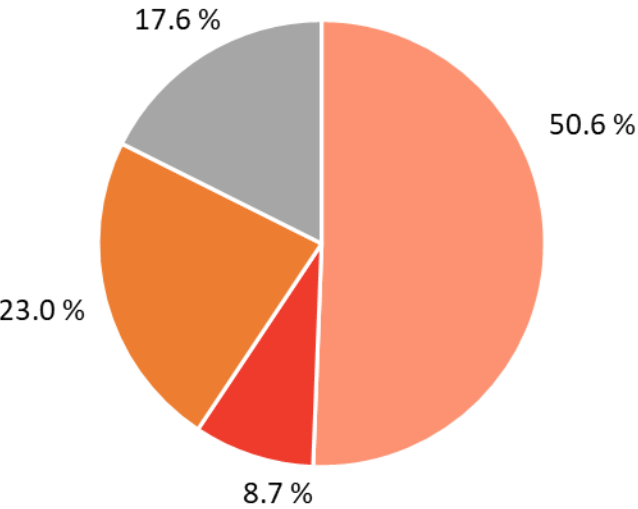
Bar and line graphs

- Easy creation, visualization of absolute values or percentages



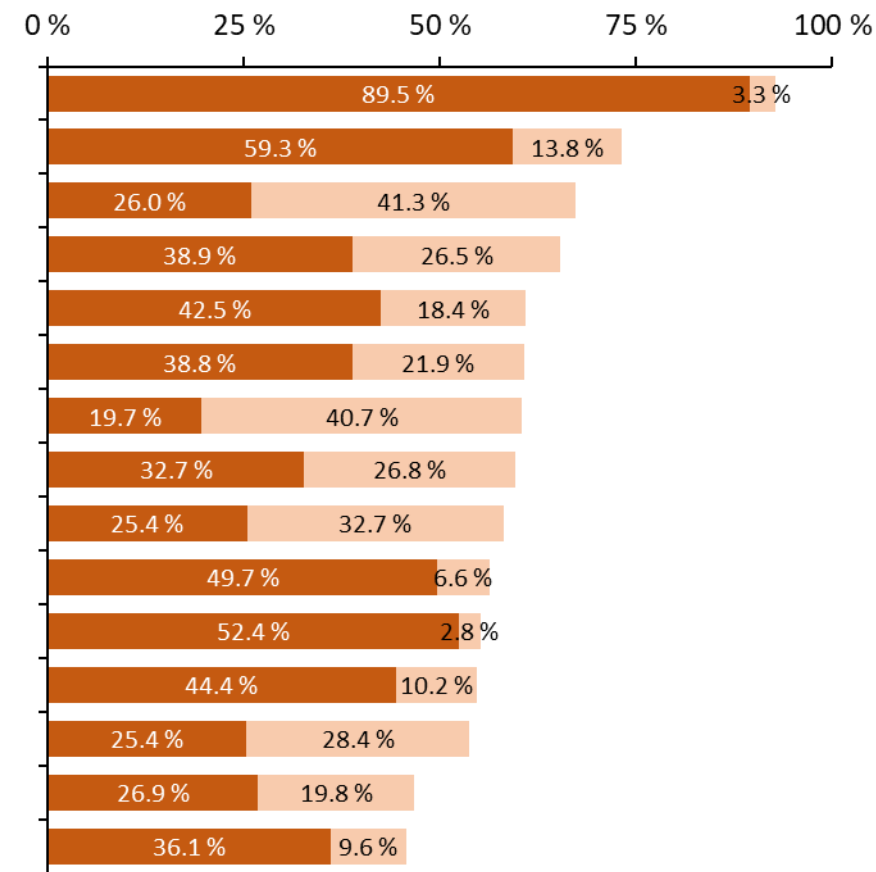
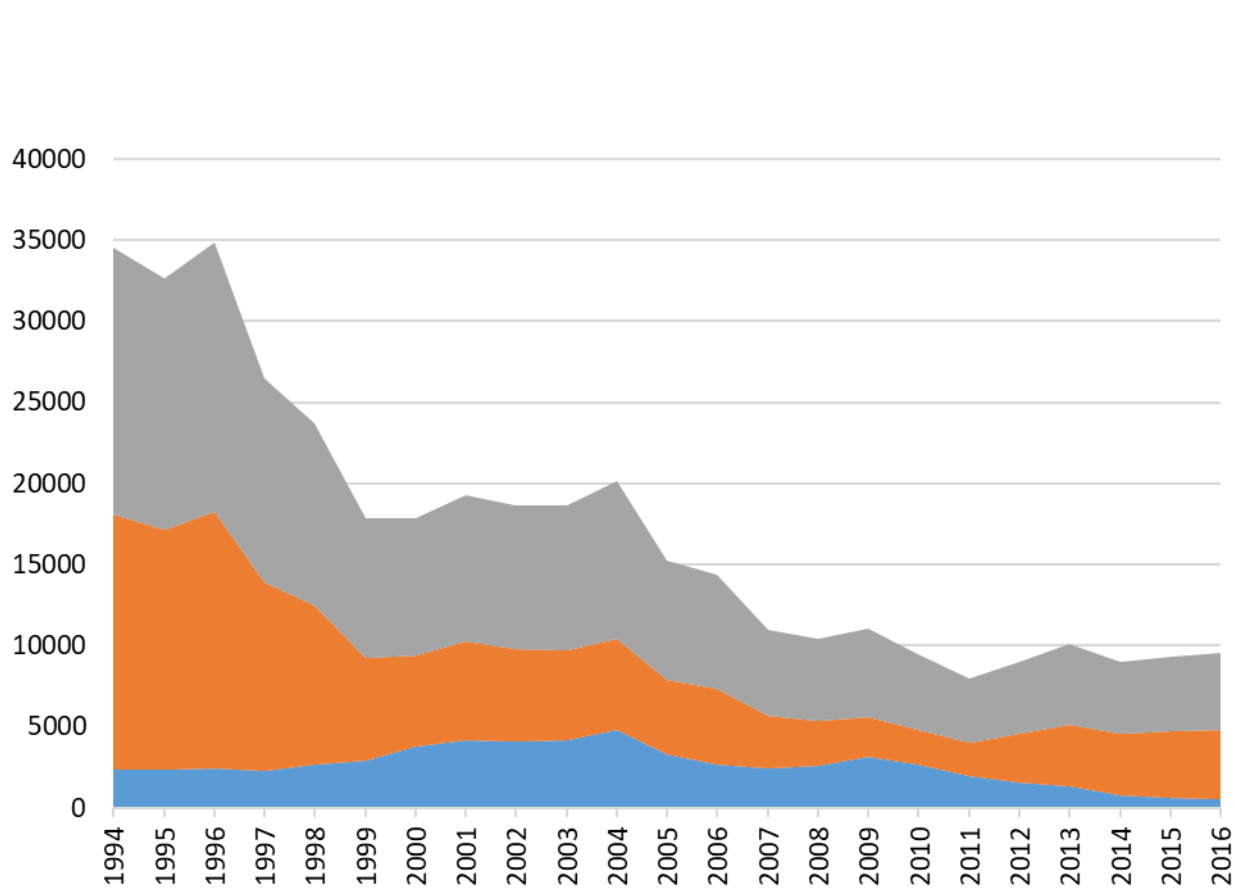
Pie and bar charts

- Simple creation, visualization percentages



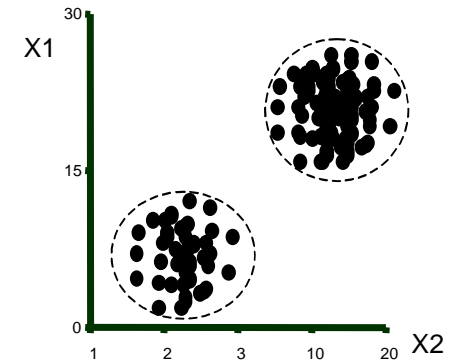
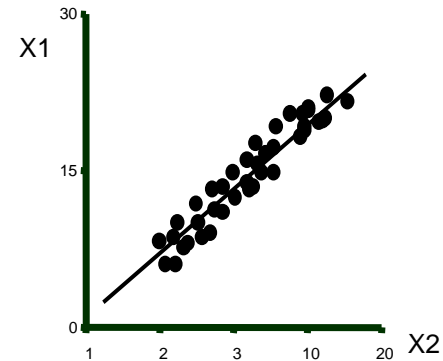
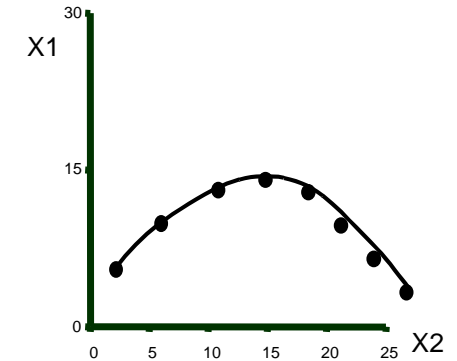
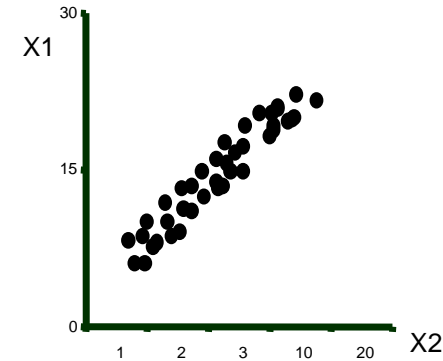
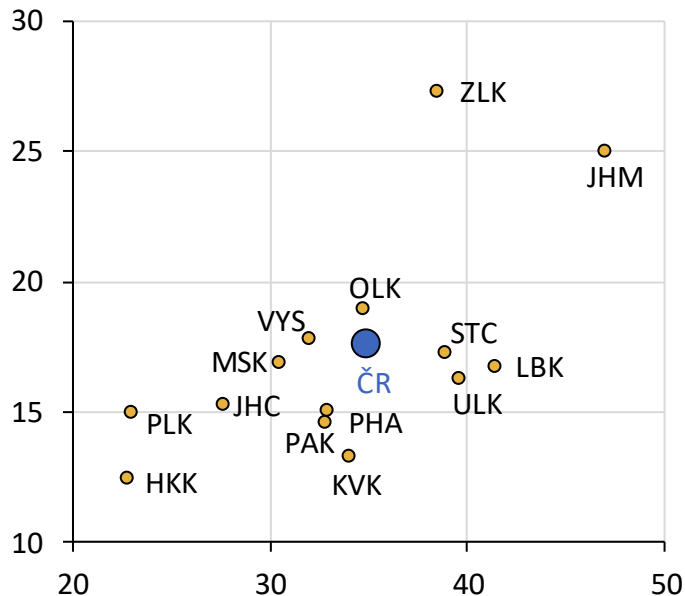
Folded charts

- Cumulative view more information



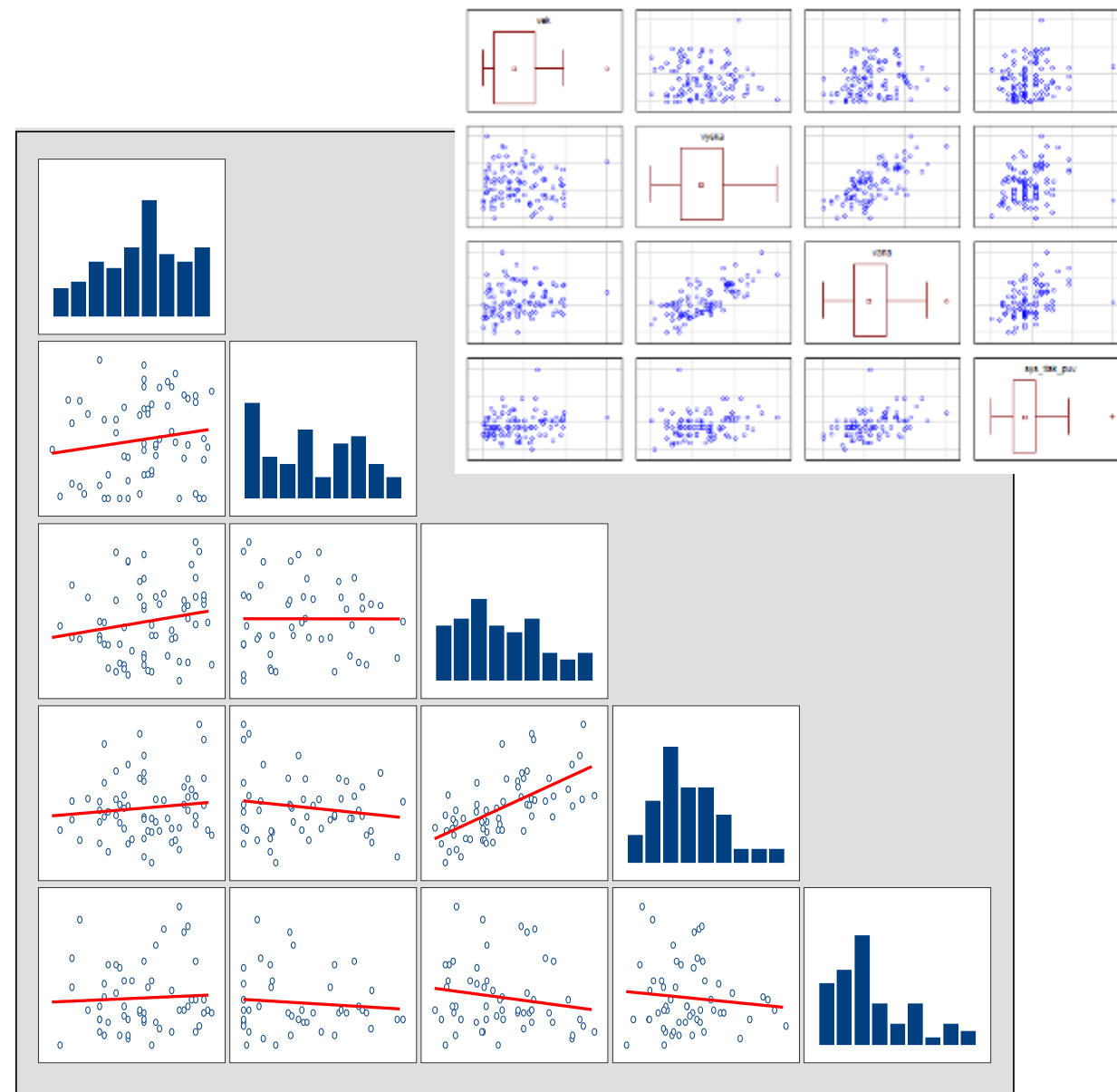
XY chart (scatter plot)

- Description of the relationship between two continuous variables
- Option to categorize and describe points
- Interleaving models into graphs
- Basic chart for viewing data before correlation and regression analysis



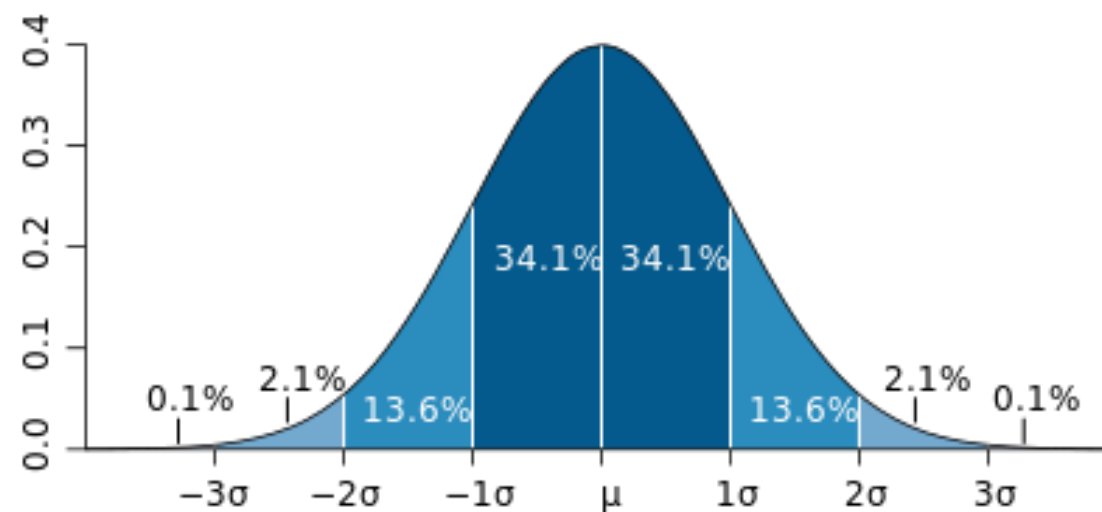
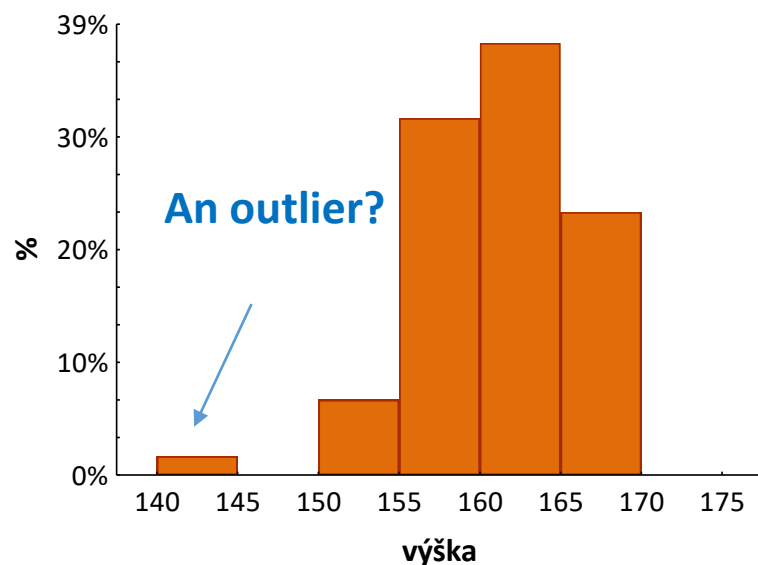
Matrix chart

- Extension of xy charts in statistical software
- Simultaneous visualization of the distribution of values (diagonal) and interrelationships of a large number of continuous variables
- Different variants
 - Set of variables each with each
 - Two sets of variables against each other
 - Addition of calculation of correlation coefficients
- Basic visualization tool before multivariate analysis



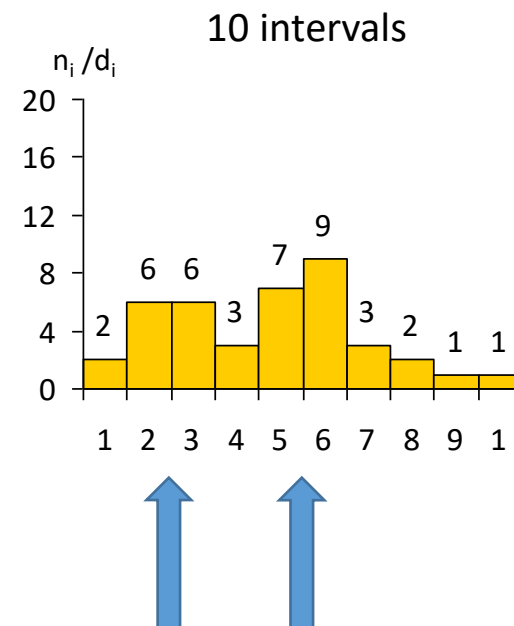
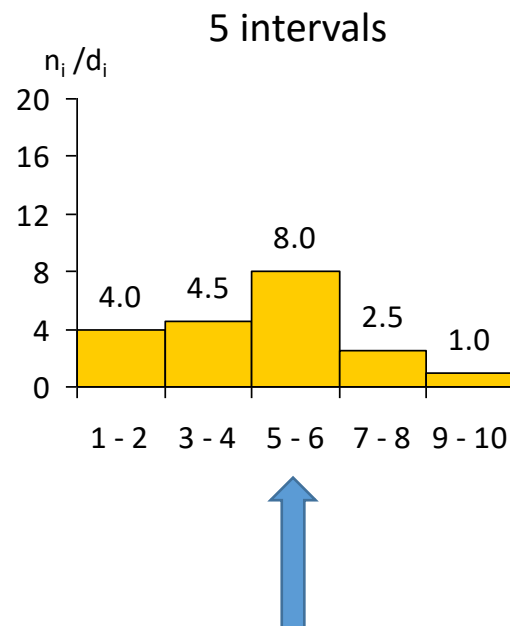
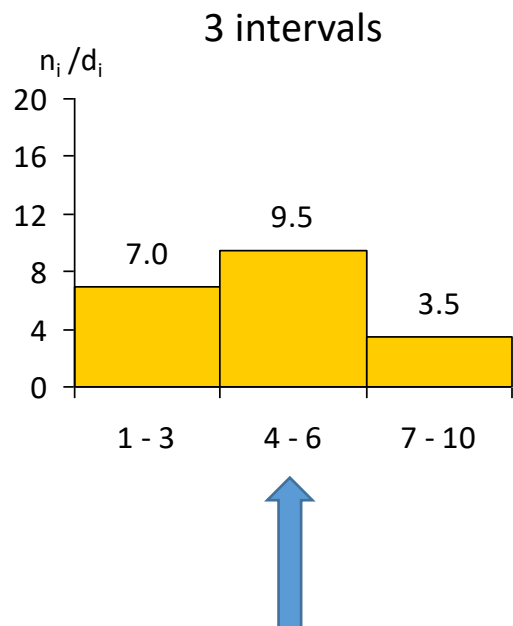
Histogram

- Graph summarizing the distribution of values of continuous variables, closely related to the theory of statistical distributions
- In classic form similar (but not identical) to a bar chart
- In practice, a bar chart is often hidden under the name histogram (acceptable if it does not lead to misinterpretation of the data)
- One of the basic charts for assessing the distribution of data



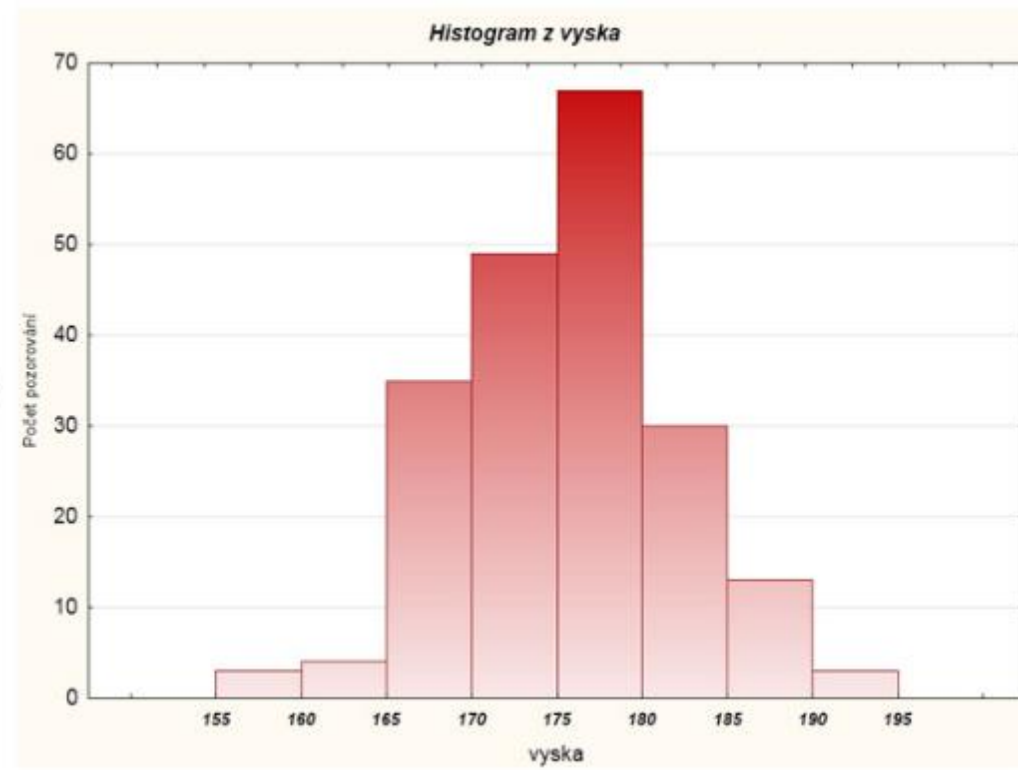
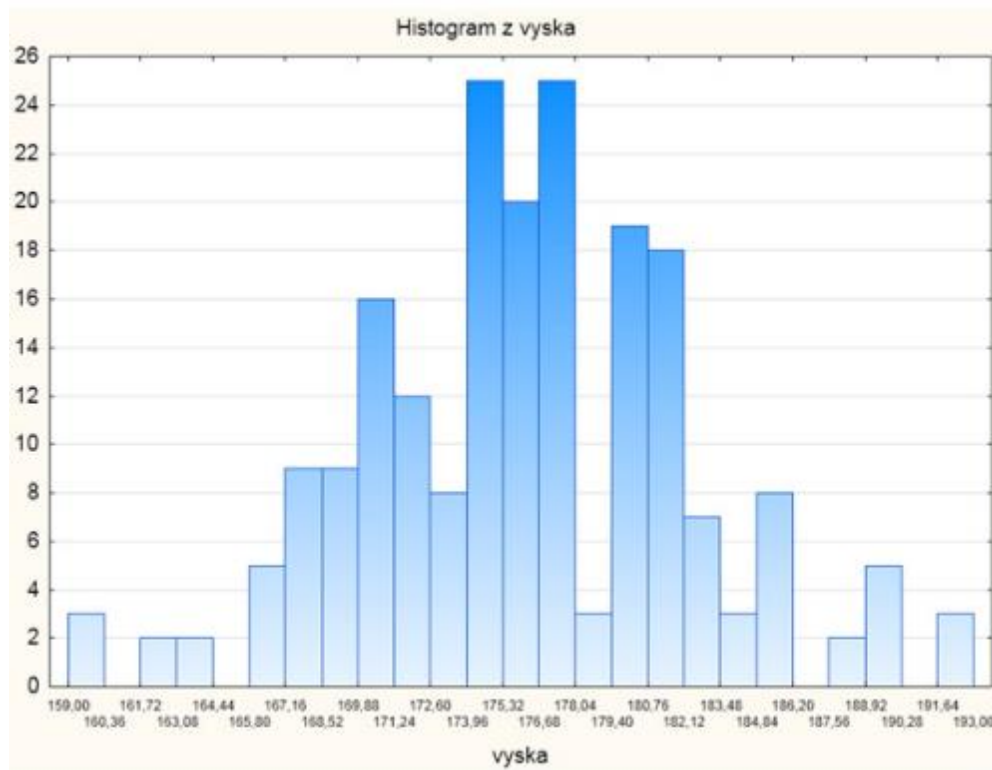
Histogram: effect of data categorization

- The number of selected intervals in the histogram determines how it will look. If the number is small we may miss important elements in the data, if it is large the information may be fragmented.



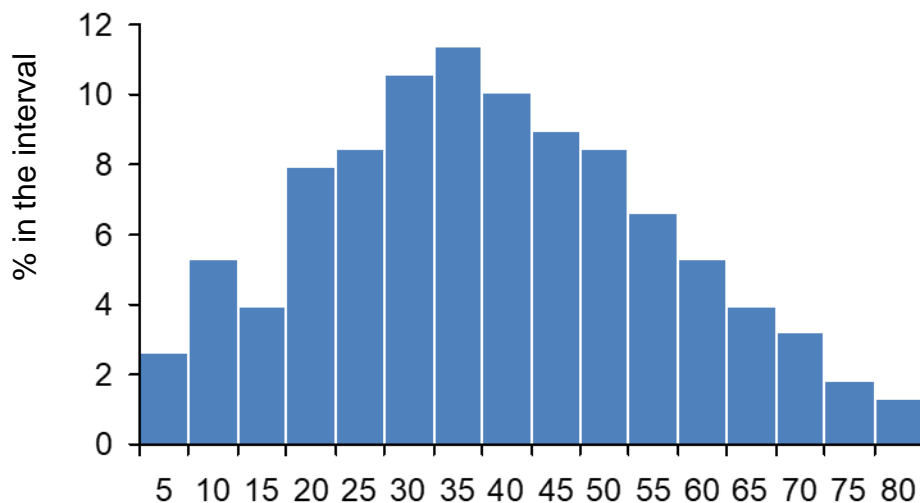
Histogram: effect of data categorization

- Choosing the number of categories - important for interpretation
- Manual or automatic selection - different algorithms (depends on sample size and data variability)

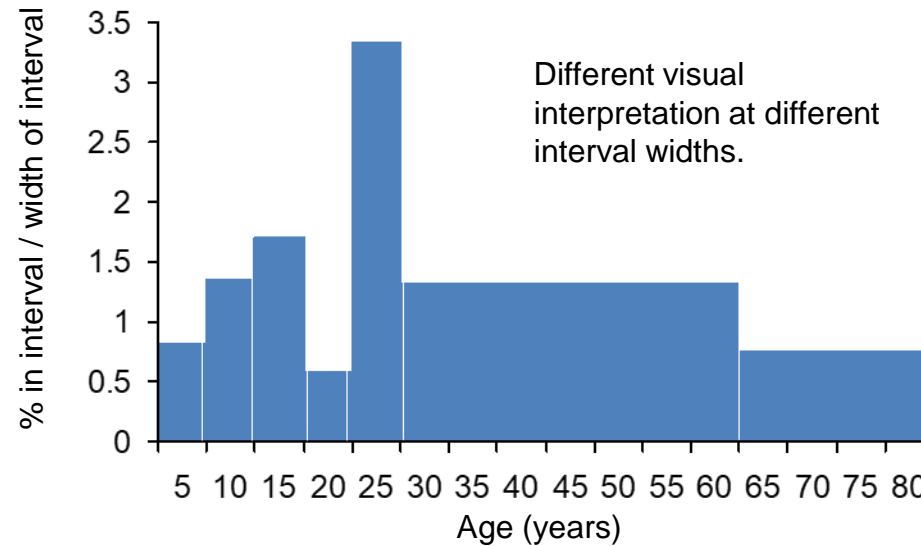
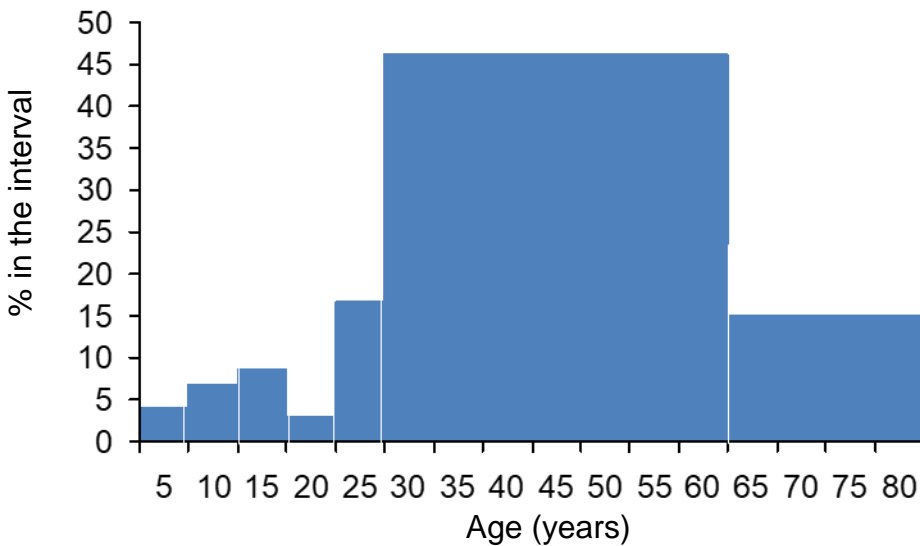
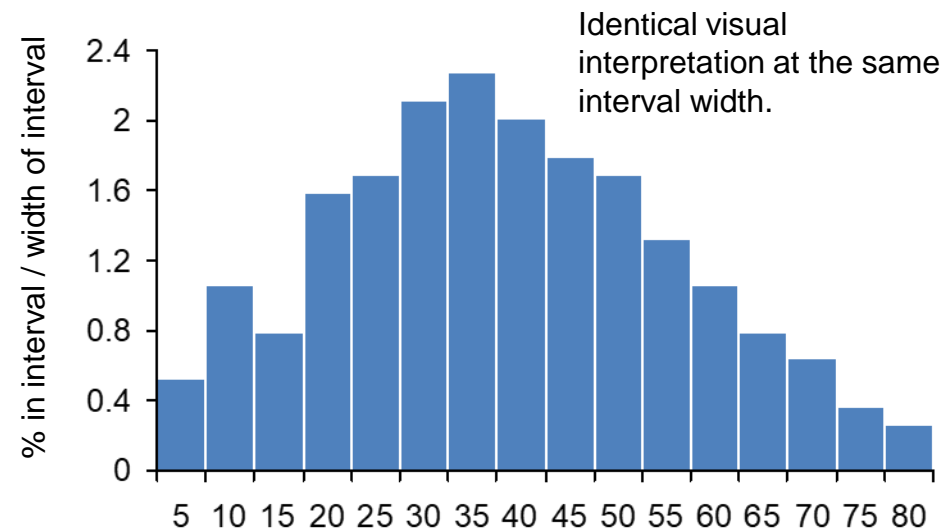


Histogram and bar chart

Bar chart

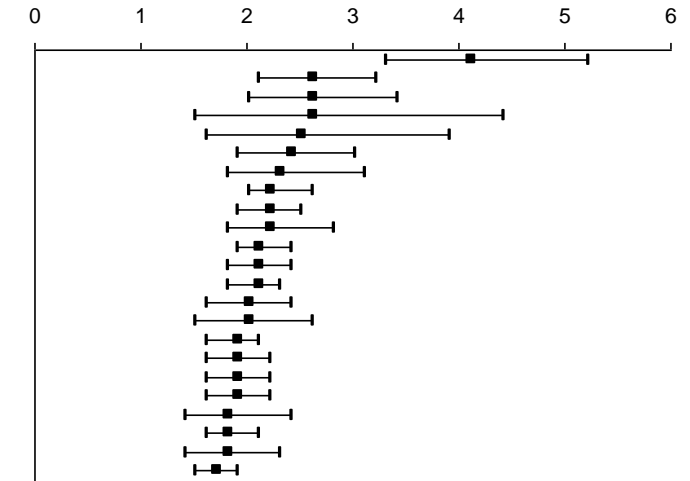
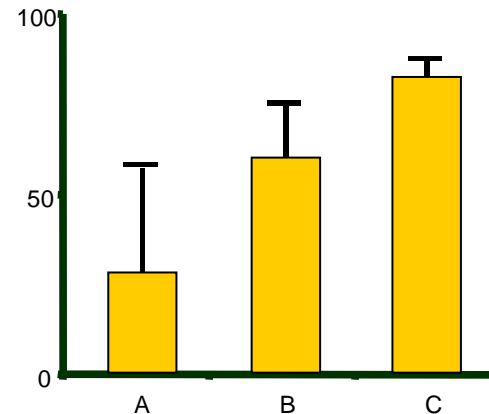
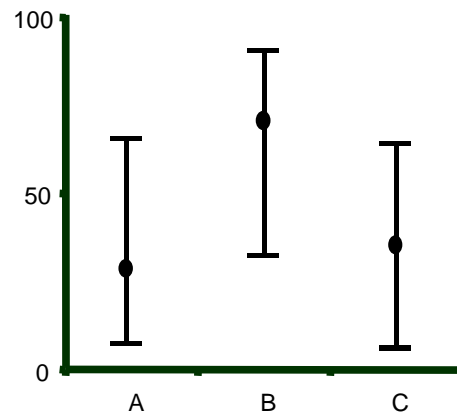
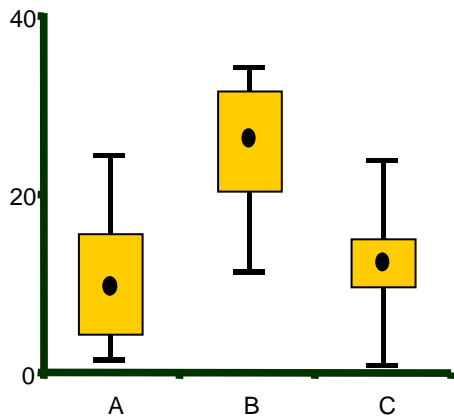


Histogram

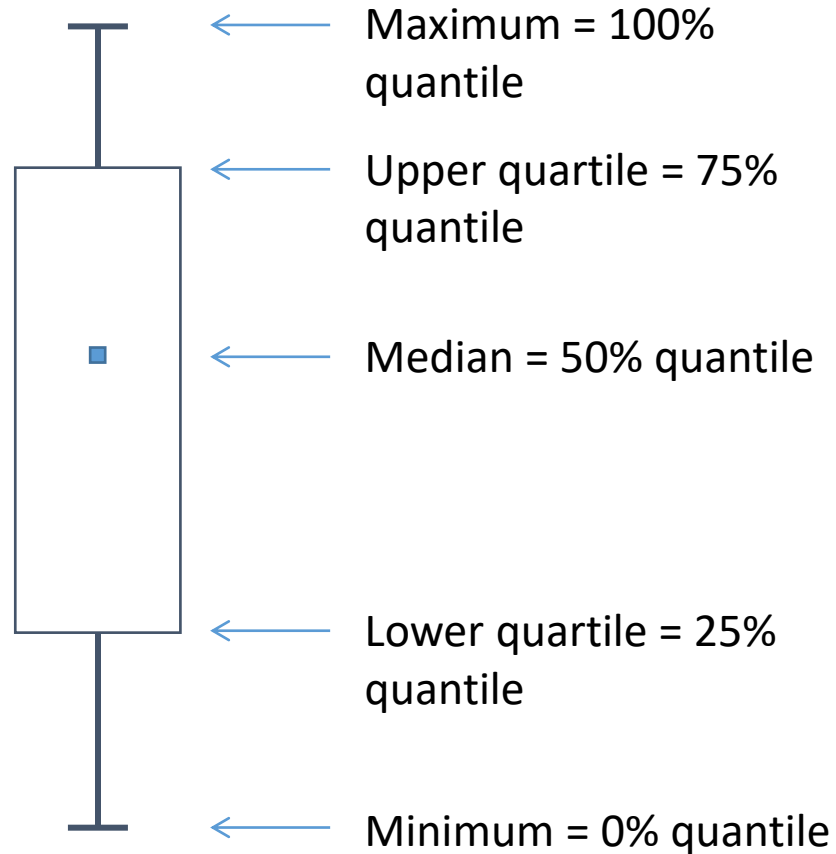


Box and whisker plot: what is it?

- A popular chart type in data analysis that allows simple comparison of multiple groups of objects and evaluation of data distribution
- Most common for describing continuous data, but usable for any type of data that can be described by mean and variability (percentages, regression coefficients, odds ratios, risk ratios, hazard ratios, etc.)
- Huge number of variants



Box and whisker plot: an example of one possible variant

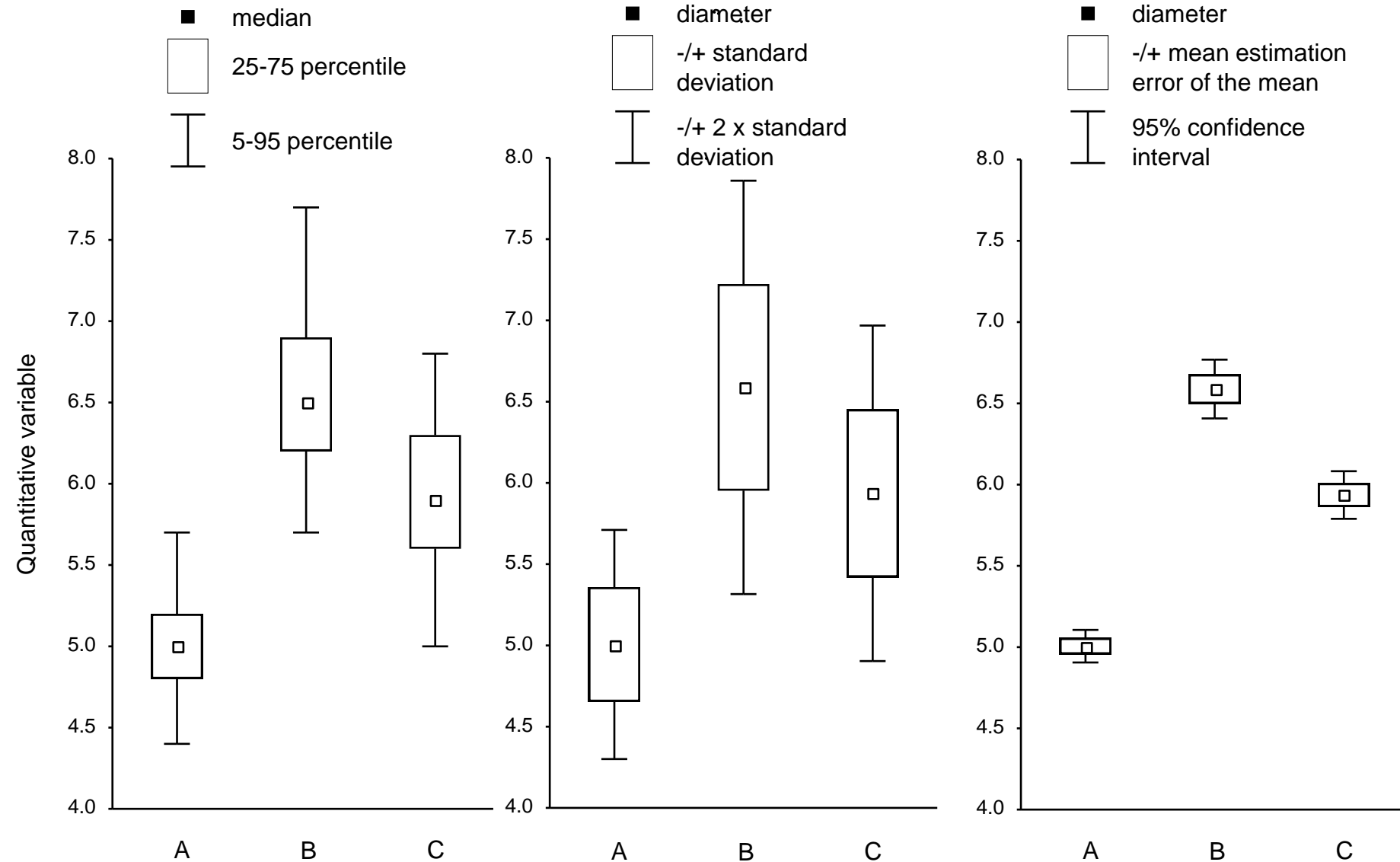


The individual points of the graphs can contain any descriptive statistics - means, standard deviations, confidence intervals, odds ratios, hazard ratios, etc.

The number of data points in the graph can be from three to e.g. nine.

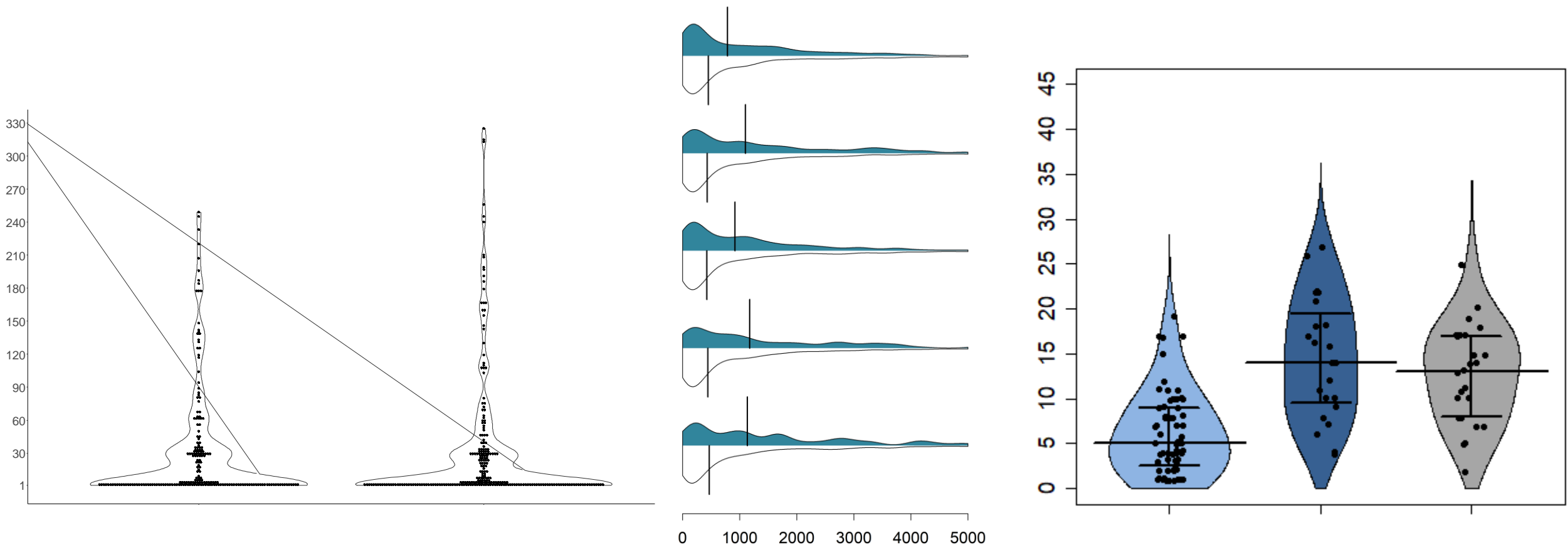
Box and whisker plot and its different variants I

- It is essential to read the labels
- Different variations of the chart may have completely different interpretations



Box and whisker chart and its different variants II: Violin plot and Beanplot

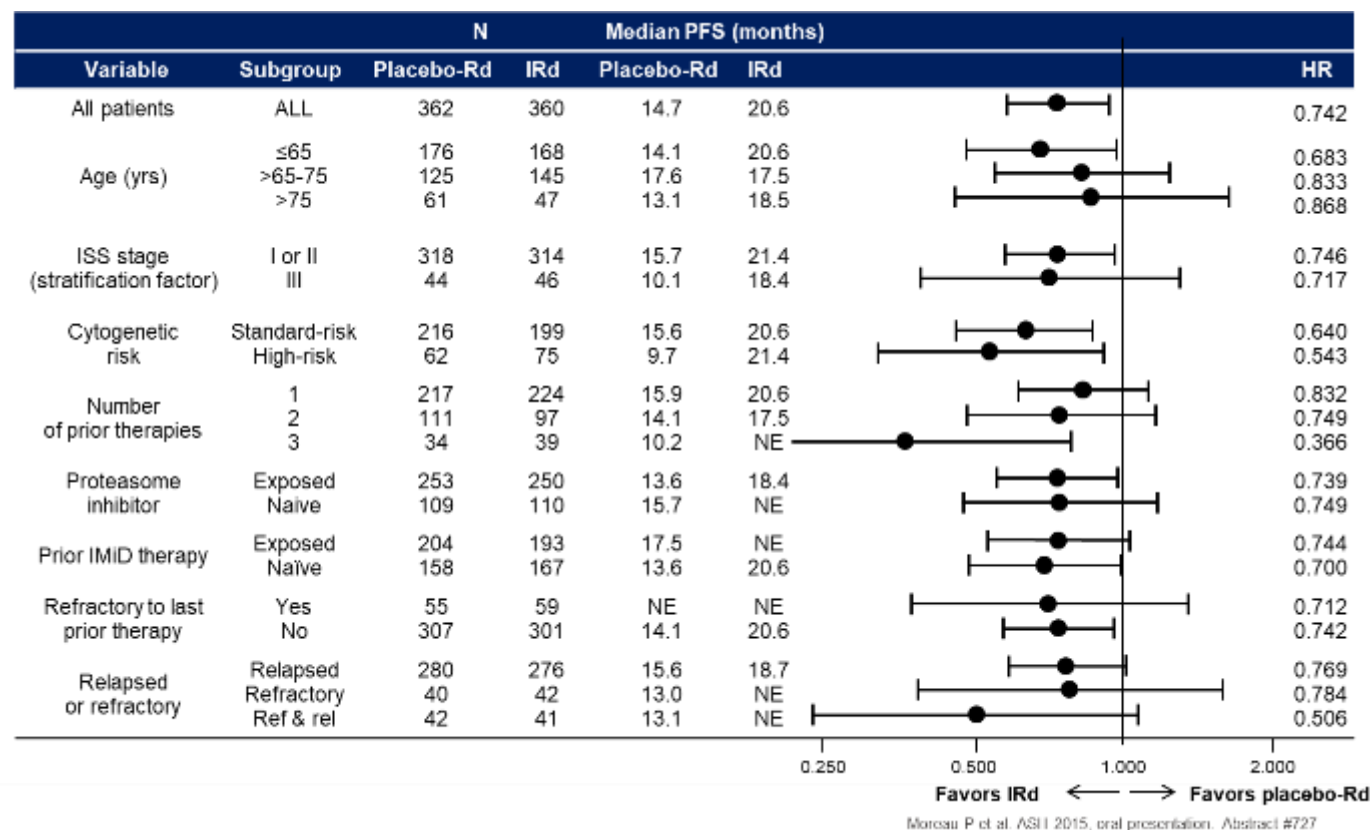
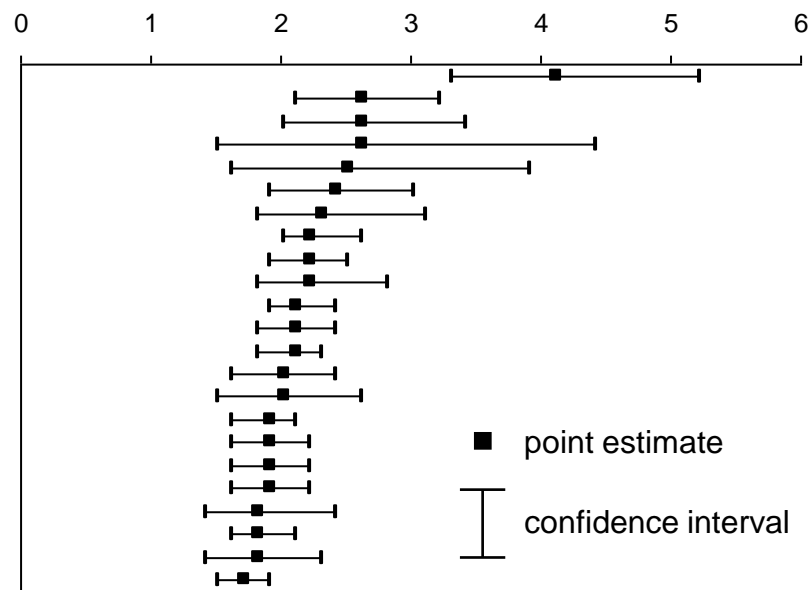
- Combination of histogram and box plot or dot plot
- Available in R - e.g. beanplot and ggplot2 libraries



Box and whisker plot and its different variants III: Forest plot

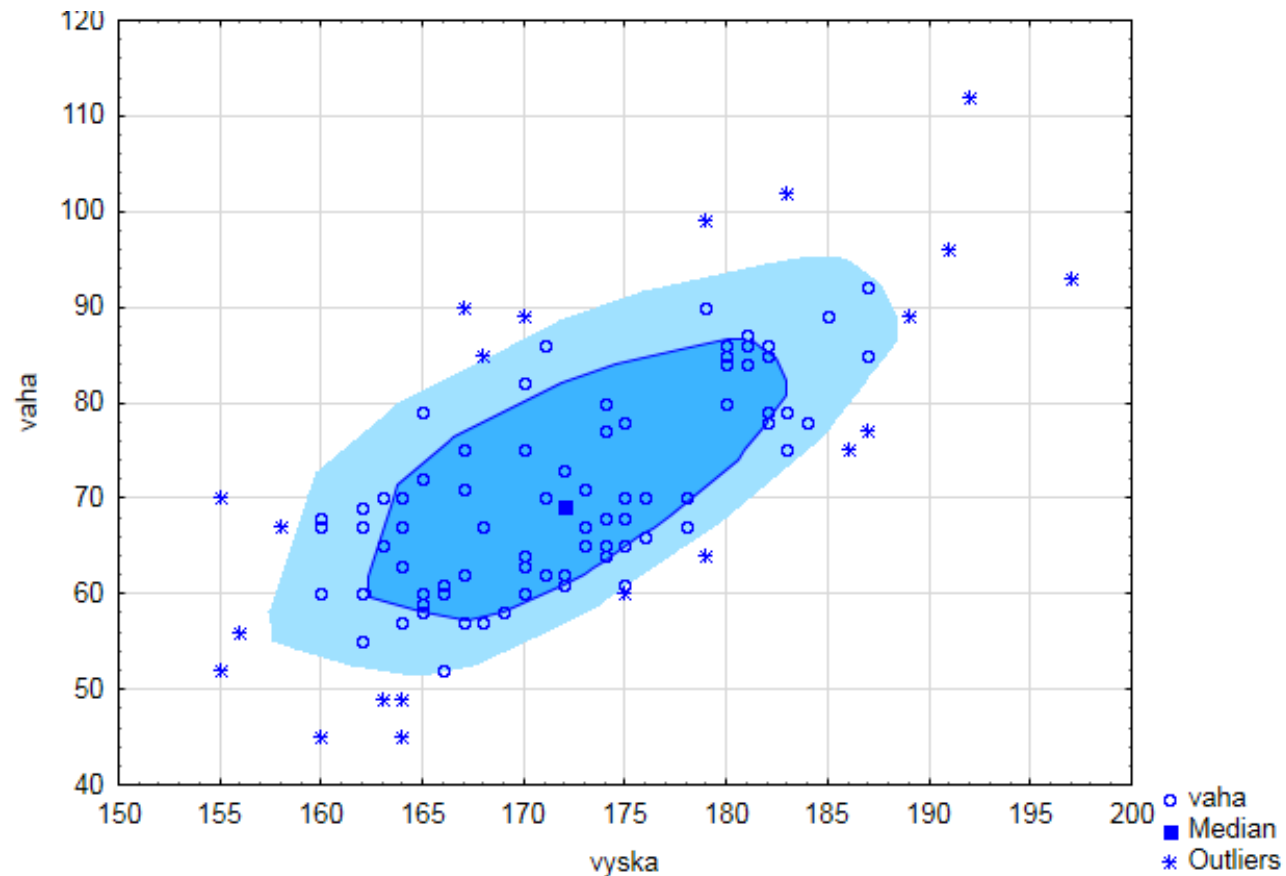
- Variant of box and whisker fence
- Often used to display regression coefficients or odds/risk/odds ratios

Characteristic assessed (mean, proportion, odds ratio, relative risk, risk ratio)



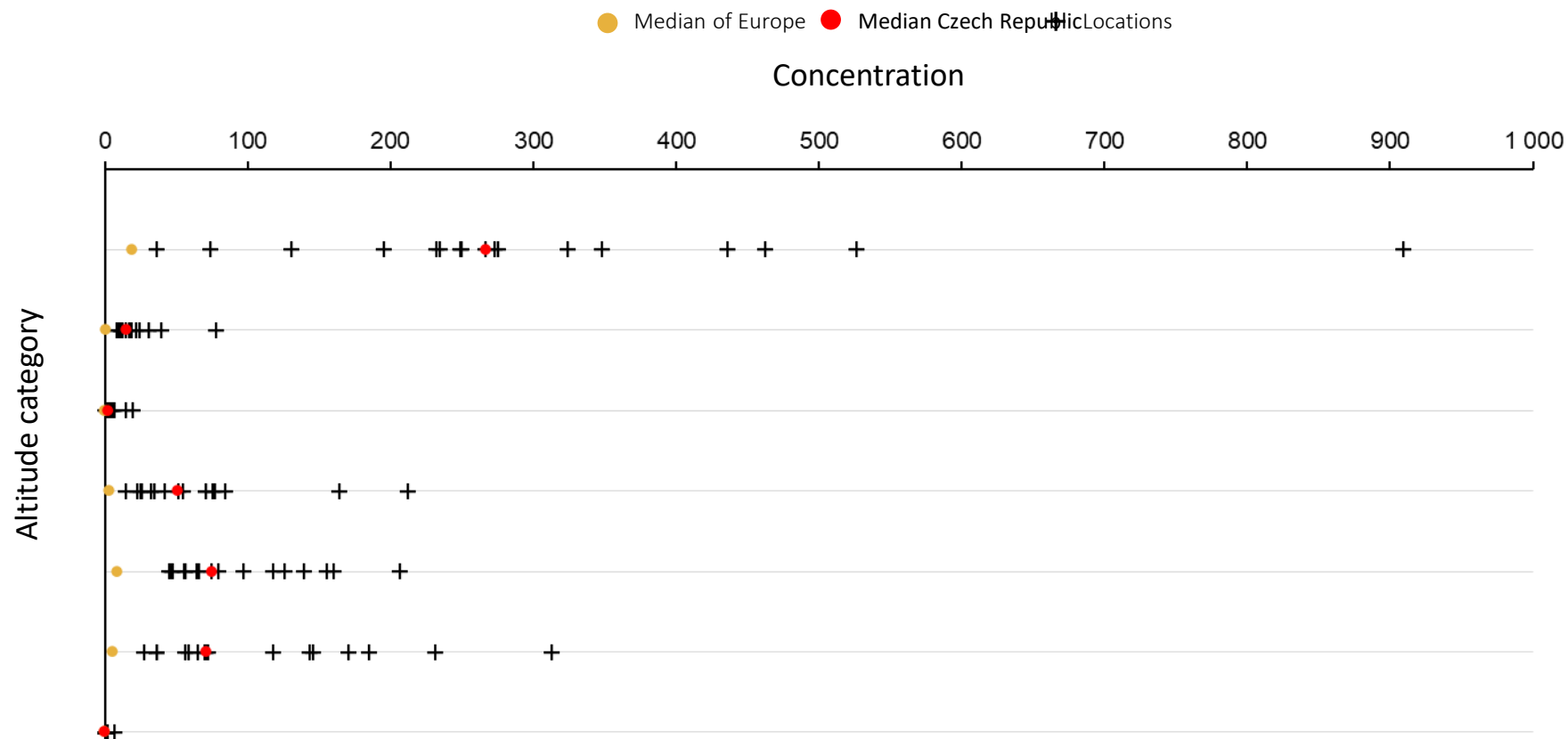
Box and whisker chart and its different variants IV: Bagplot

- Bagplot = "bivariate boxplot" (i.e. "two-dimensional box plot")



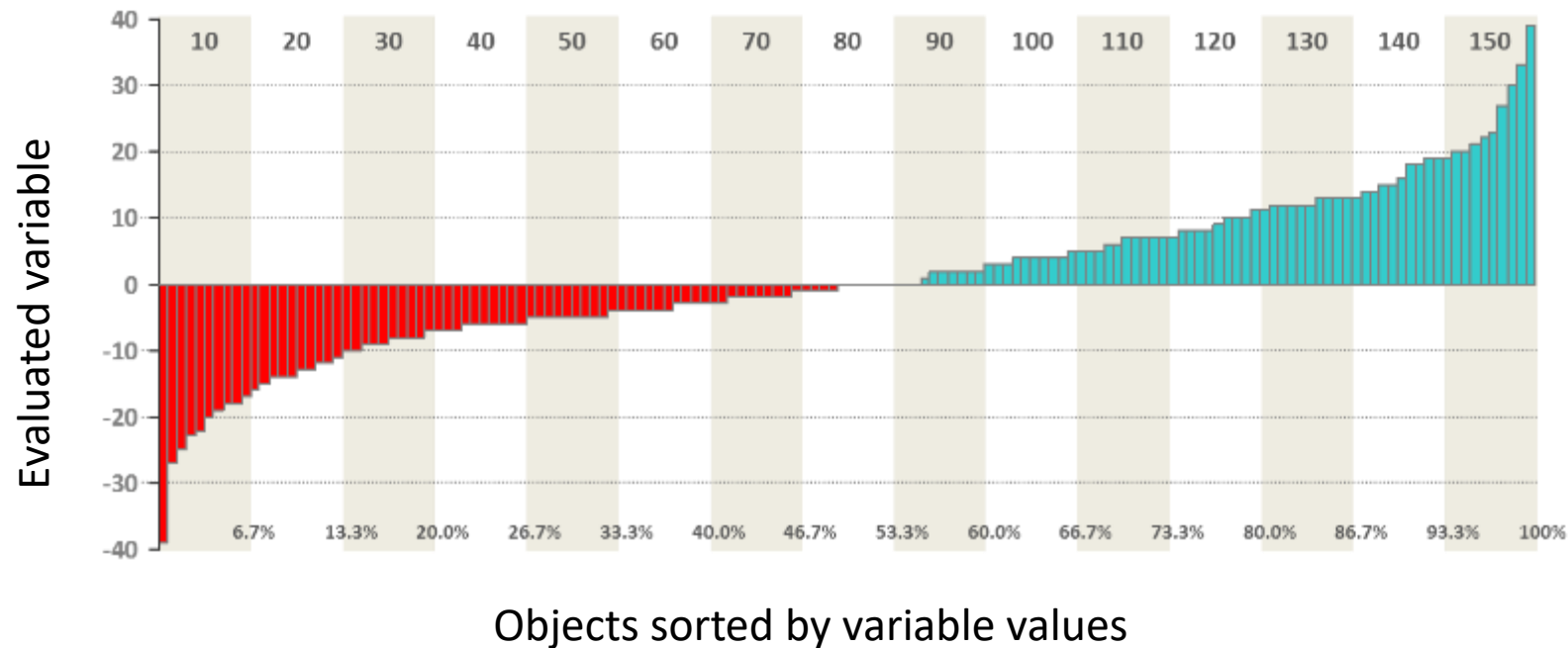
Inventive use of simple graphs: the Bead Graph

- Can be created from XY chart in MS Office
- Large amount of information in a small area



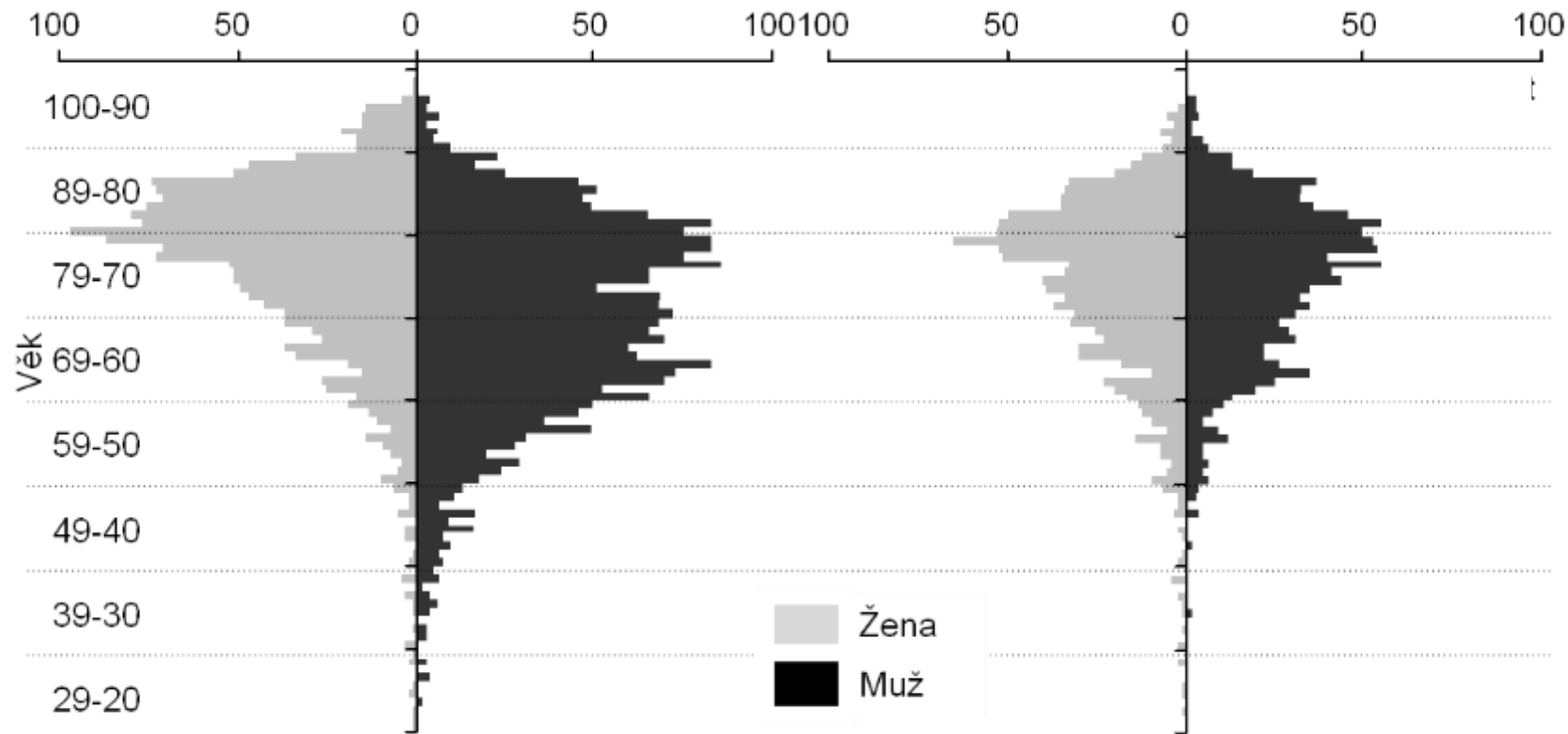
Inventive use of simple charts: waterfall plot

- Visualisation of individual object results, often for variables describing changes
- The values are sorted by size in the chart
- It can be supplemented with values of standards, percentages of objects in categories of the standard, etc.



Inventive use of simple graphs: the Demographic Pyramid

- Simple linear bar chart
- Attractive visualization for comparison of two groups of objects



Excel - conditional formatting as charts

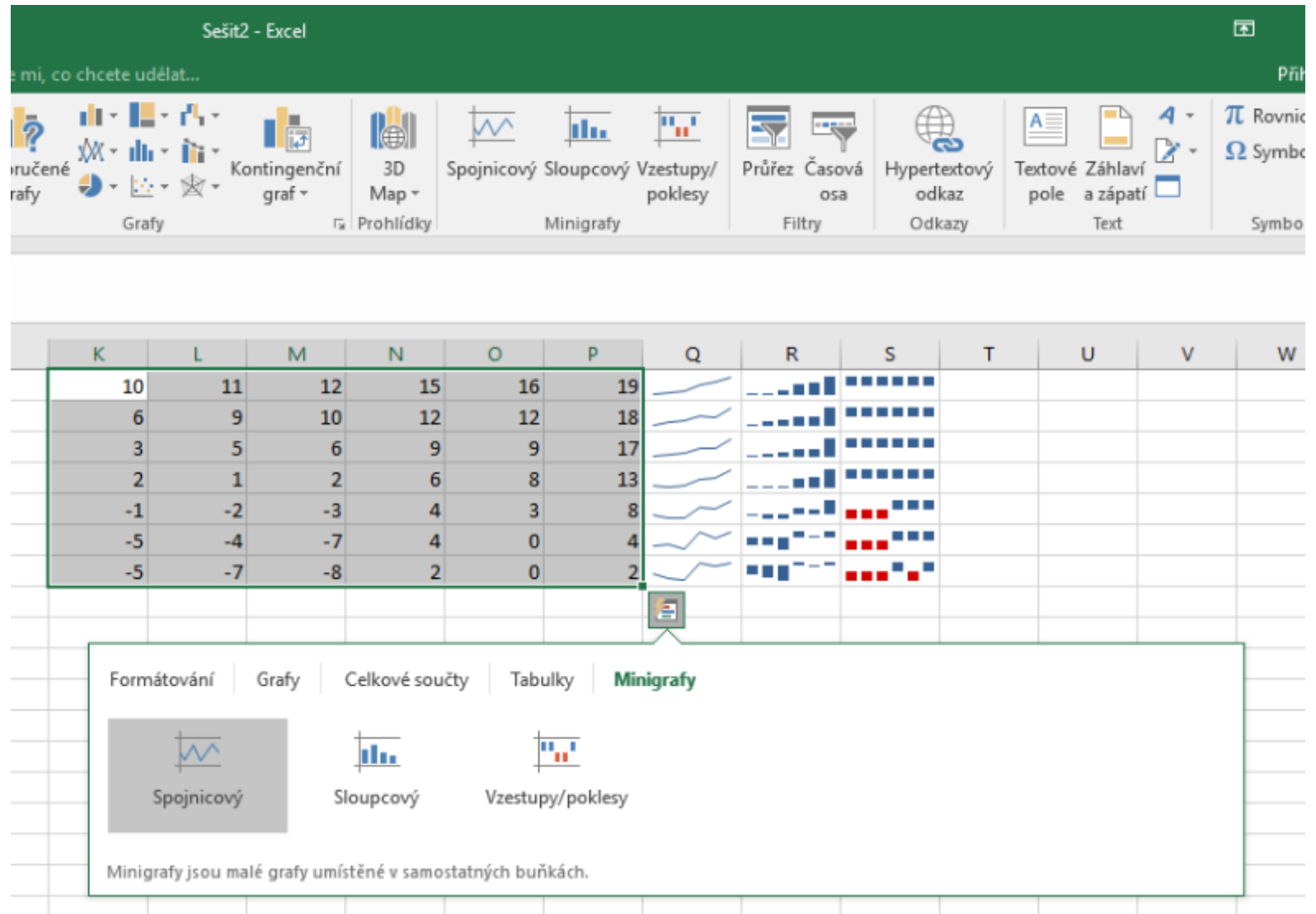
- To make excel tables clearer it is possible to use graphic elements in its cells
- Data bars and colour ranges

The image consists of three panels illustrating Excel's conditional formatting features. The left panel shows the 'Podmíněné formátování' (Conditional Formatting) menu with options like 'Pravidla zvýraznění buněk' (Cell Rules), 'Pravidla pro nejvyšší či nejnižší hodnoty' (Top or Bottom Rules), 'Datové pruhy' (Data Bars), 'Barevné škály' (Color Scales), and 'Sady ikon' (Icon Sets). The middle panel shows a data table with columns M through U and rows 1 through 6. The table is formatted with color scales: column M has blue data bars, column N has a red-to-yellow color scale, column O has a yellow-to-green color scale, column P has a green-to-blue color scale, column Q has a blue-to-red color scale, column R has a red-to-yellow color scale, column S has a yellow-to-green color scale, column T has a green-to-blue color scale, and column U has a blue-to-red color scale. The right panel shows the 'Podmíněné formátování' menu with the 'Barevné škály' (Color Scales) option selected, showing various color scale options.

	M	N	O	P	Q	R	S	T	U
1	10		1	2	3	4	5	6	
2	15		2	3	4	5	6	7	
3	1		3	4	5	6	7	8	
4	5		4	5	6	7	8	9	
5	6		5	6	7	8	9	10	
6	7		6	7	8	9	10	11	
7	1								
8	22								

Excel - charts in cells

- To make excel tables clearer it is possible to use graphic elements in its cells
- Several types of graphs allowing to visualize in one cell data series
- Basic axis and appearance editing options



Heatmap

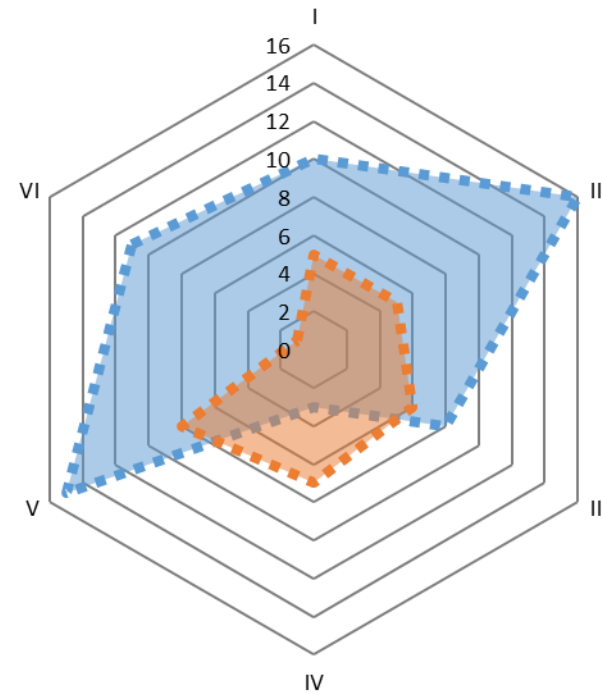
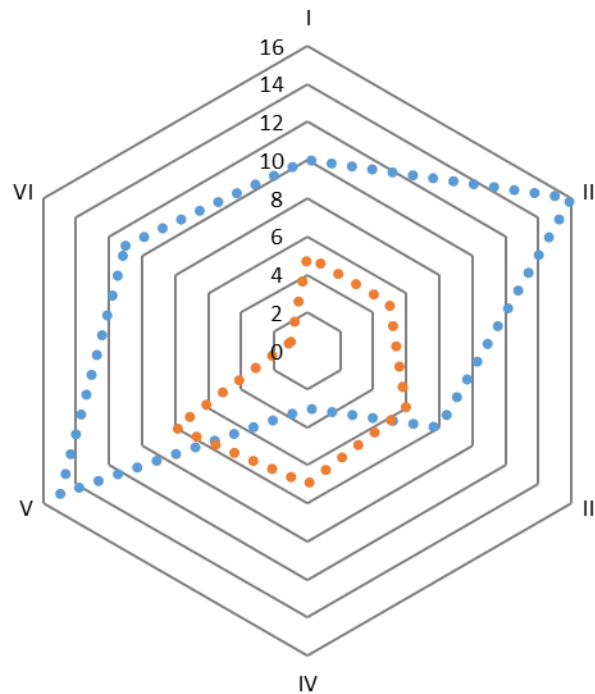
- Type of 3D graph - axes are two variables, color is the third variable
- Can be created in excel using conditional formatting
- Often in multivariate analysis to visualize association matrices

Occurrence of indicator organism in relation to two variables

Hloubka v cm vs. Koncentrace polutantu	< 60	60-69	70-74	75-79	80-84	85-89	90-94	95-99	100-109	110-119	120+
<= 30	29.8%	29.2%	27.9%	23.0%	20.5%	19.9%	20.6%	22.1%	22.1%	22.9%	23.3%
31-35	29.4%	28.2%	26.5%	22.0%	20.0%	19.5%	20.4%	21.6%	21.8%	22.6%	23.1%
36-39	18.5%	16.3%	15.8%	13.2%	12.9%	14.1%	15.3%	18.2%	20.4%	23.9%	28.4%
40-44	14.6%	14.3%	12.9%	12.0%	14.3%	20.2%	24.5%	22.2%	21.3%	20.2%	25.0%
45-49	12.6%	11.7%	13.0%	15.0%	17.9%	21.4%	22.5%	19.6%	20.3%	21.1%	30.0%
50+	12.2%	11.4%	13.6%	17.5%	22.0%	25.6%	25.9%	20.4%	19.9%	20.3%	31.3%

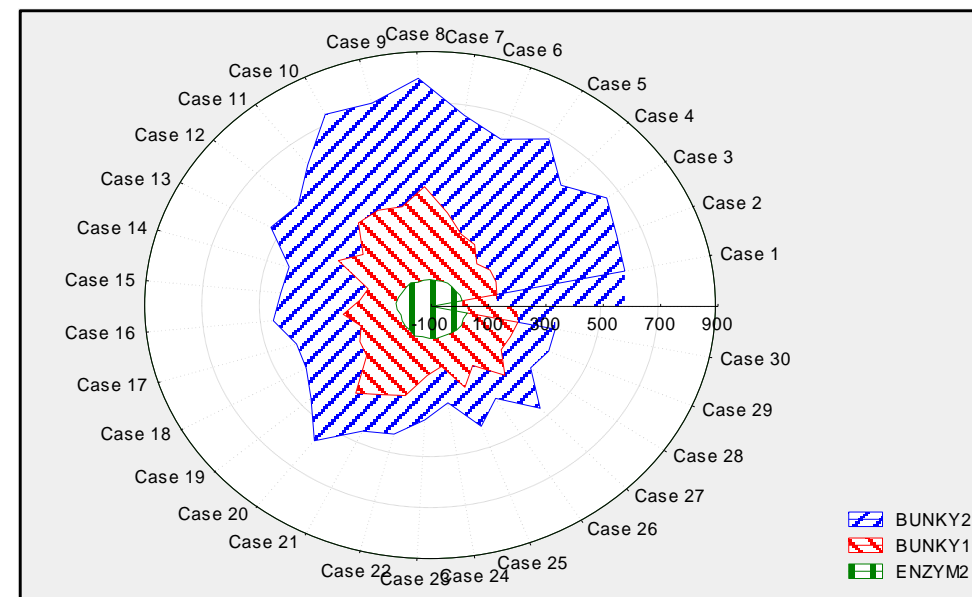
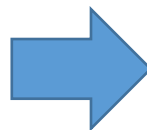
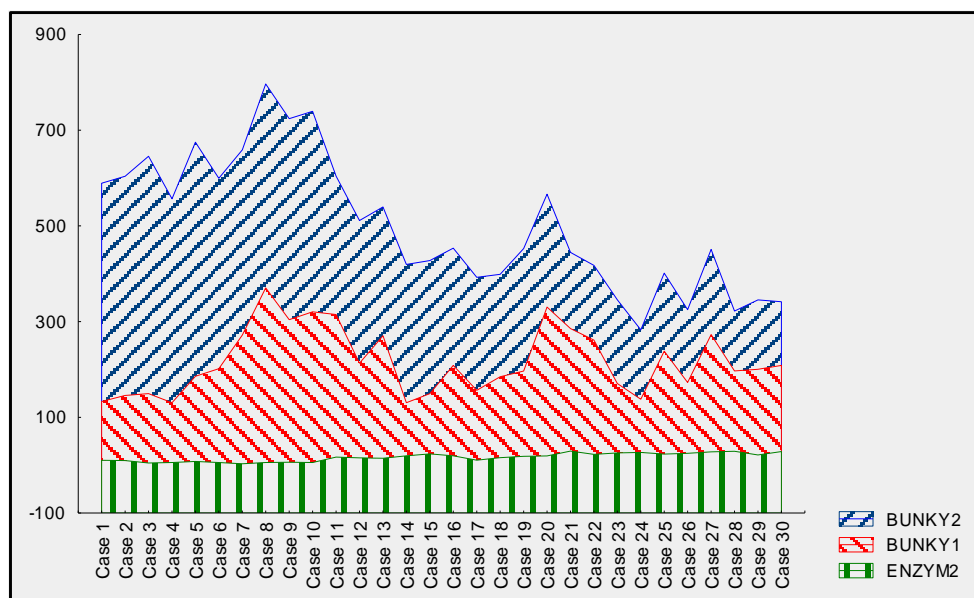
Spider / Ray Graphs

- Suitable for comparing object profiles or groups of objects using multiple variables
- Different graphic form



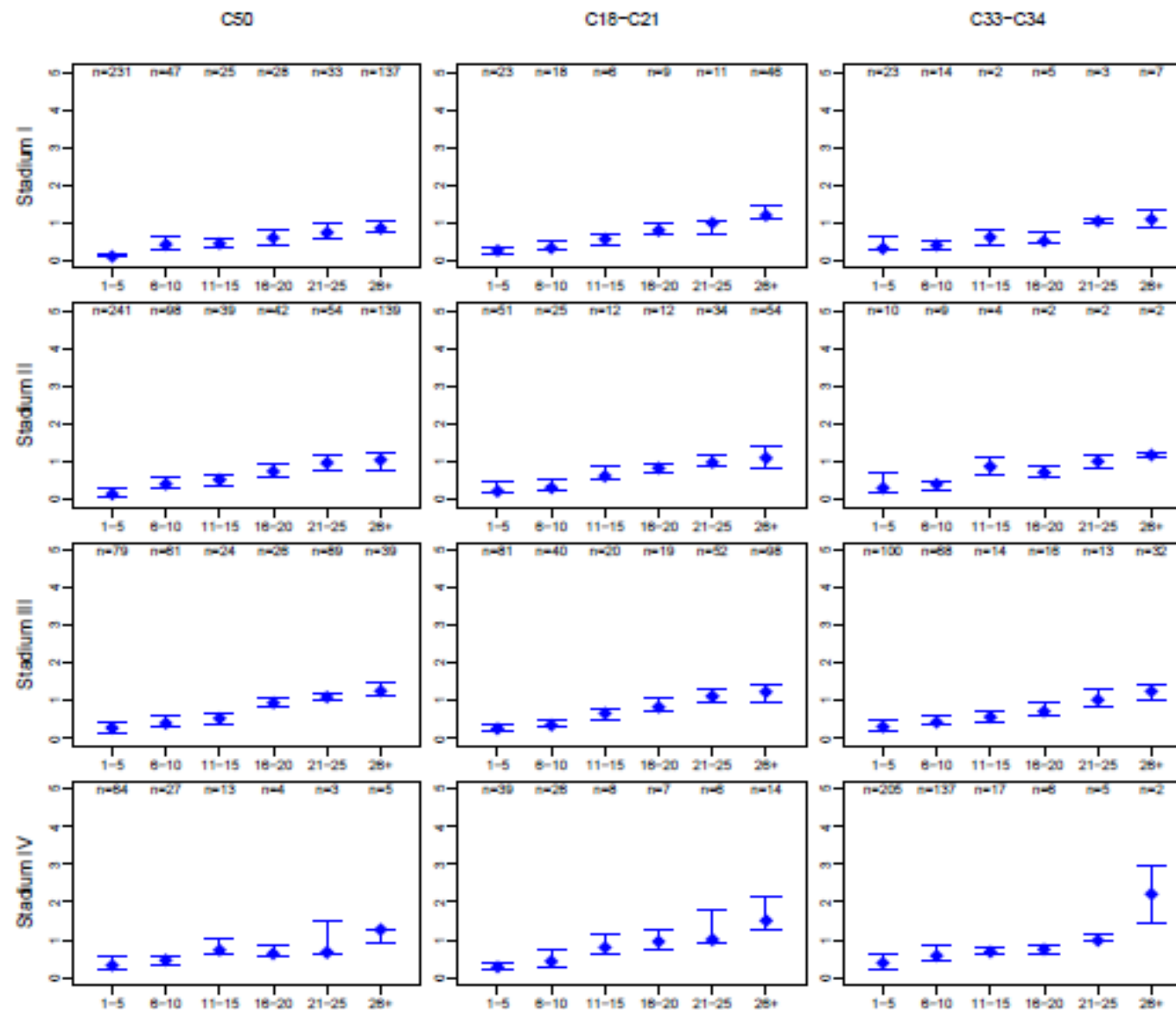
Polar graph

- Analogous to a line, bar or area chart with the X-axis plotted on a circle
- Suitable for cyclical data (circadian rhythms, seasonality, directional statistics of animal movement)



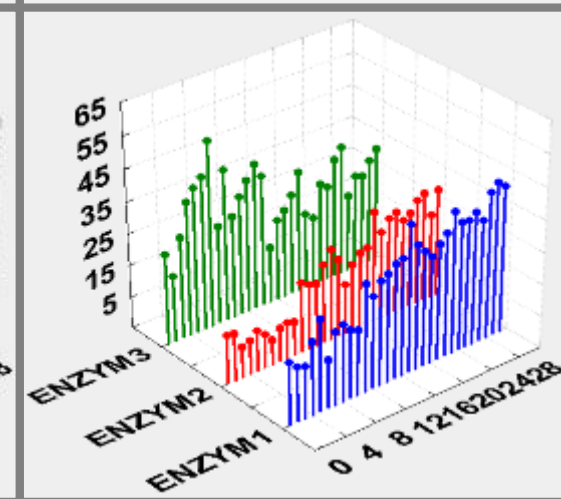
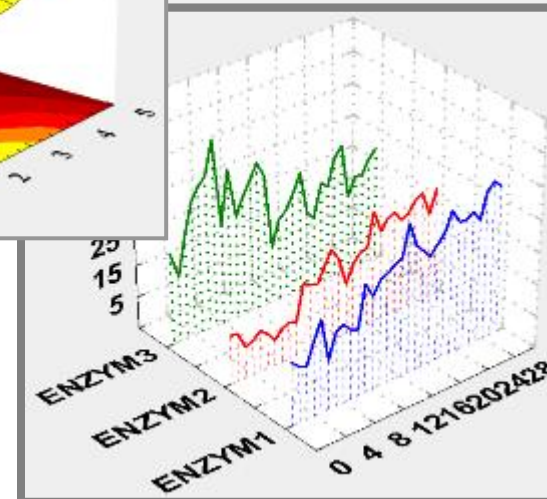
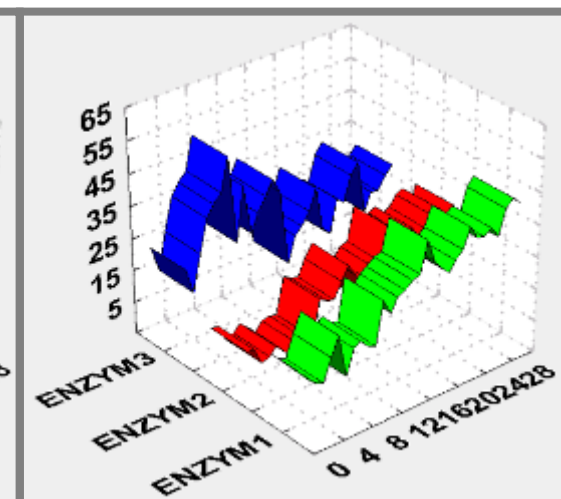
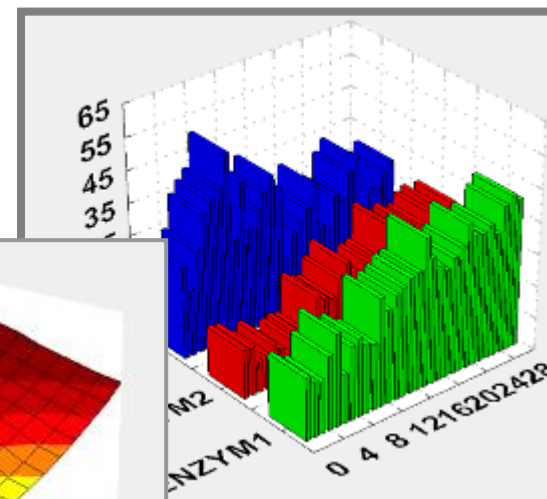
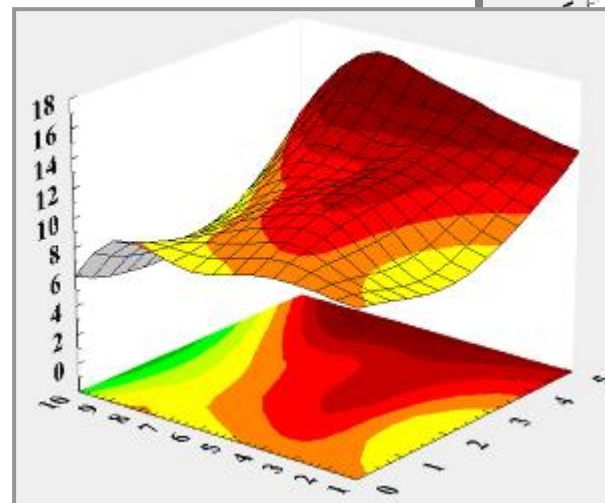
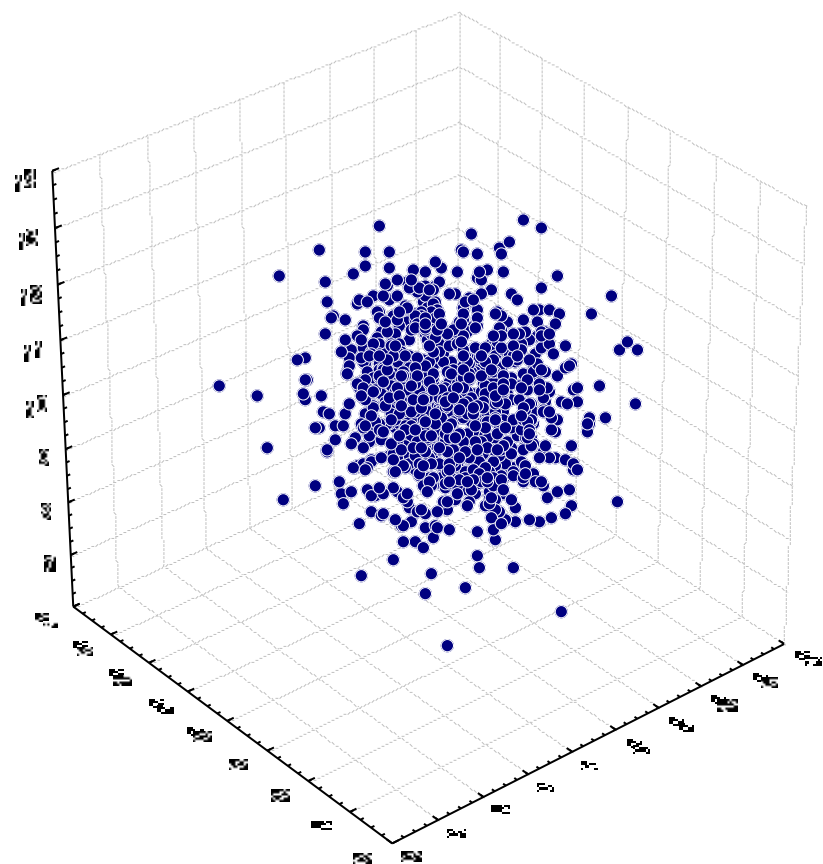
Graphic boards

- Multiple charts forming a graphic board
- Can be composed of different charts of one or more types
- Presentation of large amounts of data in a small space



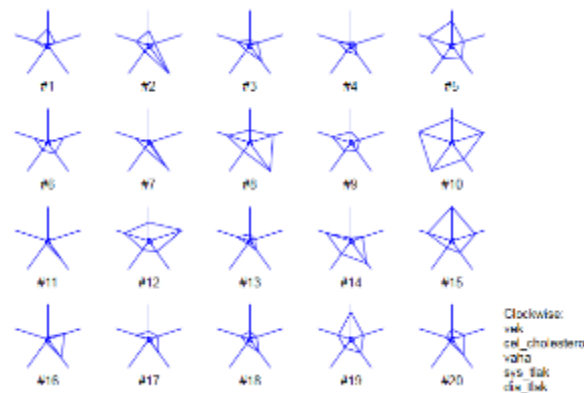
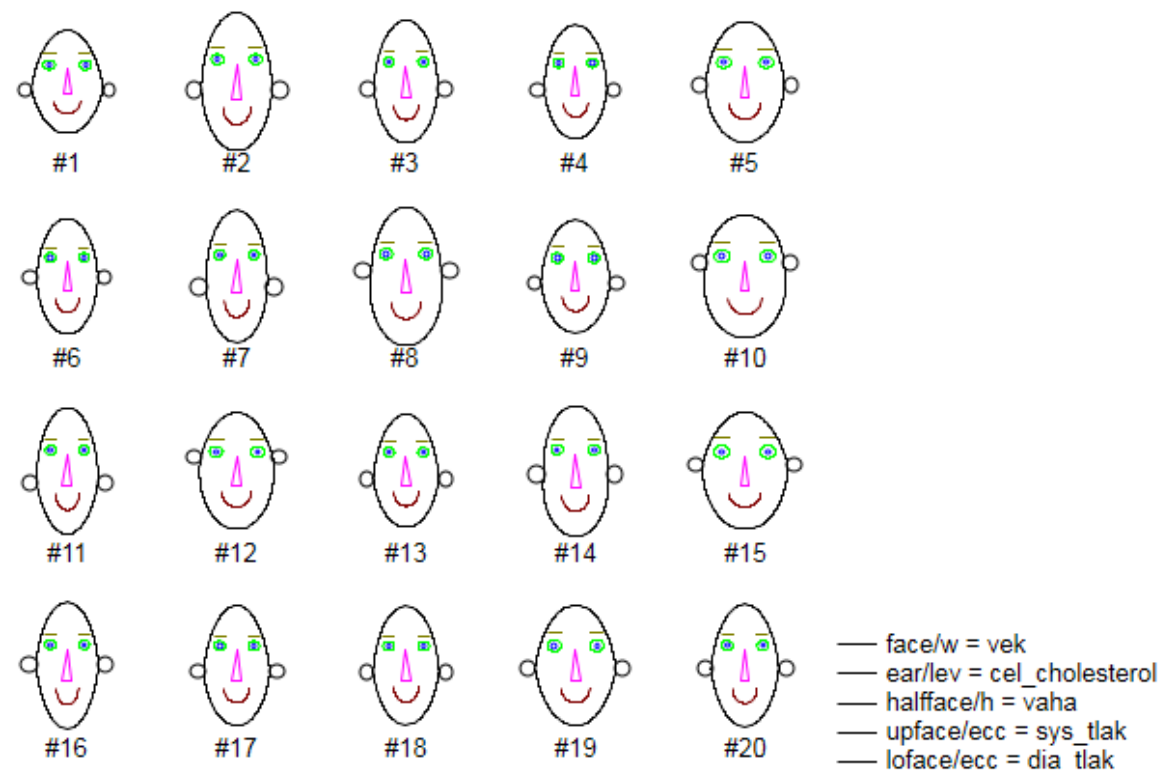
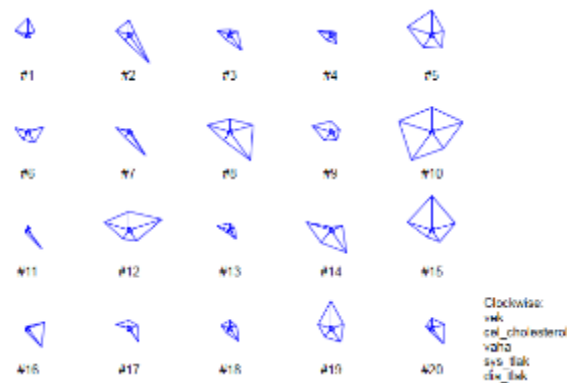
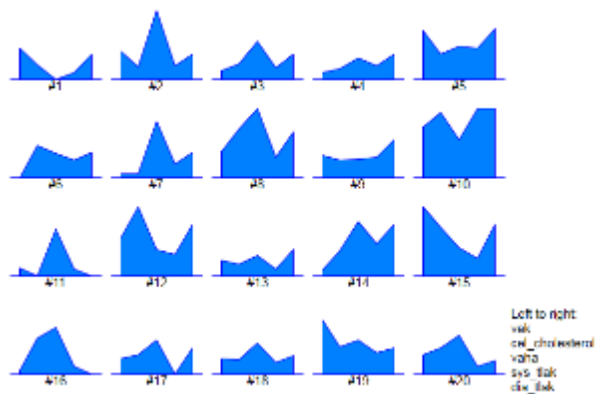
3D charts

- Many types
- Great emphasis should be placed on interpretability and meaningfulness



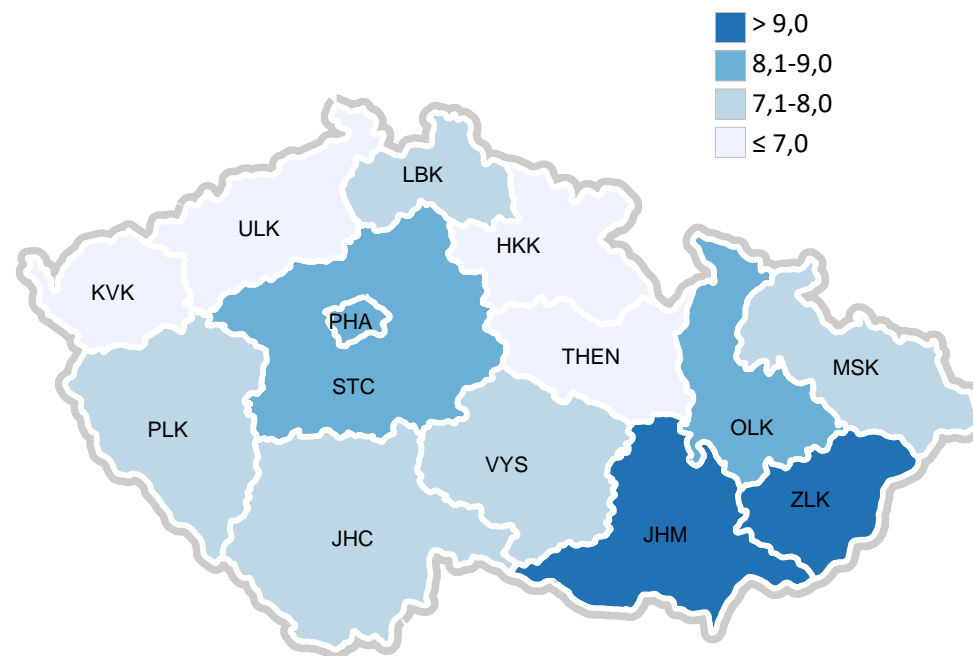
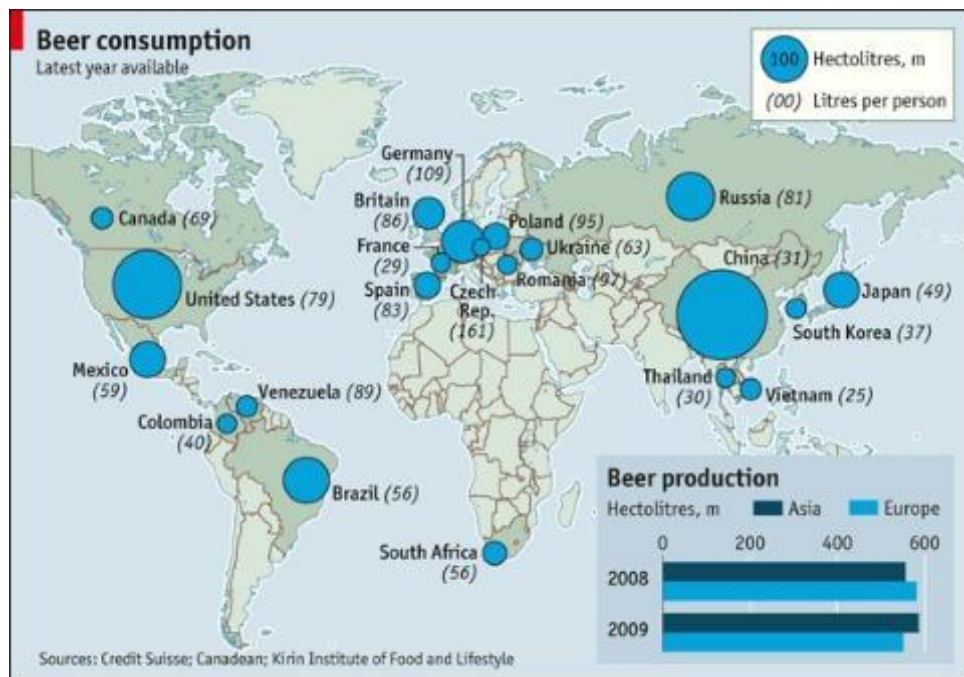
Chernoff's faces (icon charts)

- Individual variables are displayed as facial features
- It belongs to the so-called icon charts
 - character values represented as geometric shapes or symbols
 - each object (subject) corresponds to one pattern composed of the following geometric shapes or symbols
 - allows to visually compare which objects (subjects) are similar



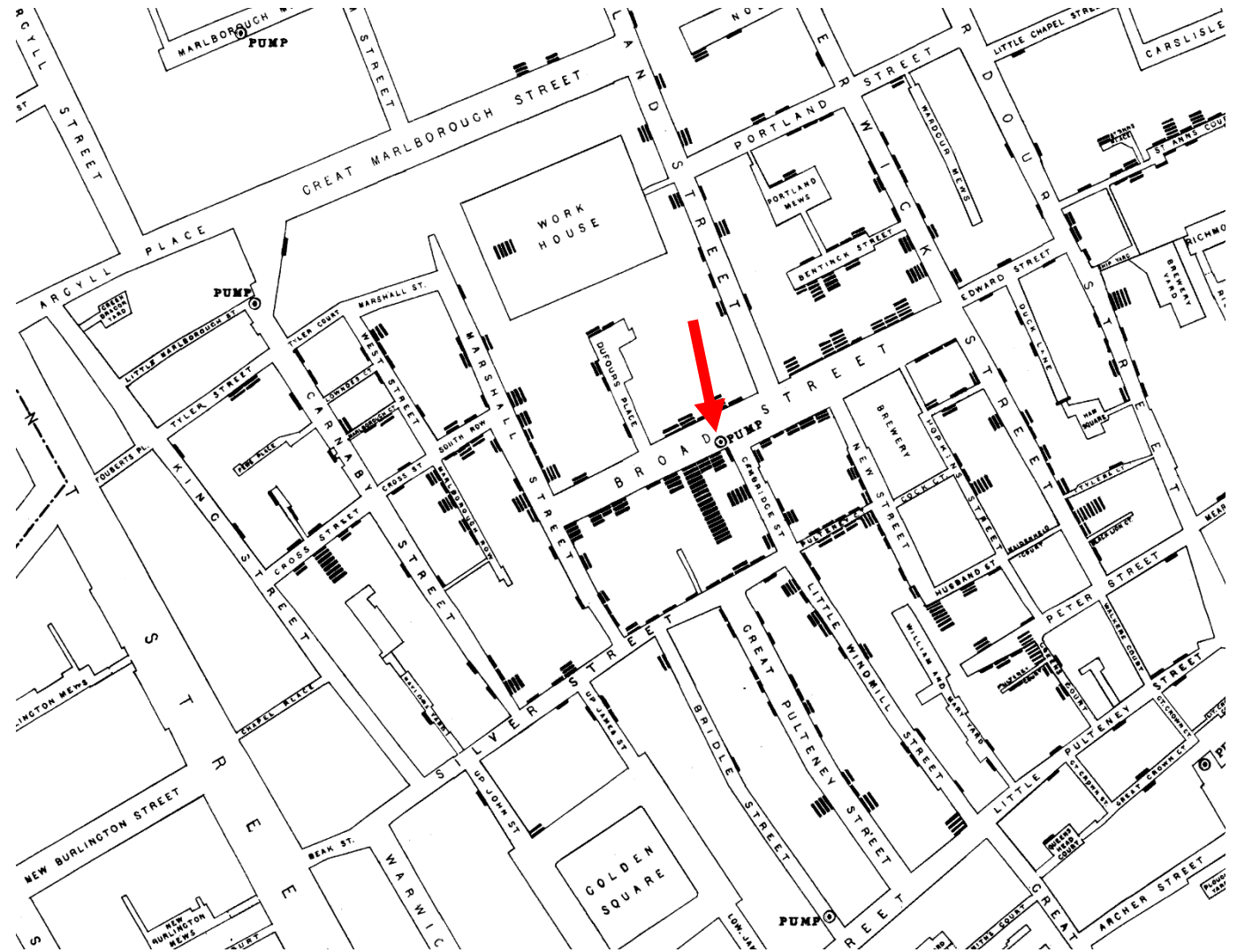
Maps are also charts

- A separate chapter on data visualisation
- Coloring regions in the map according to the analysis results or directly inserting graphs into the maps (bar, pie, etc.)
- ArcGIS - another software available on inet.muni.cz

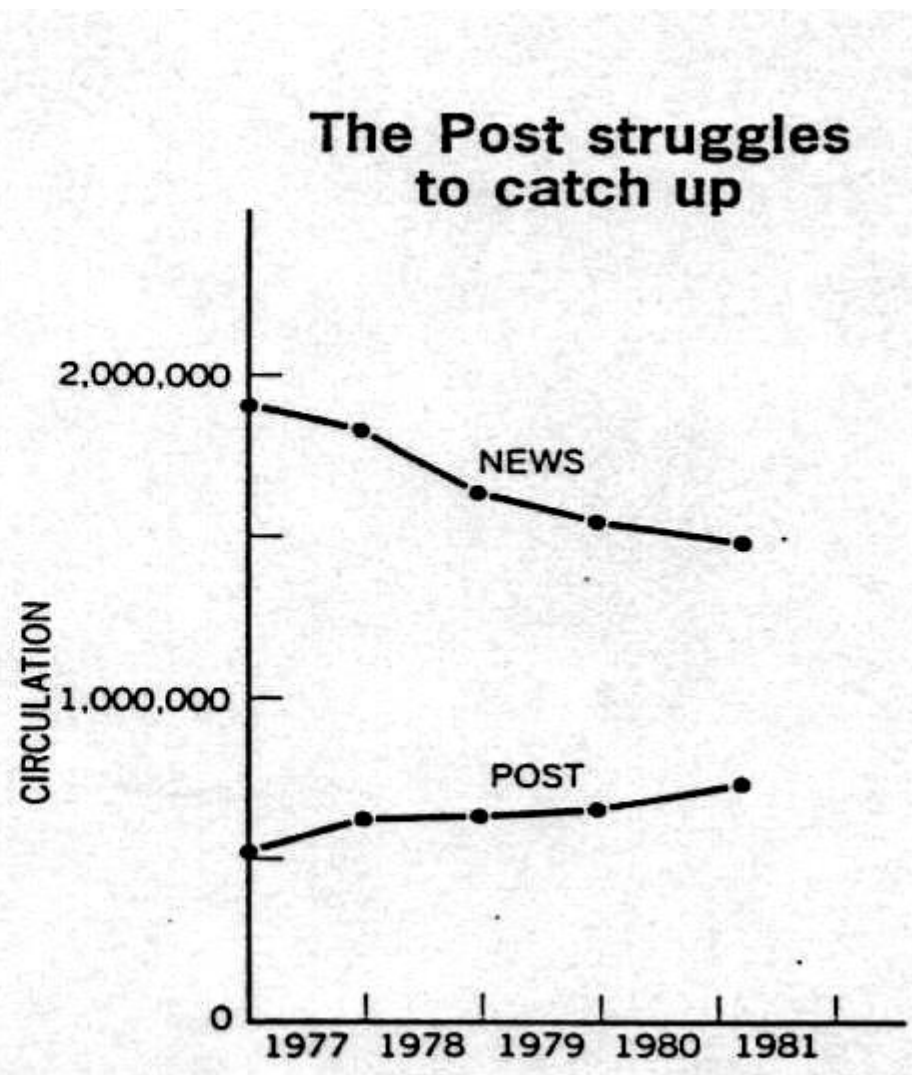
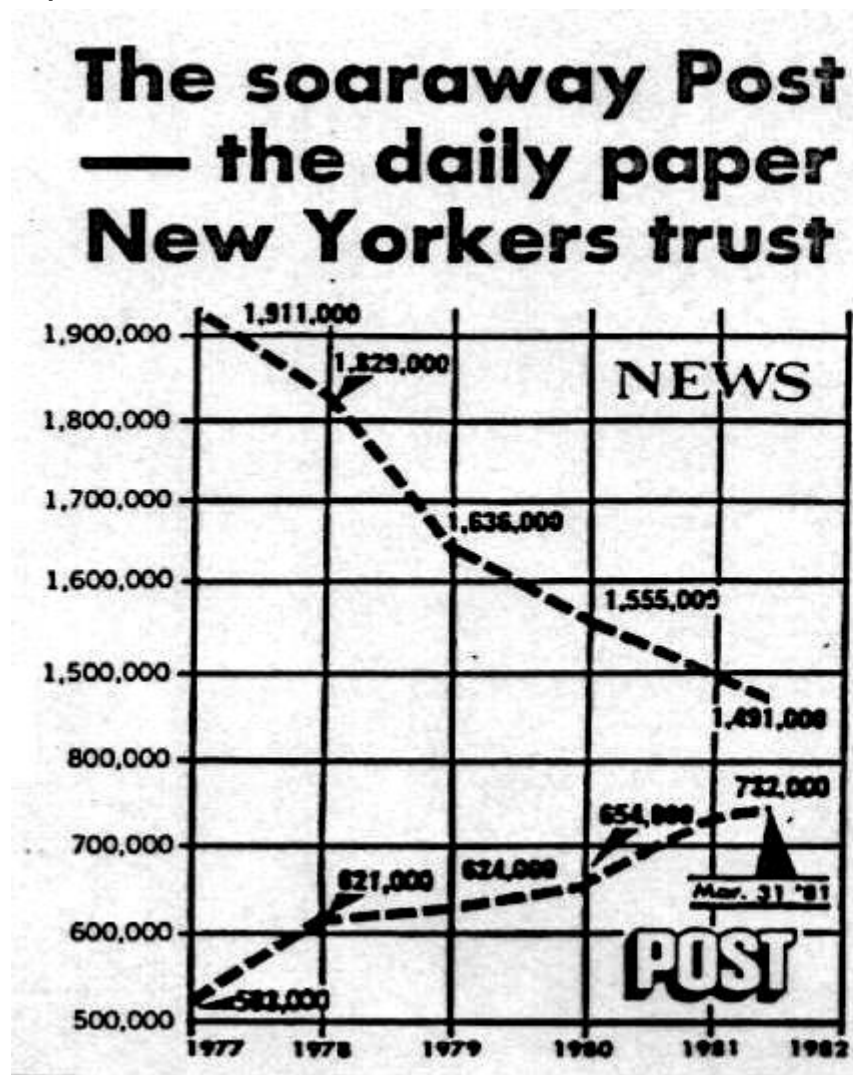


Famous maps of John Snow - Cholera in London

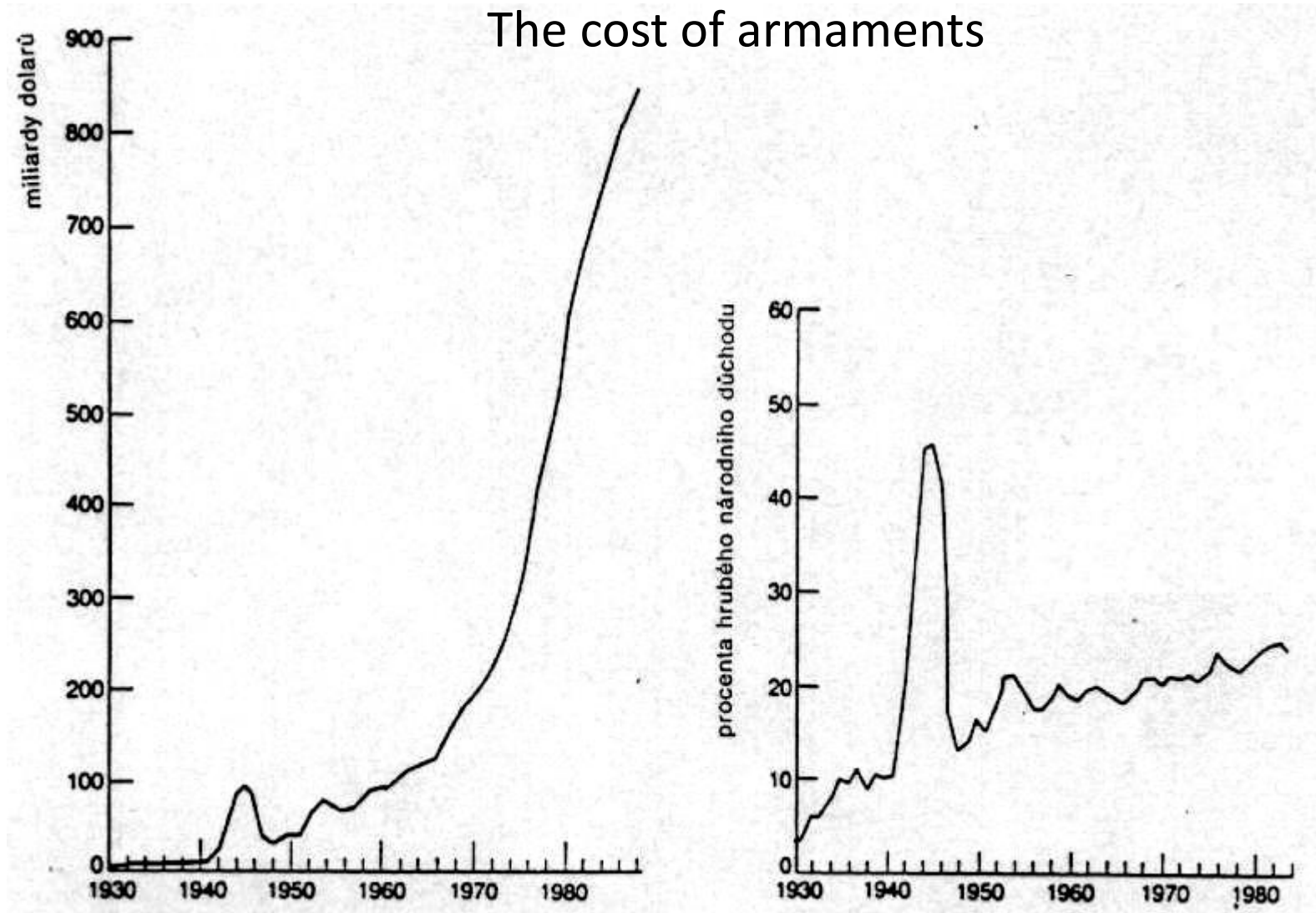
- 1854 Broad Street cholera outbreak
- Numbers of cases plotted as black columns by victims' residence
- Identification of the source of infection - contaminated well
- One of the first examples of spatial data analysis and epidemiological mapping



Incorrect use of charts: range of axes ("we don't know how to draw")



Incorrect use of graphs: standardization of axes ("we don't know what we are drawing")



Lecture 3

Information and data distribution

How information is created

Data distribution

Annotation

- The basic principle of statistics is the probability of an event occurring.
- Through sampling, we try to estimate the true probability of events.
- The key issue is sample size, the larger the sample, the greater the chance of the actual probability of the phenomenon occurring.

The origin of information: concepts I

Reality



The Observer



Phenomenon - a subset of all possible outcomes of an experiment/agency, about which it can be said whether it occurred or not

Phenomenon field - the class of all phenomena that we choose or are able to observe

Reality + Phenomenal field = Measurable space

The origin of information: concepts II

- **Experimental unit** - the object on which the investigation is carried out
- **Population** - set of experimental units (object)
- **Character** - property monitored on the object
- **Random variable** - a numerical value expressing the result of a random experiment



- A trait becomes an **observed random variable** if its value is determined **by drawing (sampling)** an object from a **population**

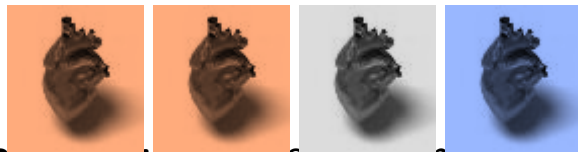
Origin of information: sampling

Statistics speak about reality through a selection from the target population

Statistical assumptions for correct sampling must be met

Random selection from the target population

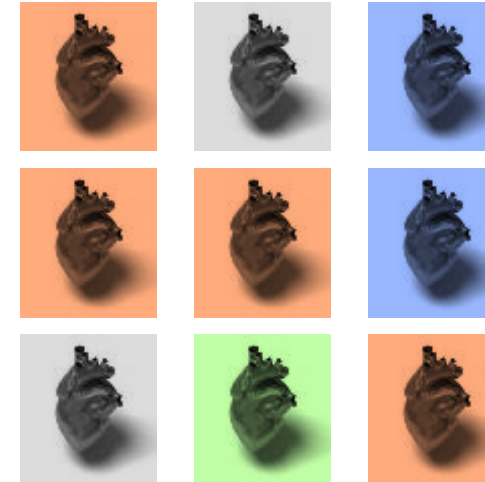
Representativeness: the sample design must reflect reality as much as possible



Independence: multiple sampling of the same object provides no new information from a statistical point of view



Target population



Sampling example

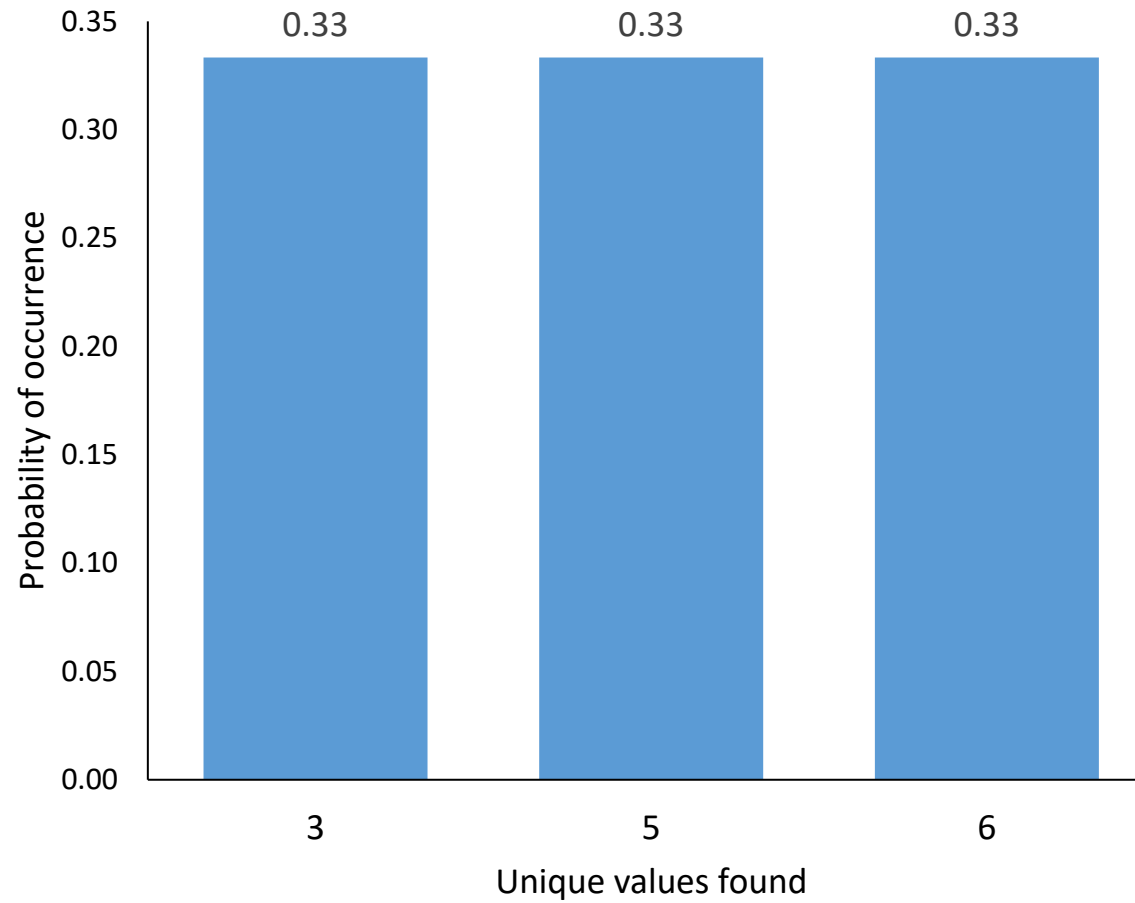
- Based on sampling, we want to find out the properties of a phenomenon
- Our target population will be dice rolls with unknown characteristics



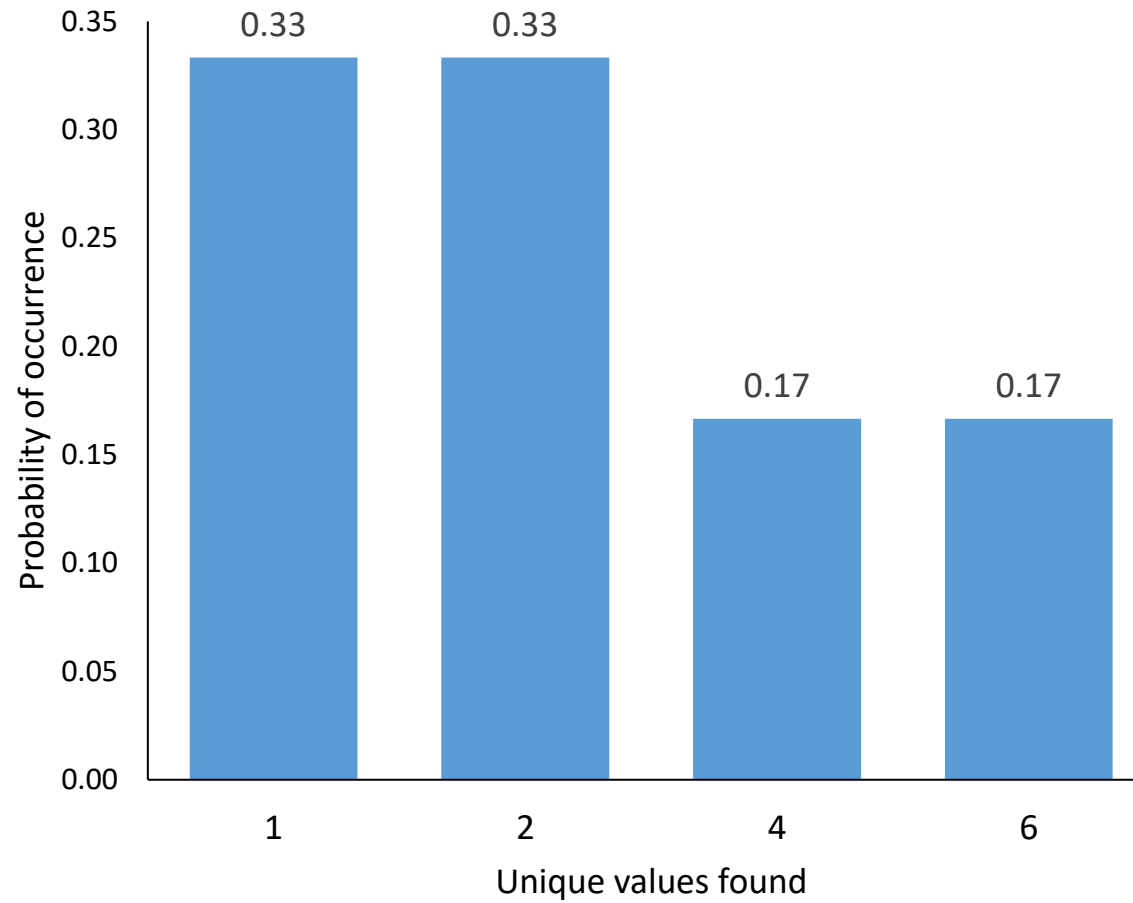
- We want to find out the properties of an unknown used cube



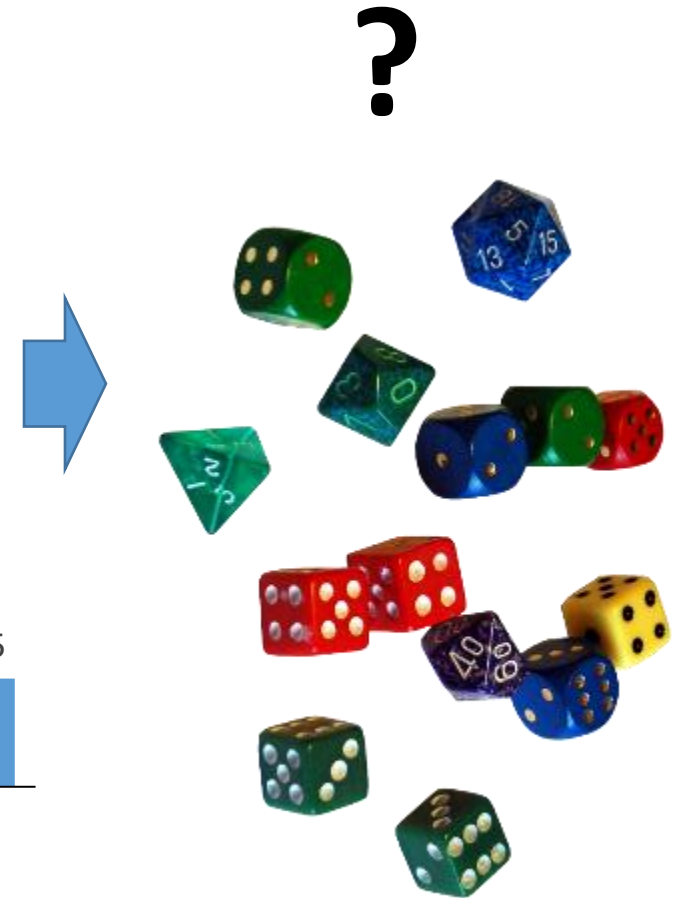
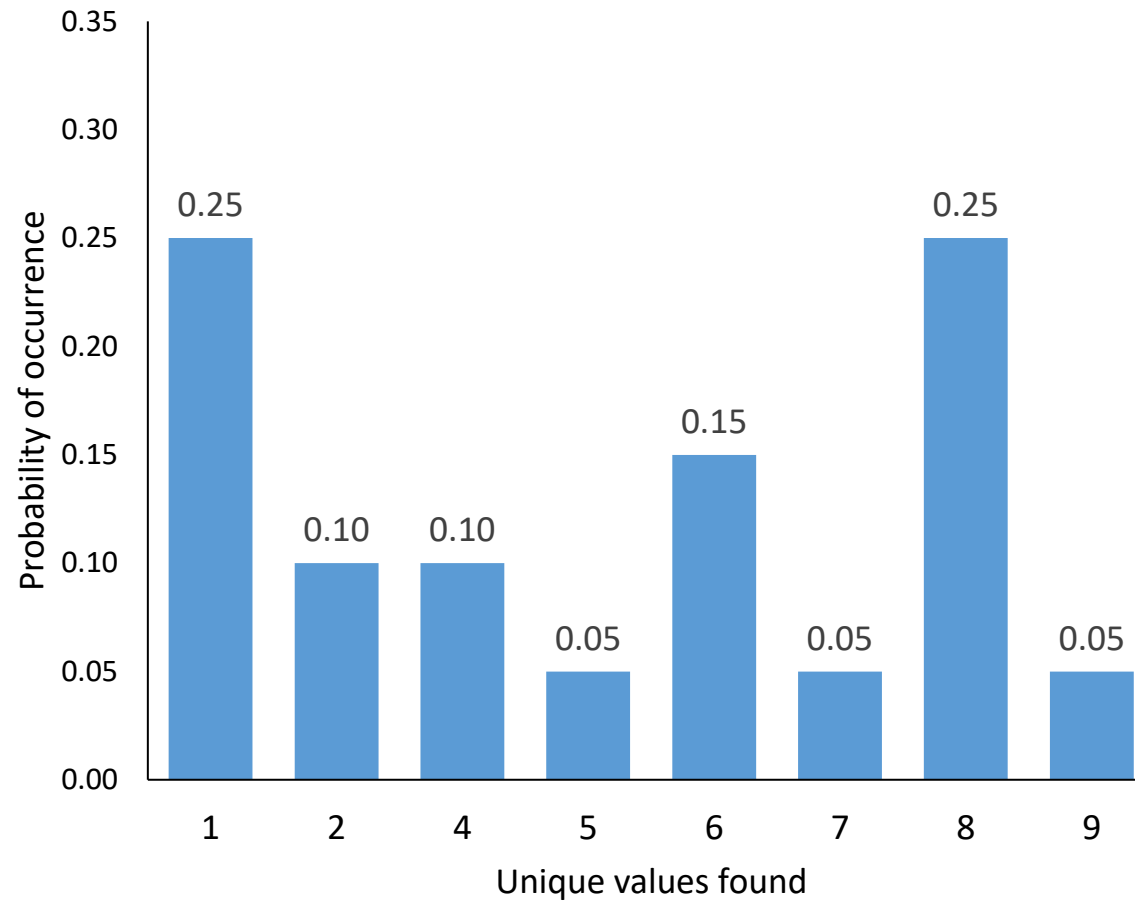
Sampling example: $n=3$



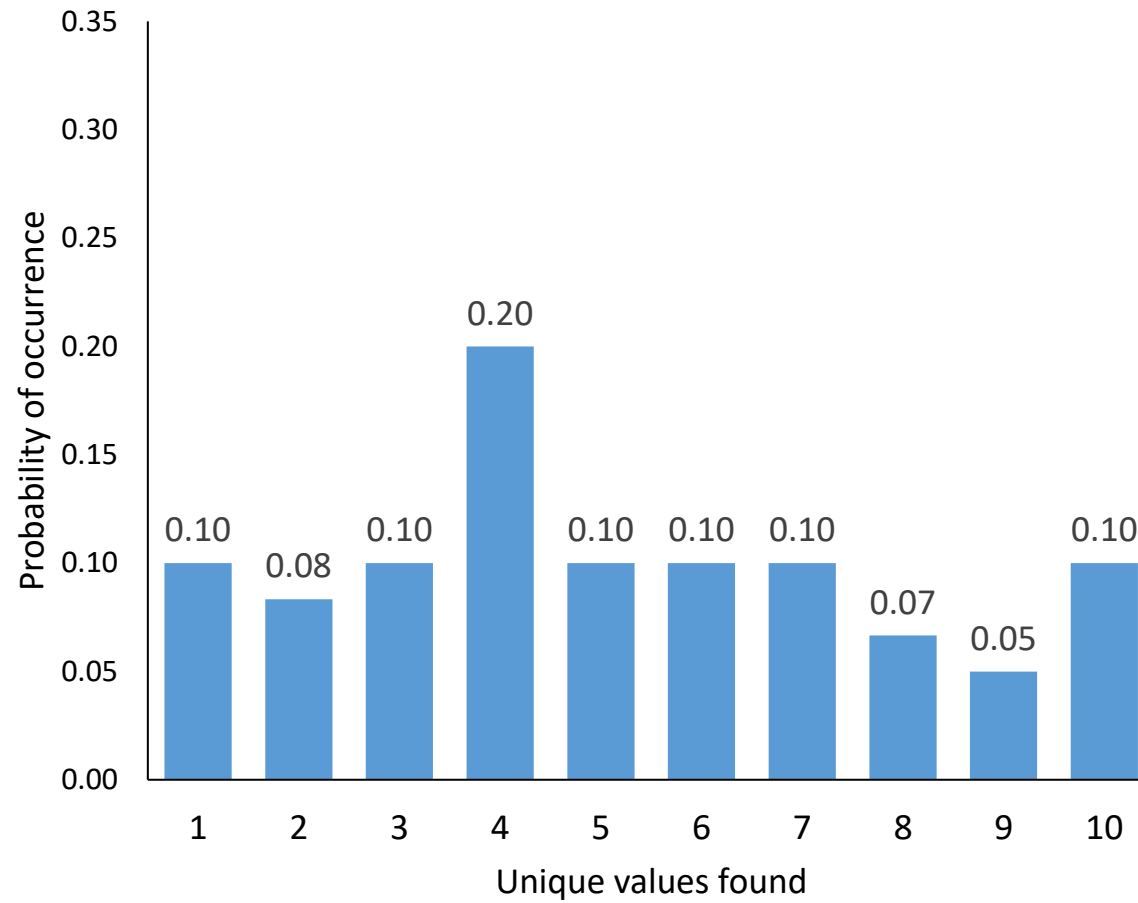
Sampling example: $n=6$



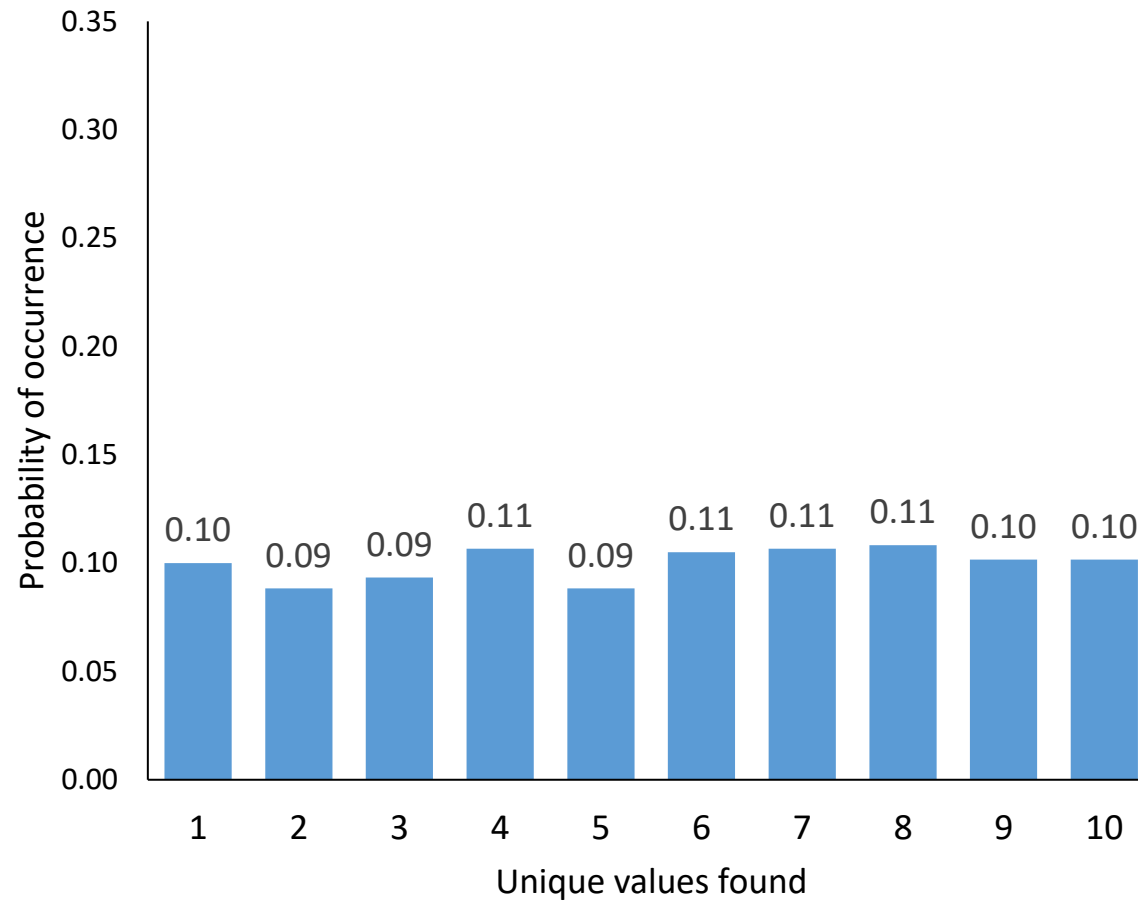
Sampling example: $n=20$



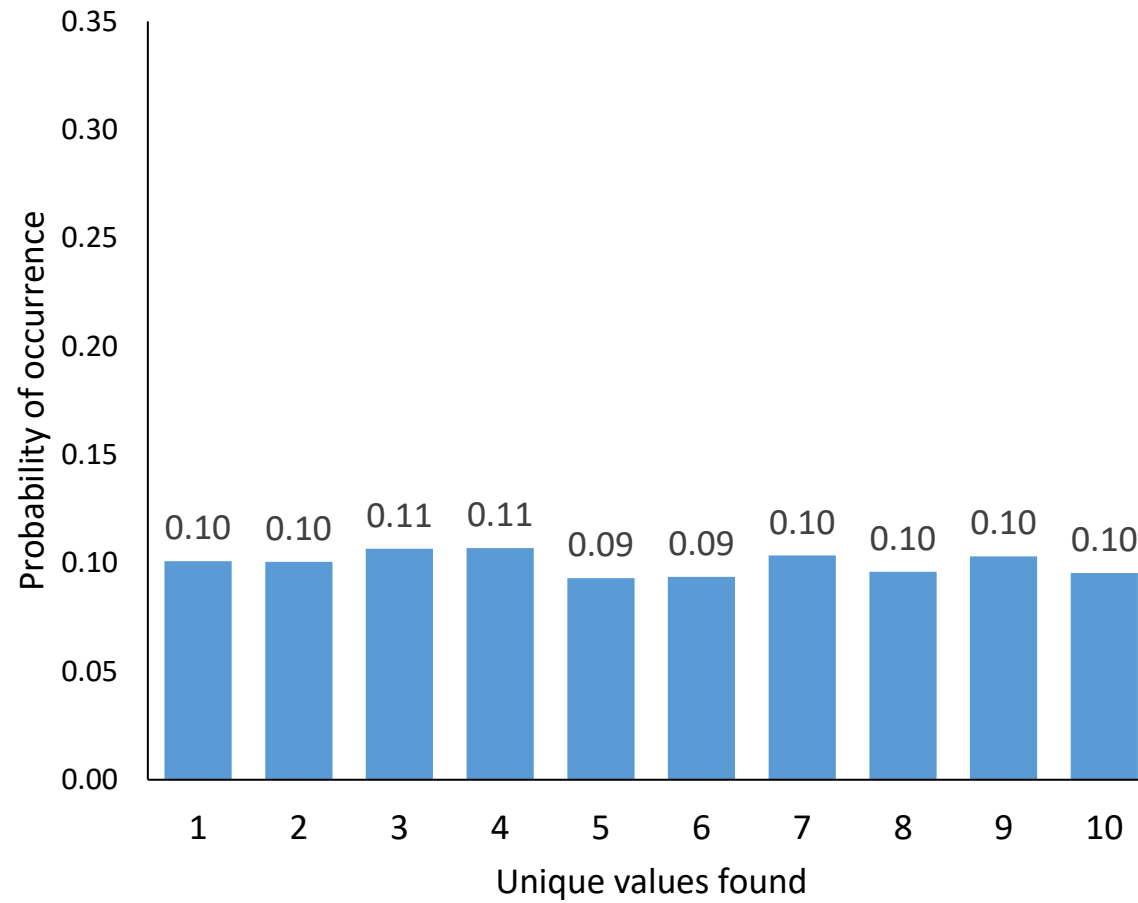
Sampling example: $n=60$



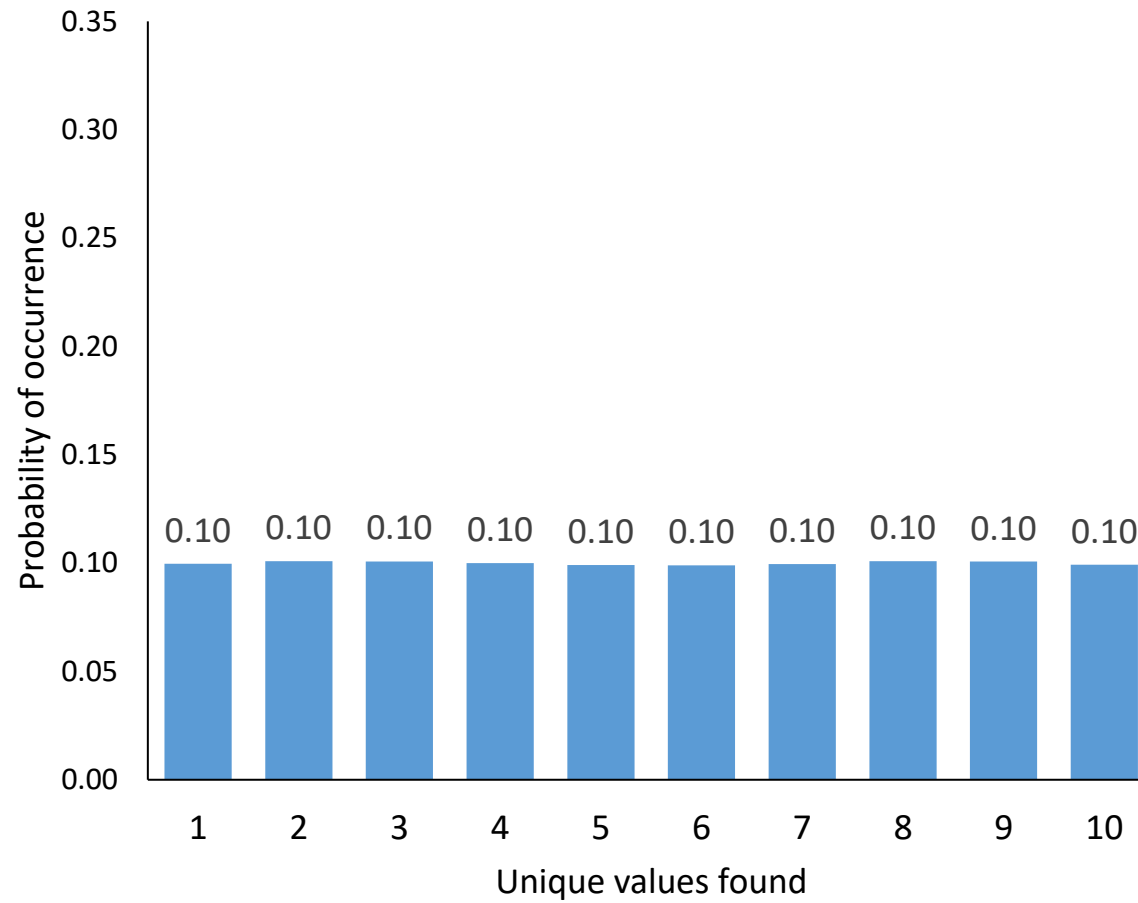
Sampling example: $n=600$



Sampling example: $n=6\ 000$



Sampling example: $n=60\ 000$



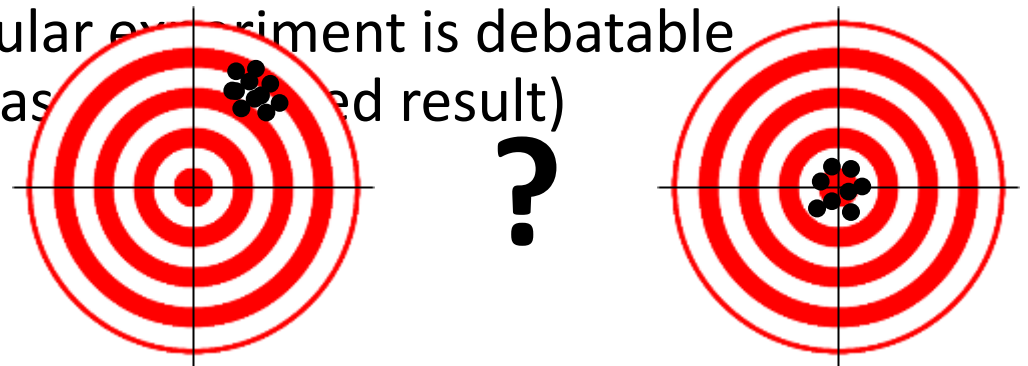
?

Sampling example: conclusion

- The phenomenon under observation is probably in the shape of a ten-sided cube

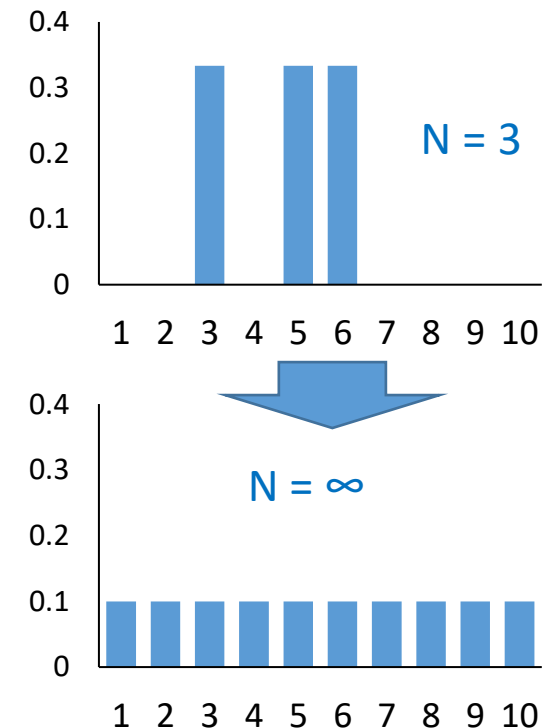


- For complex stochastic systems, the truth is only obtained after a considerable amount of experimental work has been done: we have to give the system a chance to manifest itself
- When running a random experiment, the true knowledge of the system increases with increasing number of repetitions (the results become more stable and reliable)
- However, the degree of generalization of a particular experiment is debatable (reliability and stability of results is not the same as a single result)



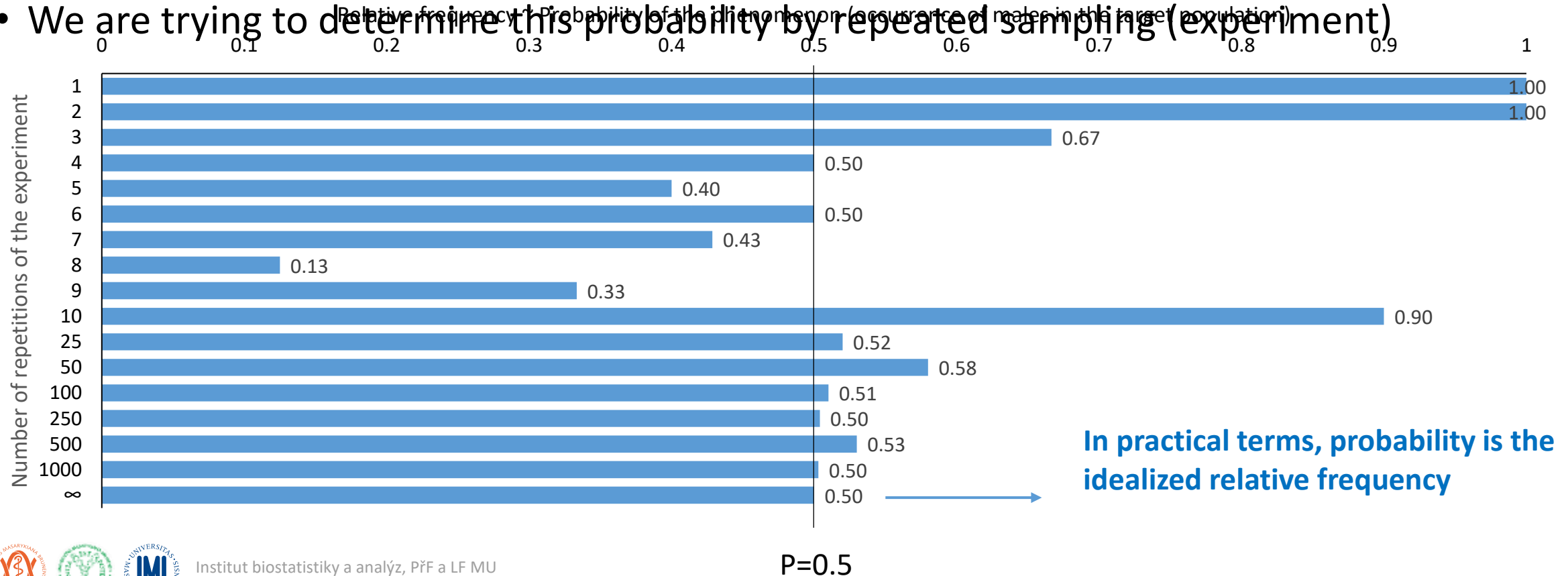
Empirical law of large numbers

- When the same random experiment is performed again independently, the proportion of occurrences of the observed phenomenon among all the previous realizations usually settles around a constant.
- A probability is any real function defined on a phenomenon field A (e.g., dice rolls) that assigns to each phenomenon A (e.g., sides of a die) a non-negative real number $P(A)$ from the interval 0 - 1.
- **In practical terms, probability is the idealized relative frequency**
- $P(A) = 1$ phenomenon certain
- $P(A) = 0$ phenomenon impossible
- $P(A \cap B) = P(A) \cdot P(B)$ independent phenomena
- $P(A \cap B) = P(A) \cdot P(B/A)$ dependent phenomena
- $P(A / B) = P(A \cap B) / P(B)$ conditional probability



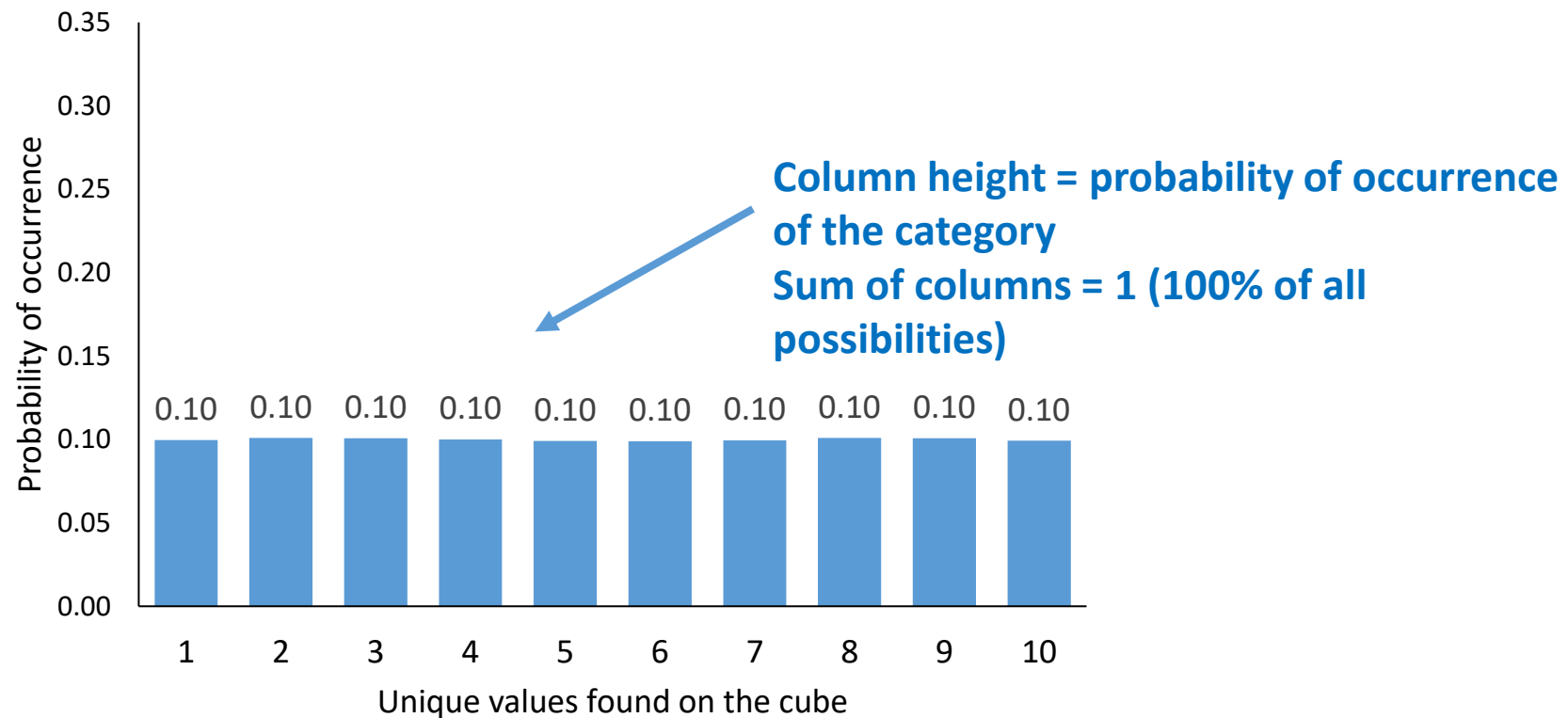
Empirical law of large numbers: an example

- We assess the prevalence of men in the population of interest (the phenomenon of "male prevalence")
- The true probability of the observed phenomenon is $p=0.5$ (but we don't actually know this)
- We are trying to determine this probability by repeated sampling (experiment)



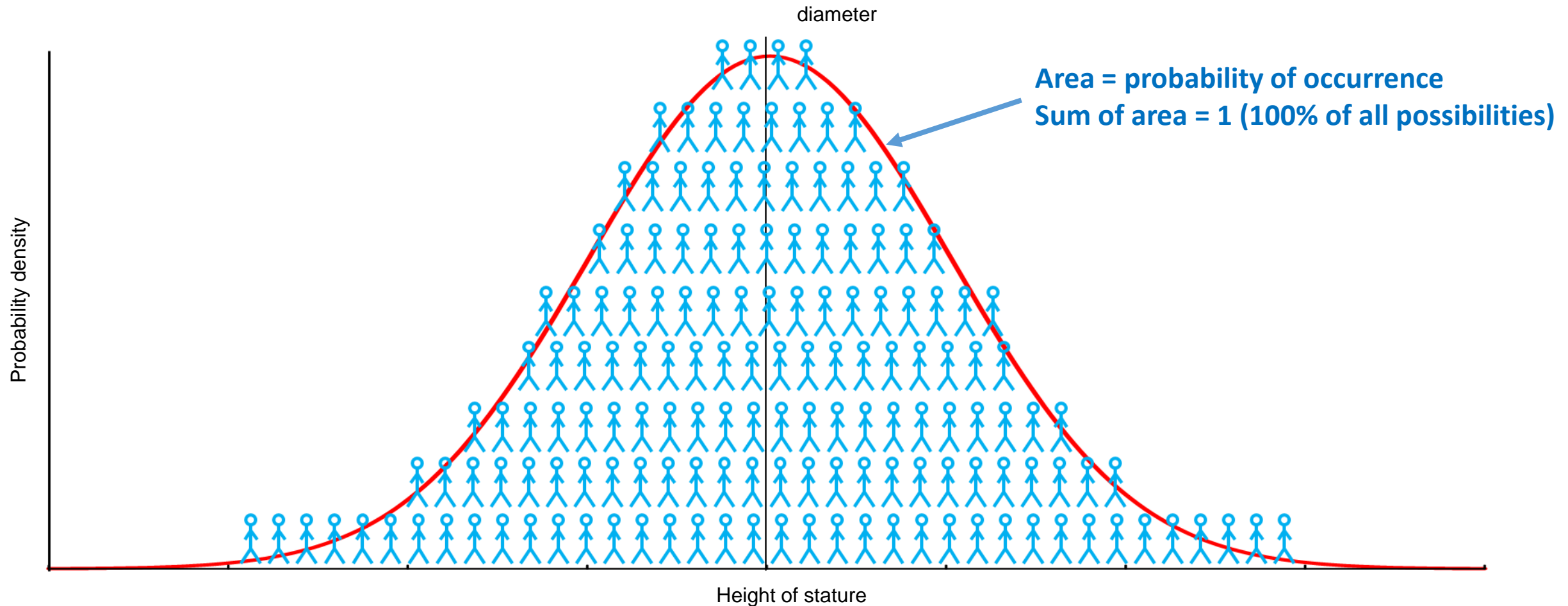
Probability of occurrence - distribution of categorical data

- there is a probability of occurrence of the phenomena (non-deterministic inferences)
- "anything is possible": only a phenomenon with probability 0 will never occur



Probability of occurrence - distribution of continuous data

- there is a probability of occurrence of the phenomena (non-deterministic inferences)
- "anything is possible": only a phenomenon with probability 0 will never occur



Basic data types

Continuous and categorical data

Basic descriptive statistics

Graphical description of data

Annotation

- Reality can be described by different types of data, each with specific characteristics, advantages, disadvantages and its own set of usable statistical methods
- From binary to categorical, ordinal to continuous data, the level of information contained in them grows.
- The basic approach to descriptive data analysis is the creation of frequency tables and their graphical representations - histograms.

How is the data created?

- A record of reality...



How is the data created?

- A record of reality...

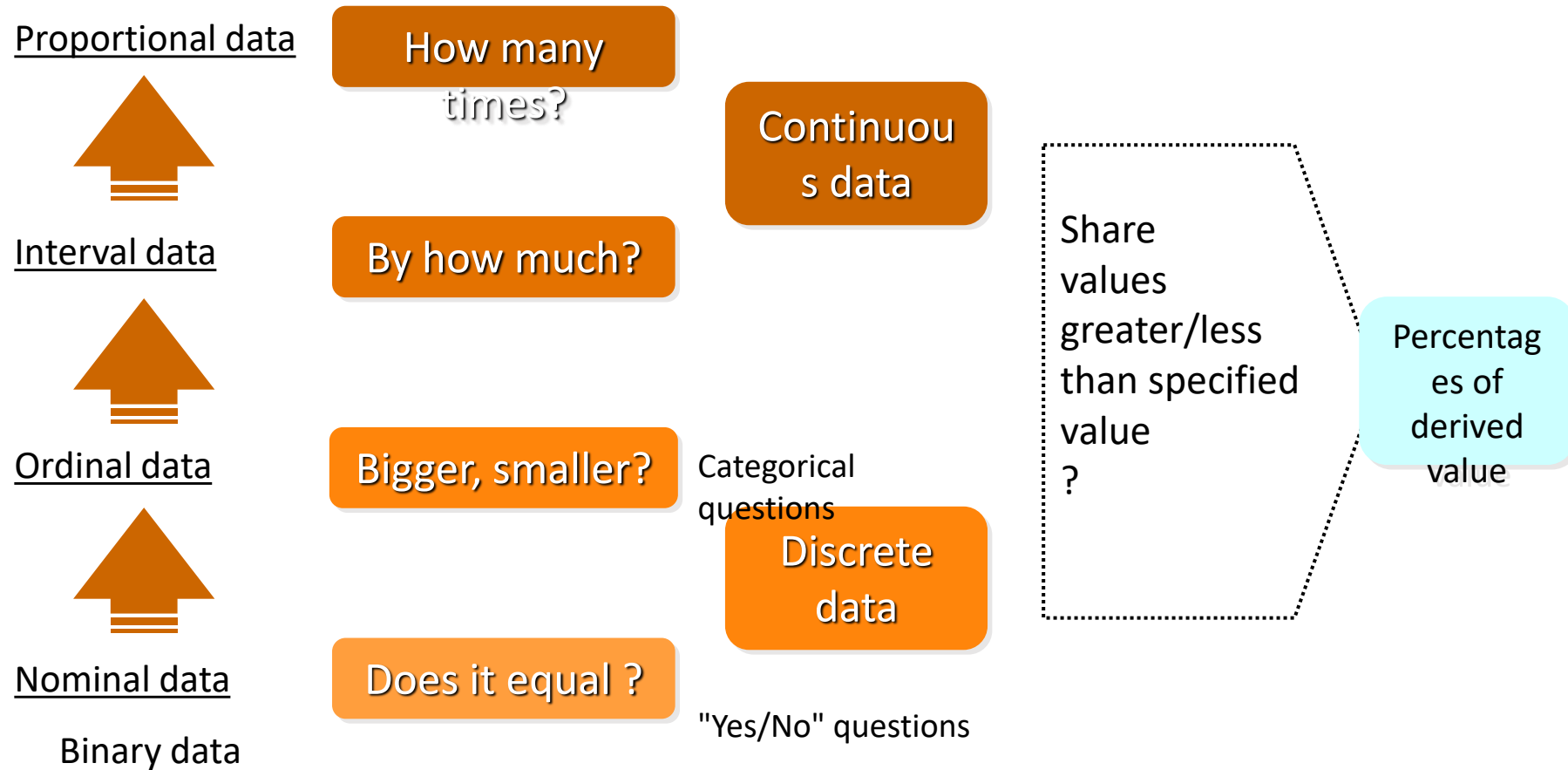
... which we want to study further → meaningfulness?

(pollutant concentration × altitude, blood pressure, glycaemia × number of hearts, number of houses)

... more or less perfect → quality?

(variability = information + error)

How information is created - different types of data mean different information



However, knowing the data type alone is not enough to achieve the information

Types of data and their information value

- Statistics are useful at all times 😊
- Even in the Ice Age
- The shaman sits in front of the cave and thinks:
 - Winter is coming and you need to stock up for the winter
 - But I have to figure out how to **properly** describe what we actually caught for supplies
 - Or starve to death



Target population

- We sample 3 categories of the observed variable prey



Prey

Squirrel

Deer

Mammoth



Binary data - did we catch something?

- The least information-intensive data are binary

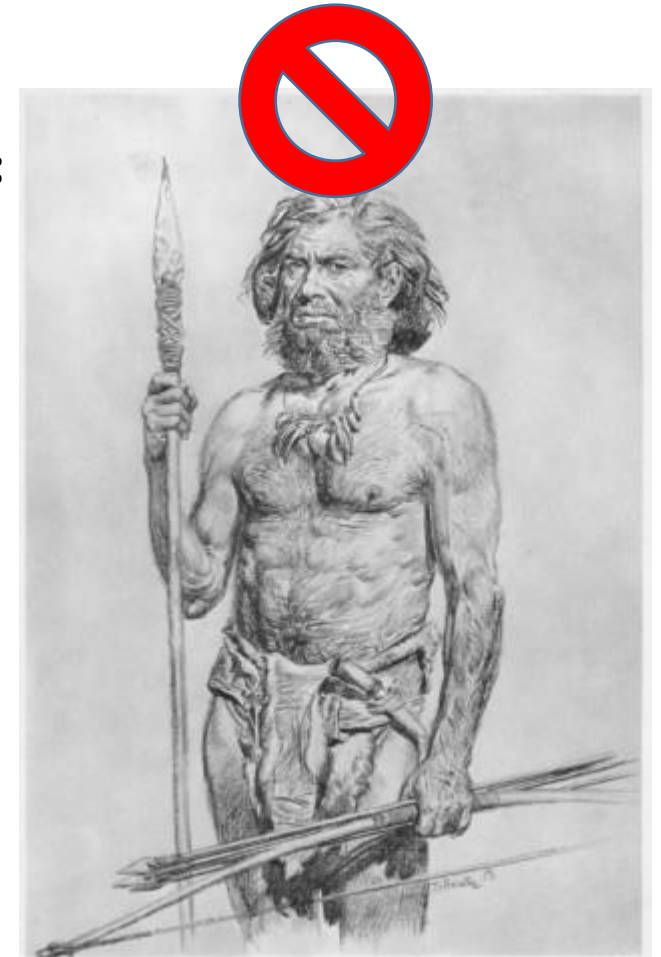


We evaluate two possible states:

Brought x did not bring booty

How can we describe:

?



Binary data - did we catch something?

- The least information-intensive data are binary



We evaluate two possible states:

Brought x did not bring booty

How can we describe:

Total number of hunts (**evaluation basis**)



Number of catches (**absolute frequency**)



Percentage of successful hunts (**relative abundance**) or most abundant category (**modu**)



Is binary data sufficient in all circumstances?

Categorical data - what did we catch?

- More information can be obtained from the category

We evaluate several possible states:

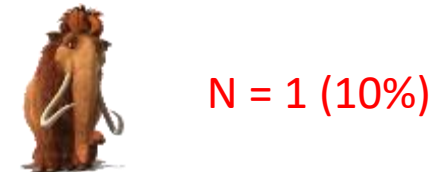


How can we describe:

Total number of hunts (evaluation basis)



Number of different catch categories
(absolute frequency)



Percentage of successful catches of
different catch categories (relative
abundance) or most abundant category
(modus)



Are categorical data sufficient in all circumstances?



Are the categories sortable?



- Sortable categories = ordinal data
- Ordinal data can be described in the same way as categorical data + **median can** be calculated for adjustable data

Are categorical data sufficient in all circumstances?

Note the median for ordinal data

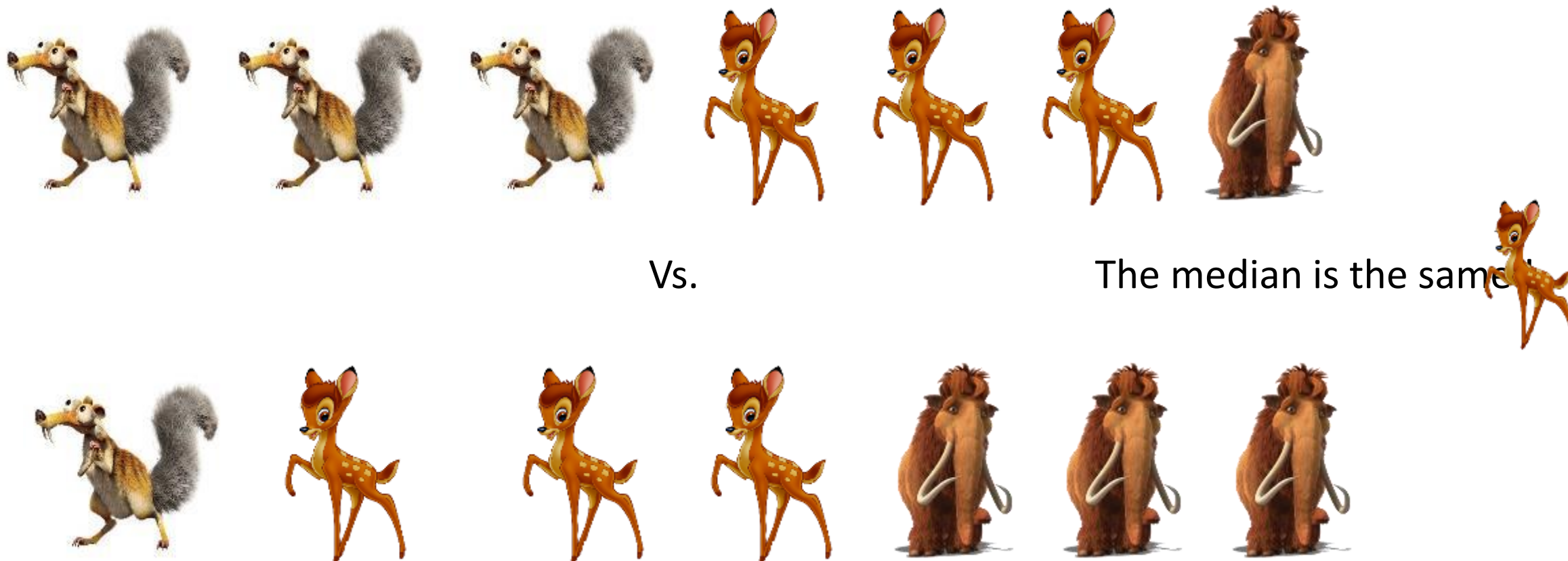
- Is the median always a good indicator of the mean of ordinal data?



Vs.



Note the median for ordinal data



- The median is the same, but the interpretation of the data is different
- The possibility and formal correctness of calculating a statistic does not mean that it is an appropriate method.

Quantitative data - what is the volume of prey ?

- The most informative data are quantitative
- To describe them, it is necessary to assess their distribution
 - Average
 - Median
 - Standard deviation
 - Minimum, maximum
 - Percentiles
 - Etc.



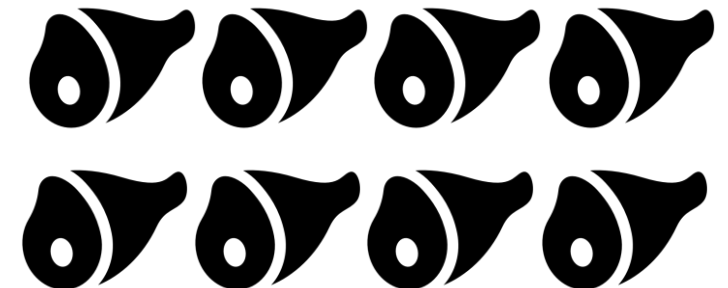
=



=



=



Data types: summary

- Qualitative variable (categorical) - it can be categorized, but it cannot be quantified, or it does not make sense to assign a numerical expression to each category.
- Examples: gender, HIV status, drug use, hair colour
- Quantitative variable (numeric) - we can assign it a numeric value. There are two types of quantitative variables:
 - Continuous: can take any value within a certain range.
 - Examples: height, weight, distance, time, temperature.
 - Discrete: can only take on a countably many values.
 - Examples: number of blood cells, number of hospitalizations, number of bleeding episodes per year, number of children in the family.

Qualitative data can be further divided into

- Binary data - only two yes/no categories.
- Nominal data - multiple categories that cannot be ranked against each other.
 - There is no point in asking for a bigger/smaller relationship.
- Ordinal data - multiple categories that can be sorted together.
 - It makes sense to ask about the bigger/smaller relationship.

Qualitative data - examples

- Binary data
 - diabetes (yes/no)
 - gender (male/female)
- Nominal data
 - blood groups (A/B/AB/O)
 - EU country (Belgium/.../Czech Republic/.../UK)
- Ordinal data
 - degree of pain (mild/medium/major/unbearable)
 - cigarette consumption (non-smoker/ex-smoker/occasional smoker/regular smoker)
 - stage of malignant disease (I/II/III/IV)

How information is created - description of different types of data

Wednesday statistics

Proportional data



AVERAGE

Continuous data

Interval data



MEDIAN

Ordinal data



MODUS

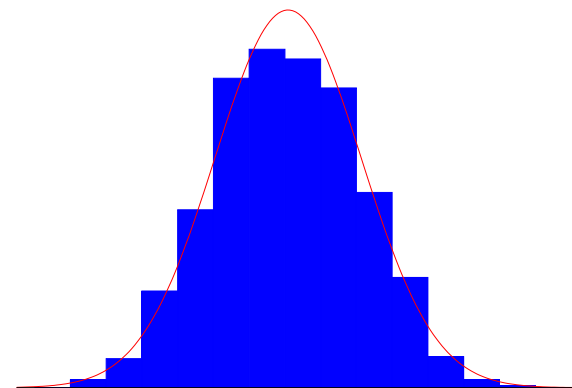
Discrete data

Nominal data

Binary data

Absolute and relative frequencies

- Quantitative data - frequency of distribution values in each interval.

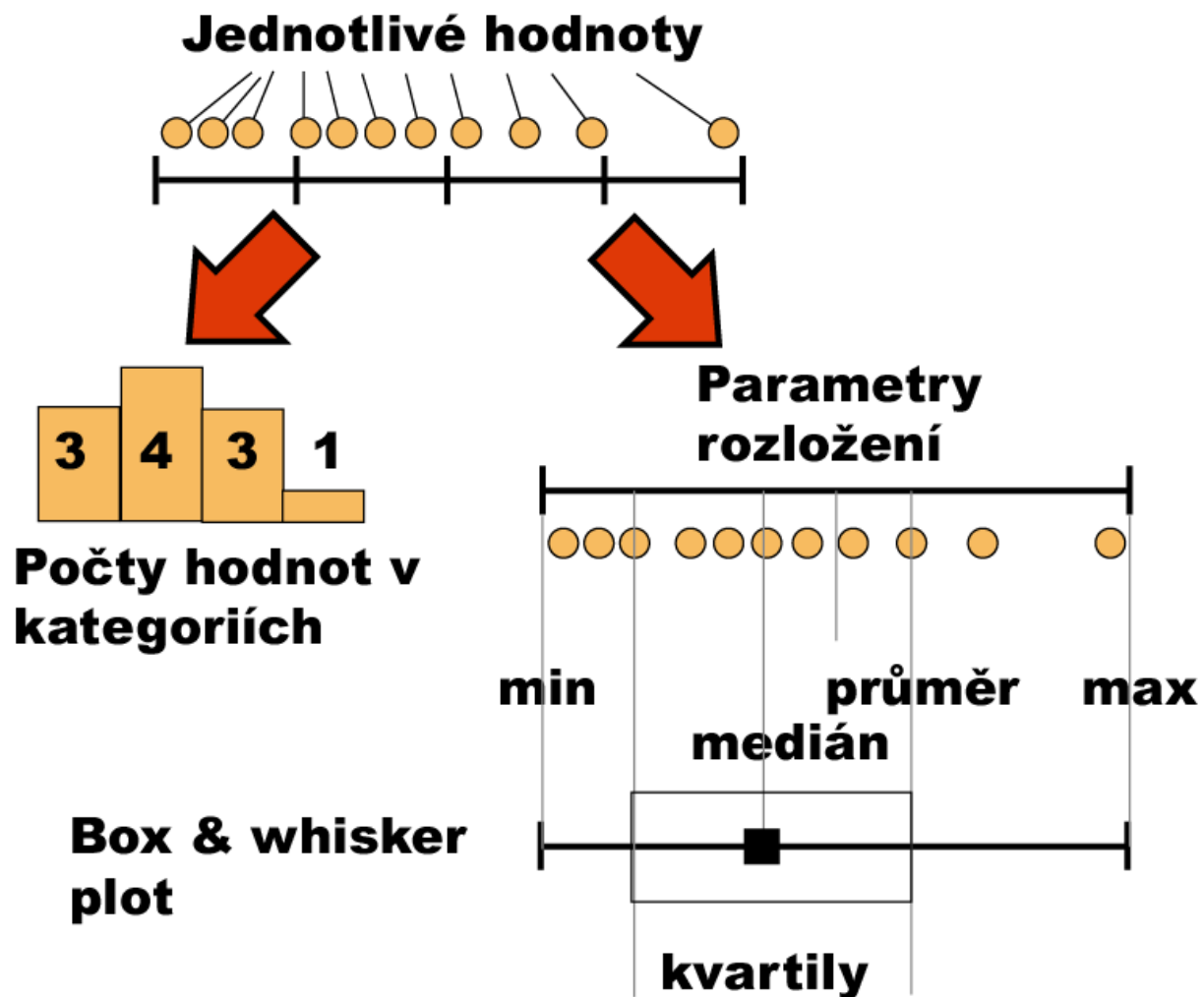


- Qualitative data - table with frequency of each category.

Category	Frequency
B	5
C	8
D	1

Data series and its properties

- There are often several possible ways to describe the data in the analysis
- The selection criterion is not only formal mathematical correctness, but also the meaningfulness and informational value of the descriptive statistics used in the given situation

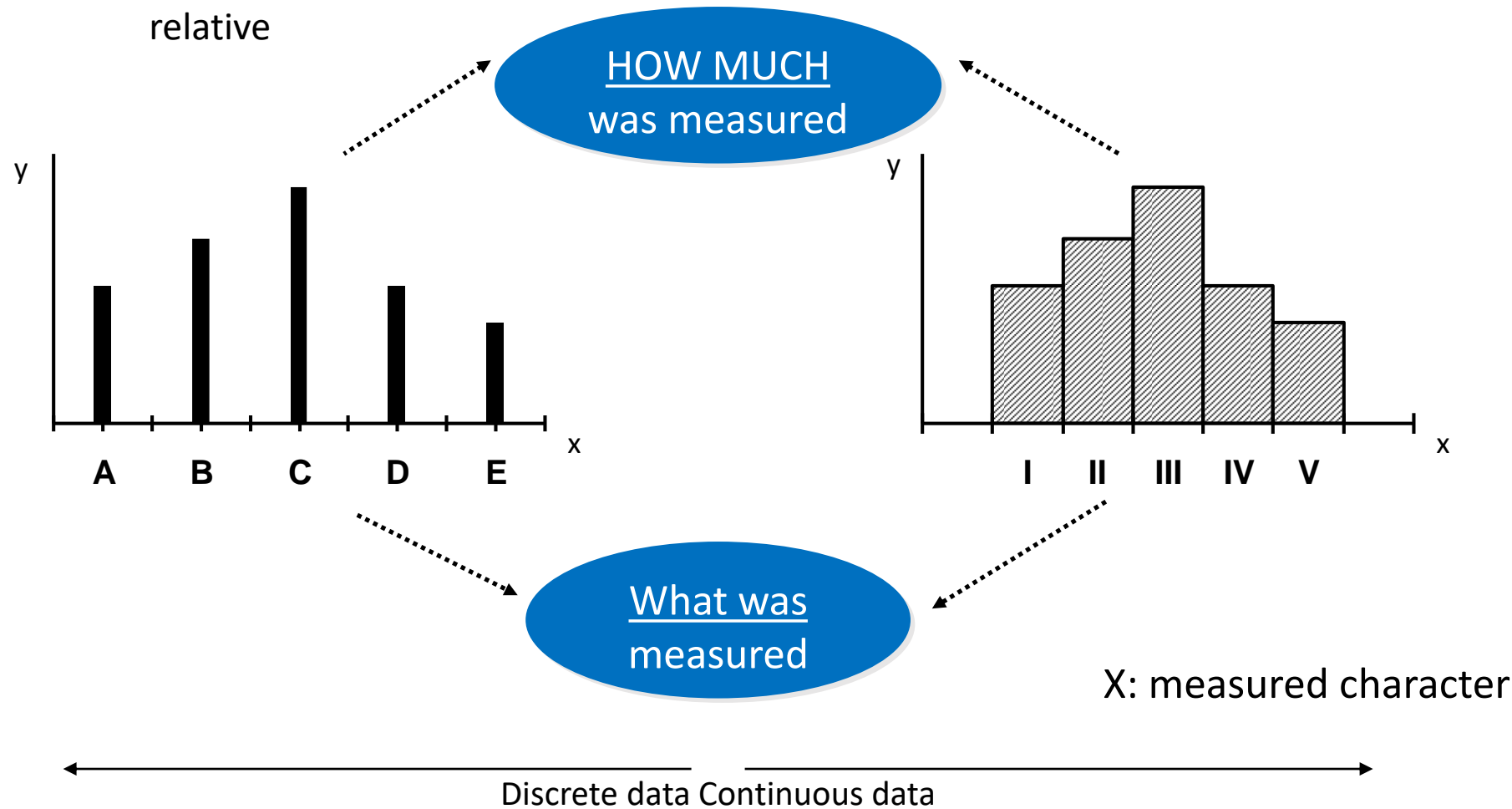


Derived data: beware of derived indices

- X: Average number of products in the shop
- Y: Estimate of the average space offered for product display
- Described by diameter and min-max range
 - X: 1,2 : (1,15 - 1,24) \longrightarrow + / - 3,8 %
 - Y: 1,8 : (1,75 - 1,84) \longrightarrow + / - 2,5 %
 - $\frac{X}{Y} = 0,667 : \left(\frac{1,15}{1,84} - \frac{1,24}{1,75} \right) \longrightarrow$ + / - 6,2 %
- The new quantity has a different range width than the ones from which it is derived

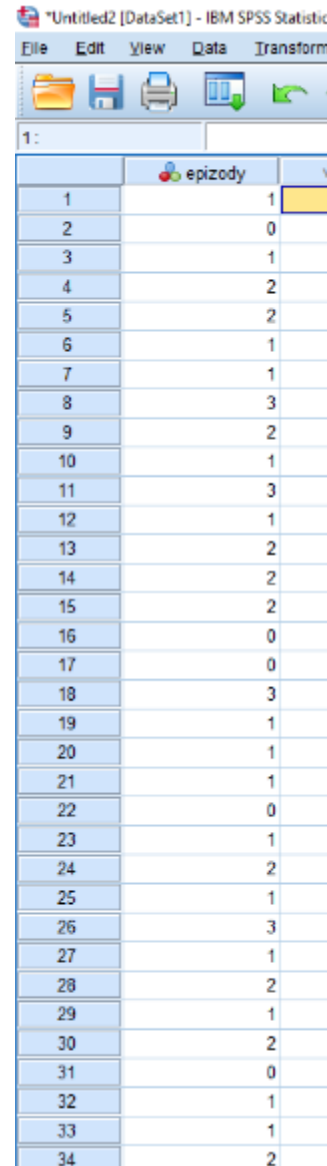
Information generation: repeated measurements inform the distribution of values

Y: frequency - absolute / relative

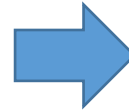


Frequency summarization - basic tool for data description: qualitative data

- The aim of summarisation is to simplify the data into a clear form
- N = 100 patients with haemophilia
- The variable assessed is the number of bleeding episodes per month
- The simplest summary is the frequency table



1:	epizody
1	1
2	0
3	1
4	2
5	2
6	1
7	1
8	3
9	2
10	1
11	3
12	1
13	2
14	2
15	2
16	0
17	0
18	3
19	1
20	1
21	1
22	0
23	1
24	2
25	1
26	3
27	1
28	2
29	1
30	2
31	0
32	1
33	1
34	2

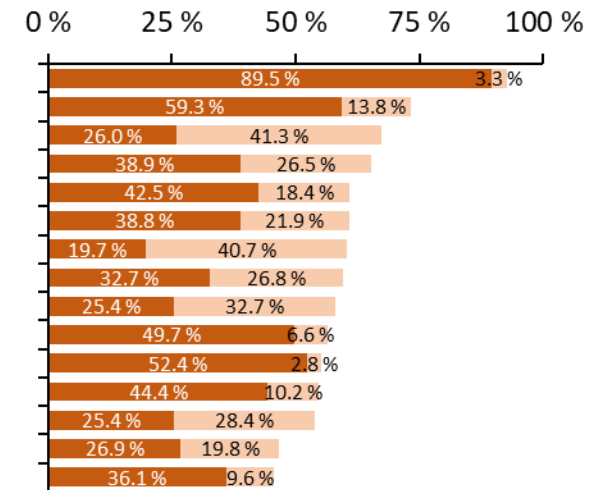
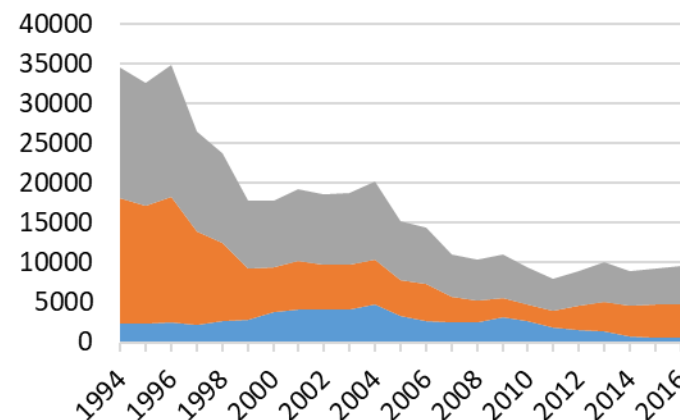
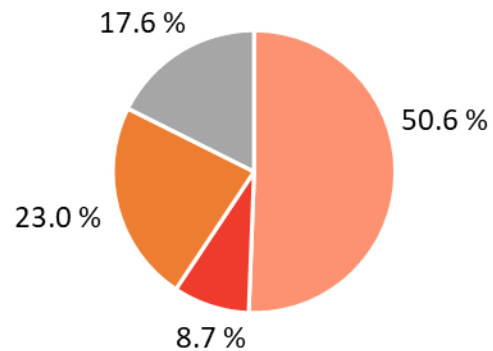
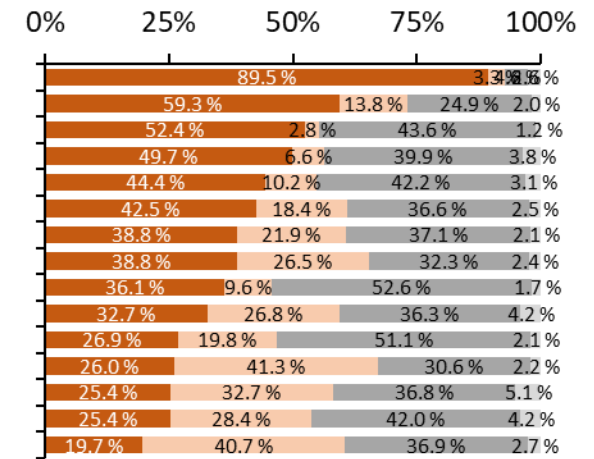
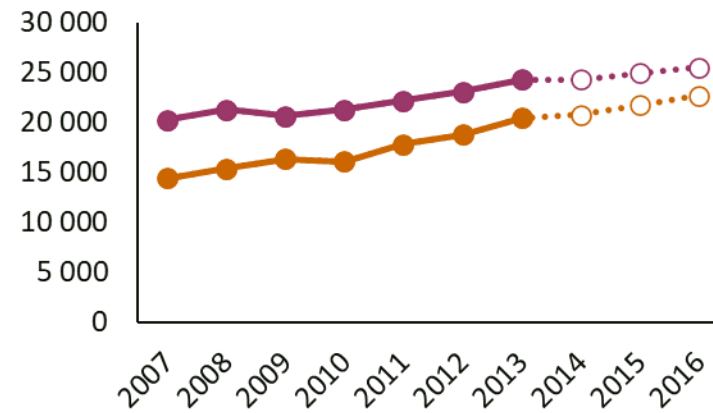
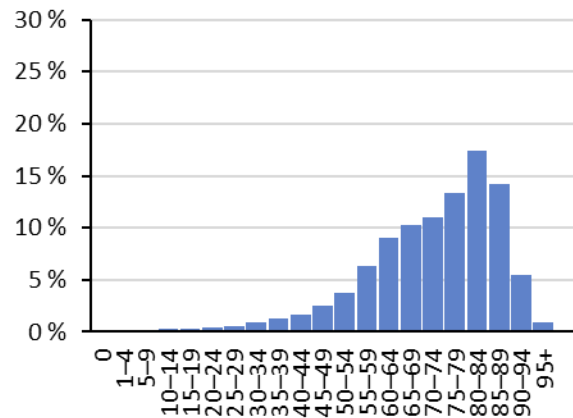


		epizody			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	22	22,0	22,0	22,0
	1	27	27,0	27,0	49,0
	2	29	29,0	29,0	78,0
	3	22	22,0	22,0	100,0
	Total	100	100,0	100,0	

- The table shows the unique values in the data
- **Frequency** = number of values in the category (absolute frequency)
- **Percent** = percentage of category (relative frequency)
- **Valid percent** = percentage of the category (not including missing values)
- **Cumulative percent** = cumulative percentage of categories up to a given category (cumulative relative frequency; only makes sense for ordinal data, similarly there is a cumulative absolute frequency)

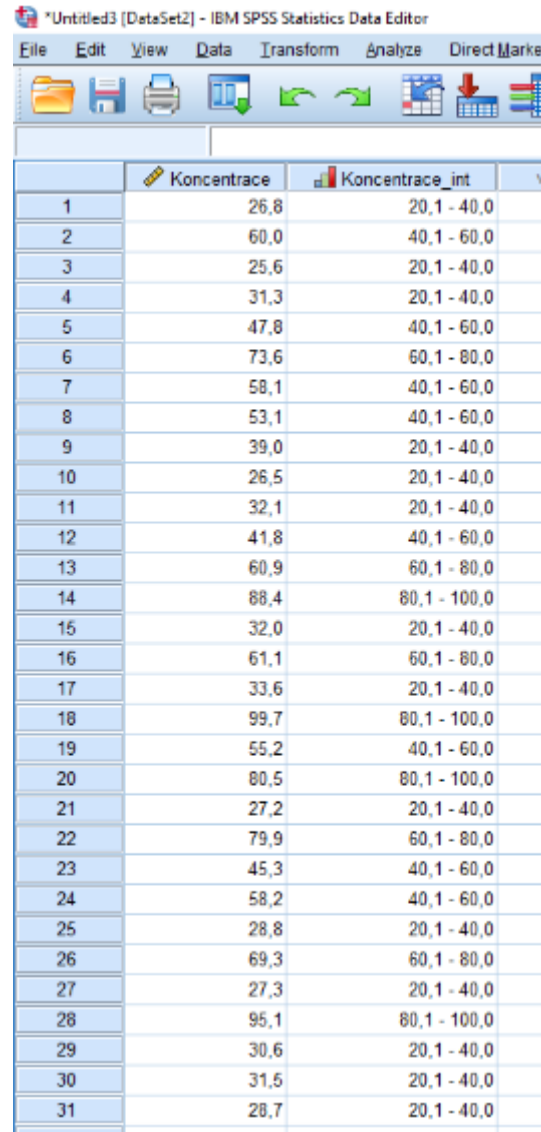
Visualization of frequency table of qualitative data

- Any graphs that allow visualization of counts and percentages (pie, bar, bar, line)



Frequency summarization - basic tool for data description: quantitative data

- The aim of summarisation is to simplify the data into a clear form
- N = 100 patients with
- The variable assessed is the concentration of the substance in the blood
- The simplest summary is again the frequency table
- Another option is to calculate proxy summary statistics (mean, median, etc.)



	Konzentrace	Konzentrace_int
1	26,8	20,1 - 40,0
2	60,0	40,1 - 60,0
3	25,6	20,1 - 40,0
4	31,3	20,1 - 40,0
5	47,8	40,1 - 60,0
6	73,6	60,1 - 80,0
7	58,1	40,1 - 60,0
8	53,1	40,1 - 60,0
9	39,0	20,1 - 40,0
10	26,5	20,1 - 40,0
11	32,1	20,1 - 40,0
12	41,8	40,1 - 60,0
13	60,9	60,1 - 80,0
14	88,4	80,1 - 100,0
15	32,0	20,1 - 40,0
16	61,1	60,1 - 80,0
17	33,6	20,1 - 40,0
18	99,7	80,1 - 100,0
19	55,2	40,1 - 60,0
20	80,5	80,1 - 100,0
21	27,2	20,1 - 40,0
22	79,9	60,1 - 80,0
23	45,3	40,1 - 60,0
24	58,2	40,1 - 60,0
25	28,8	20,1 - 40,0
26	69,3	60,1 - 80,0
27	27,3	20,1 - 40,0
28	95,1	80,1 - 100,0
29	30,6	20,1 - 40,0
30	31,5	20,1 - 40,0
31	28,7	20,1 - 40,0

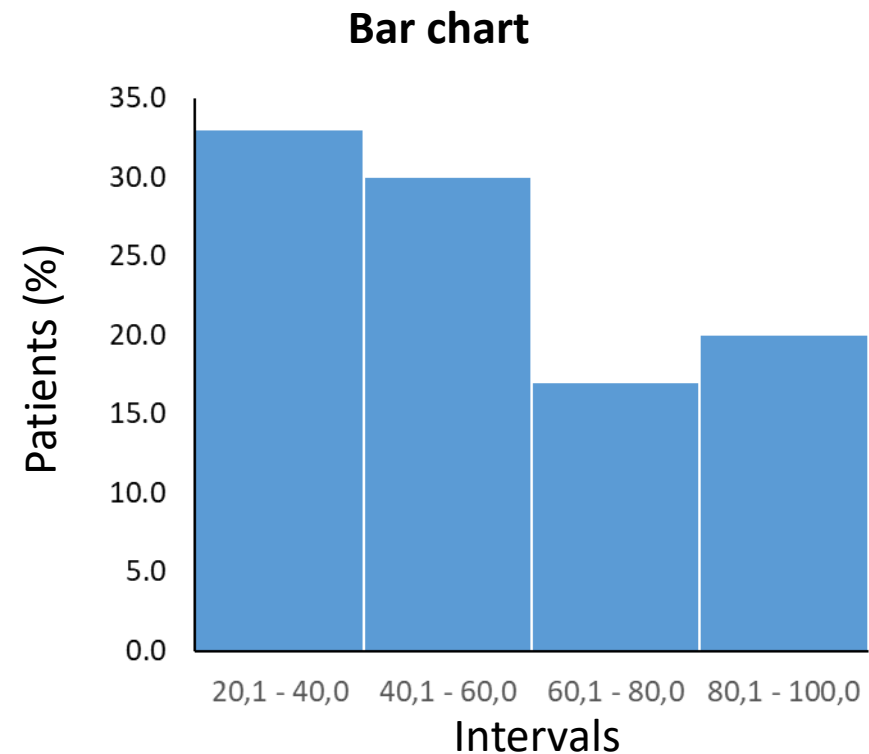
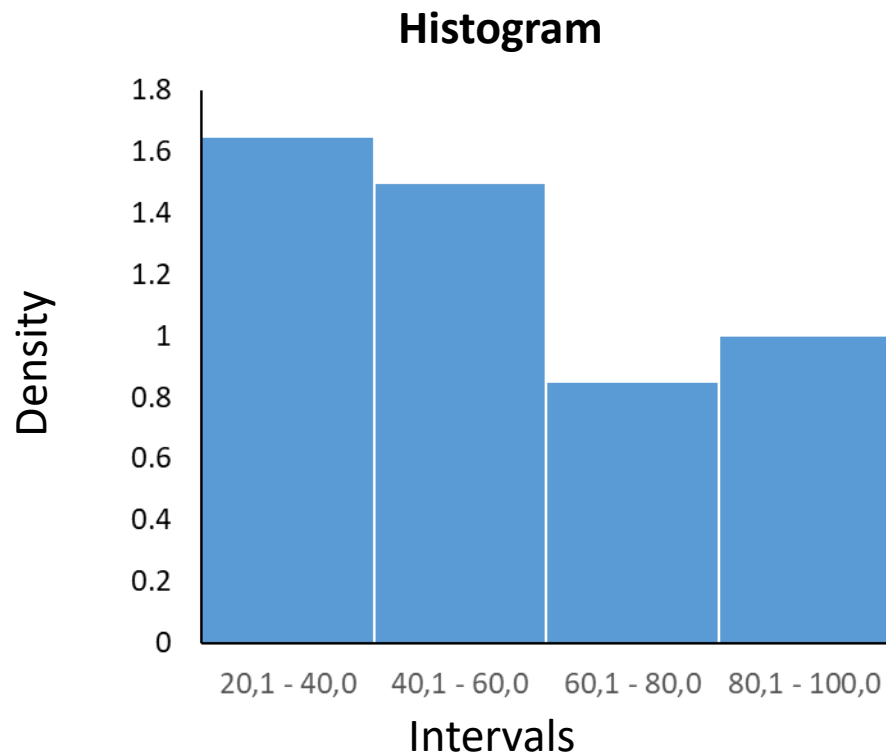


Konzentrace intervaly					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	20,1 - 40,0	33	33,0	33,0	33,0
	40,1 - 60,0	30	30,0	30,0	63,0
	60,1 - 80,0	17	17,0	17,0	80,0
	80,1 - 100,0	20	20,0	20,0	100,0
	Total	100	100,0	100,0	

- The table shows the unique values in the data
- In contrast to qualitative data, it is essential for the meaningfulness of the output to define intervals (of equal or different widths) in the data
- **Frequency** = number of values in the category (absolute frequency)
- **Percent** = percentage of category (relative frequency)
- **Valid percent** = percentage of the category (not including missing values)
- **Cumulative percent** = cumulative percentage of categories up to a given category (cumulative relative frequency; similarly, there is a cumulative absolute frequency)

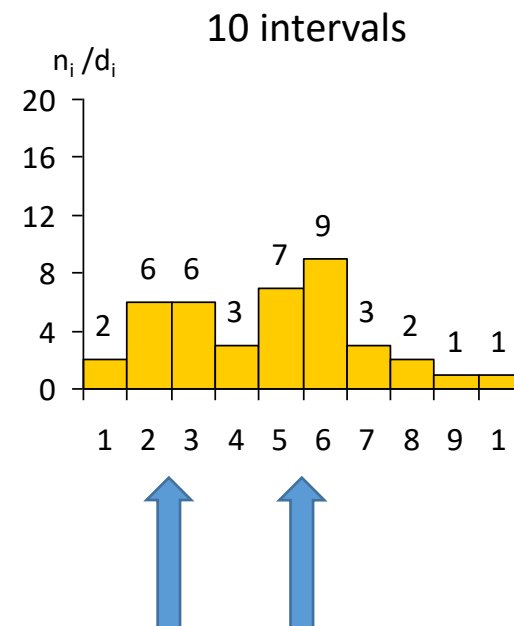
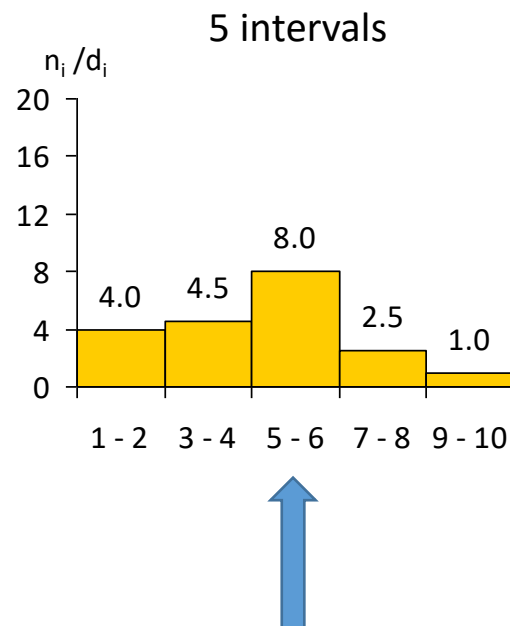
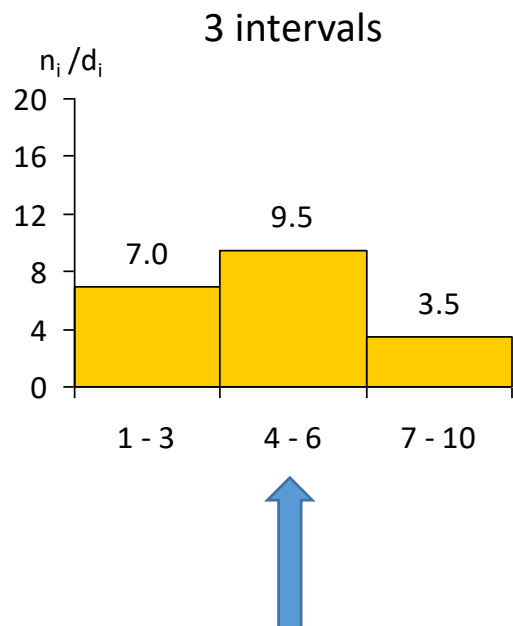
Visualization of frequency table of quantitative data

- The basic tool for visualizing continuous data based on a frequency table is the histogram
- Unlike a bar chart, the visualized value is the area of the bar, not its height



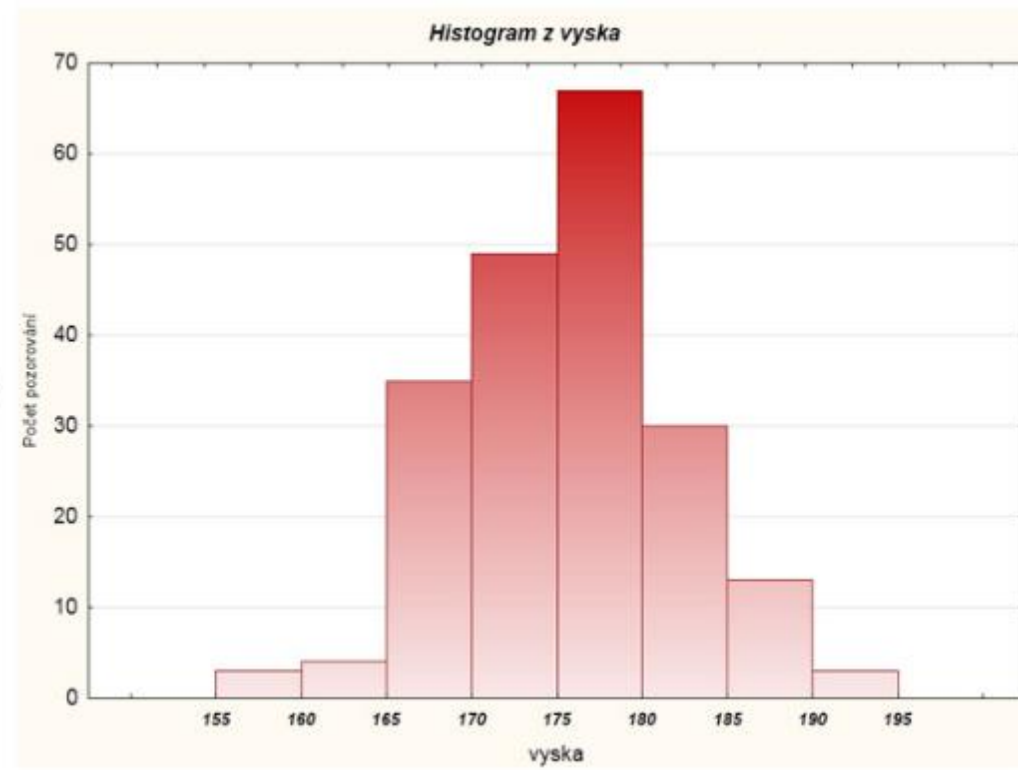
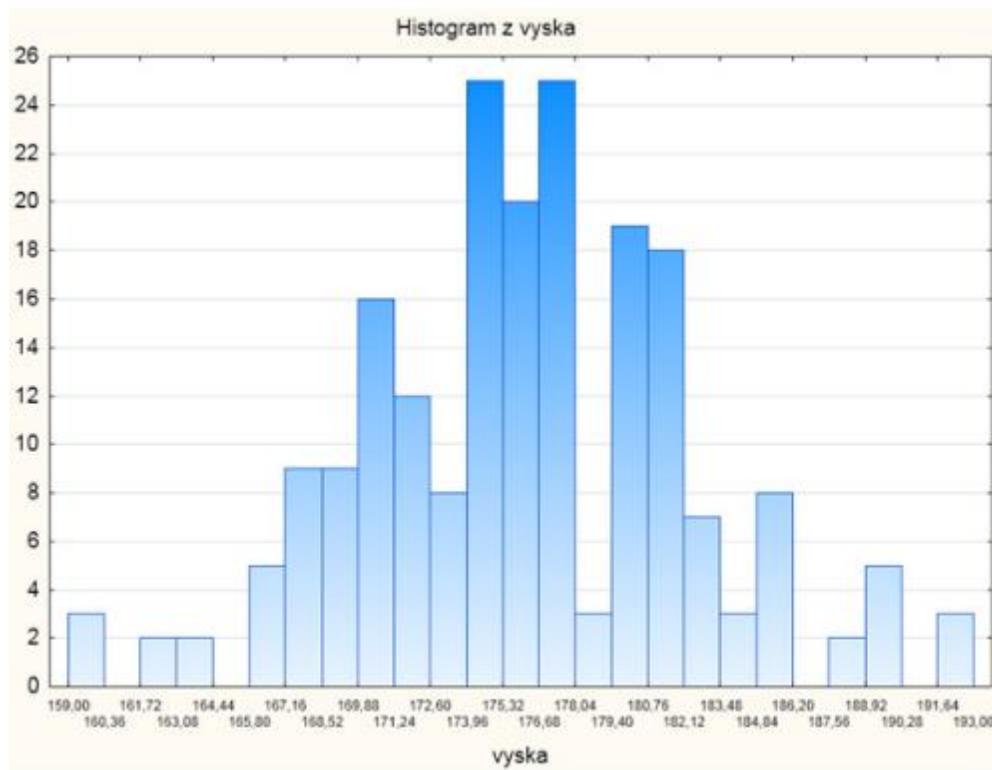
Histogram: effect of data categorization

- The number of selected intervals in the histogram determines how it will look. If the number is small we may miss important elements in the data, if it is large the information may be fragmented.



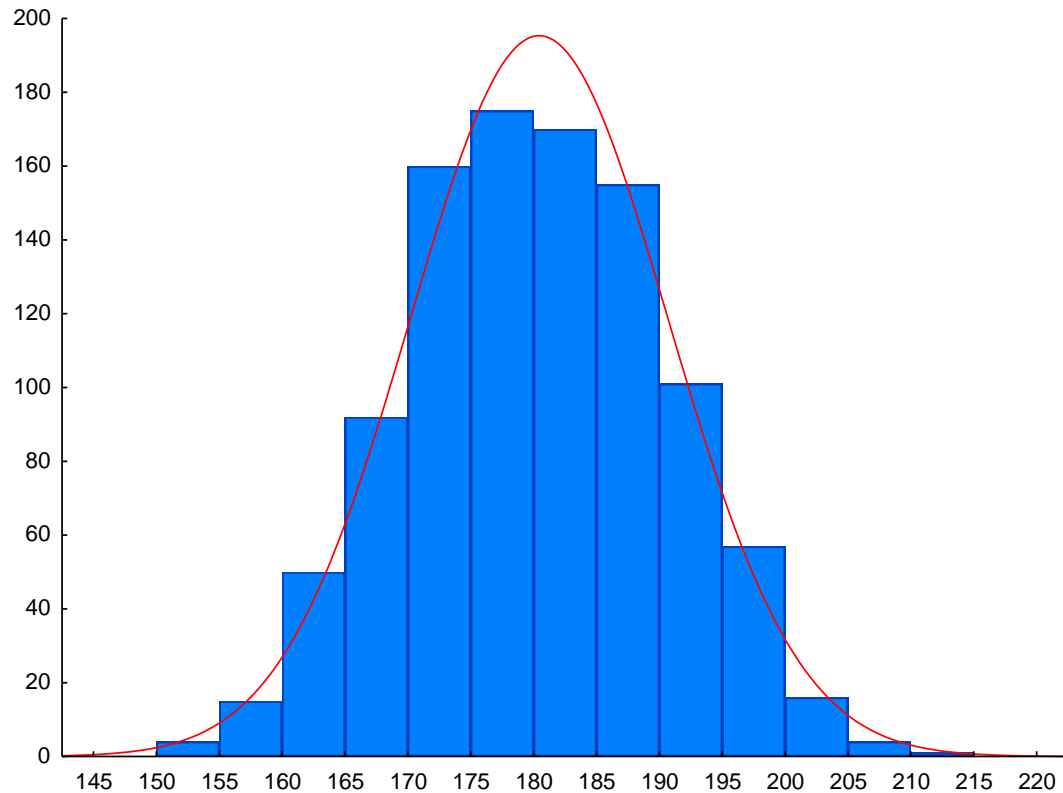
Histogram: effect of data categorization

- Choosing the number of categories - important for interpretation
- Manual or automatic selection - different algorithms (depends on sample size and data variability)

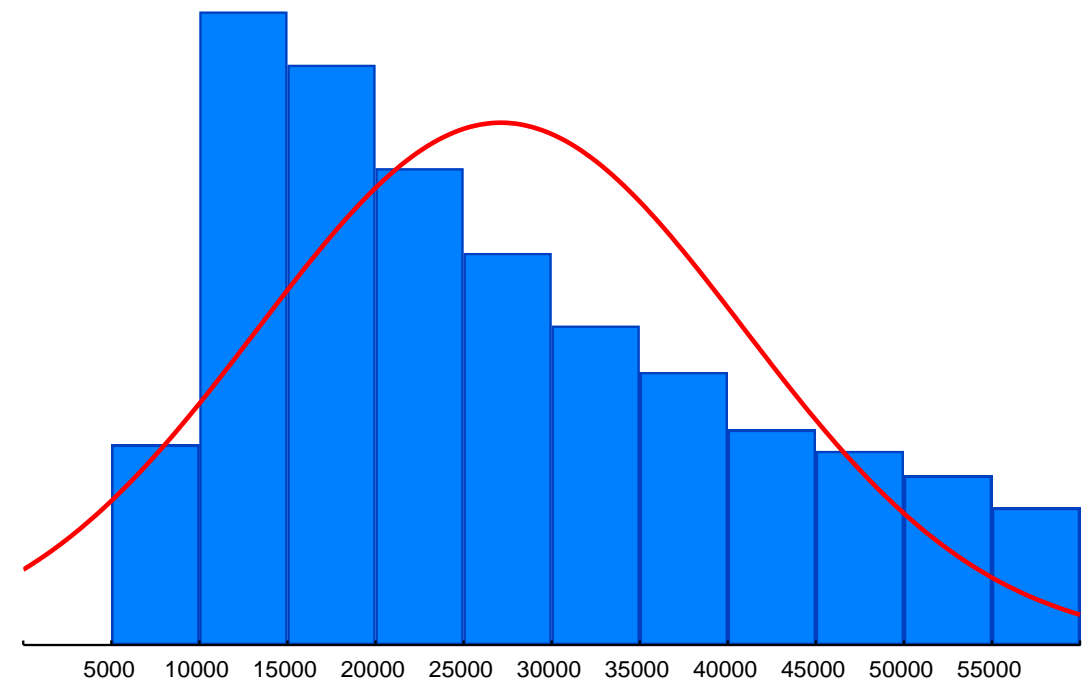


Histogram: a tool to assess the distribution of data

- Histogram of real data is related to the model distribution



?



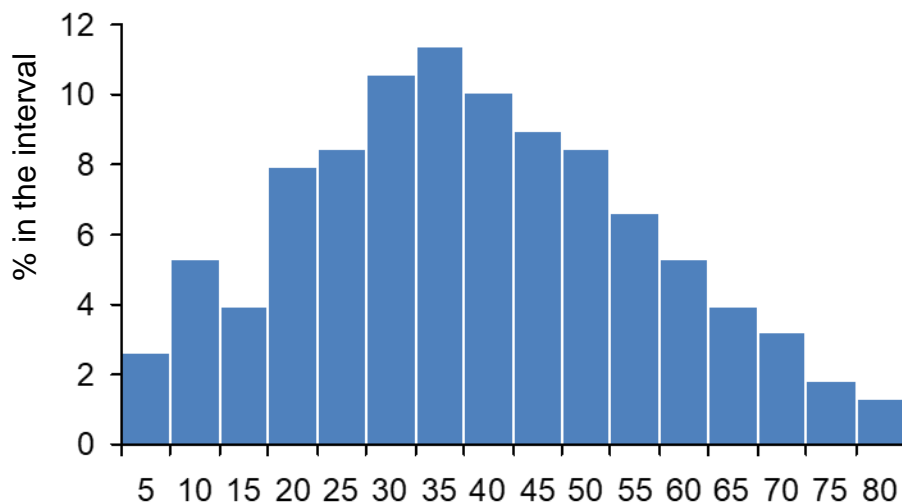
Why it is important to know what a real histogram is I

- Most people think visually - data visualisation is so crucial for first perception and interpretation of data
- Due to the different visual interpretation of the histogram and bar chart when using differently wide intervals, using a bar chart may be misleading in some situations
- In practice, however, a bar chart is often used instead of a "true" histogram (even by statistical software manufacturers)
- In the case of the same interval width, the interpretation problem does not arise (with different interval widths, SW disables some options = settings for advanced users)

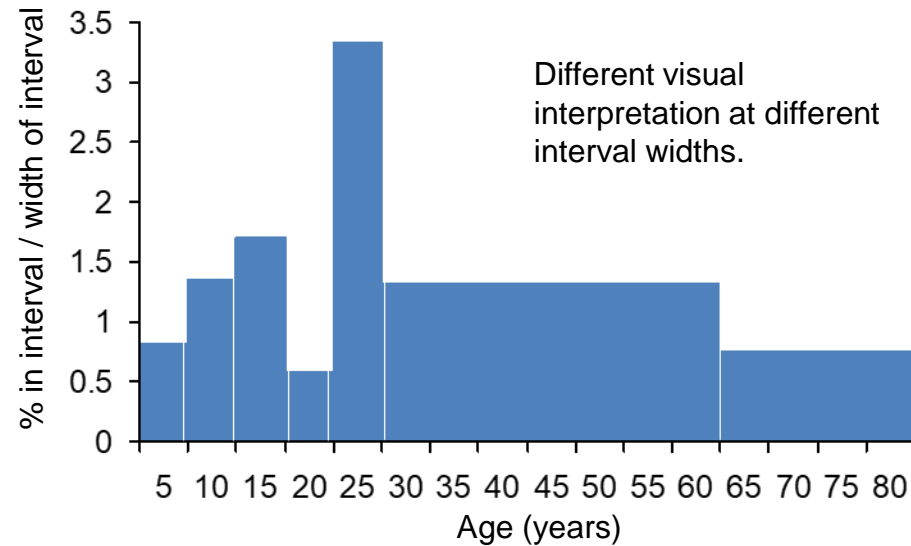
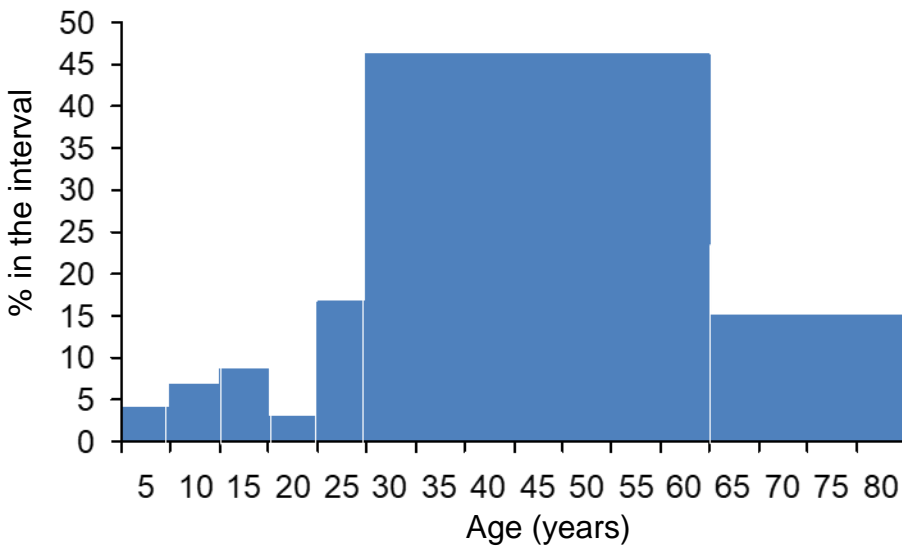
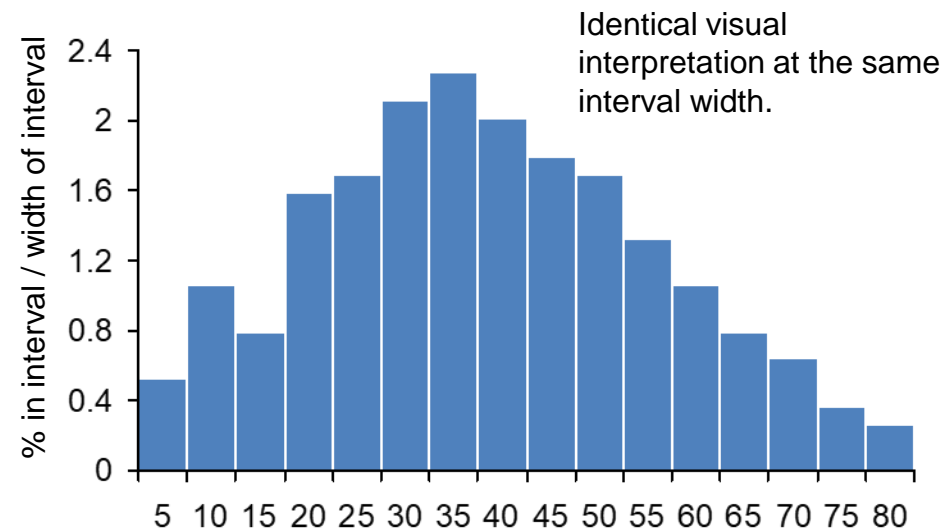


Histogram and bar chart

Bar chart

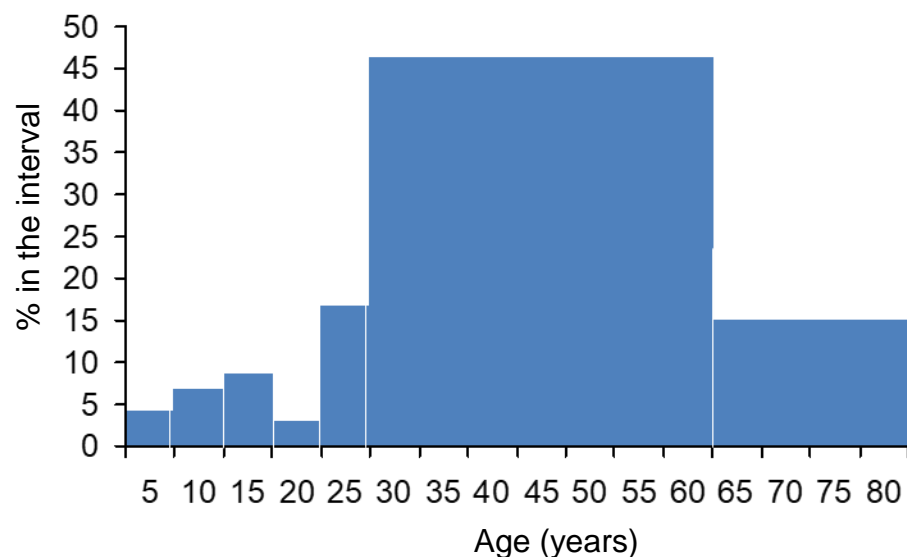


Histogram

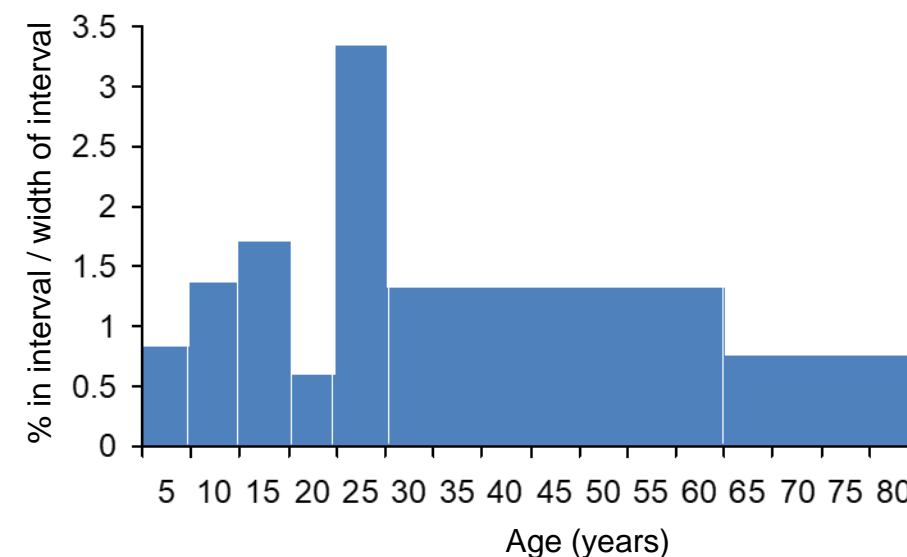


Example: age of those involved in serious road accidents

- The ages of those involved in serious road accidents in one London borough were analysed
- Does the interpretation of data visualized with a bar graph and a histogram differ?
- Which interpretation makes more sense to you and why?



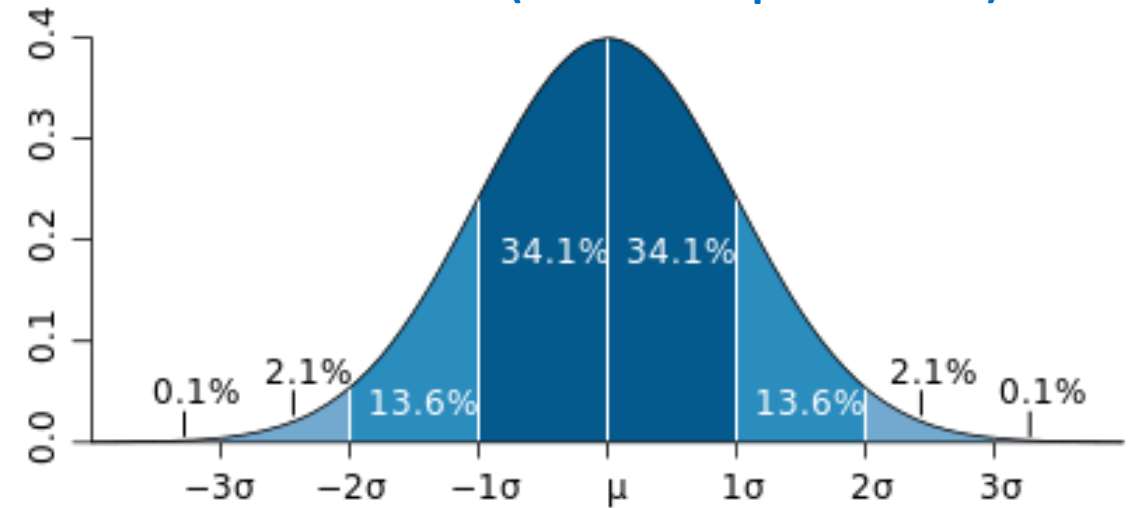
Age	N	%
0 - 4	28	4,1%
5 - 9	46	6,7%
10-15	58	8,5%
16 - 19	20	2,9%
20 - 24	114	16,6%
25 - 59	316	46,1%
> 60	103	15,0%



Why it is important to know what a real histogram is II

- Statistical analyses are based on model distributions, which we use in calculations as a proxy for measured data (if the real data fits the model distribution, we can use the model in calculations instead)
- The models describe the probability density distribution of the occurrence of a given value = the probability of occurrence of values is given by the area of the graph
- **Distribution** = real data
- **Division** = model

Area = probability of occurrence
Sum of area = 1 (100% of all possibilities)



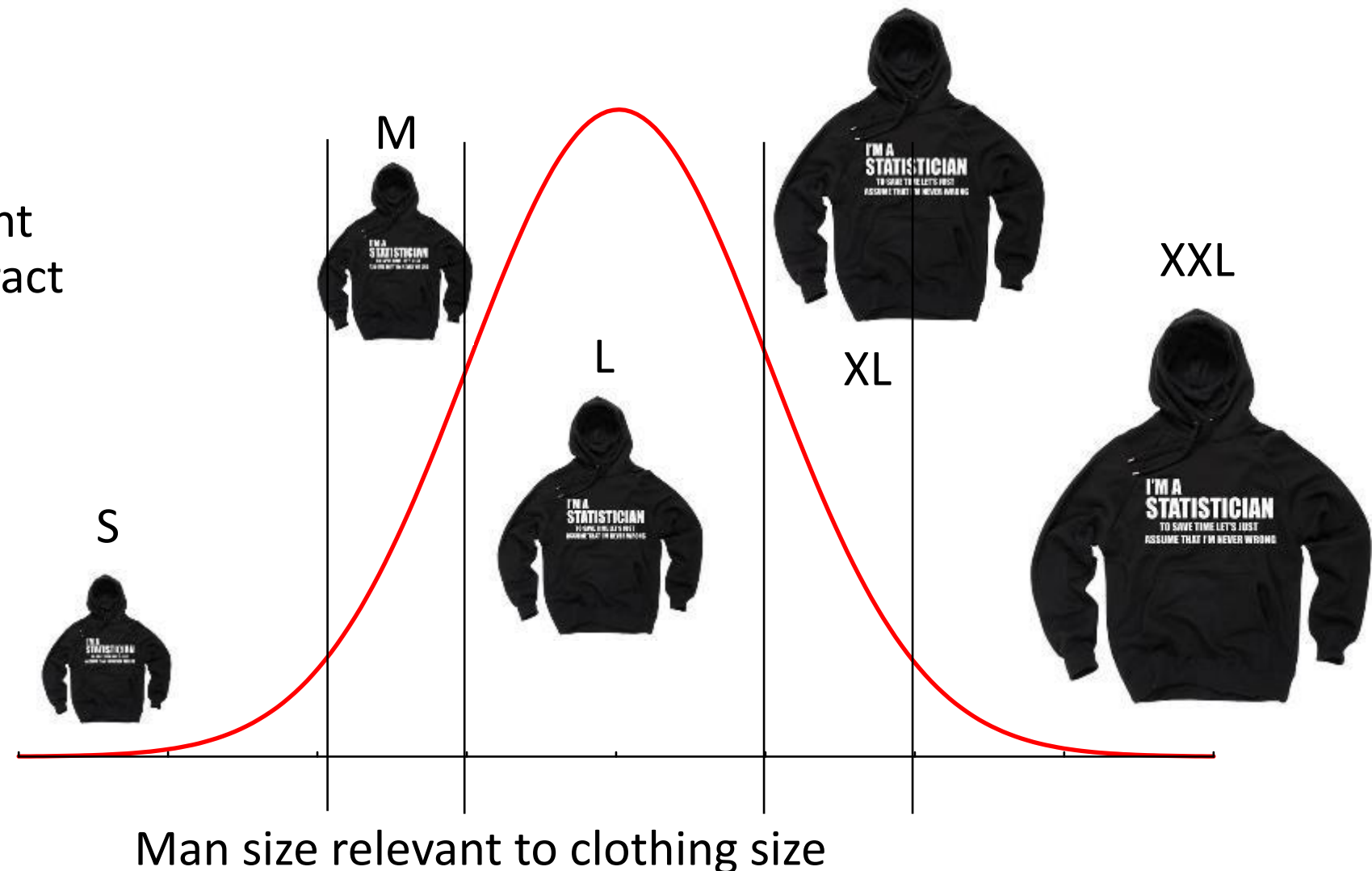
Example: optimising clothing inventories

- Imagine you own a clothing store and you want to optimize the stock of different sizes of clothes = you need to find out what % of people in the population need what clothes
- What is the size distribution of people in the population?
- Uniform, normal, lognormal ????



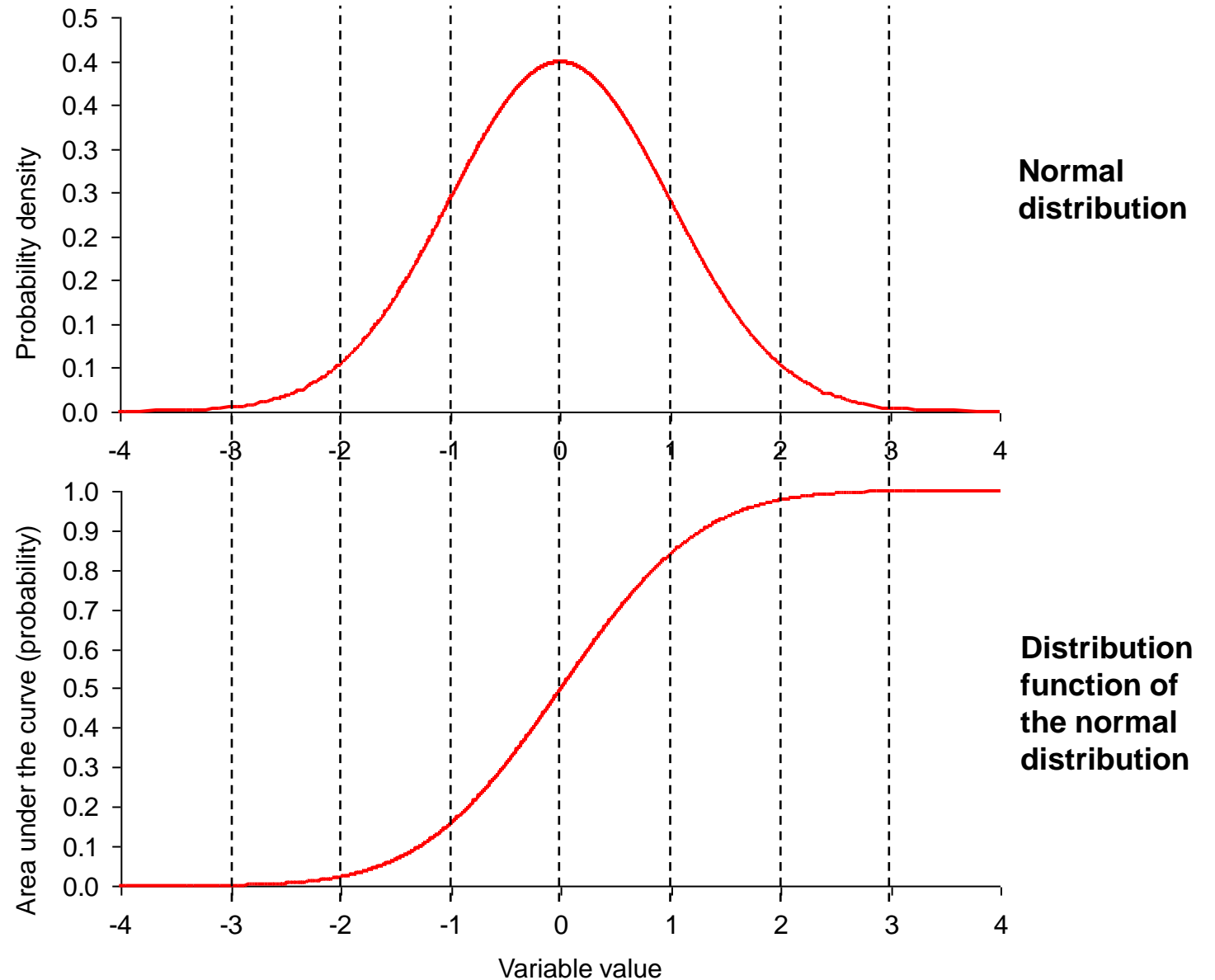
Example: optimising clothing inventories

- It can be assumed that the size of people is normally distributed
- If we are able to determine ranges of values for different clothing sizes, we can subtract the stock ratios from the normal distribution curve
- Integrate?
- Could it be easier?



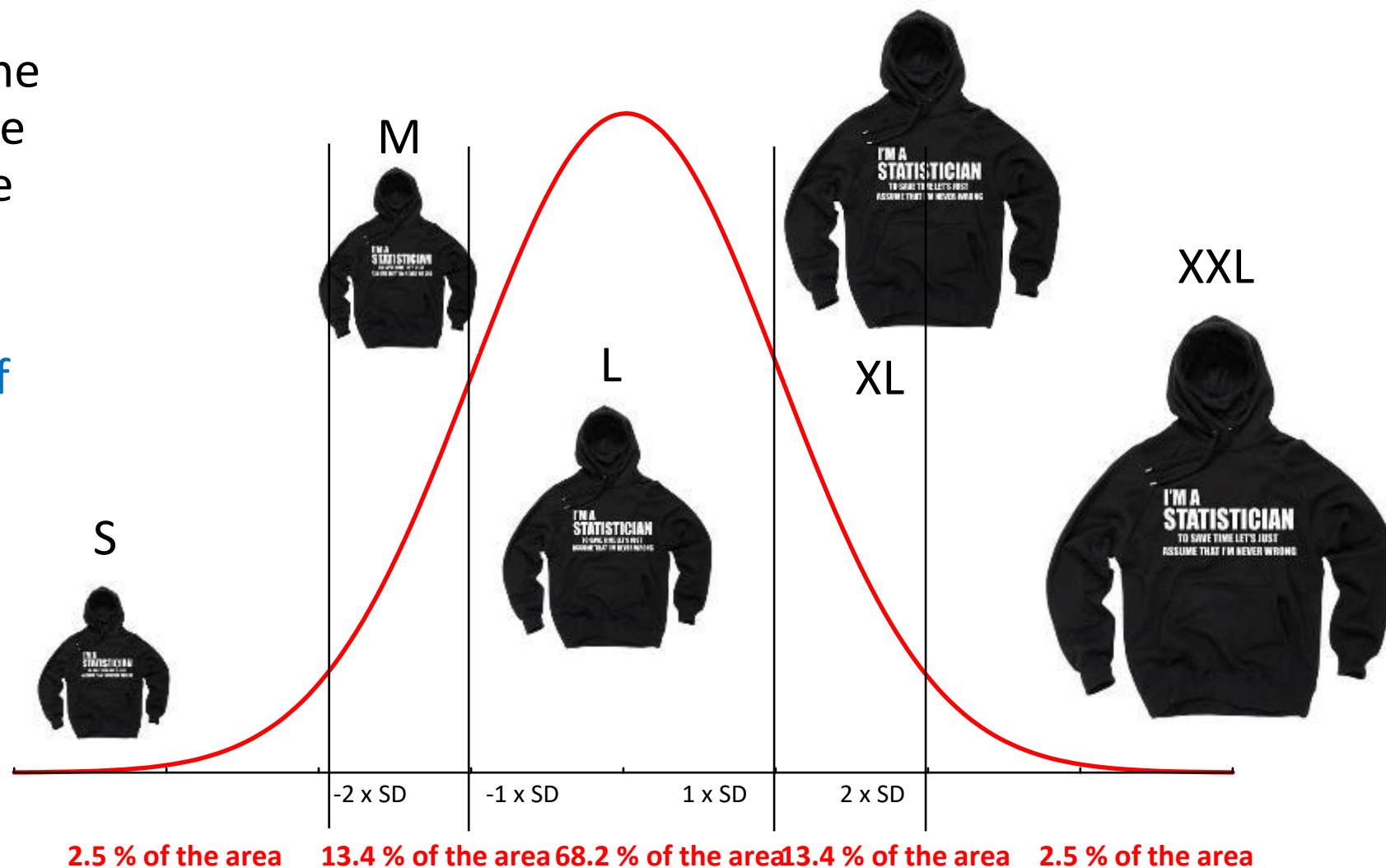
Normal distribution and its distribution function

- There are distribution functions for the model distributions
- For a given value of the distribution, they give the area (=probability) under the curve up to that value
- A basic tool in many statistical calculations
- **Quantile of the model distribution:** the value to which a given area under the distribution curve corresponds (e.g. 95% quantile is the value of the variable under which 95% of all values lie)



Example: optimising clothing inventories

- We derive the solution to the example by knowing the size distribution of people in the target population and its distribution function
- Approximate proportions of different clothing sizes:
 - S: 2.5%
 - M: 13.4%
 - L: 68.2%
 - XL: 13.4%
 - XXL: 2.5%



Man size relevant to clothing size