

Experimental Humanities II (HUMB002) 2016
STATISTICAL ANALYSIS

Lecture 1

VARIABLES, FREQUENCIES,
DISTRIBUTIONS
MEASURES OF CENTRAL TENDENCY AND
VARIABILITY

Pavla Linhartová

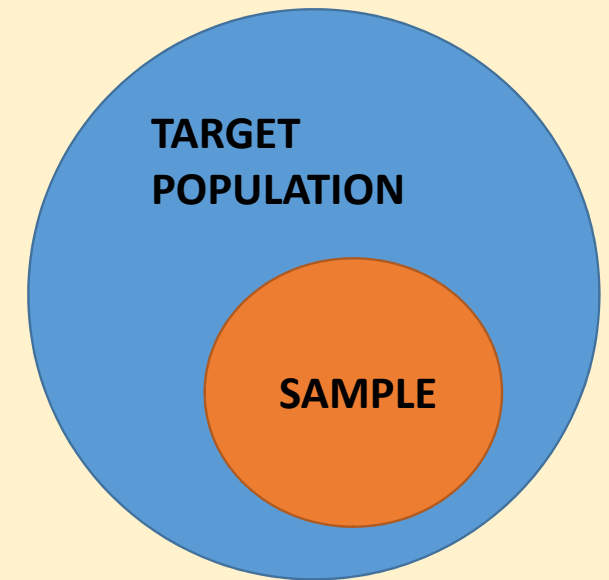
The lectures and exercises are based on the lectures from the subject PSY117 – Statistical analysis by Stanislav Ježek and Jan Širůček from Department of Psychology, Faculty of Social Studies MU Brno

Course introduction

- Lectures + Exercises (basic computations, Excel, SPSS, Statistica)
- 2 tests (20 + 20 points)
- 1 seminar thesis (15 points)
- Final exam project (45 points)

Data and measurement

- Target population X Sample
- Sample must match population characteristics, so that we can infer data obtained from the sample on the population
- Data are some information measured in the sample
- Variables are coded data
- There is always some measurement error:
$$Y = T + e$$
 (obtained score = true score + error)
- There are various measurement error sources



Statistics for (experimental) research

- Descriptive statistics (data description and visualisation)
 - Get to know your data in detail
 - Present your data clearly and correctly
 - What can we say about the explored phenomena?
 - What can we say about relations between them?
- Statistical inference (inference on population from a sample)
 - How does your model (assumption) fit the data?
 - Hypothesis testing, deriving estimates

Statistics for (experimental) research

- Statistics is „only“ a tool (a powerful one)
 - GARBAGE IN, GARBAGE OUT: output quality depends on input quality
 - STATISTICS IN NOT ENOUGH: quality input = quality research methodology (valid and reliable methods, study design, data acquisition and sample selection)
 - CORRECT USE OF STATISTICAL TOOLS: tests assumptions, correct choice of statistical tests, correct analysis process
 - RESULTS INTERPRETATION: understanding the results, correct results presentation, awareness of results (analysis, data) strengths and weaknesses

Statistics for life

- Critical thinking
- Probability thinking
- Understanding measurement and presentation biases
- Interpreting statistical reports around you (news, television, internet...)
- ...statistical literacy

VARIABLE TYPES AND LEVELS OF MEASUREMENT

Variables: Levels of measurement

	Level of measurement	Possible operations	Examples
1	NOMINAL (nominální)	$= \neq$	colour, tram numbers
2	ORDINAL (ordinální, pořadová)	$= \neq > <$	school grades, agreement
3	INTERVAL (intervalová)	$= \neq > < + -$	temperature, IQ, year
4	RATIO (poměrová)	$= \neq > < + - \times \div$	weight, frequency, age

1 + 2 = categorical, qualitative, 3 + 4 = metric, cardinal, quantitative

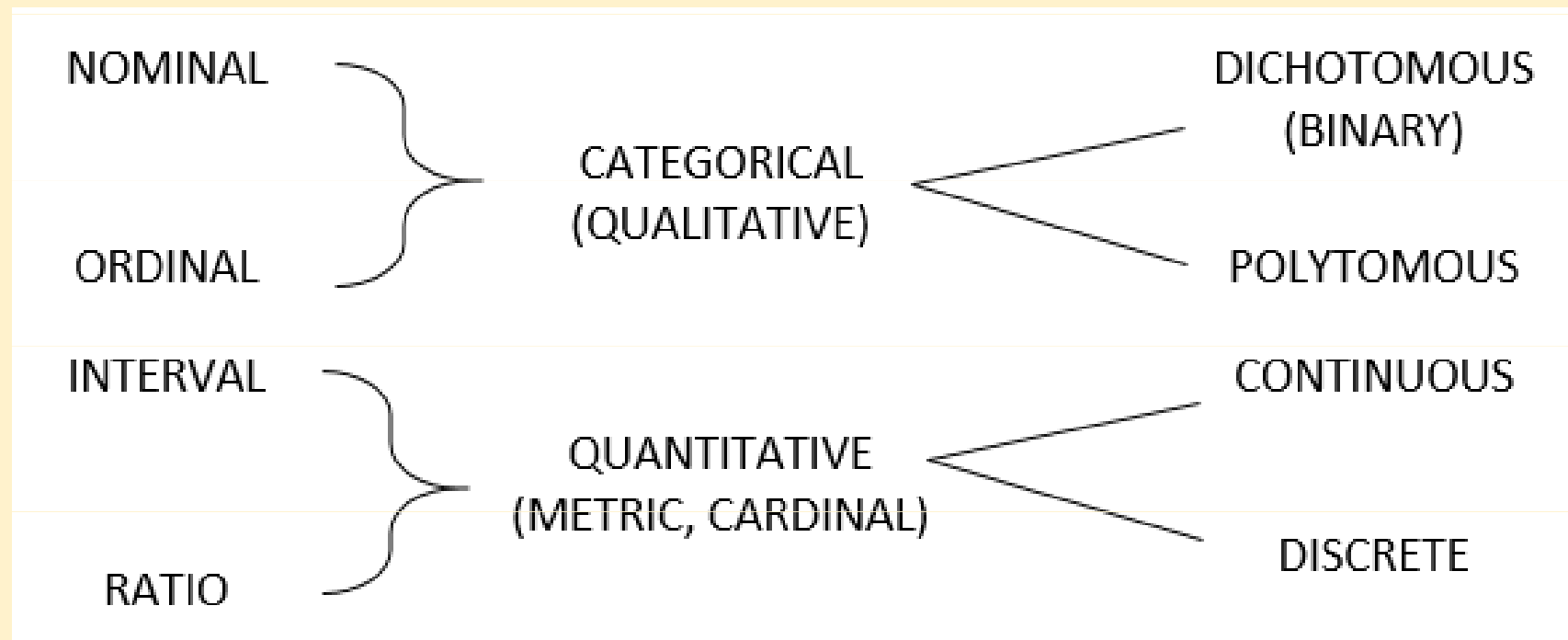
Considering level of measurement

- When you consider variable measurement level, following question can be useful. Let's take as an example „eye colour“:
 1. Does George have the same eye colour as Caroline?
 2. Does George have bigger eye colour than Caroline?
 3. How much bigger is George's eye colour than Caroline's?
 4. How many times bigger is George's eye colour than Caroline's?



Variables according to number of possible values

- **Continuous variables:** infinite number of values (real numbers)
- **Discrete variables:**
 - (infinite) number of values, but only some values (typically integers)
 - we usually treat them as continuous
 - Only a few values
 - Dichotomous (binary) – only two possible values
 - Polytomous – several values

Variables: summary

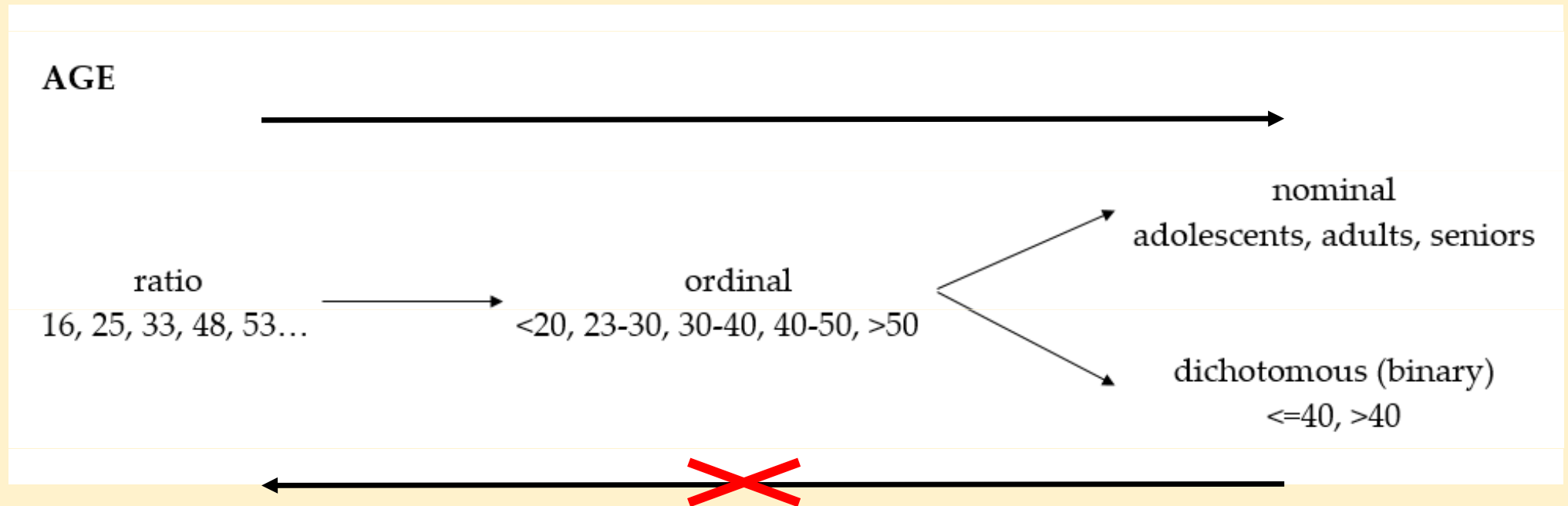


Level of measurement conversion



	Level of measurement	Possible operations	Examples
1	NOMINAL (nominální)	$= \neq$	colour, tram numbers
2	ORDINAL (ordinální, pořadová)	$= \neq > <$	school grades, agreement
3	INTERVAL (intervalová)	$= \neq > < + -$	temperature, IQ, year
4	RATIO (poměrová)	$= \neq > < + - \times \div$	weight, frequency, age

Level of measurement conversion



Treating variable types

- Variable type depends on HOW we measure, not what we measure!
- Measure on the highest level possible, you can always transfer to the lower levels, not the other way...
- Variables measured on discrete ordinal polytomous scale are in psychology and other sciences usually considered as interval continuous variables, e.g. attitude scales:
 - 1=totally disagree, 2=somewhat disagree, 3=neither agree, nor disagree, 4=somewhat agree, 5=totally agree
- We usually try to achieve interval variable type (if it's reasonable)
 - > easier statistics

Data matrix

		VARIABLES		
SUBJECTS	ID	Gender	Age	...
	1	male	25	
	2	male	34	
	3	female	32	
	4	female	29	
	5	male	32	
	

- 1 column = 1 variable, 1 row = 1 subject (object)
- Variable has either string, or numeric coding
- Numeric coding + labels
- Short variable names + labels

Coding issues

- Repeated measurement
 - pre-test and post-test scores have to be coded as individual variables, not individual cases!
- Forced choice answer formats
 - 1 question = 1 variable, 1 answer possibility = 1 value
- Multiple choice answer formats
 - answer possibilities have to be coded as individual variables with values 1 = checked, 0 = not checked

View Data matrix handout and Coding rules in study materials

FREQUENCIES AND DISTRIBUTIONS

Frequencies

- How many values we have in different variables?
- How many people answered this way?
- What was the most frequent answer? etc.
- Frequency tables and frequency graphs

Frequency tables

Value (Interval)	Absolute frequency	Cumulative absolute frequency	Relative frequency (%)	Cumulative relative frequency
Minimum (Interval 1)				
Value 2 (Interval 2)				
Value 3 (Interval 3)				
...				
Maximum (the last interval)				
Missing		N		100%
Total	N		100%	

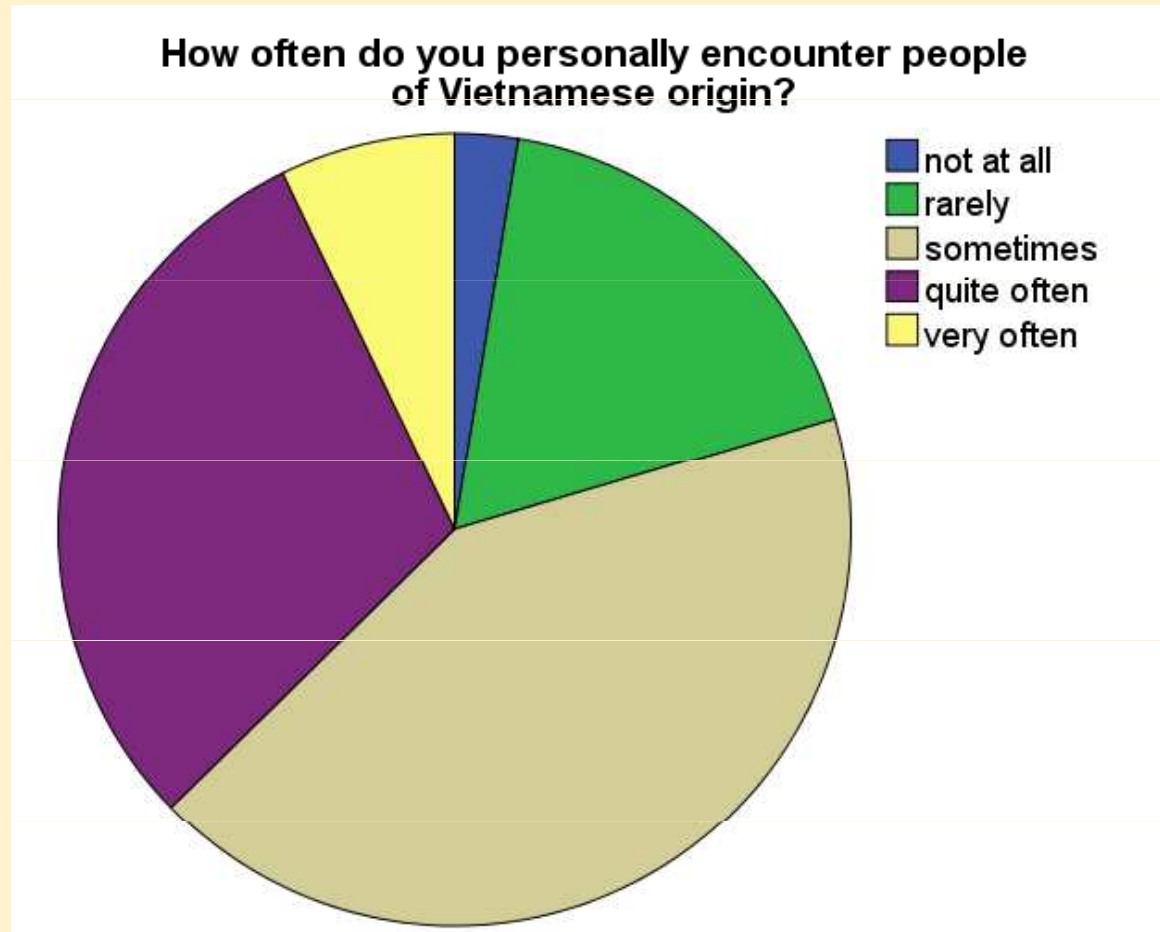
Frequency tables

- Ordered from lowest to highest values
- Usually include missing values
- Suitable for categorical variables or metric variables with a few values
- We can use intervals instead of values
- It is better to use graphs for displaying frequencies of variables with many values
- Frequency table has to always include absolute and relative frequencies and the last summing row

Frequencies visualisation

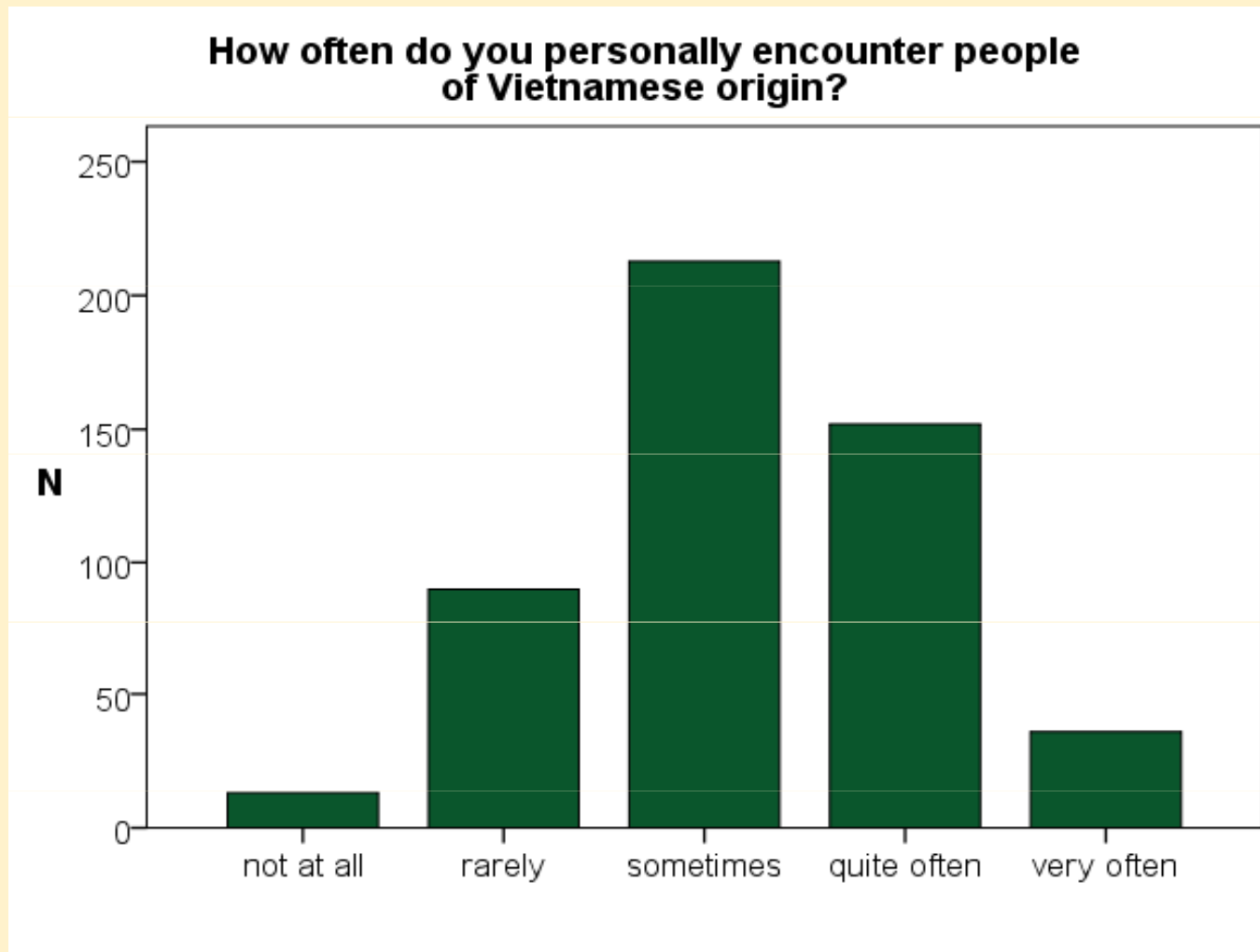
- Categorical data
 - Pie chart – used rarely
 - **Bar chart**
- Metric data
 - **Histogram**
 - **Boxplot**
 - Stem-and-leaf diagram

Categorical data: Pie chart



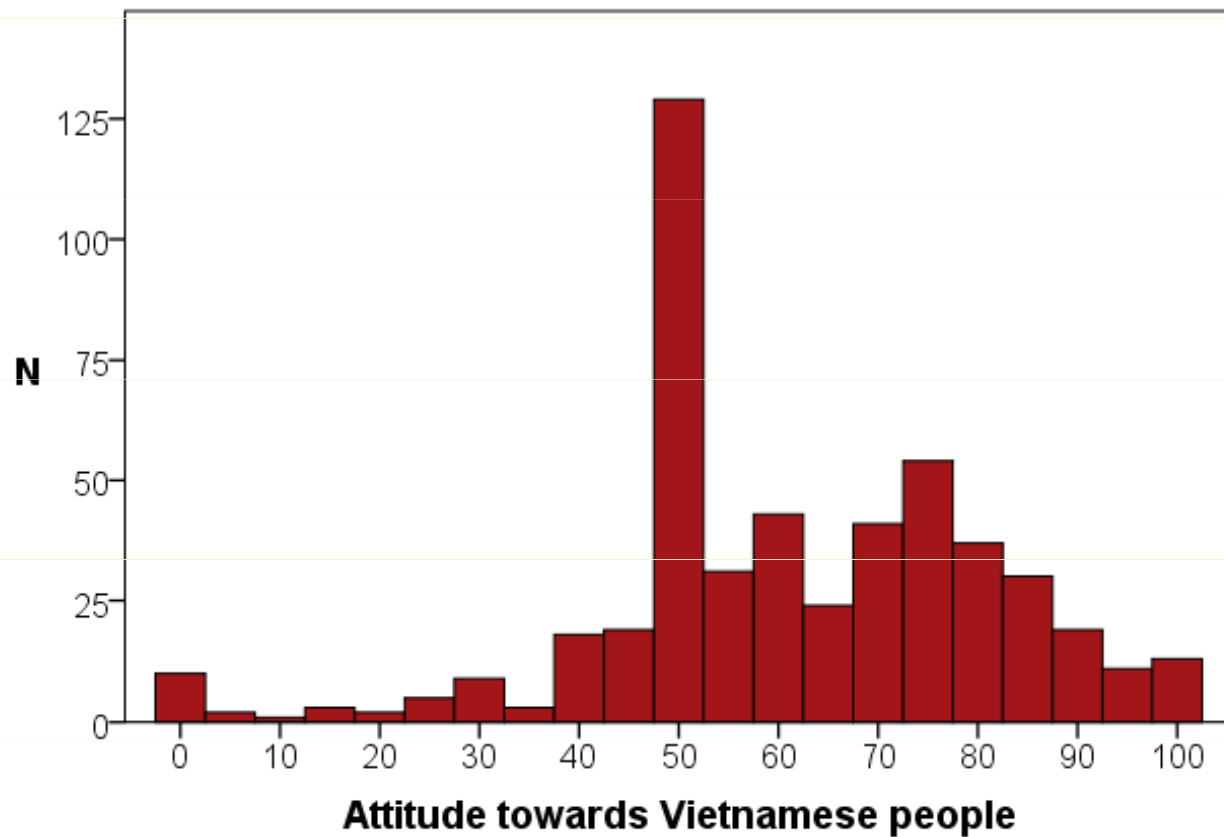
- Used rarely – doesn't show distribution

Categorical data: Bar chart



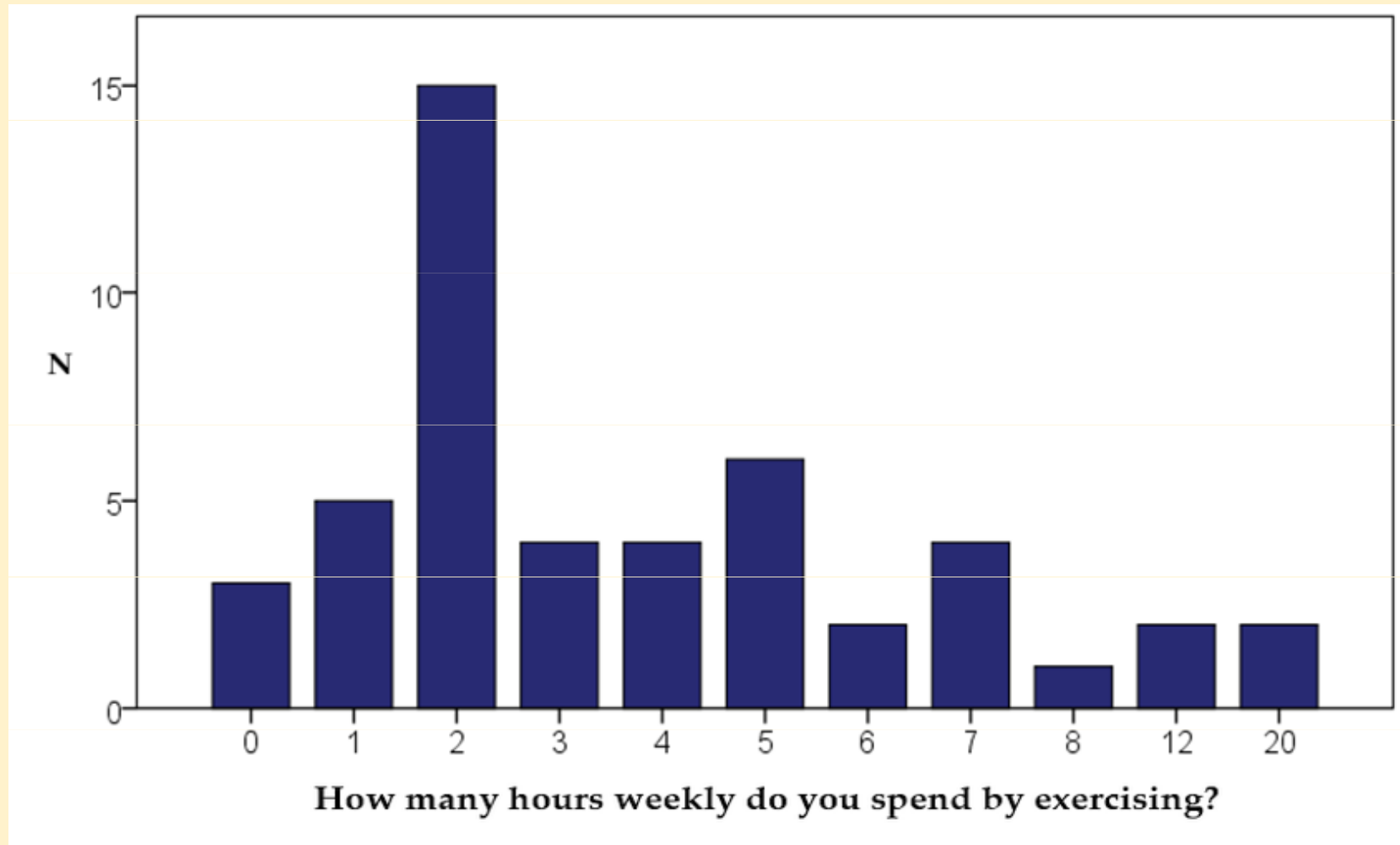
Metric data: Histogram

Express your personal attitude towards Vietnamese people on a thermometer scale ranging from 0 to 100, where 0 means cold, 50 mean neutral and 100 meand warm.

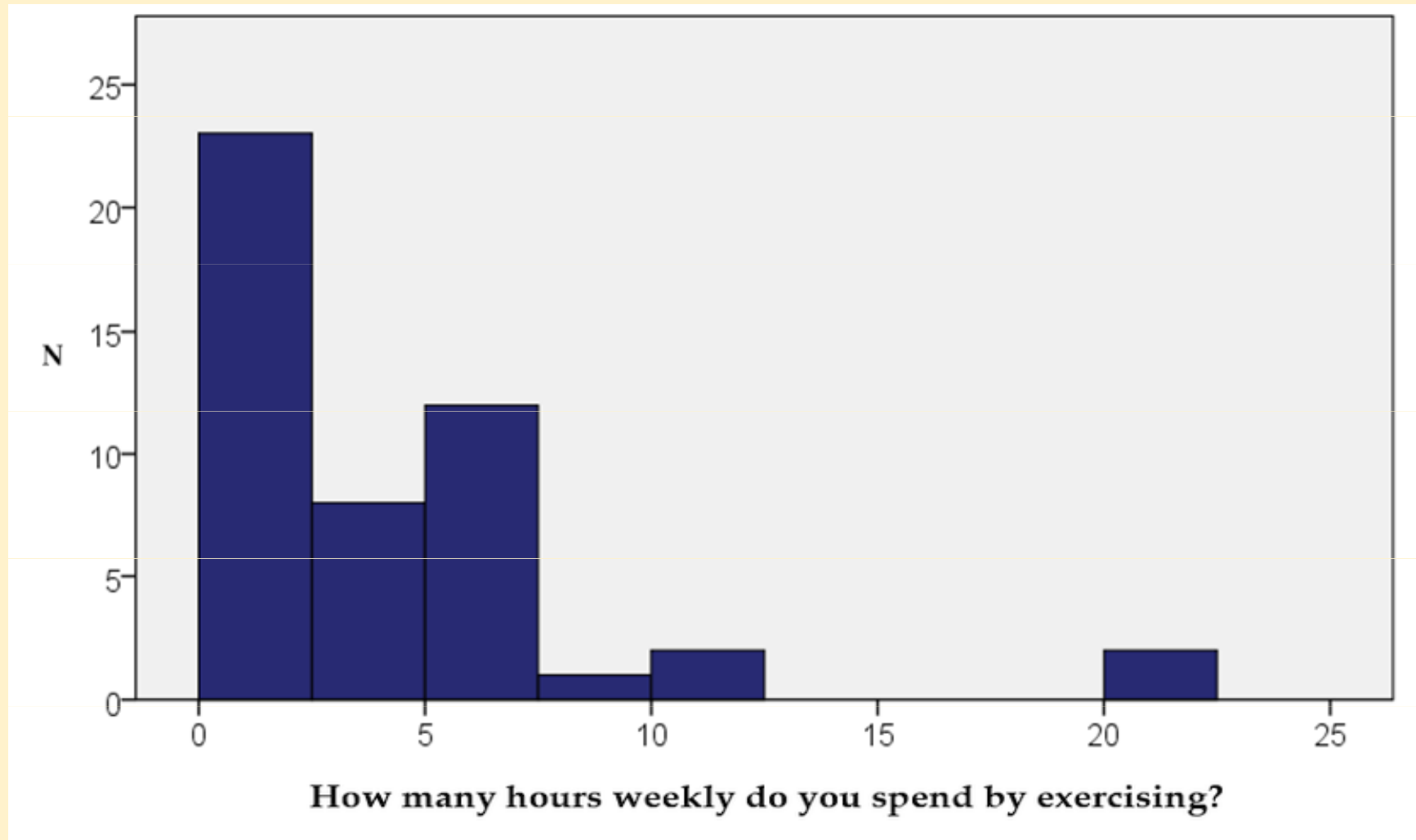


- Similar as Bar chart, but with real values on X axis

Bar chart vs. Histogram



Bar chart vs. Histogram



Correct frequencies visualisation

- Each graph and table has to be described enough so that it's understandable even without reading the source text
- Use headings, scale labels, categories labels
- Use reasonable scales ranging

agentúra Polis - prieskum preferencií

#siet'

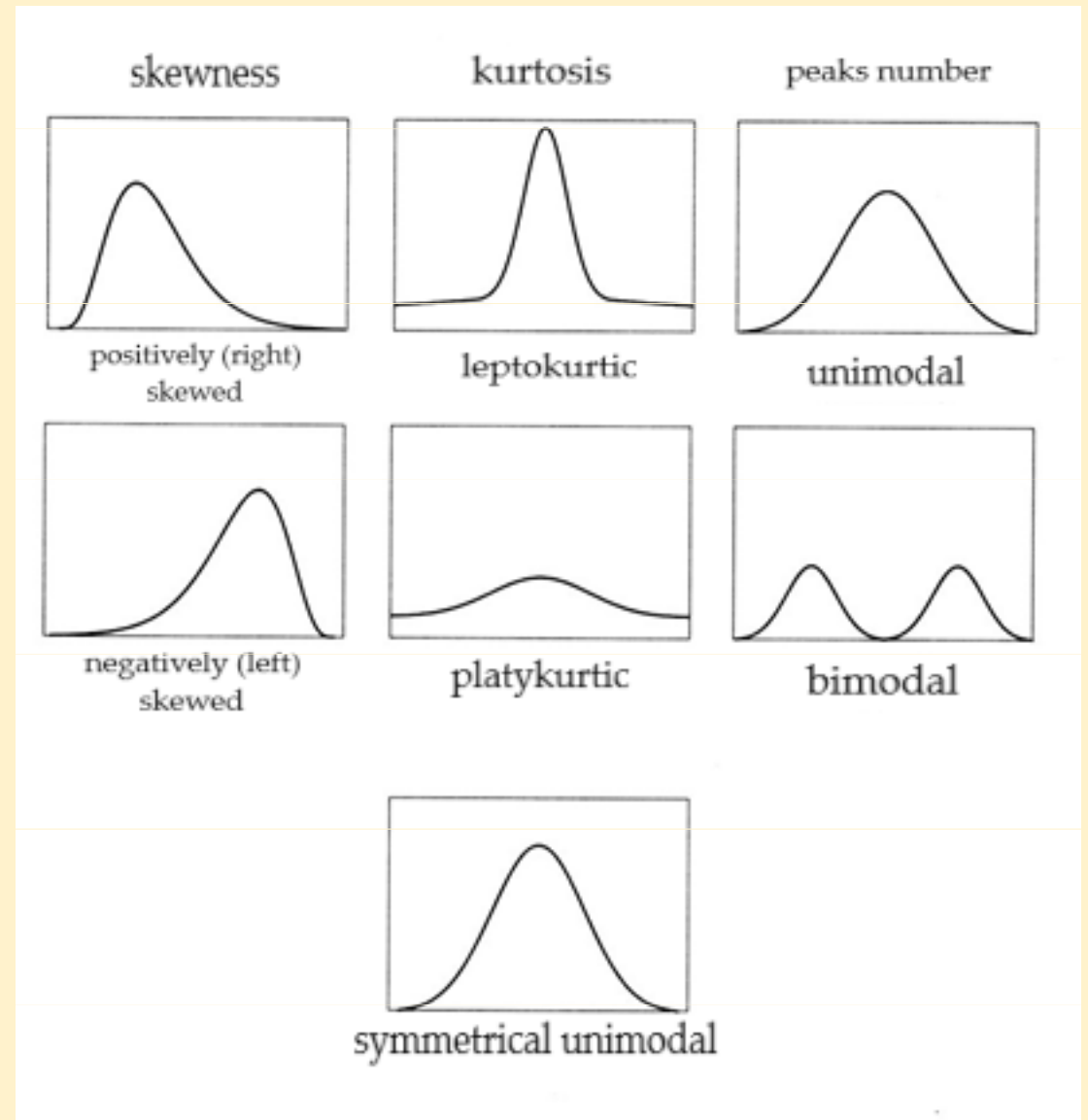


Frequency distribution

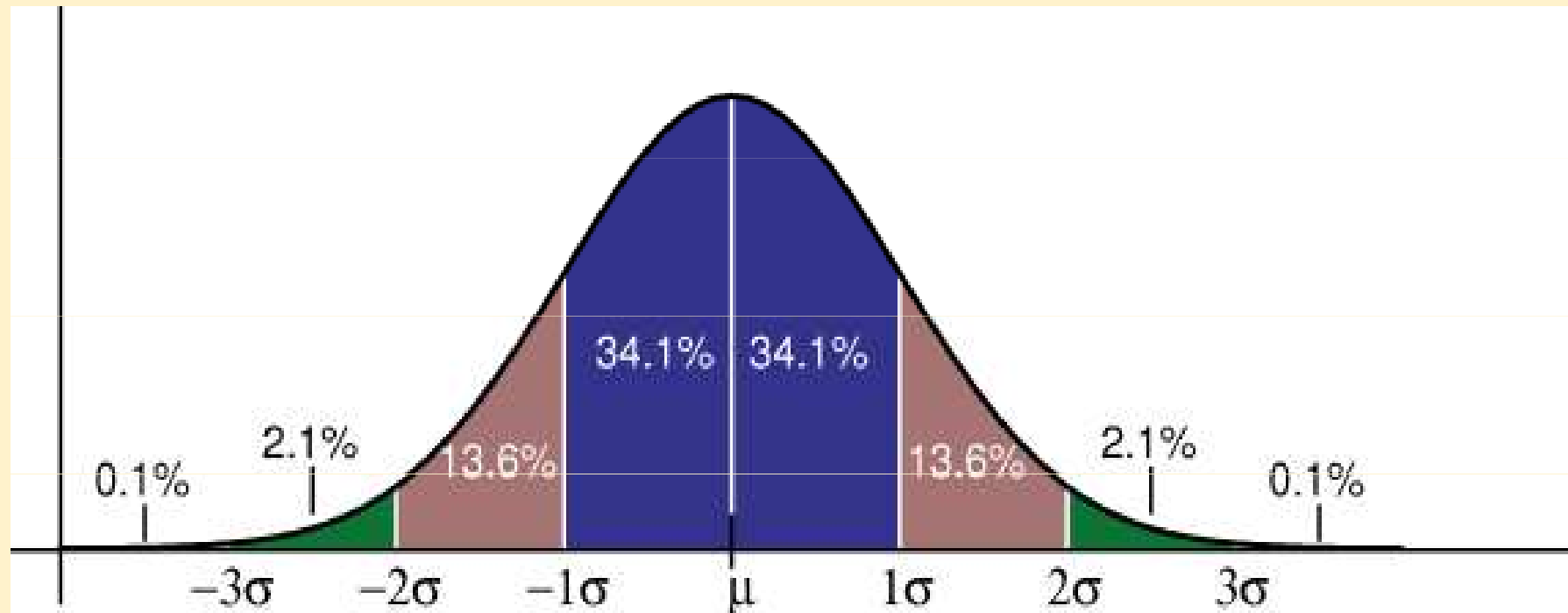
- Each variable has some frequency distribution, which follows from the nature of the variable
- Population frequency distribution X Sample frequency distribution
- We need samples representative to population, so that the sample frequency distribution is close to population frequency distribution

Frequency distribution

- Normal
- Uniform
- Number of peaks: unimodal, bimodal, multimodal
- Skewness:
 - positively skewed (right-skewed, floor effect)
 - negatively skewed (left-skewed, ceiling effect)
- Kurtosis: leptokurtic, platykurtic

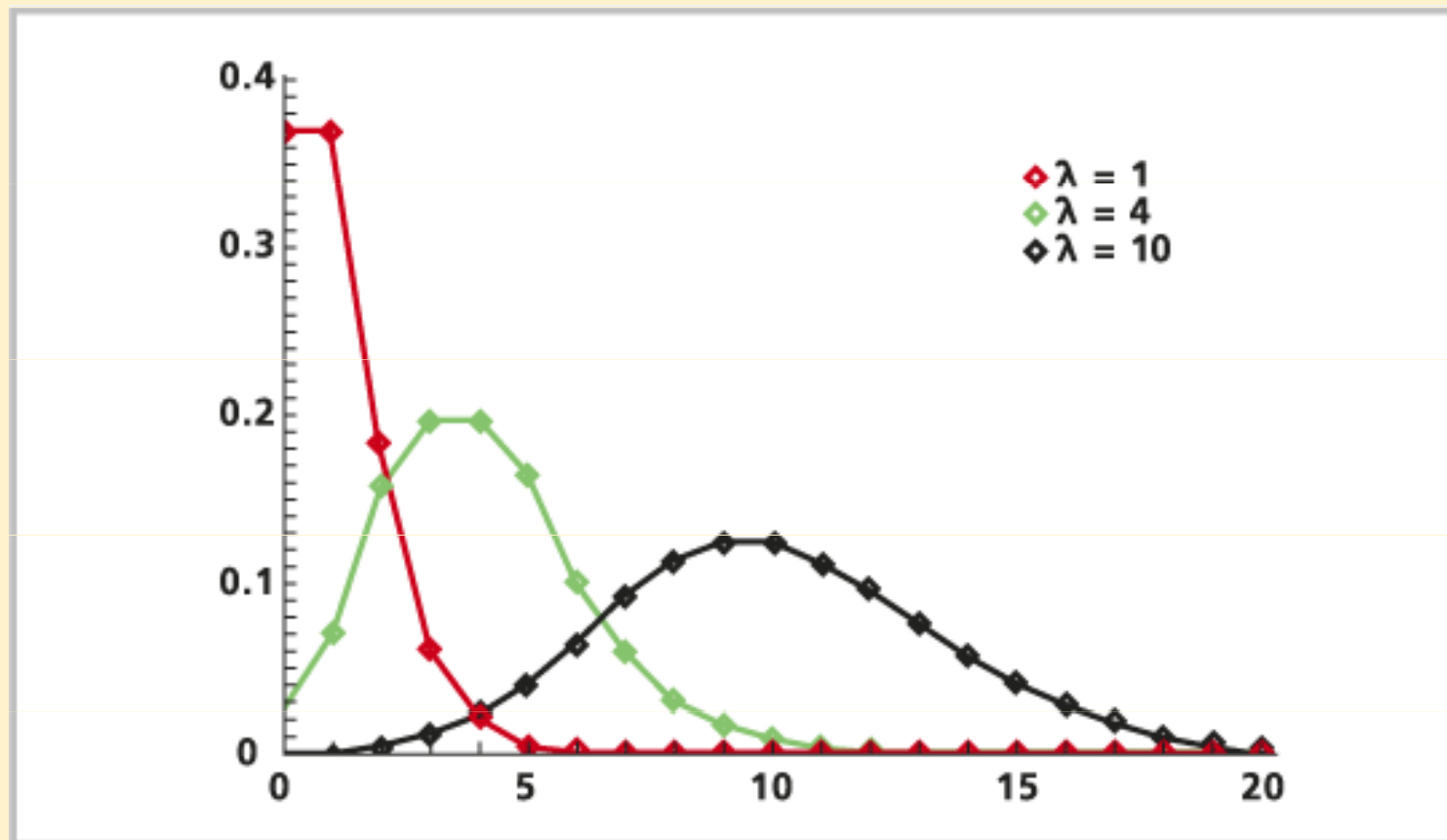


Normal (Gaussian) distribution (bell curve)



Poisson distribution

- Distribution of infrequent events, lambda = number of events per time unit



MEASURES OF CENTRAL TENDENCY AND VARIABILITY

Central tendency

- One number describing variable distribution
 - One number: both advantageous and dangerous...
- Central tendency indicates mean, typical, representative, expected value
- Central tendency indicates where on the scale are the data located



Mode, median, mean

- **MODE (MODUS):** categorical typical value \hat{X}, Mo
 - The most frequent value, the value with the highest frequency
 - The only possibility for categorical data
- **MEDIAN:** ordinal measure of central tendency \tilde{X}, Md
 - Value of the element in the middle of the rank-ordered sample
 - 50. percentile (P_{50})
 - If the total number of values is even, median is the centre of the interval between the two middle values
 - Can be used for ordinal data and higher measurement levels
- **ARITHMETIC MEAN:** deviation measure of central tendency \bar{X}, M, m
 - Only for interval and ratio data
 - Easily biased by extreme values

Counting mode, median and mean

- Mode
 - Read from frequency table
 - Excel: MODE(data)
- Median
 - Categorical variables: read from frequency table (best from cumulative relative frequencies)
 - For odd N, Me is X_k (k^{th} element of rank-ordered sequence of variable values), where $k = (N+1)/2$
 - For even N, Me is the mean of X_k and X_{k+1} , where $k = N/2$
 - Excel: MEDIAN(data), PERCENTIL(data;0.5)
- Mean
 - Excel: PRŮMĚR(data), AVERAGEA(data)

Measures of central tendency: examples

- Data A: 1, 2, 3, 3, 3, 4, 5, 5, 6, 8, 10 (N=11)
 - $M = 4.55$, $Md = 4$, $Mo = 3$
- Data B: 1, 2, 3, 3, 3, 4, 5, 5, 6, 8, **100** (N=11)
 - $M = 12.72$, $Md = 4$, $Mo = 3$
- Data C: 1, 2, 3, 3, 4, 5, 5, 6, 8, 10 (N=10)
 - $Md = ?$, $Mo = ?$
 -  $Md = 4.5$, $Mo = 3$ and 5
- Data D:
 - $Md = ?$
 - 
 - $Md = 10.5$

X	f	cum f
16–20	5	30
11–15	10	25
6–10	11	15
1–5	4	4

Group mean X Weighted mean

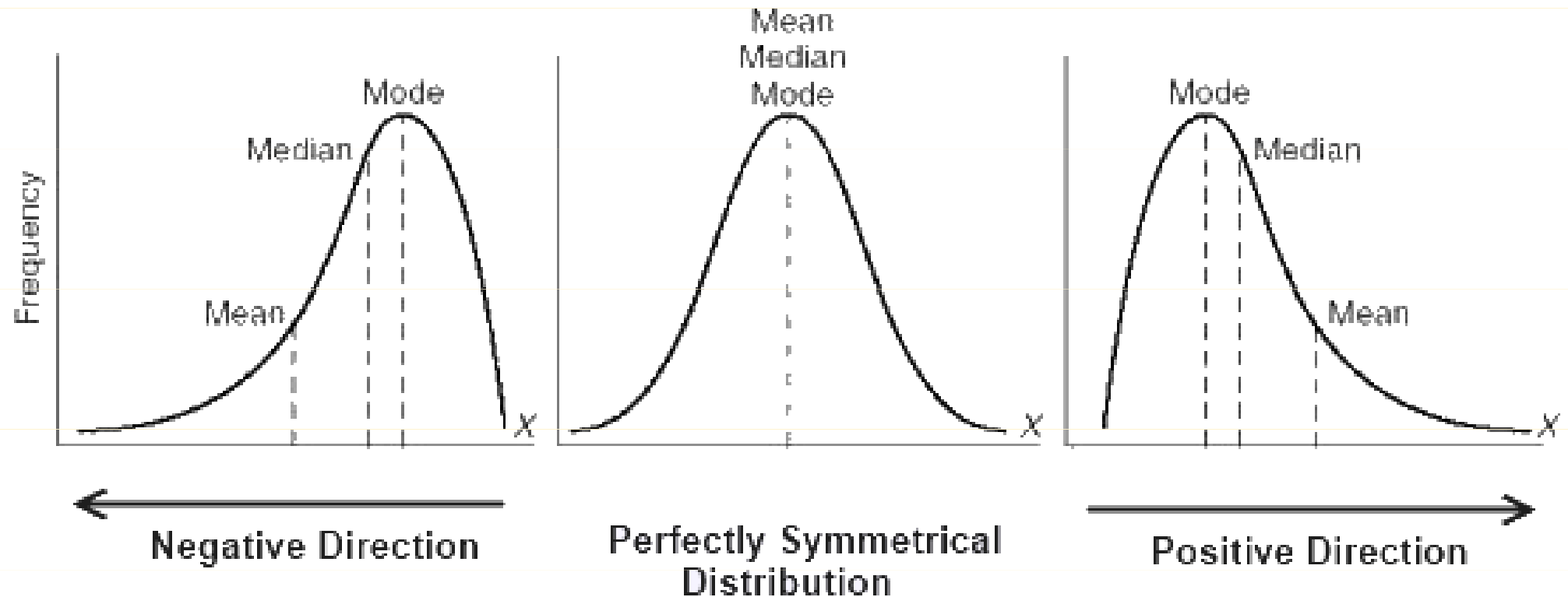
- **Group mean:** A group of female teachers has average salary of 45 000 CZK and a group of male teacher has average salary of 50 000 CZK. What is the average salary of both groups?
 - $(45\ 000 + 50\ 000) / 2 = 47\ 500$
- **Weighted mean:** If the average salary of 10 female teachers is 45 000 CZK and the average salary of 40 male teachers is 50 000 CZK, what is the average salary of all 50 teachers?
 - $(10 \times 45\ 000) + (40 \times 50\ 000) / (10 + 40) = 49\ 000$
- More on estimating mean, median and mode from group frequencies with examples here:
<http://www.mathsisfun.com/data/frequency-grouped-mean-median-mode.html>

Relations between mode, median and mean

(a) Negatively skewed

(b) Normal (no skew)

(c) Positively skewed



Measures of variability

- The second number for distribution description
- Indicates how little or how much are the data on the scale distributed
- Low variability = most values are the same or close to each other
- High variability = values are very diverse (or bimodal distribution)

Range, variance, standard deviation

- Ordinal measures of variability
 - Range: $X_{\max} - X_{\min}$ (extremely biased by end values)
 - Interquartile range: $Q_3 - Q_1$ ($P_{75} - P_{25}$), IQR
- Deviation measures of variability
 - **based on deviations from the mean: $x = X - M$**
 - **variance** = average squared deviation, s^2 , $\text{VAR}(X)$
 - population: (Sx^2 / n) vs. sample $(Sx^2 / (n - 1))$
 - Sx^2 = sum of squared deviations = **sum of squares**
 - **standard deviation – s, SD**
 - square root of the variance, return to the original units

Counting measures of variability

- $IQR = Q_3 - Q_1$
 - $Q_1 = X_k$ (k^{th} element of rank-ordered sequence of variable values), where $k = (N+1) * 0.25$ rounded down
 - $Q_3 = X_k$, where $k = (N+1) * 0.75$ rounded down
 - Excel: PERCENTIL(data, 0.25), resp. PERCENTIL(data, 0.75)
- Variance, Standard deviation
 1. Count deviation score for every value: $x_i = X_i - M$
 2. Count squared deviations
 3. Sum squared deviation and divide by $N-1$
 4. For SD, we square-root the variance
 - Excel: VAR.P(data), VAR.S(data), STDEVPA(data) ~ SMODCH.P(data), STDEVA(data) ~ SMODCH.VÝBĚR.S(data)

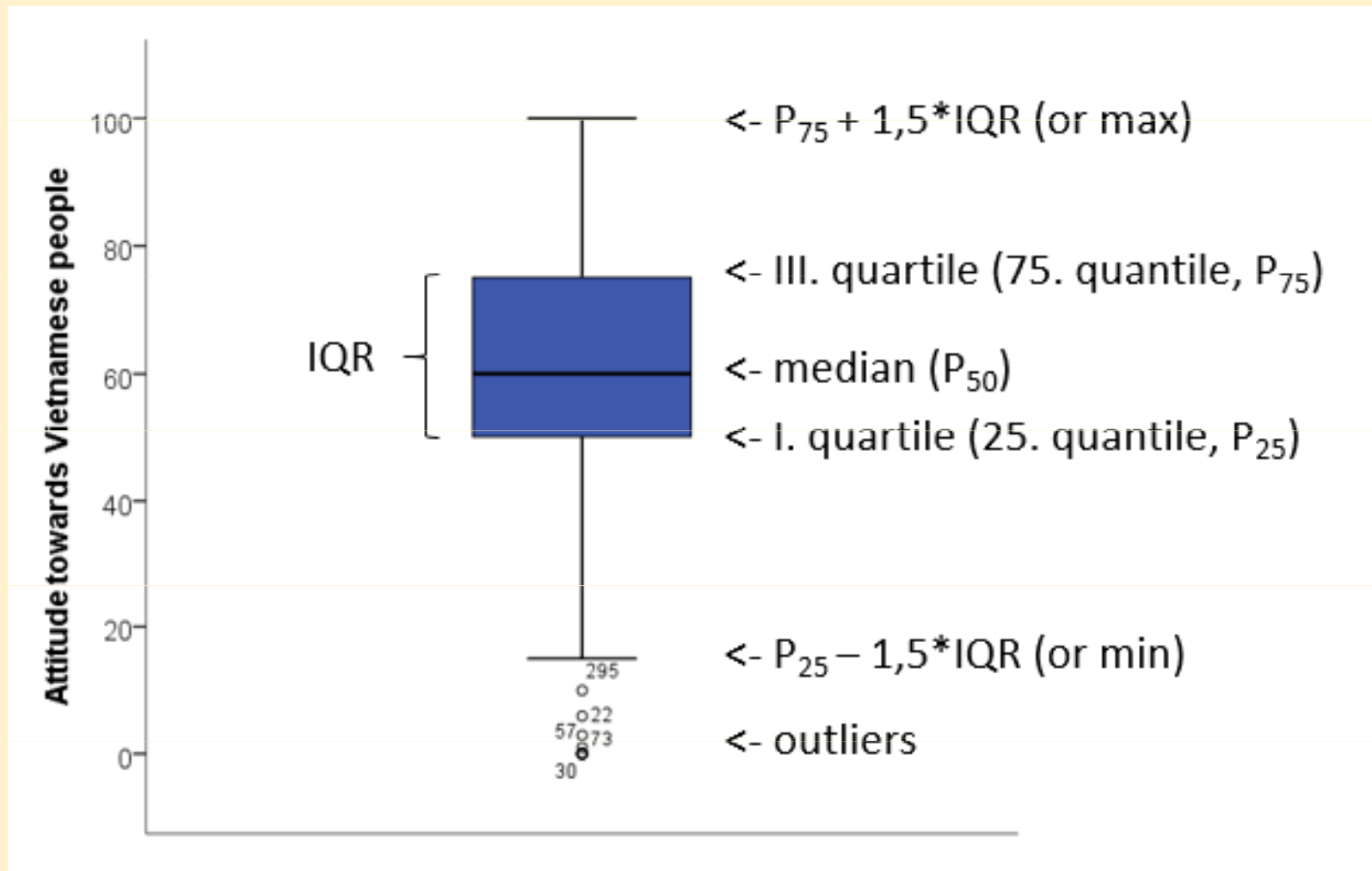
Statistics vs. Parameters

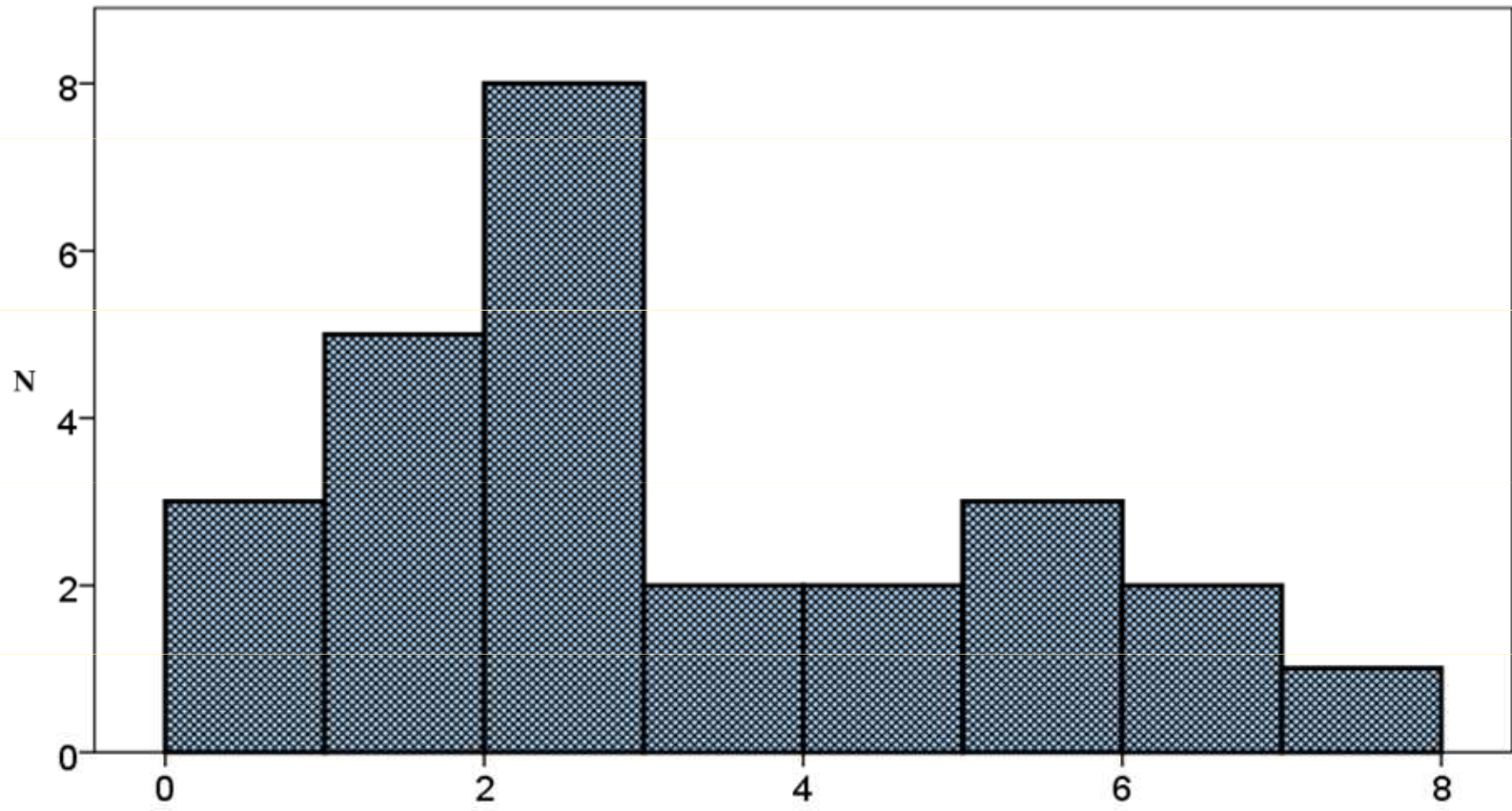
- Population \sim Sample, Parameters \sim Statistics
- Statistics: m , s^2 , s
- Parameters: μ , σ^2 , σ
- A teacher is interested in math knowledge among students in his class (3rd grade). The class is his population.
- We are interested in math knowledge of students in third grade in Czech Republic. We collect data from several classes randomly picked from all primary schools in Czech Republic. The classes are our sample, our target population are all third grade student in Czech Republic.

Measures of central tendency and variability

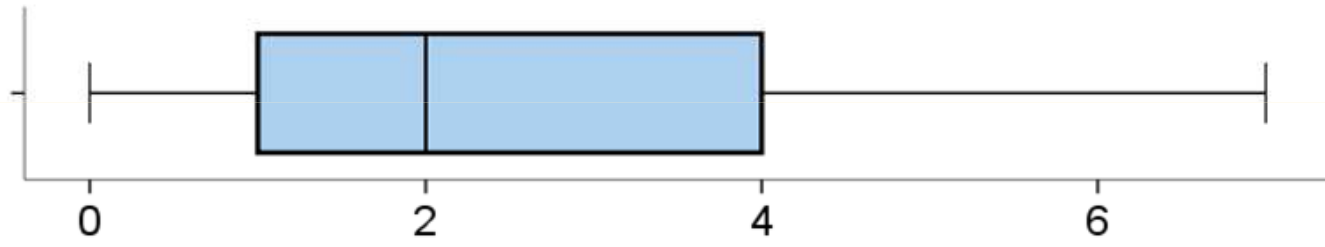
- It is necessary to be able to count measures of central tendency and variability by hand
- It is necessary to be aware of how does adding a constant or multiplying by a constant change different measure of central tendency and variability
- Think about when is it appropriate to use different measures and why
- How does the distribution shape influence the measure and vice versa: what can we say about the data distribution from the measures of central tendency and variability
- View exercises

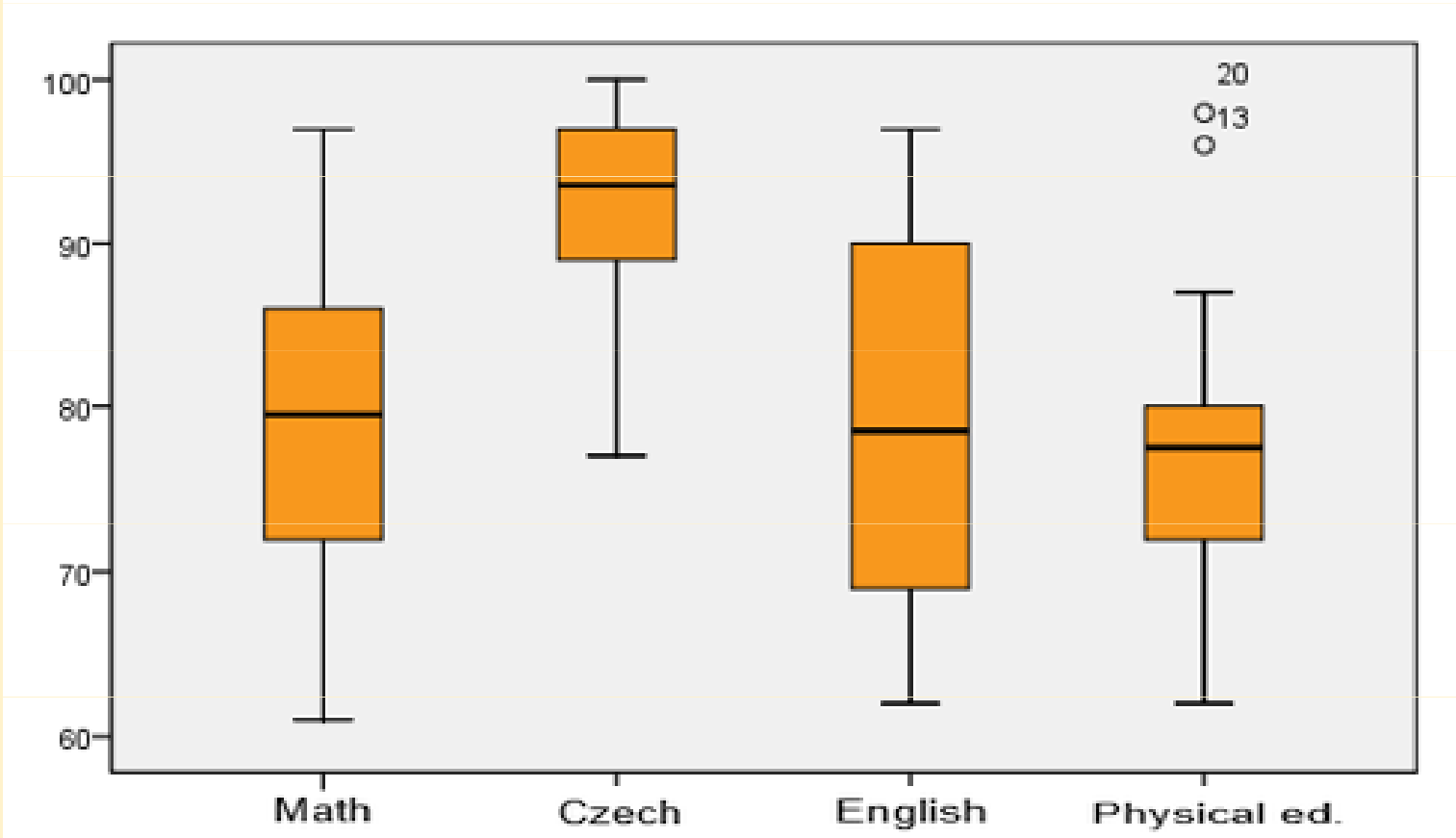
Boxplot (cz: krabicový graf s anténami)





How many hours weekly do you spend by exercising?





Choice of descriptive statistics

- We consider:
 - level of measurement
 - distribution shape – symmetrical? normal?
 - the aim of the study: description X inference, comparison
- For (distribution) description only:
 - we use mainly ordinal statistics (we can also add deviation statistics)
 - N, min, Q_1 , median, Q_3 , max
 - boxplot
 - percentiles for individual scores
- For statistical inference, comparison etc.:
 - we use deviation statistics (if we have sufficient measurement level)
 - N, M, SD
 - distribution description, z-scores for individual scores

Descriptive statistics in research studies

- We always present descriptive statistics
 - We usually present descriptive statistics only for variables we work with (we interpret)
 - Minimum: N, M, SD (or ordinal equivalents Q_1 , Md, Q_3 , IQR)
 - Usually more appropriate: N, X_{\min} , X_{\max} , M, SD
 - If needed: skewness, kurtosis, interesting quantiles...
- Distribution description
 - we usually don't present frequency tables and graphs (if statistical description isn't the aim of the study, e.g. diagnostic test manual)
 - If needed, we usually mention the distribution shape only verbally (approximately normal, negatively skewed etc.)
 - We usually assess normality and deal with deviation from normal distribution
- Usually one or two decimals
- Czech in Czech text, English in English text (beware: decimal point/comma etc.)
- Usually there are norms for tables presentation, e.g. APA 6th manual