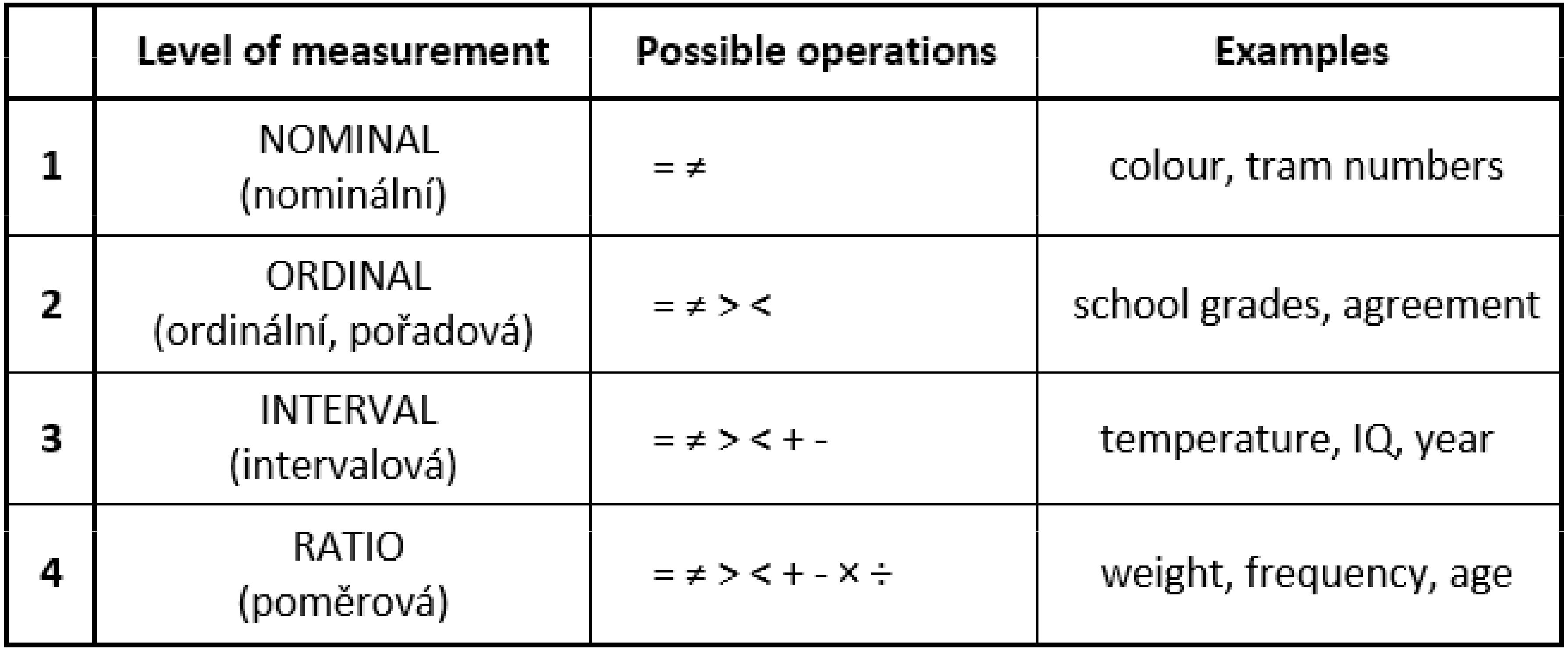
Experimental Humanities II (HUMB002) 2016
STATISTICAL ANALYSIS

REVISION AND EXERCISES

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

Variables: Levels of measurement



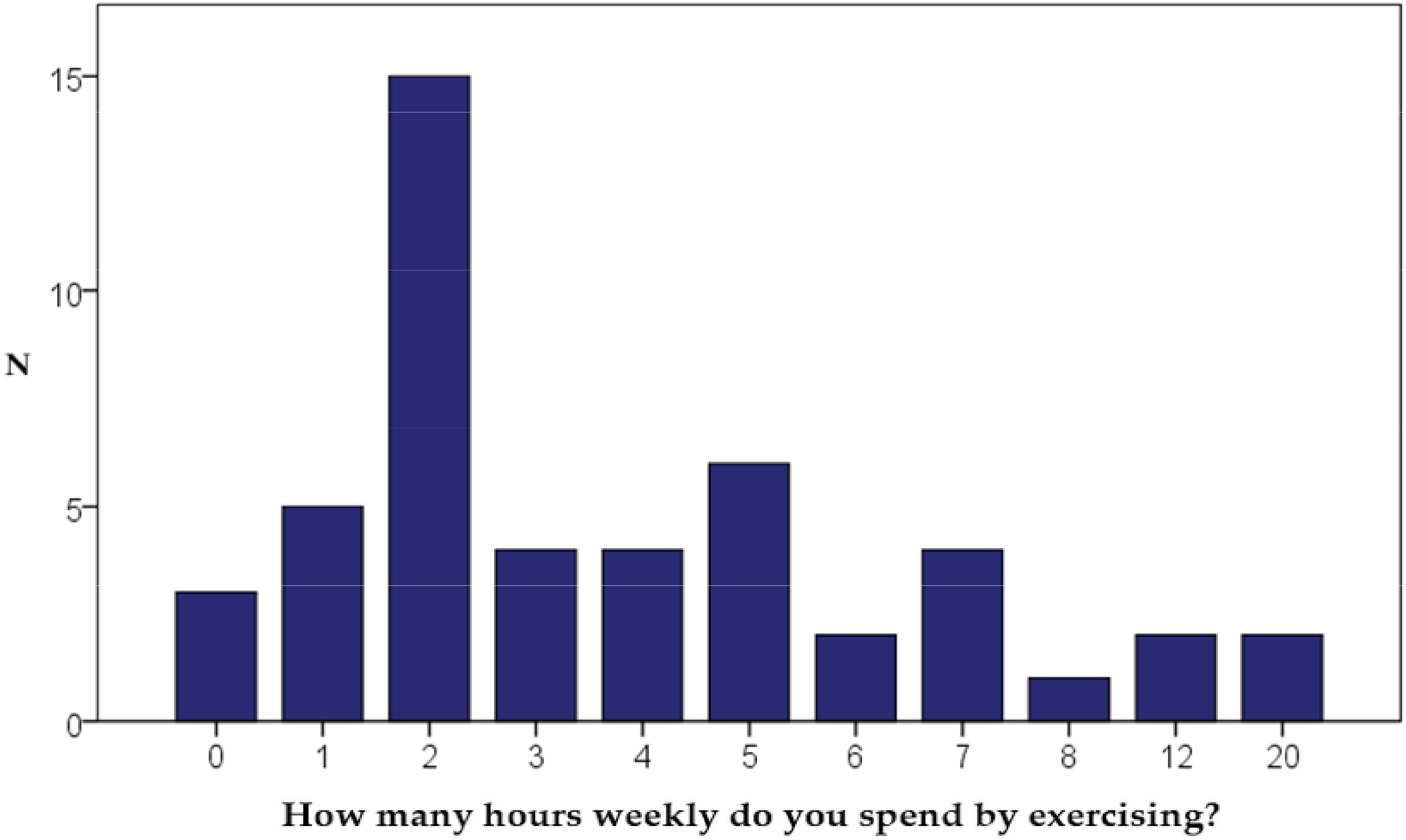
	Level of measurement	Possible operations	Examples
1	NOMINAL (nominální)	= ≠	colour, tram numbers
2	ORDINAL (ordinální, pořadová)	= ≠ > <	school grades, agreement
3	INTERVAL (intervalová)	= ≠ > < + -	temperature, IQ, year
4	RATIO (poměrová)	= ≠ > < + - × ÷	weight, frequency, age

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

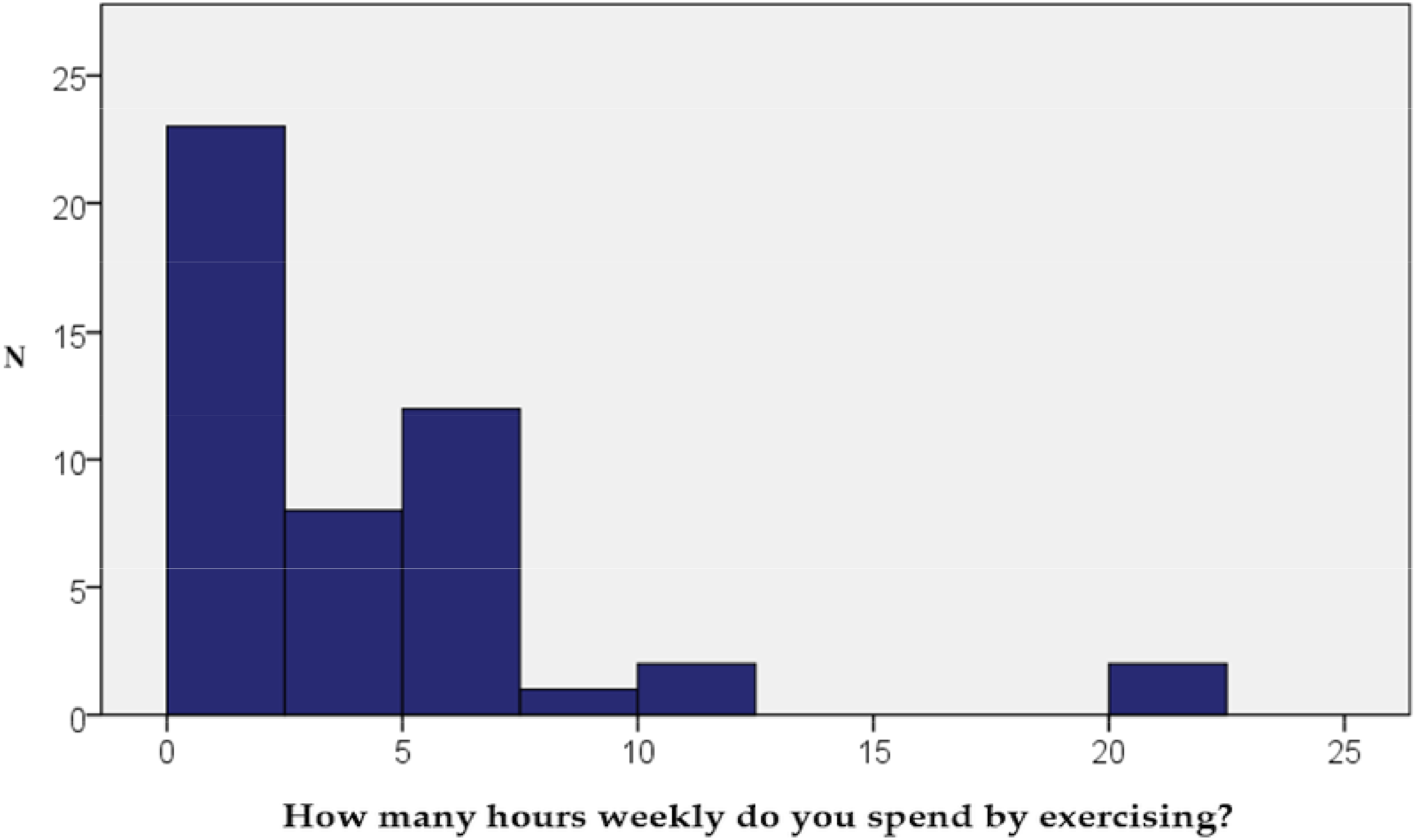
Frequency tables

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%	

Bar chart vs. Histogram

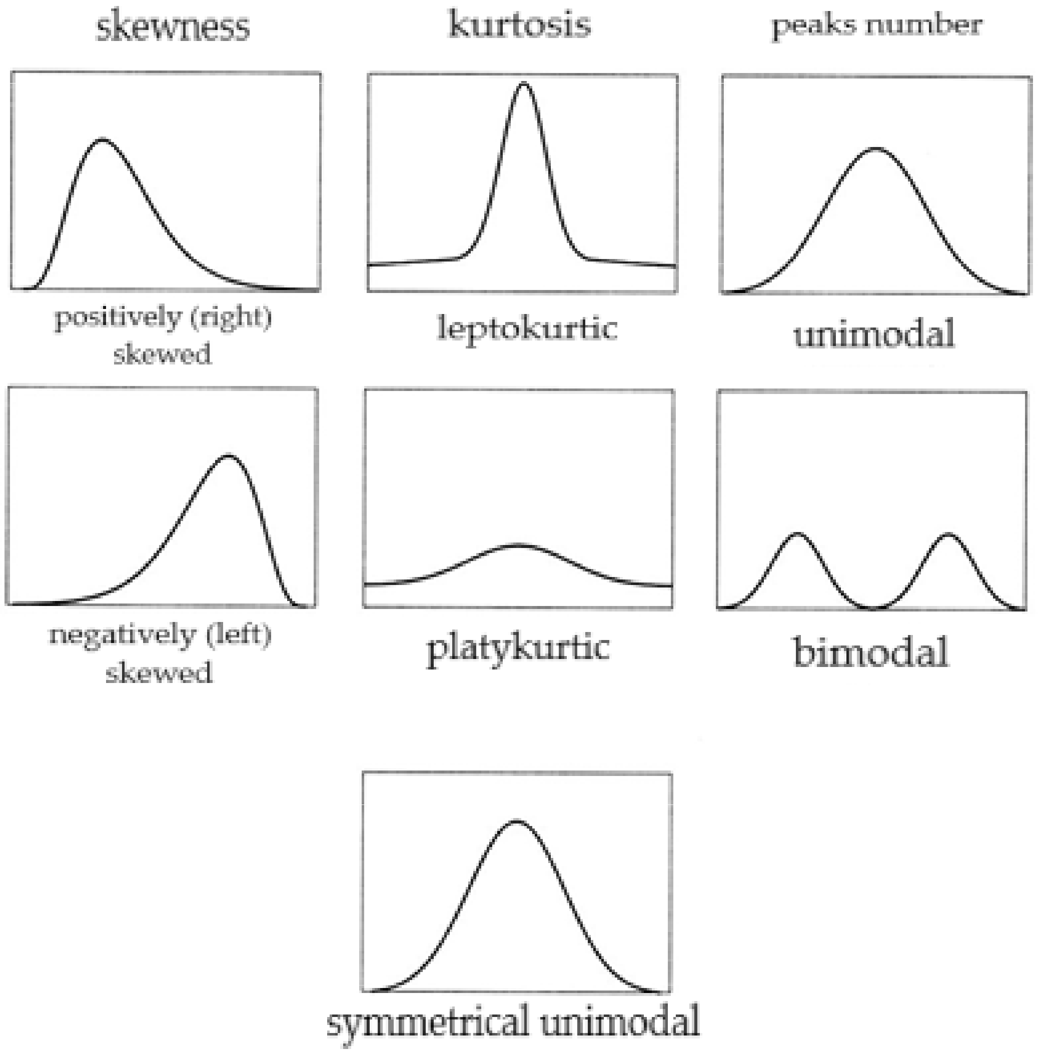


Bar chart vs. Histogram



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



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 $\widetilde{X}, Md$
  - 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 $\overline{X}, M, m$
  - Only for interval and ratio data
  - Easily biased by extreme values

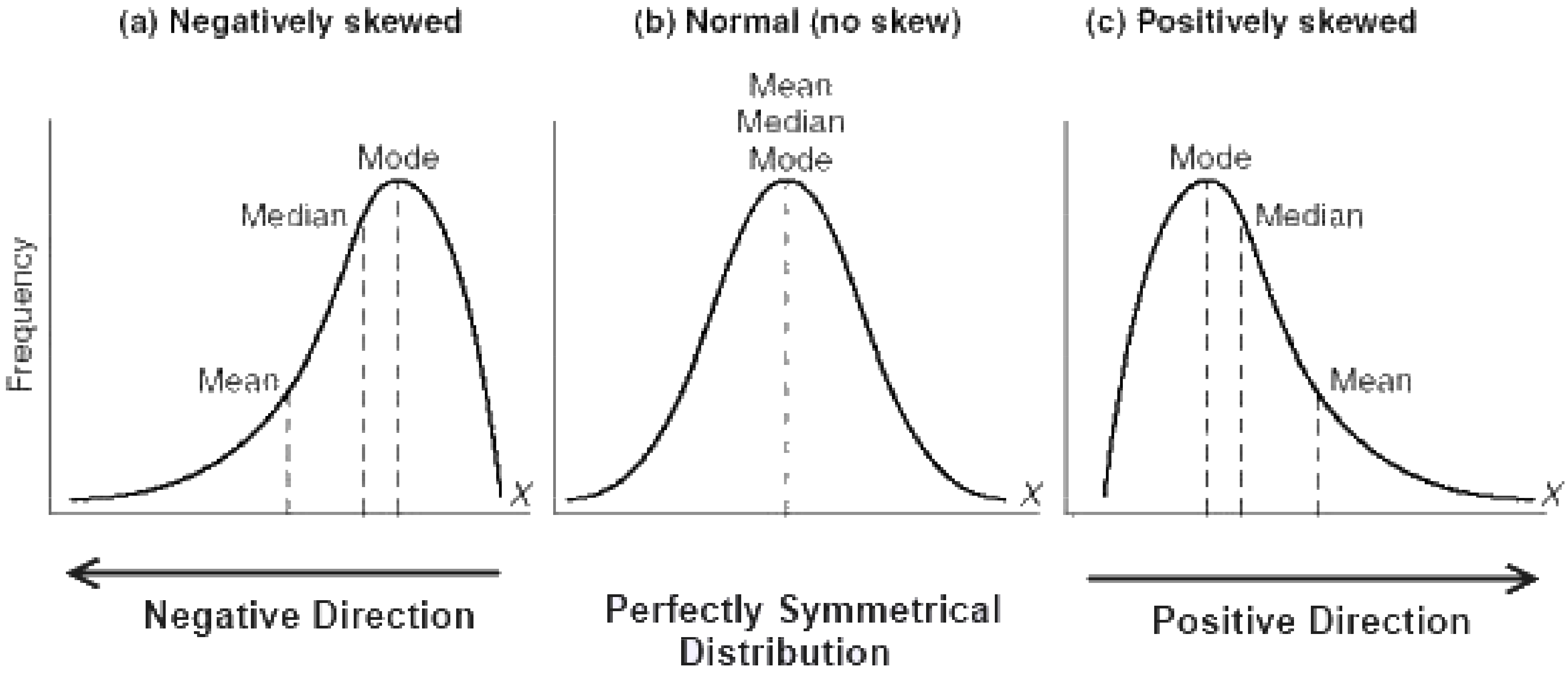
Counting mode, median and mean

- 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)

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 x 45 000) + (40 x 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



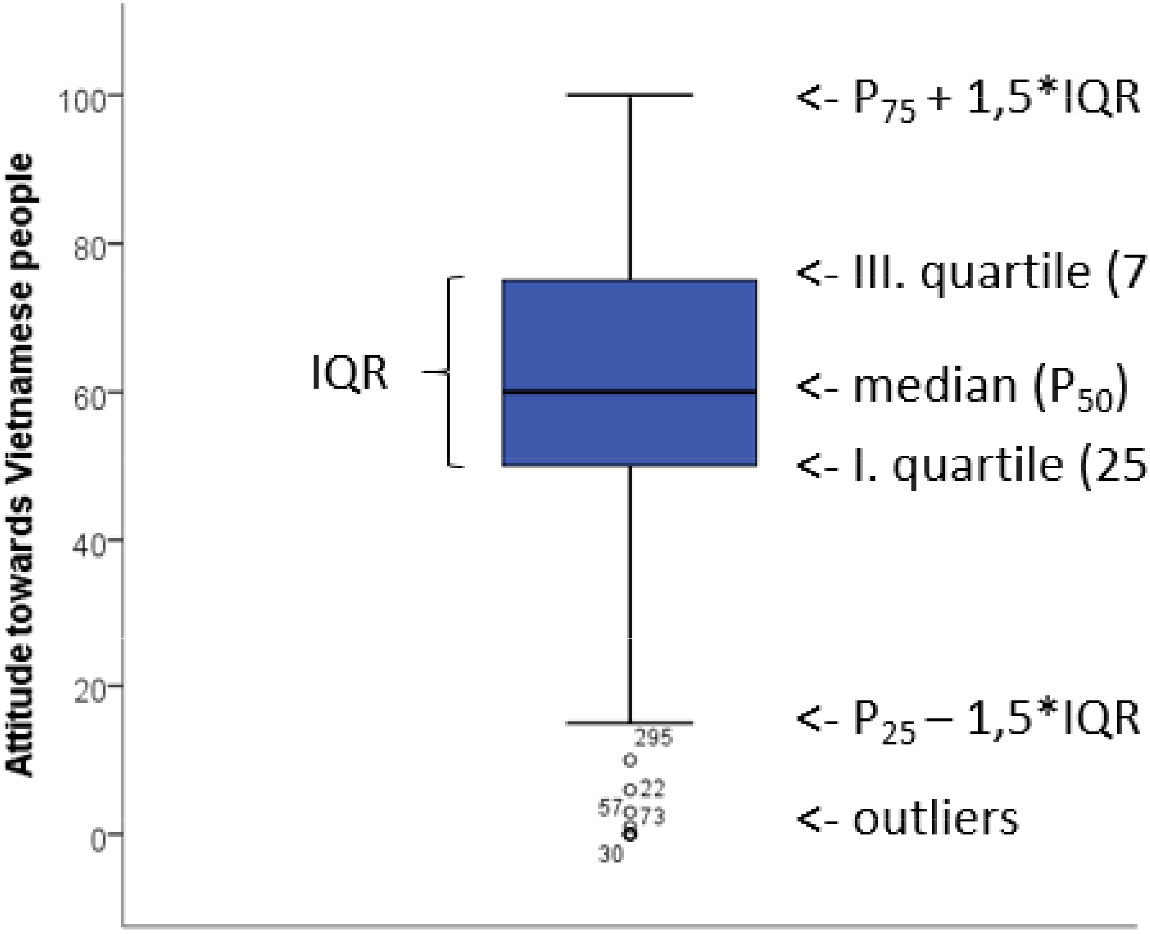
Range, variance, standard deviation

- 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, 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)

Boxplot (cz: krabicový graf s anténami)



<- P_{75} + 1,5*IQR (or max)

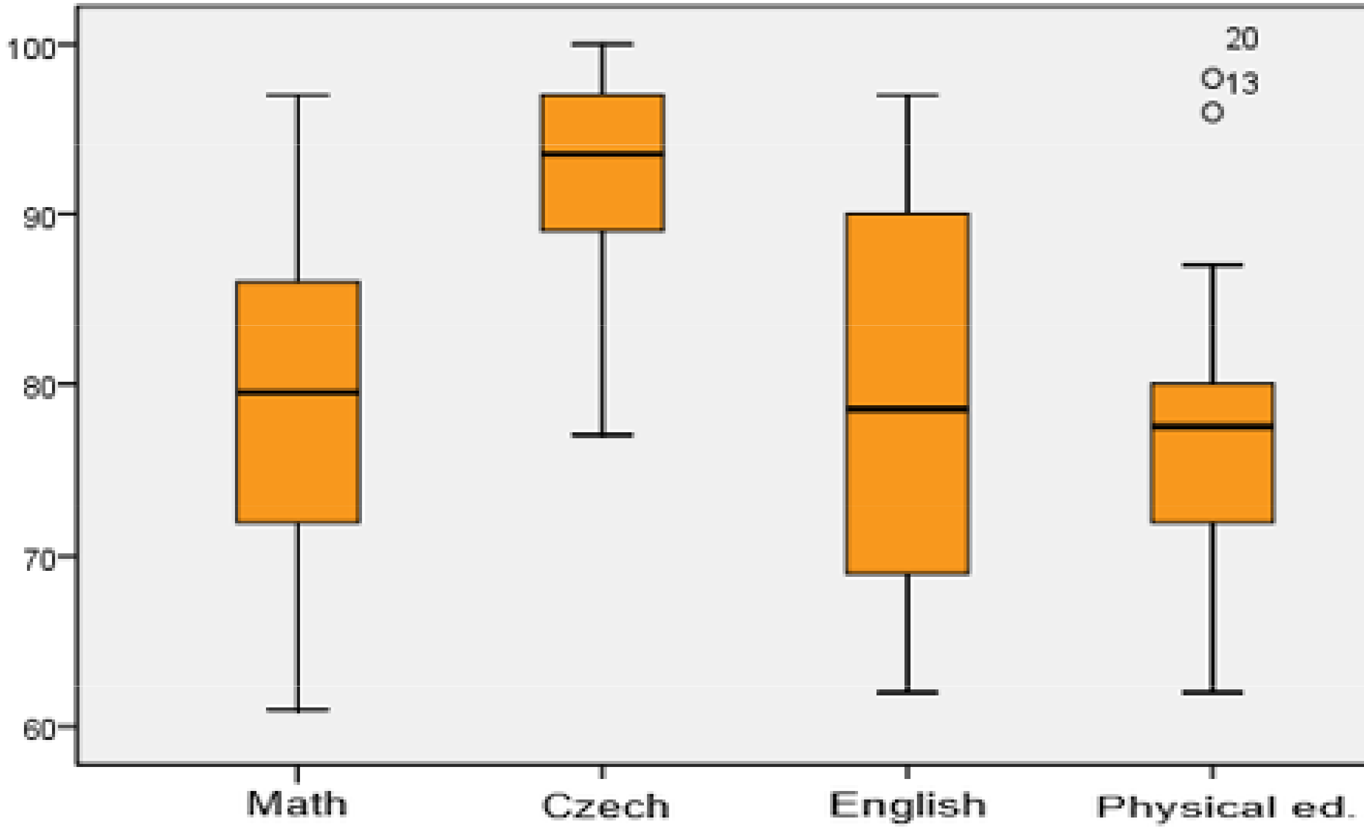
<- III. quartile (75. quantile, P_{75})

<- median (P_{50})

<- I. quartile (25. quantile, P_{25})

<- P_{25} – 1,5*IQR (or min)

<- outliers

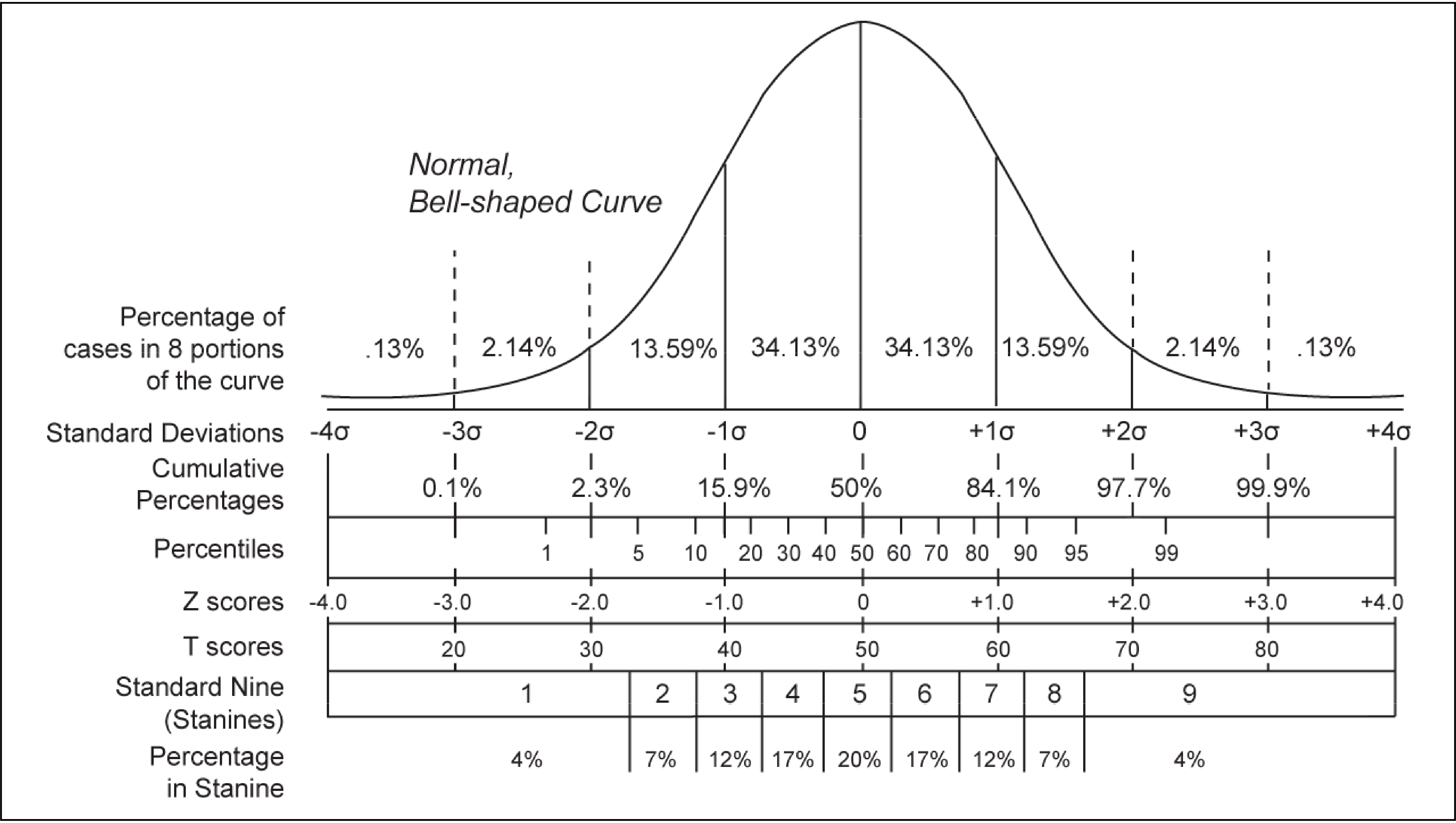

100
90
80
70
60
20
13
Math
Czech
English
Physical ed.

Scores transformation

- We often transform observed score for easier understanding and interpretation
- Making interpretation easier – linear transformations
  - e.g. multiplying by 10 or 100 for getting rid of the decimals
  - distribution shape remains unchanged
  - descriptive statistics will change in a predictable way
  - possibility of standardization
- Change of distribution shape – nonlinear transformations
  - log/exp functions, square(root)…
- Change of measurement level – ordinal transformation (ranking)

Linear transformation – standardization

- The most usual transformation: **standardization (z-scores)**
  - scores transformation, so that **M=0, SD=1**
  - measurement unit becomes SD – we can easily compare scores from different scales (but differences in distribution remain!)
  - **$z_i = (X_i – M) / SD$**
- Scores derived from z-scores:
  - T scores: M=50, SD=10; **$T_i = 50 + 10z_i$**
  - IQ scores: M=100, SD=15
  - Stens (Standard TENs, cz: steny): M=5.5, SD=2; $Sten_i = 2z_i + 5.5$
  - Stanines (STAndard NINEs, cz: staniny): M=5, SD=2; $Stanine_i = 2z_i + 5$
- Normal distribution is always required for correct standardized scores interpretation!


Normal,
Bell-shaped Curve
Percentage of cases in 8 portions of the curve
.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
.13%
Standard Deviations
-4σ
-3σ
-2σ
-1σ
0
+1σ
+2σ
+3σ
+4σ
Cumulative Percentages
0.1%
2.3%
15.9%
50%
84.1%
97.7%
99.9%
Percentiles
1
5
10
20
30
40
50
60
70
80
90
95
99
Z scores
-4.0
-3.0
-2.0
-1.0
0
+1.0
+2.0
+3.0
+4.0
T scores
20
30
40
50
60
70
80
Standard Nine (Stanines)
1
2
3
4
5
6
7
8
9
Percentage in Stanine
4%
7%
12%
17%
20%
17%
12%
7%
4%

Counting quantiles in Excel

- NORM.S.DIST(z;1) – returns corresponding percentile for given z-score (=how many people have the same or lower z-score)
  - Percentage of people between two given z-scores: NORM.S.DIST(higher z;1) minus NORM.S.DIST(lower z;1)

- NORM.S.INV(p) – returns corresponding z-score for given percentile

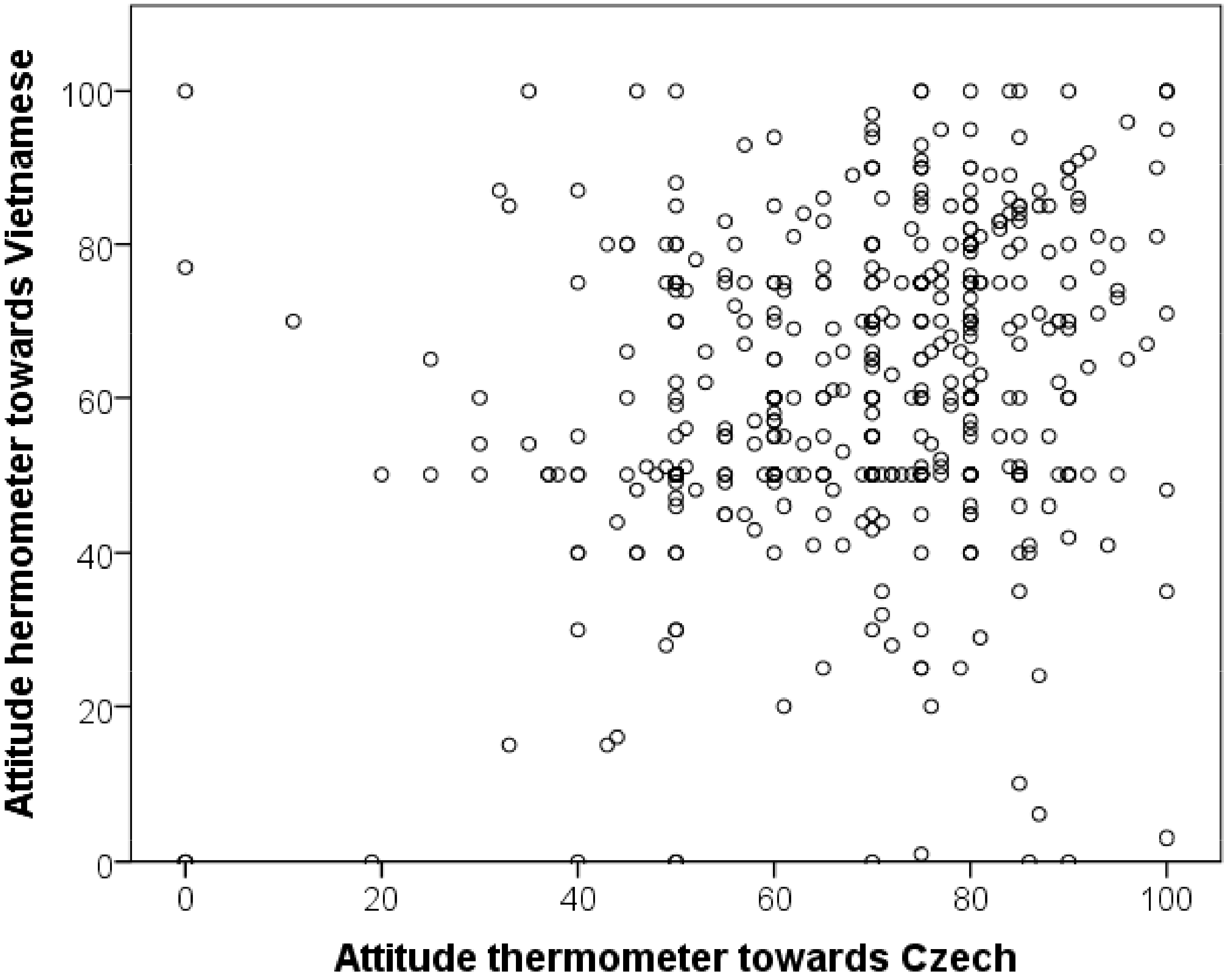
Contingency table

		Math grades					**Total**
		1	2	3	4	5	
Czech language grades	1	82	40	8	1	0	131
	2	71	200	73	17	0	361
	3	4	75	109	25	0	213
	4	1	7	23	24	1	56
	5	0	0	2	1	2	5
Total		158	322	215	68	3	766

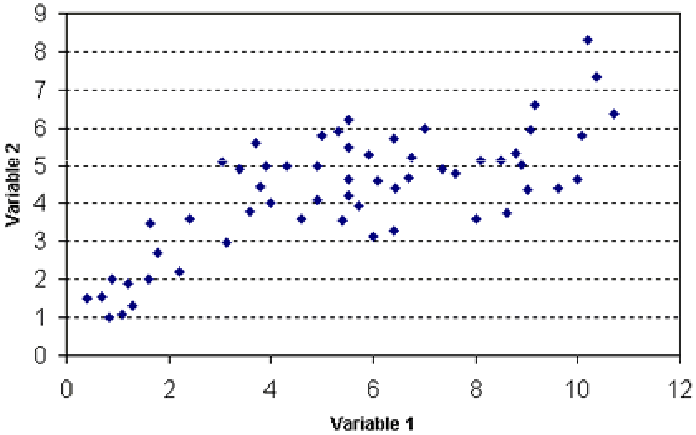
- For any variables, but most suitable for discrete variables with not many values
- The cells can contain both absolute or relative frequencies (row, column and total frequencies)
- The last row and column contain so called row/column marginal frequencies
- Graphical representation of contingency table is 3D bar chart or 3D histogram
- High frequencies in diagonals indicate linear relationship between variables

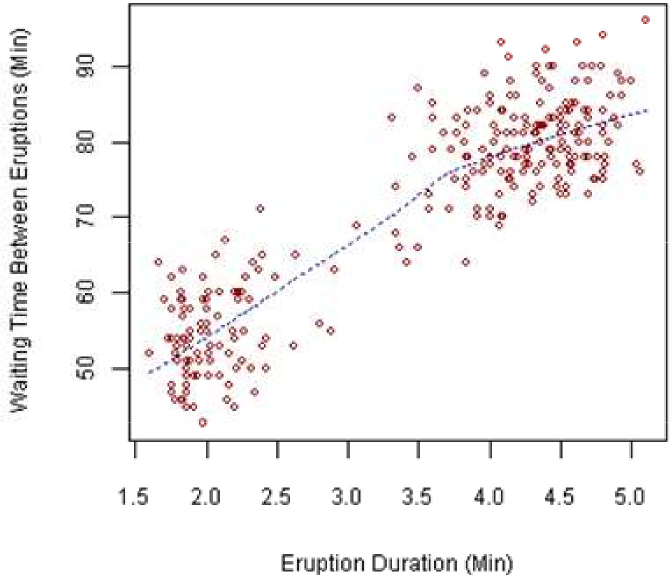
Scatterplot

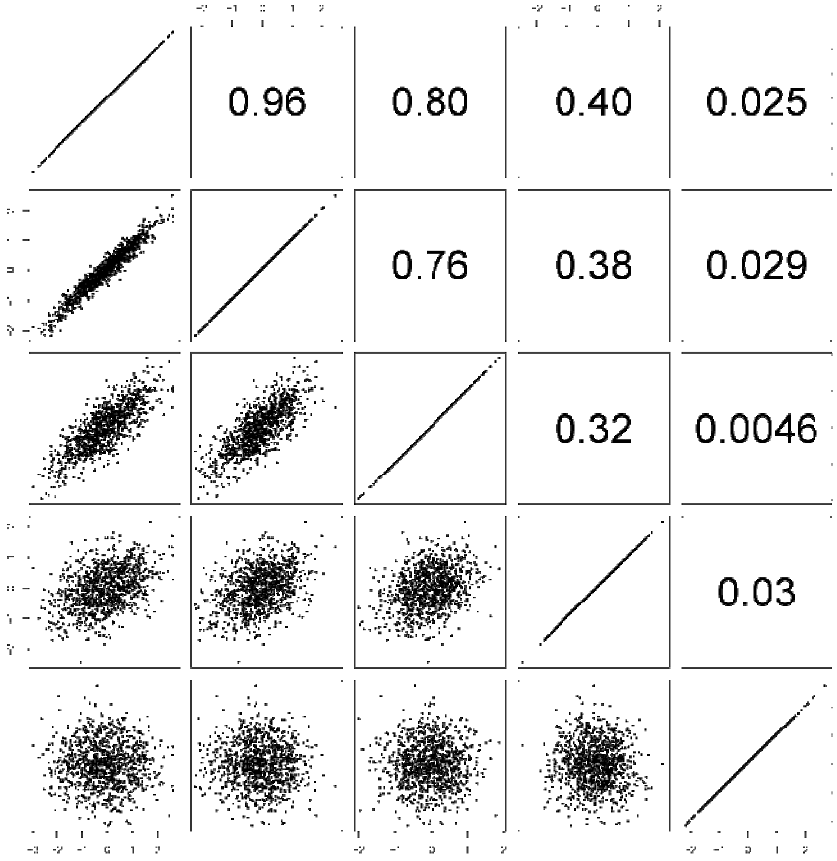
- Substitutes contingency table for continuous variable
- Each axis represents one variable
- Each point represents one subject (unit)
- Frequency of the same values can be represented e.g. by the dot size



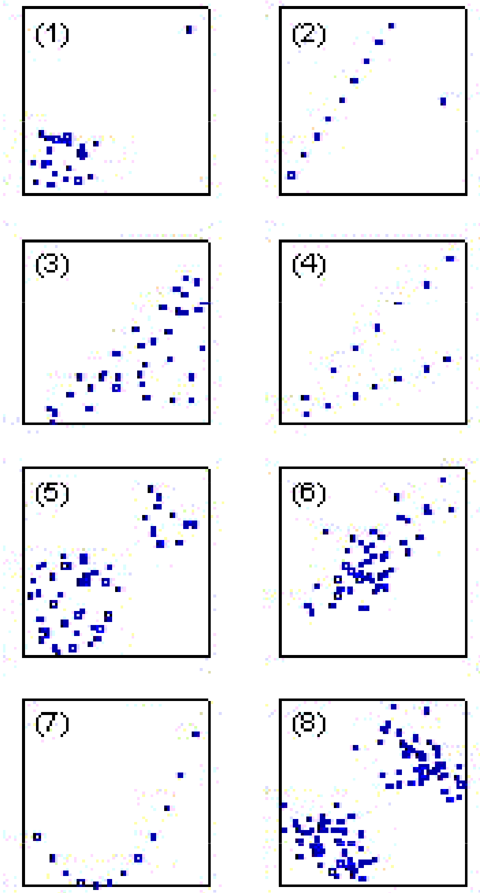
Different forms of associations







LINEAR RELATIONSHIPS



Correlation (=standardized shared variance)

- For better interpretation, we standardize the covariance – same as with z-scores, we divide the deviation score by standard deviation

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - m_x}{s_x}\right)\left(\frac{Y_i - m_y}{s_y}\right) = \frac{c_{xy}}{s_x s_y}$$

- We already know the circled part: that's z-score transformation, so to make it easier:

$$r_{xy} = \frac{\sum_{i=1}^{n} z_{X_i} z_{Y_i}}{n-1}$$

= Pearson's product-moment correlation
(cz: součinový, momentový koeficient korelace)

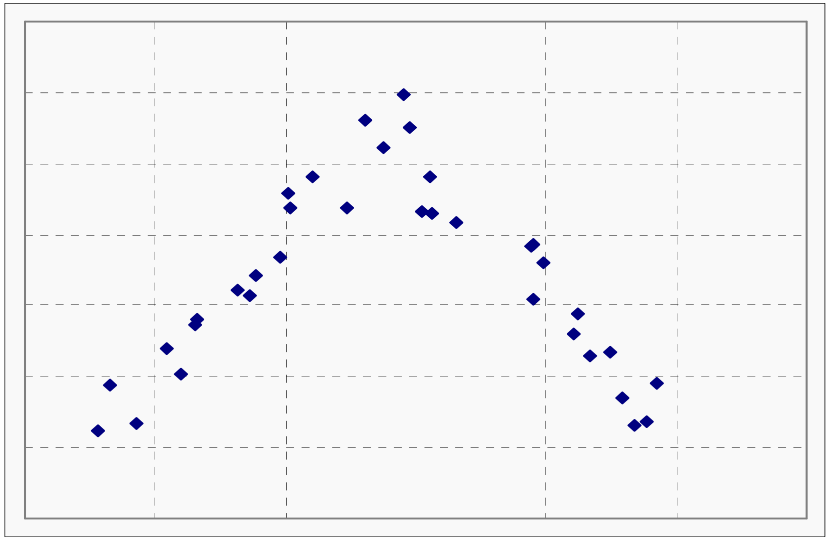
Characteristics of Pearson’s correlation coefficient

- A deviation statistics:
    - interval or higher measurement level required
    - great impact of extreme values
    - suitable for normally distributed variables description (approximately normal distribution of both variables required)
    - expresses only the association strength, not causality!!!

- Takes values between -1 a +1
    - 0 = no association
    - +1(-1) = perfect positive (negative) association = variables identity
      = direct (indirect relationship)

Characteristics of Pearson's correlation coefficient

- r^2 = coefficient of determination (R^2, D) = proportion of shared variance
- Consequence: r 0,3 – r 0,1 ≠ r 0,7 – r 0,5 (0,09-0,01=0,08; 0,49-0,25=0,24)
- r=0 doesn't mean there isn't any relationship between variables, it only means there is no linear association between them



Computing correlation

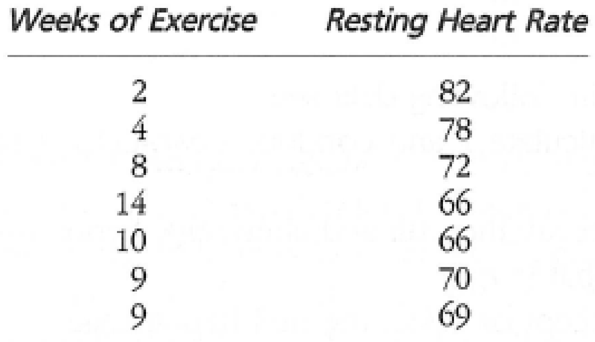
1. Check assumptions: interval or higher measurement level, normal distribution of both variables, extreme values, assumption of linear relationship **(plot <u>scatterplot</u> and histograms)**
2. Compute z-scores for all observed scores – you will need M and SD for both groups: **z_i = (X_i – M) / SD**

   Excel: =AVERAGEA(data), =PRŮMĚR(data), =STDEVA(data), =SMODCH.VÝBĚR.S(data), =STANDARDIZE(X;M;SD)
3. Compute correlation:

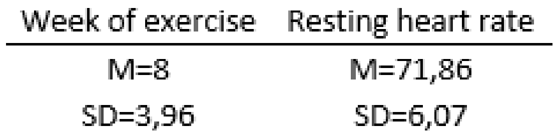
$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - m_x}{s_x}\right)\left(\frac{Y_i - m_y}{s_y}\right) = \frac{c_{xy}}{s_x s_y}$$

r_{xy} = [(-1,52*1,67) + (-1,01*1,01) + (0,00*0,02) + (1,52*-0,97)
+ (0,51*-0,97) + (0,01*-0,31) + (0,25*-0,47)] / (7-1) = **-0,94**

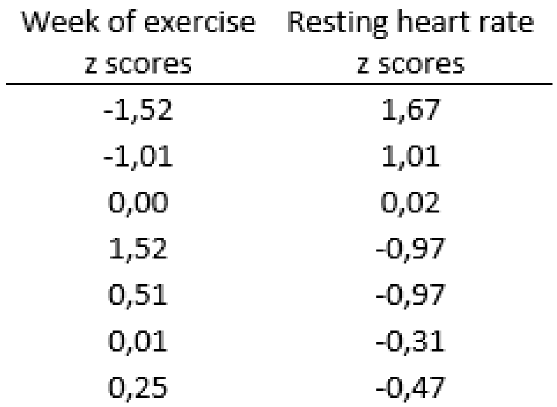
Excel: =COVARIANCE.P(var1, var2), =COVARIANCE.S(var1, var2),
=CORREL(var1, var2)



Weeks of Exercise	*Resting Heart Rate*
2	82
4	78
8	72
14	66
10	66
9	70
9	69



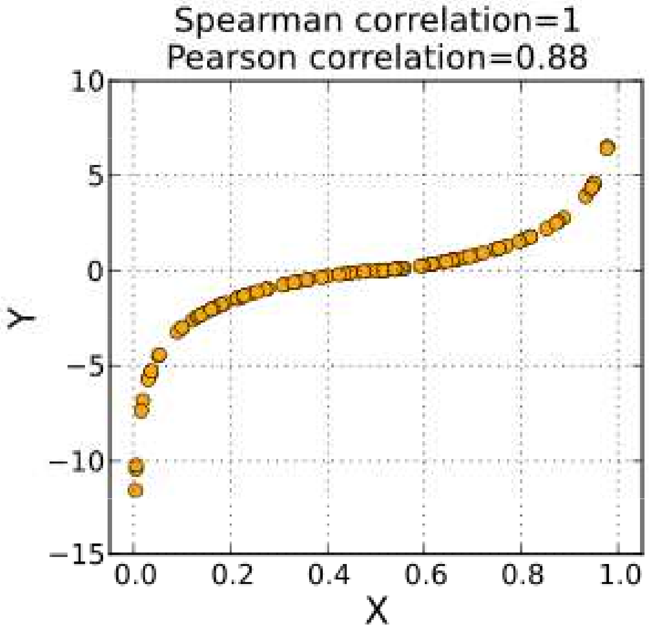
Week of exercise	Resting heart rate
M=8	M=71,86
SD=3,96	SD=6,07



Week of exercise z scores	Resting heart rate z scores
-1,52	1,67
-1,01	1,01
0,00	0,02
1,52	-0,97
0,51	-0,97
0,01	-0,31
0,25	-0,47

Rank (ordinal) correlation coefficients

- Suitable not only for ordinal data, but also for interval data with deviations from normal distribution
- Capture also nonlinear monotonic relationships
- To what extent are ranking of the two correlated variables the same
- Spearman rho coefficient - ρ, r_s
  - Based on differences magnitude in rankings
  - Ordinal equivalent for Pearson's correlation
  - r^2 can be interpreted
  - Usually used as a more resistant variant of Pearson's r
  - Calculated in the same way as Pearson's, but on rankings
- Kendall tau coefficient – t (+ „b“ and „c“ variants)
  - Based on number of values „out of order“
  - No effect of outliers
  - b and c variants deal with more values of the same ranking



Statistical prediction

- **Example 1:** Imagine that all students have exactly the same grades from math and physics. The variables would be identical:
  - **What would be the value of correlation between the variables?**

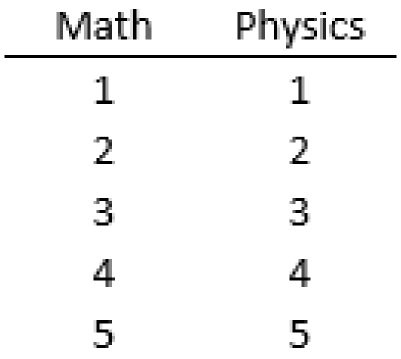
    $r = 1$
  - **What would be the value of coefficient of determination?**

    r^2 (R^2, D) = $1^2 = 1$
  - **What would be the proportion of shared variance? What does it mean?**

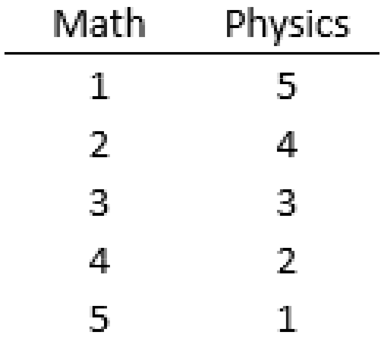
    $R^2 * 100 = 1 * 100 = 100\%$

    It means that we can predict 100% of math grades values correctly from physics values (or the other way).
  - **What of the information above would change if all students had exactly opposite math grades than physics grades?**

    $r = -1$, $R^2 = -1^2 = 1$, $R^2 * 100 = 100\%$
- Of course, usually we can predict with much less precision, but we’re trying to predict with the highest precision. For that, we need correlates with high correlation with the predicted variable.



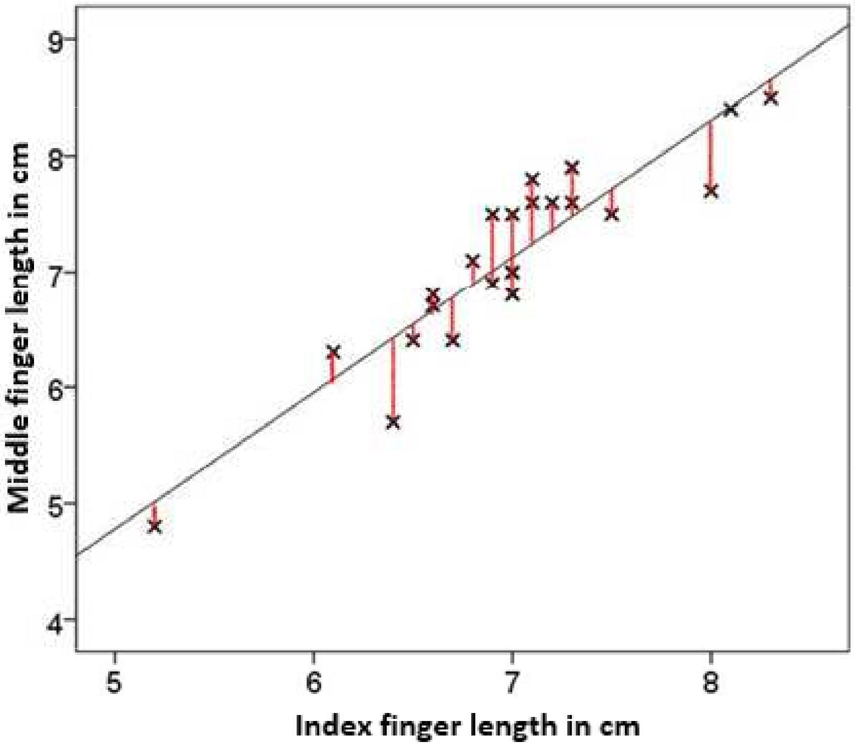
Math	Physics
1	1
2	2
3	3
4	4
5	5



Math	Physics
1	5
2	4
3	3
4	2
5	1

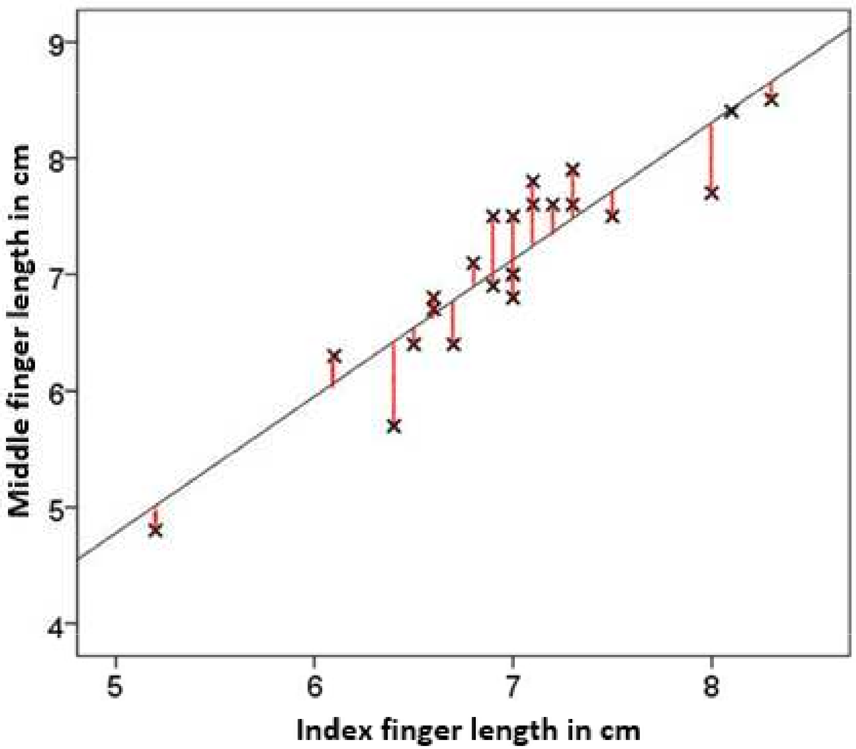
Linear regression

- If Pearson correlation well describes relationship between two variables, we can express the relationship by linear function:
- Y' = a + bX; *Y = Y' + e = a + bX + e*
- a = intercept (cz: průsečík), b = slope (směrnice)
- How can we find the best regression line?
    - Estimate by **least squares estimation** – we are trying to minimize the sum of residual squares
- b = $r_{xy}(SD_y/SD_x)$
- a = $M_y - bM_x$
- If the values of X and Y are in z-scores, then b = r_{xy}
- a, b – correlation coefficients
- The line goes through values M_x and M_y
- The sum of residuals is zero, the sum of squared residuals is the least possible



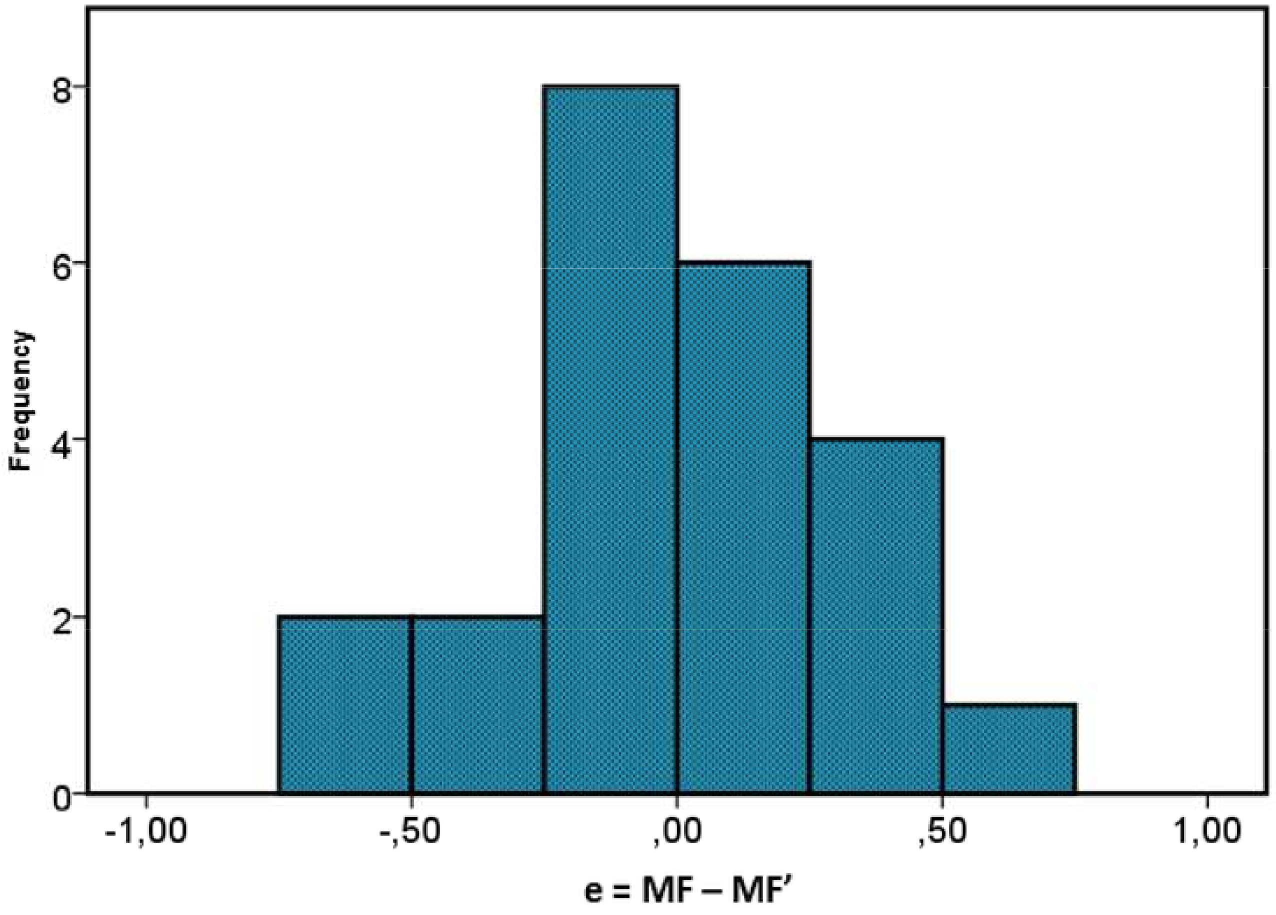
Predicted values

IF	MF	MF'
6,5	6,4	6,5413
7	7	7,1291
7,5	7,5	7,7169
5,2	4,8	5,0130
6,6	6,7	6,6589
6,6	6,8	6,6589
7	7	7,1291
6,8	?	



Y´ = 1,175*X – 1,096 = 1,175*6,8 – 1,096 = 6,894

Distribution of residuals



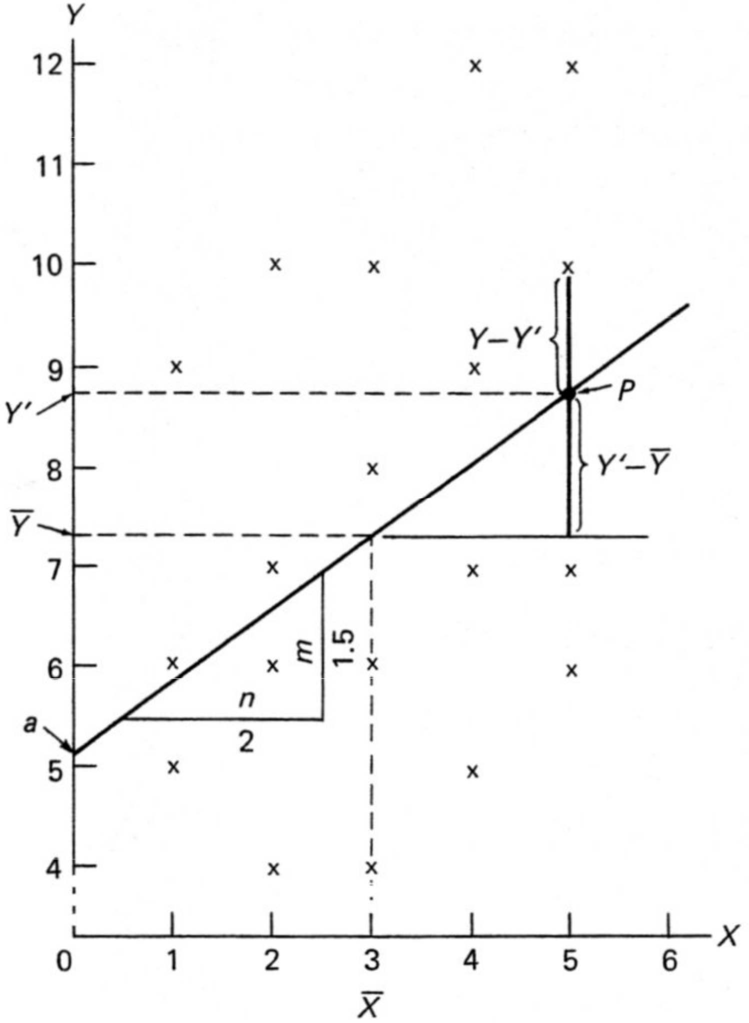
- M_e = 0
- SD_e = 0,337

Linear regression: model fit

$$s_{reg}^2 = \frac{\sum (m_y - Y')^2}{n-1} \qquad s_{res}^2 = \frac{\sum (Y - Y')^2}{n-1}$$

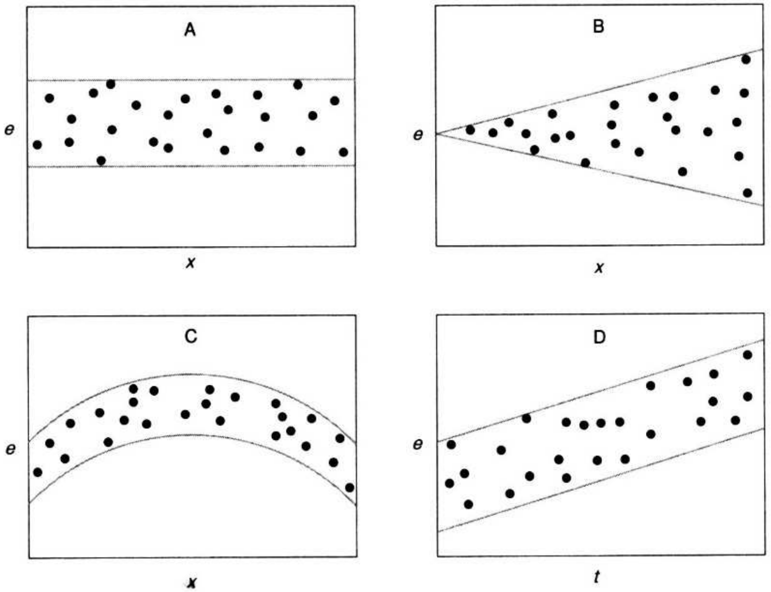
$$s_y^2 = \frac{\sum (Y - m_y)^2}{n-1}$$

- $s_y^2 = s_{reg}^2 + s_{res}^2$
- $R^2 = s_{reg}^2 / s_y^2$ ... $s_{res}^2 = s_y^2(1-R^2)$
- Coefficient of determination: R^2
    - Exaplained variance proportion
    - Measure of model fit with the data (regression success)
- **For simple linear regression it applies: $R^2 = r^2$**



Linear regression: assumptions

- Assumptions are the same as for Pearson correlation:
  - The basic assumption: the relationship really is linear
  - The residuals have normal distribution with M=0 and SD = S_{res}
  - It means the 95% of estimation residuals lie approx. between $-2s_{res}$ a $+2s_{res}$
  - **homoscedascity** (cz: homoskedascita): residuals independency = the residual variance won't change with increasing X

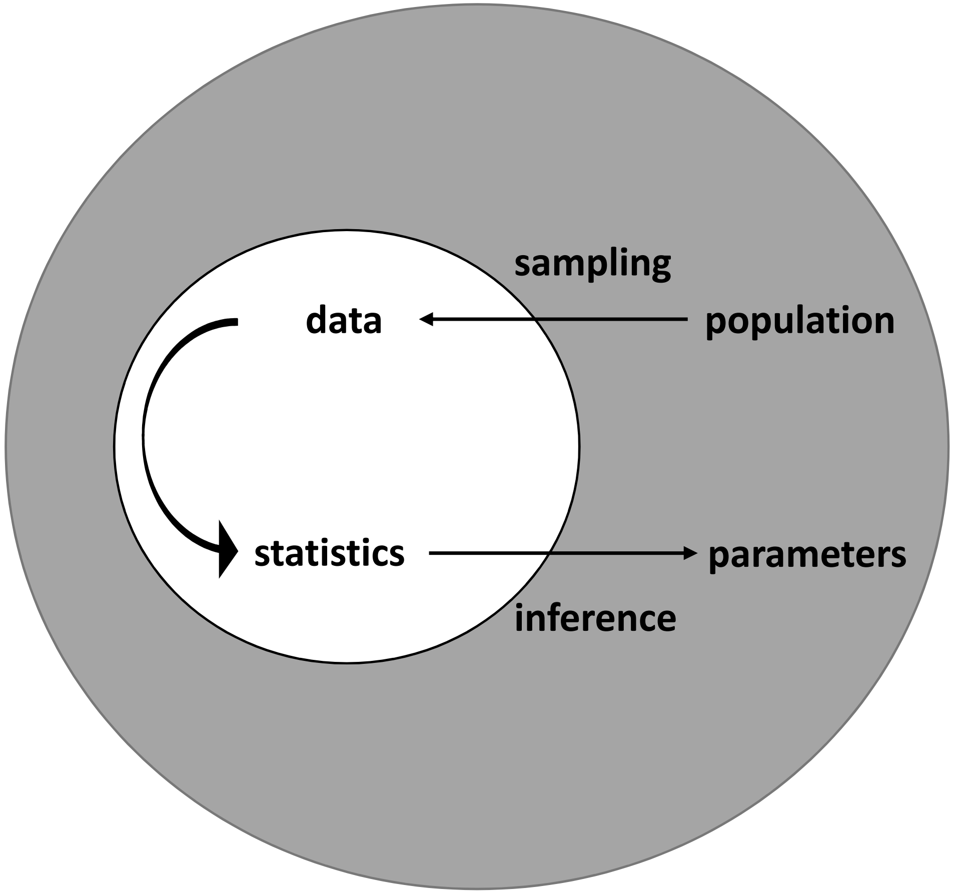


  - The model validity depends on data from which was the model extrapolated
  - Watch out for extreme values (as with all deviation statistics)

Other regression types

- Simple linear regression: one independent and one dependent variable
- Multiple linear regression: more independent variables (predictors)
  - $Y = a + b_1X_1 + b_2X_2 + \ldots + b_mX_m$
  - complicated by relationships between the predictors
- Logistic regression:
  - Dependent variable is dichotomous (nominal)
  - Prediction of dependent variable values probability
- If the relationship isn‘t linear:
  - We can try to transform the variables, so that the relationship becomes linear
  - We can divide the sample into subgroups in which the relationship is linear

From description to inference



- Data description, parameters estimation
- Statistical inference (cz: usuzování, inference, indukce)
- Random sampling
  - every subject has the same probability of being included in the sample
  - If we don’t have random sampling, how is our sample different?

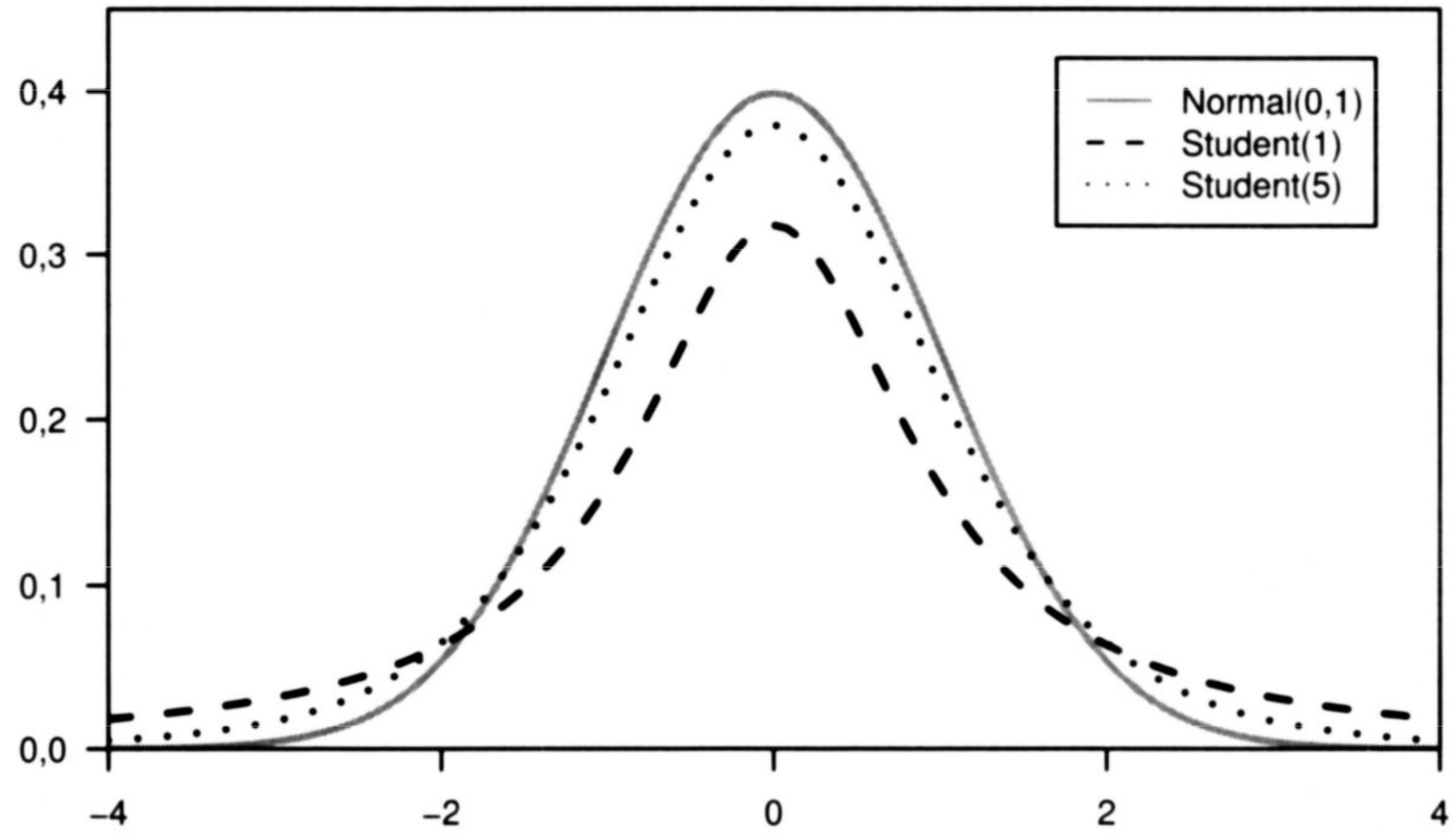
Sampling distribution and standard error

- If we compute the same statistics on many independent samples from a population, we get many different parameter estimates
- These estimates have some distribution – **sampling distribution (cz: výběrové rozložení)**
- **http://onlinestatbook.com/stat_sim/sampling_dist/index.html**
- Sampling distribution can be described by:
  - Mean – sampling distribution mean is close to parameter
  - Standard deviation – in sampling distribution called **standard error (cz: směrodatná chyba,** také střední chyba či výběrová chyba)
  - The higher is the number of samples (statistics estimates), the lower is the standard error

Sampling distribution of mean (estimate)

- Mean estimate has approximately **normal distribution**
  - With mean μ and standard error: $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$
  - This applies even when the distribution is not normal – thanks to **central limit theorem**
  - The problem is that we usually don't know σ
- If we don't know σ, we have to use *s*
  - Mean is still μ and standard error is now: $$s_{\bar{x}} = \frac{s}{\sqrt{N}}$$
  - The sampling distribution is not normal, but Student's t-distribution

Student's t-distribution



Sampling distribution of other statistics

- For every statistics we need to know it theoretical sampling distribution
  - Relative frequencies: approximately normal distribution
  - Pearson’s r – after Fisher transformation normal distribution

Point vs. Interval estimates

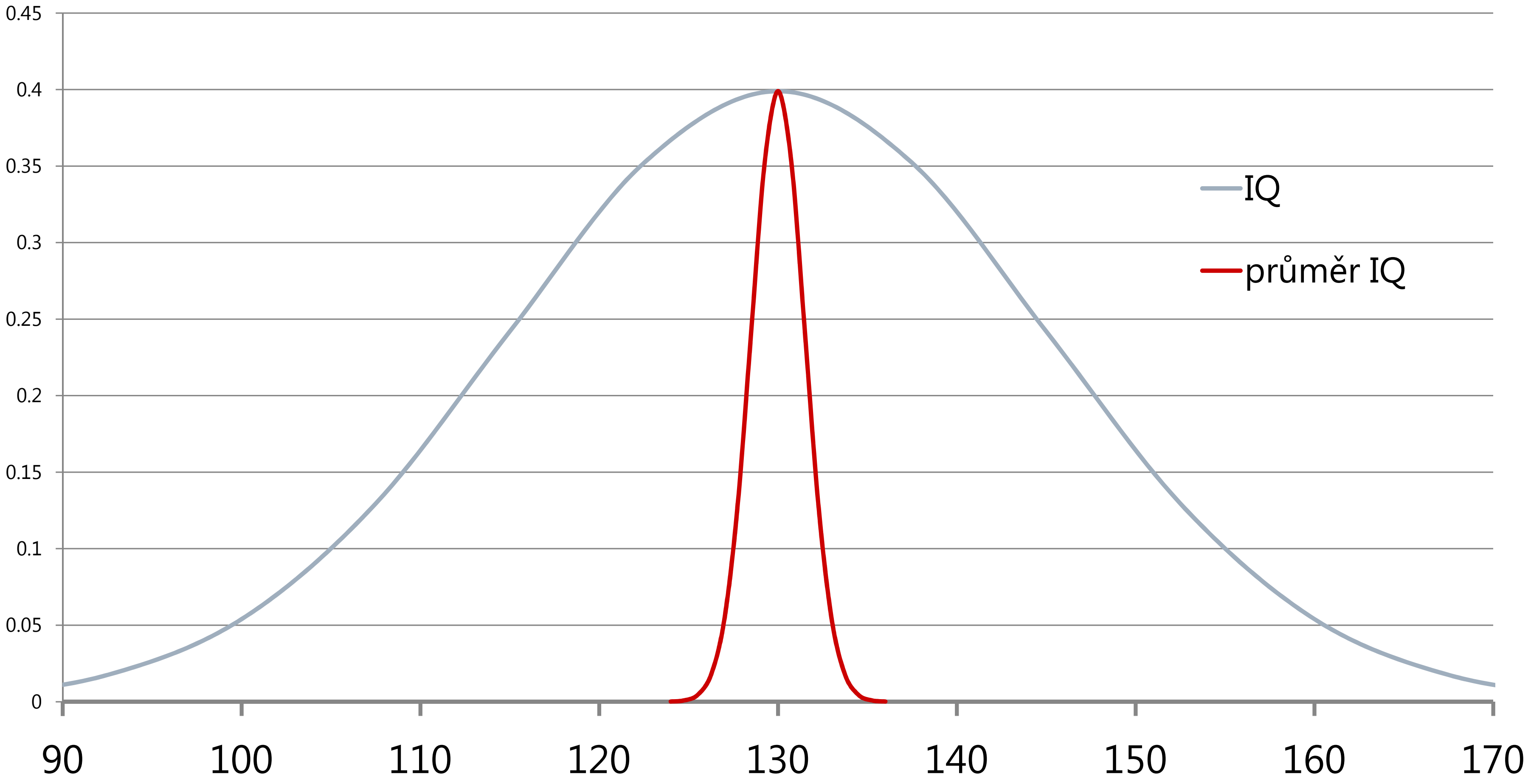
- Parameter can be estimated by:
  - Point estimate – that we’re trying to estimate the parameter value itself (e.g. mean)
  - Interval estimate – estimating an interval in which the parameter lies with certain probability
    - The result is **confidence interval (cz: interval spolehlivosti), CI**
    - Confidence interval can be computed from point estimate and its sampling distribution (point ± deviation)
    - Interval estimate is better – we have more information

$$_{(1-\alpha)}CI = \overline{X} \pm {}_{1-\alpha/2}z\,\sigma_{\overline{X}}$$

    - α – error probability, (1 – α) is **confidence level (cz: hladina spolehlivosti)**
    - We typically use 95% or 99% confidence level, then it means that the parameter lies in the confidence interval with 95% probability (where α is 0.05 = 5% error probability, (1 – α) = (1 – 0.05))

Computing confidence interval for mean 1

- In a sample of 100 children with multicolored eyes we computed mean IQ 130 and we know that σ =15.
  - Point parameter estimate (μ) is 130
  - Interval parameter estimate:
    - **Sampling distribution of mean is normal**…
    - …with centre in μ. We don't know μ, so we use our **point interval m = 130**.
    - …with **standard error of mean $s_m = \sigma/\sqrt{N}$** $= 15/\sqrt{100} =$ 1,5.
    - We **choose our confidence level**: $1-\alpha = 95\%$
    - Then we find **z-score between which lies 1-α % of normal distribution**:
      95% of normal distribution lies between z-scores -1.96 and 1.96
      in other words: ${}_{1-\alpha/2}z = {}_{0,975}z = 1.96$
      Excel: **=NORM.S.INV(0.975)**
    - Confidence interval: **$(m - 1.96s_m; m + 1{,}96s_m)$** = (130 – 1.96*15; 130 + 1.96*15) = (127.1 ; 132.9)
    - **That is: with 95% probability $127.1 \leq \mu \leq 132.9$**


0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
90
100
110
120
130
140
150
160
170
IQ
průměr IQ

Computing confidence interval for mean 2

- In a sample of 100 children with multicolored eyes we computed mean IQ 130 and s = 15.
  - Point parameter estimate (μ) is 130
  - Interval parameter estimate:
    - We don't know σ, so the sampling distribution of mean is not normal, but **Student's t-distribution with df = N-1 = 99**
    - The distribution centre will again be our **point interval m = 130**.
    - **Standard error of mean** is $s_m = s / \sqrt{n} = 15 / \sqrt{100} = 1,5$
    - We **choose our confidence level**: $1-\alpha = 95\%$
    - Then we find **t-score between which lies 1-α % of t-distribution**: 95% of t-distribution with df=99 lies between t-scores -1.98 and 1.98 in other words: ${}_{1-\alpha/2}t\,(df) = {}_{0,975}t\,(99) = 1.98$
    - Excel: **=T.INV(p;df), here =T.INV(0.975;99)**
    - Confidence interval: **(m – 1.98s_m; m + 1,98s_m)** = (130 – 1.98*15; 130 + 1.98*15) = (127.0 ; 133.0)
    - **That is: with 95% probability 127.0 $\leq \mu \leq$ 133.0**

Sampling distribution of relative frequencies *p*

- …is approximately normal with mean *p* and standard error $\sqrt{p(1-p)/n}$
- (1-α)% confidence interval thus is:

$$(p - z_{1-\alpha/2}\sqrt{p(1-p)/n};\ p + z_{1-\alpha/2}\sqrt{p(1-p)/n})$$

Sampling distribution of Pearson's correlation *r*

- We don't know sampling distribution of correlation…
- …but we know sampling distribution of correlation after Fisher transformation: in Excel: *Z* = FISHER(r)
- Sampling distribution of the Fisher *Z* is approximately normal with mean *Z* and standard error s_Z=1/√(n-3)
- (1−α)% CI for *Z*: $\left(Z - z_{1-\alpha/2}s_Z ; Z + z_{1-\alpha/2}s_Z\right)$
- Then, we need to transform Fisher *Z* back to Pearson's *r:* in Excel: =FISHERINV(z)

$$\left(FISHERINV\ \left(Z - z_{1-\alpha/2}s_Z\right); FISHERINV\ \left(Z + z_{1-\alpha/2}s_Z\right)\right)$$

Confidence interval for correlation

On a sample of 20 children we found correlation between number of hours spend by reading per week and score in a creativity test r=0.45. Compute 90% confidence interval for the correlation.

- Transform Pearson's correlation to Fisher Z in Excel: =FISHER(0.45) = 0.48
- Sampling distribution of Fisher Z is approximately normal
- Compute standard error for Fisher Z: s_Z = 1/√(20-3) = 0.24
- Compute border z-scores for 90% confidence interval:
  =NORM.S.INV(1-0.1/2) =NORM.S.INV(0.95) = 1.64
- 90% CI for Fisher Z: (0.48 – 1.64*0.24; 0.48 + 1.64*0.24) = (0.09; 0.88)
- Tranform the results back to Pearson's r:
  90% CI for Pearson's r: (=FISHERINV(0.09); =FISHERINV(0.88)) = (0.09; 0.71)

General procedure for computing CIs

1. **Determine sampling distribution of given statistics:**
   - for mean with known σ: normal distribution (z-scores)
   - for mean with unknown σ: Student's t-distribution (t-scores) with df=N-1
   - for relative frequency: normal distribution (z-scores)
   - for Pearson's correlation: normal after Fisher transformation (then z-scores)
2. **If needed, transform the statistics:**
   - from the above only for correlation, Excel: =FISHER(r)
3. **Determine standard error for given sampling distribution:**
   - for mean with known σ: $s_m = \sigma / \surd N$
   - for mean with unknown σ: $s_m = s / \surd N$
   - for relative frequency: $s_p = \surd(p(1-p)/N)$
   - for Fisher Z: $s_Z = 1 / \surd(N-3)$

General procedure for computing CIs

4. Determine point estimate for given sampling distribution:

- m, p, r

5. Choose confidence level – typically 95% or 99% (theoretically any):

- for 95%: α = 0.05 = 5% error probability
  1 – α = 1 – 0.05 = 0.95 = 95%,
  1 – α/2 = 1 – 0.025 = 0.975

6. Find boundary scores between which lies (1 – α) % of given distribution:

- for normal distribution (in Excel): =NORM.S.INV(1 – α/2)
- for t-distribution (in Excel): =T.INV(1 – α/2; df)

5. Compute confidence interval:

- CI = point estimate ± boundary score*standard error
- normal distribution: CI = point estimate ± ${}_{1-\alpha/2}$z*standard error
- t-distribution: CI = point estimate ± ${}_{1-\alpha/2}$t(df)*standard error

$$\frac{\sum x^2}{N} \qquad \frac{\sum x^2}{N\ -1} \qquad z_i = \frac{x_i - M}{SD} \qquad T_i = 50 + 10z_i$$

$$c_{xy} = \frac{1}{N-1}\sum_{i=1}^{n} x_i y_i \qquad r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - m_x}{s_x}\right)\left(\frac{Y_i - m_y}{s_y}\right) = \frac{c_{xy}}{s_x s_y}$$

$$r_{xy} = \frac{\sum_{i=1}^{n} z_{X_i} z_{Y_i}}{n-1} \qquad b = r_{xy}\left(\frac{SD_y}{SD_x}\right) \qquad a = M_y - bM_x$$

$$s^2_{reg} = \frac{\sum (m_y - Y')^2}{n-1} \qquad s^2_{res} = \frac{\sum (Y - Y')^2}{n-1} \qquad s^2_y = \frac{\sum (Y - m_y)^2}{n-1} \qquad R^2 = \frac{S^2_{reg}}{s^2_y}$$

$$s^2_y = s^2_{reg} + s^2_{res}$$

$$s^2_{res} = s^2_y\,(1 - R^2)$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \qquad s_{\bar{x}} = \frac{s}{\sqrt{N}} \qquad s_z = \frac{1}{\sqrt{(N\ - 3)}} \qquad s_p = \sqrt{\frac{p(1-p)}{N}}$$

A researcher is interested in the influence of media on fear of immigrants. He included ten people in his research. He asked every respondent how many articles about immigrant he read during the past week. Each respondent filled in a questionnaire which measured their fear of immigrants. Based on the questionnaire the researcher assigned to every respondent a number between 1 and 8, where 1 means no fear of immigrants and 8 means extreme fear of immigrants. The following data are hypothetical:

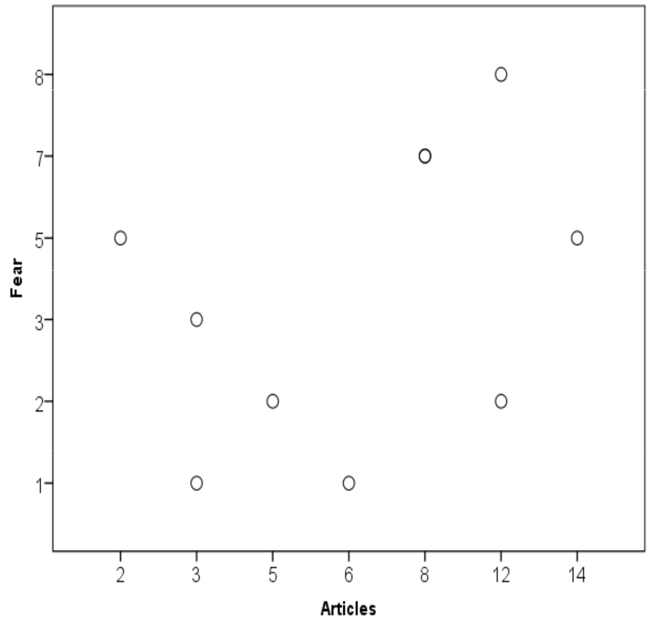
Number of articles: 5, 8, 12, 3, 12, 6, 8, 14, 2, 3

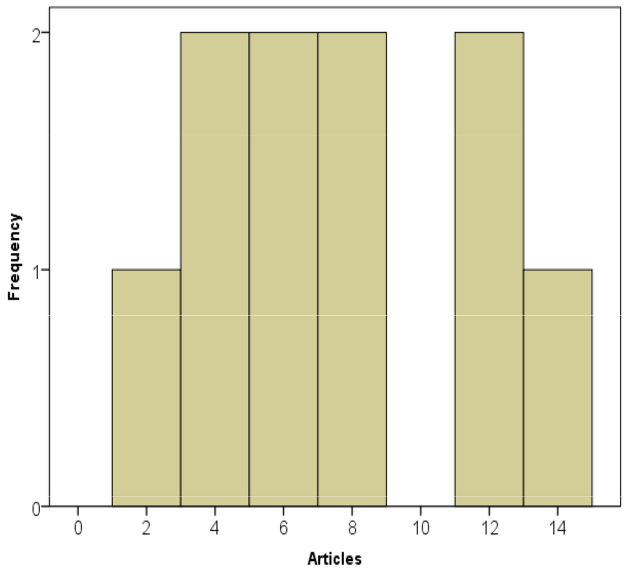
Fear of immigrants: 2, 7, 2, 1, 8, 1, 7, 5, 5, 3

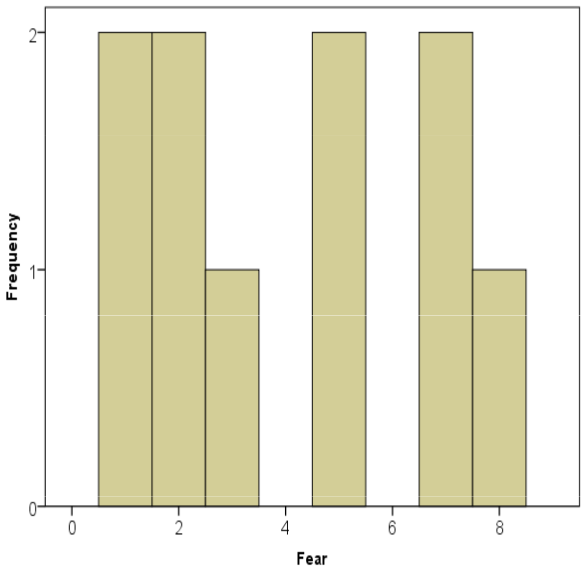
1. Create frequency tables for both variables. Draw histograms for both variables.
2. Draw scatterplot. What would you say about the relationship of the variables from the scatterplot?
3. Compute regression equation for prediction of fear of immigrants from number of articles read per week.
4. How successful is our prediction? Compute percentage of explained variance. Compute residual variance. Predict fear of immigrants for a person who reads 10 articles per week.
5. Compute 95% confidence interval for mean of number of articles read per week.
6. Compute 99% confidence interval for mean of fear from immigrants.

Articles	Frequency	Percent	Cumulative percent
2	1	10	10
3	2	20	30
5	1	10	40
6	1	10	50
8	2	20	70
12	2	20	90
14	1	10	100
Total	10	100	

Fear	Frequency	Percent	Cumulative percent
1	2	20	20
2	2	20	40
3	1	10	50
5	2	20	70
7	2	20	90
8	1	10	100
Total	10	100	







Although we have a small dataset, from the scatterplot we can assume on moderate positive linear relationship

To compute regression equation, we have to means and standard deviations for both variables and their correlation. We want to predict fear of immigrants from number of articles read. So, fear is our dependent or predicted variable Y and articles are our predictor (or independent variable) X.

$\mathbf{M_x}$ (articles) = (5+8+12+3+12+6+8+14+2+3)/10 = **7.3**

$\mathbf{M_y}$ (fear) = (2+7+2+1+8+1+7+5+5+3)/10 = **4.1**

$S^2_x = [(-2.3)^2 + (0.7)^2 + (4.7)^2 + (-4.3)^2 + (4.7)^2 + (-1.3)^2 + (0.7)^2 + (6.7)^2 + (-5.3)^2 + (-4.3)^2] / 9 =$

$= 18.01$

$\mathbf{SD_x} = \surd 18.01 = \mathbf{4.24}$

$S^2_y = [(-2.1)^2 + (2.9)^2 + (-2.1)^2 + (-3.1)^2 + (3.9)^2 + (-3.1)^2 + (2.9)^2 + (0.9)^2 + (0.9)^2 + (-1.1)^2] / 9 =$

$= 6.99$

$\mathbf{SD_y} = \surd 6.99 = \mathbf{2.64}$

Compute z-scores by $z_i = (X_i – M) / SD$:

$\mathbf{r_{xy}} = [(-0.54)*(-0.79) + 0.16*1.10 + 1.11*(-0.79) + (-1.01)*(-1.17) +$

$+ 1.11*1.48 + (-0.31)*(-1.17) + 0.16*1.10 + 1.58*0.34 + (-1.25)*0.34 +$

$+ (-1.01)*(-0.42)] / 9 = \mathbf{0.403}$

X	x
5	-2,3
8	0,7
12	4,7
3	-4,3
12	4,7
6	-1,3
8	0,7
14	6,7
2	-5,3
3	-4,3

Y	y
2	-2,1
7	2,9
2	-2,1
1	-3,1
8	3,9
1	-3,1
7	2,9
5	0,9
5	0,9
3	-1,1

Zx	Zy
-0.54	-0.79
0.16	1.10
1.11	-0.79
-1.01	-1.17
1.11	1.48
-0.31	-1.17
0.16	1.10
1.58	0.34
-1.25	0.34
-1.01	-0.42

Regression for prediction fear of immigrants from number of articles read:

b = 0.403*(2.64/4.24) = 0.25

a = 4.1 – 0.25*7.3 = 2.28

Y’ = 2.28 + 0.25X

Predicted fear for a person who reads 10 articles per week:

Y’ = 2.28 + 0.25*10 = 4.78

Percentage of explained variance:

R^2 in simple regression = r^2 = 0.403^2 = 0.16

Residual variance:

$s_{res}^2 = \ s_y^2\,(1 - R^2)$ = 6.99 (1 – 0.16) = 5.87

95% confidence interval for mean of number of articles read per week:

M_x (articles) = 7.3

SD_x = 4.24

$s_{mx} = SD_x$ / √N = 4.24 / √10 = 1.34

$_{1-\alpha/2}t$ = T.INV(1-α/2;df) = T.INV(0.975;9) = 2.26

$CI_{95\%} = (M_x - {}_{1-\alpha/2}t * s_{mx} ; M_x - {}_{1-\alpha/2}t * s_{mx}) = (7.3 – 2.26*1.34 ; 7.3 + 2.26*1.34) = (4.27 ; 10.33)$

99% confidence interval for mean fear of immigrants:

M_y (fear) = 4.1

SD_y = 2.64

$s_{my} = SD_y$ / √N = 2.64 / √10 = 0.83

$_{1-\alpha/2}t$ = T.INV(1-α/2;df) = T.INV(0.995;9) = 3.25

$CI_{95\%} = (M_y - {}_{1-\alpha/2}t * s_{my} ; M_y - {}_{1-\alpha/2}t * s_{my}) = (4.1 – 3.25*0.83 ; 4.1 + 3.25*0.83) = (1.40 ; 6.80)$

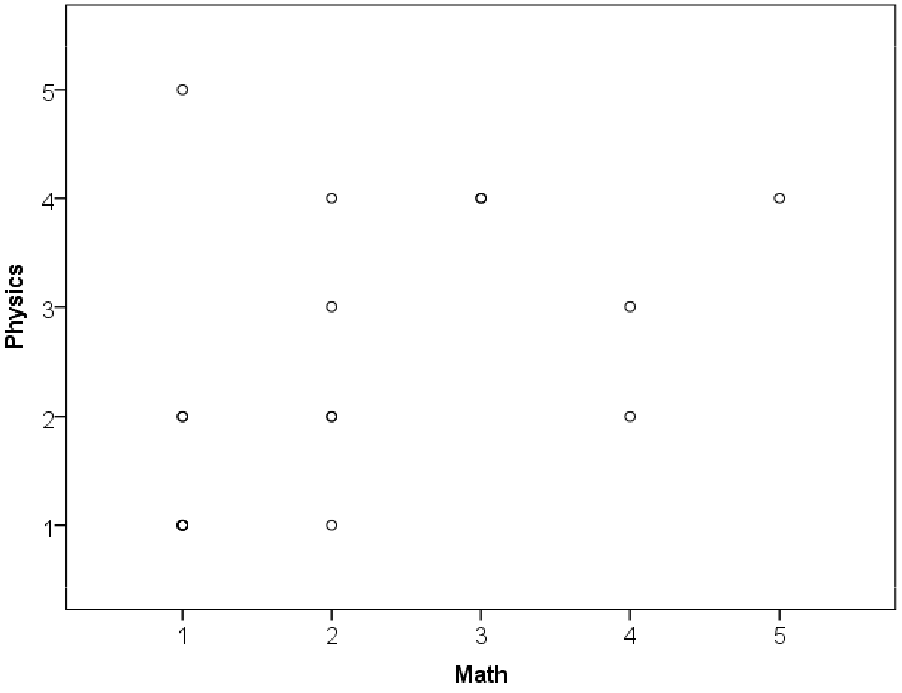
Bellow you can see grades of 18 students from math and from physics:

Math: 1, 2, 1, 4, 2, 1, 5, 2, 3, 1, 4, 1, 1, 2, 1, 3, 1, 2

Physics: 1, 3, 2, 3, 2, 5, 4, 4, 4, 1, 2, 1, 2, 2, 1, 4, 1, 1

1. Draw contingency table for the two variables, then draw scatterplot.
2. What would you judge about the relationship of the variables from the contingency table and scatterplot?
3. What descriptive statistics would you use for these variables and why?
4. Compute mode, median and interquartile range for both variables. Draw boxplots for both variables.
5. What correlation coefficient would you use for inspection of the relationship between these two variables?

		Math					Total
		1	2	3	4	5	
Physics	1	5	1	0	0	0	6
	2	2	2	0	1	0	5
	3	0	1	0	1	0	2
	4	0	1	2	0	1	4
	5	1	0	0	0	0	1
Total		8	5	2	2	1	18

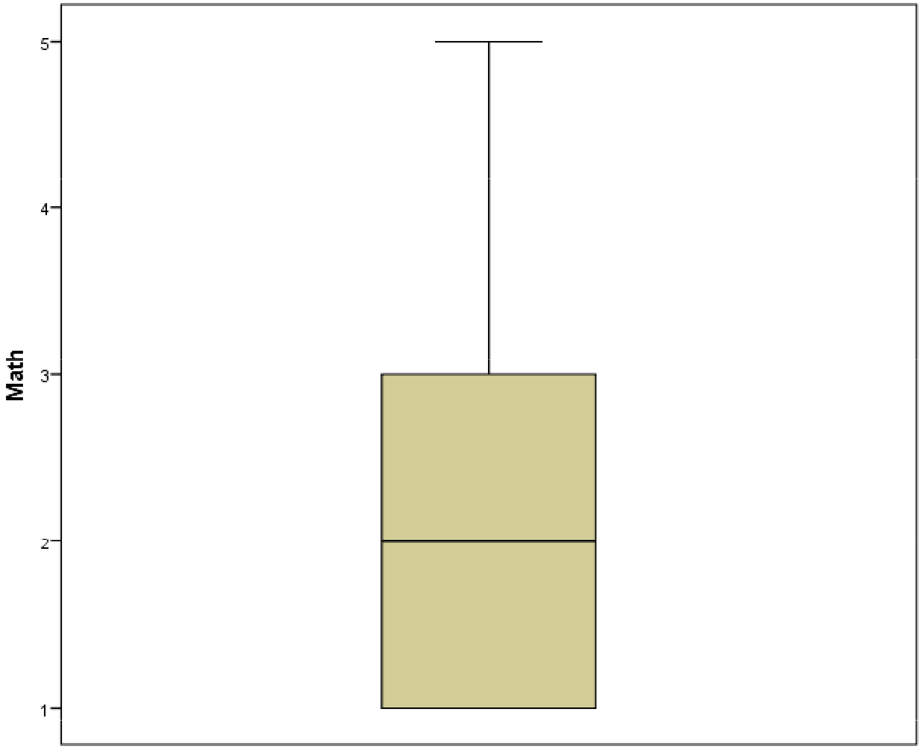


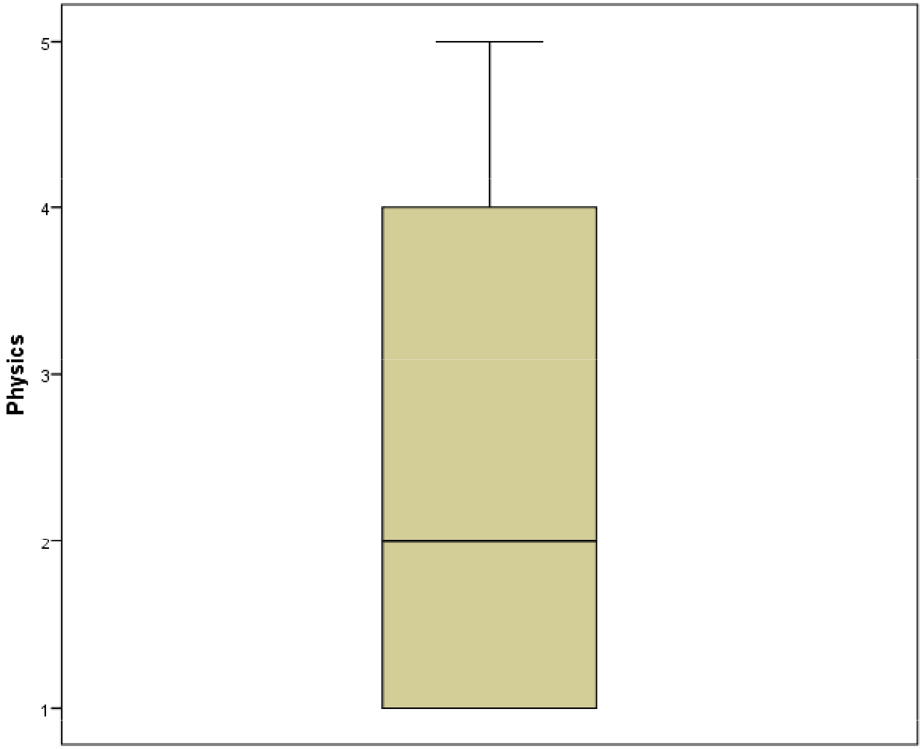
Based on the scatterplot, we can assume on low to moderate positive relationship. The student with 1 from math and 5 from physics would decrease the correlation substantially.

Because grades are ordinal variable, we would prefer to use median and inter-quartile range (and present freqeuncy tables). For finding median and quartiles, rank-order the variable values:

Math rank-ordered values:
1, 1, 1, 1, **1**, 1, 1, 1, 2, **|** 2, 2, 2, 2, **3**, 3, 4, 4, 5
The median is 2 (lies „between values 2 and 2“)
Q1 is 1 and Q3 is 3, mode is 1
(or look in frequency table)
IQR = Q3 – Q1 = 3 – 1 = 2

Physics rank-ordered values:
1, 1, 1, 1, **1**, 1, 2, 2, 2, **|** 2, 2, 3, 3, **4**, 4, 4, 4, 5
The median is 2, mode is 1
Q1 is 1 and Q3 is 4
(or look in frequency table)
IQR = Q3 – Q1 = 4 – 1 = 3





We would prefer to use Spearman rho correlation. (r_s = 0.55 – only to complete, you don’t need to compute the correlation)

We measured correlation between number of hours spent by exercising per month and depression rate r = -0.12 on a sample of 120 people. Compute 95% confidence interval for correlation.

We measured mean intelligence of 200 high school students with special program for talented children M = 128. Compute 99% confidence interval for mean intelligence of students educated with this special program if you know that population standard deviation of intelligence in σ = 15.

From a research study we know that the probability of having schizophrenia if one of your parents has it is p = 0.45. The study was conducted with 30 participants. Compute 90% confidence interval for probability of having schizophrenia if one of your parents has it.

We measured correlation between number of hours spent by exercising per month and depression rate r = -0.12 on a sample of 120 people. Compute 95% confidence interval for correlation.

$r = -0.12$

$N = 120$

Z =FISHER(-0.12) = -0.121

sz = 1/√(120-3) = 0.09

${}_{1-\alpha/2}z$ = NORM.S.INV(1-α/2) = NORM.S.INV (0.975) = 1.96

$CI_{95\%}$ = (FISHERINV(Z - ${}_{1-\alpha/2}z$ * s_z) ; FISHERINV(Z - ${}_{1-\alpha/2}z$ * s_z))
= (FISHERINV(-0.121 – 1.96*0.09) ; FISHERINV(-0.121 + 1.96*0.09))
= (FISHERINV(-0.30) ; FISHERINV(0.06)) = (-0.29 ; 0.06)

We measured mean intelligence of 200 high school students with special program for talented children M = 128. Compute 99% confidence interval for mean intelligence of students educated with this special program if you know that population standard deviation of intelligence in σ = 15.

$M = 128$

$\sigma = 15$

$\sigma_m = \sigma / \sqrt{N} = 15 / \sqrt{200} = 1.06$

${}_{1-\alpha/2}z = \text{NORM.S.INV}(1-\alpha/2) = \text{NORM.S.INV}\ (0.995) = 2.58$

$CI_{99\%} = (M - {}_{1-\alpha/2}z * \sigma_m\ ; M - {}_{1-\alpha/2}z * \sigma_m) = (128 – 2.58*1.06\ ; 128 + 2.58*1.06) = (125.3\ ; 130.7)$

From a research study we know that the probability of having schizophrenia if one of your parents has it is p = 0.45. The study was conducted with 30 participants. Compute 90% confidence interval for probability of having schizophrenia if one of your parents has it.

$p = 0.45$

$N = 30$

$s_p = \surd((p*(1-p))/N) = \surd((0.45*(1-0.45))/30) = 0.09$

${}_{1-\alpha/2}z = \mathrm{NORM.S.INV}(1-\alpha/2) = \mathrm{NORM.S.INV}\ (0.995) = 1.64$

$CI_{90\%} = (p - {}_{1-\alpha/2}z * s_p\ ;\ p - {}_{1-\alpha/2}z * s_p) = (0.45 – 1.64*0.09\ ;\ 0.45 + 1.64*0.09)$
$= (0.3\ ;\ 0.6) = (30\%\ ;\ 60\%)$