

MUNI
MED

MIAM021p(s) Analýza a management dat pro zdravotnické obory – přednáška a cvičení (jaro 2024)

MICHAL SVOBODA

Institut biostatistiky a analýz LF MU
svoboda@iba.muni.cz

Osnova

- Excel: opakování, příprava dat, základní vzorce
- Základy popisné statistiky
- Základní rozdělení pravděpodobnosti, testování hypotéz
- Parametrické testy
- Neparametrické testy
- Analýza kontingenčních tabulek
- Základy korelační analýzy a lineární regrese

Důležité informace

- Výuka: 11:00–13:30, D29/347-RCX2
- Materiály v IS
- Software: Microsoft Office - Excel, Statistica
- Pro získání zápočtu/kolokvia je třeba:
 - 1. Účast – povoleny jsou 2 absence**
 - 2. Domácí úkoly – povoleno 1 neodevzdání**
 - za účelem procvičení, dostanete zpětnou vazbu, na dalším cvičení se vrátíme, kdyby byl problém
 - 3. Závěrečný úkol** – praktické úkoly (povoleny materiály)

Organizace výuky

- 20. 2. – Excel: opakování, příprava dat, základní vzorce
- 27. 2. – Základy popisné statistiky
- 19. 3. – Základní rozdělení pravděpodobnosti, testování hypotéz
- 26. 3. – Parametrické testy
- 2. 4. – Neparametrické testy
- 9. 4. – Analýza kontingenčních tabulek, korelační analýza
- 16. 4. – Ukončení předmětu, test

**M U N I
M E D**

Základy popisné statistiky

Typy proměnných
Popisná statistika

Typy proměnných

- **Kvalitativní (kategoriální) proměnná**
Ize ji řadit do kategorií, ale nelze ji kvantifikovat
Příklad: ??
- **Kvantitativní (numerická) proměnná**
můžeme ji přiřadit číselnou hodnotu
Příklad: ??

Typy proměnných

- **Kvalitativní (kategoriální) proměnná**
Ize ji řadit do kategorií, ale nelze ji kvantifikovat
Příklad: pohlaví, HIV status, barva vlasů ...
- **Kvantitativní (numerická) proměnná**
můžeme ji přiřadit číselnou hodnotu
Příklad: výška, hmotnost, teplota, počet hospitalizací ...

Kvalitativní proměnné, znaky

- **Binární znaky:** dvě kategorie, obvykle se kódují pomocí číslic 1 (přítomnost znaku) a 0 (nepřítomnost znaku).

Příklad: ??

- **Nominální znaky:** několik kategorií (A, B, C), které nelze uspořádat.

Příklad: ??

- **Ordinální znaky:** několik kategorií, které lze vzájemně seřadit, tedy můžeme se ptát, která je větší/menší ($1 < 2 < 3$).

Příklad: ??

Kvalitativní proměnné, znaky

- **Binární znaky:** dvě kategorie, obvykle se kódují pomocí číslic 1 (přítomnost znaku) a 0 (nepřítomnost znaku).
*Příklad: Diabetes (1 – ano, 0 – ne),
Pohlaví (1 – muž, 0 – žena).*
- **Nominální znaky:** několik kategorií (A, B, C), které nelze uspořádat.
Příklad: krevní skupiny (A – B – AB – 0)
- **Ordinální znaky:** několik kategorií, které lze vzájemně seřadit, tedy můžeme se ptát, která je větší/menší ($1 < 2 < 3$).
*Příklad: stupeň bolesti (mírná – střední – velká)
stadium maligního onemocnění (I – II – III – IV)*

Popis a vizualizace kvalitativních proměnných

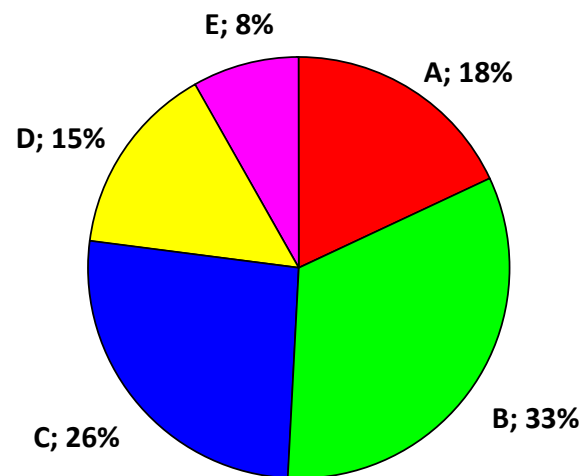
- **Popis kvalitativních dat:** četnost jednotlivých kategorií
- **Vizualizace kvalitativních dat:** koláčový nebo sloupcový graf

Příklad: Známká z biostatistiky (podzim 2014)

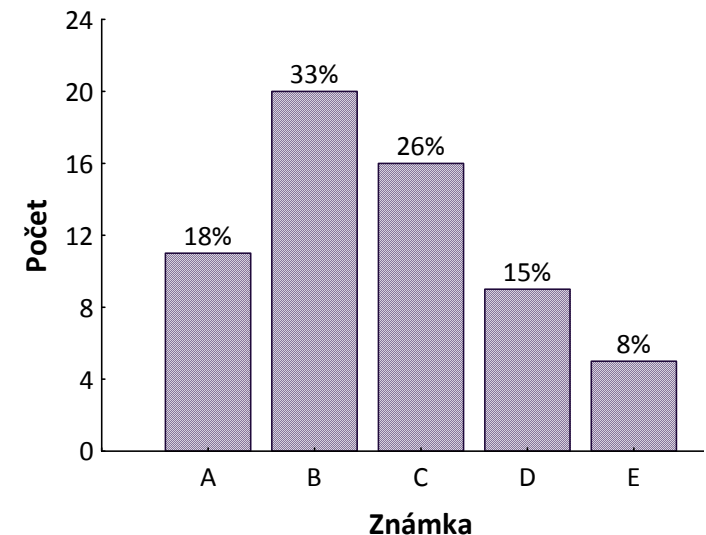
Frekvenční tabulka

Známká	n	%
A	11	18,0
B	20	32,8
C	16	26,2
D	9	14,8
E	5	8,2
F	0	0,0
Celkem	61	100,0

Koláčový graf



Sloupcový graf



Kvantitativní parametry – popisné statistiky

- **Charakteristiky polohy** (míry střední hodnoty, míry centrální tendence)

Udávají, kolem jaké hodnoty se data centrují, resp. které hodnoty jsou nejčastější; popis „těžiště“ – míry polohy.

Aritmetický průměr, medián, modus,

- **Charakteristiky variability** (proměnlivosti)

Zachycují rozptýlení hodnot v souboru (proměnlivost dat).

Rozptyl, směrodatná odchylka, kvantily, rozsah hodnot (min–max)

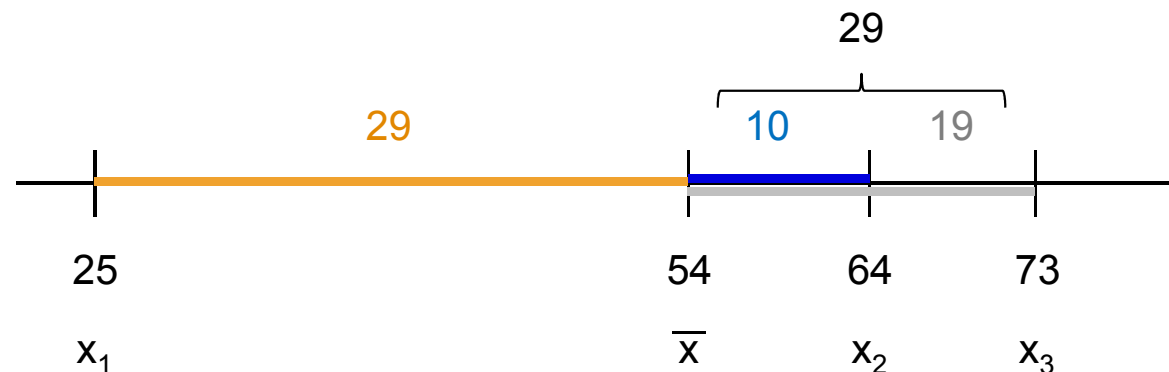
Charakteristiky polohy

- **Modus:** nejčastěji se vyskytující hodnota proměnné v souboru (u kvalitativních proměnných).
- **α -kvantil:** je-li $\alpha \in (0,1)$, pak α -kvantil x_α je číslo, které rozděluje uspořádaný datový soubor na dolní úsek, obsahující aspoň podíl α všech dat a na horní úsek obsahující aspoň podíl $1-\alpha$ všech dat.
- $x_{0,50}$ – **medián**, $x_{0,25}$ – **dolní kvartil**, $x_{0,75}$ – **horní kvartil**, $x_{0,1} \dots x_{0,9}$ – decily
- **Medián:** hodnota, jež dělí řadu podle velikosti seřazených hodnot na dvě stejně početné poloviny.

Charakteristiky polohy

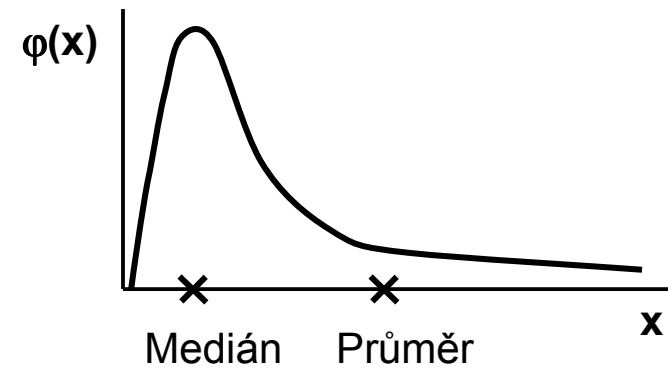
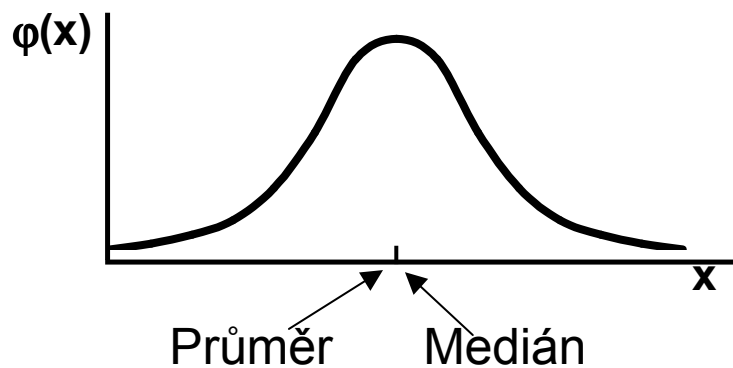
- **Aritmetický průměr:** je definován jako součet všech naměřených údajů vydělený jejich počtem,
kde x_i jsou jednotlivé hodnoty a n jejich počet
- **„Těžiště“ dat** – tzn. součet rozdílů podprůměrných hodnot od průměru je stejný jako součet rozdílů nadprůměrných hodnot od průměru

$$\bar{x} = (25+64+73) / 3 = 54$$



Průměr vs. medián

- **POZOR:** Průměr je silně ovlivněn extrémními hodnotami (tzv. odlehlá pozorování), medián jimi ovlivněn není.
- **Průměr** je vhodný ukazatel středu souboru u normálního, resp. symetrického rozložení, **medián** i v případě proměnných s neznámým rozdělením.
- V případě symetrického rozložení jsou průměr a medián v podstatě shodné, v případě asymetrického rozložení nikoliv!

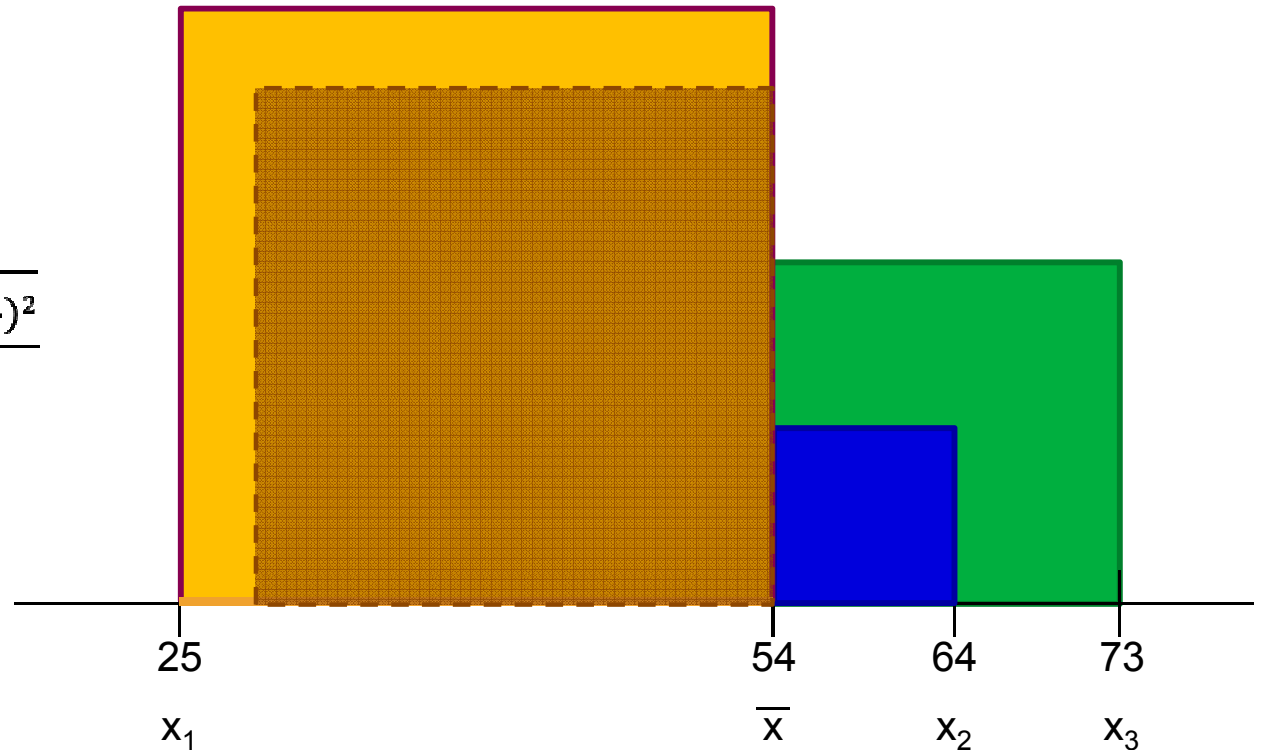


Charakteristiky variability

- **Kvartilové rozpětí:** $q = x_{0,75} - x_{0,25}$
- **Rozptyl (variance):** ukazatel šířky rozložení získaný na základě odchylky jednotlivých hodnot od průměru (jeho vypovídací schopnost je nejvyšší v případě symetrického/normálního rozložení).
- **Směrodatná odchylka (SD):** druhá odmocnina z rozptylu
- **Standardní chyba průměru (SE):** říká, jak přesný je výpočet průměru

Rozptyl, SD, SE

$$s = \sqrt{\frac{(25 - 54)^2 + (64 - 54)^2 + (73 - 54)^2}{2}}$$
$$= \sqrt{651} = 25,5$$

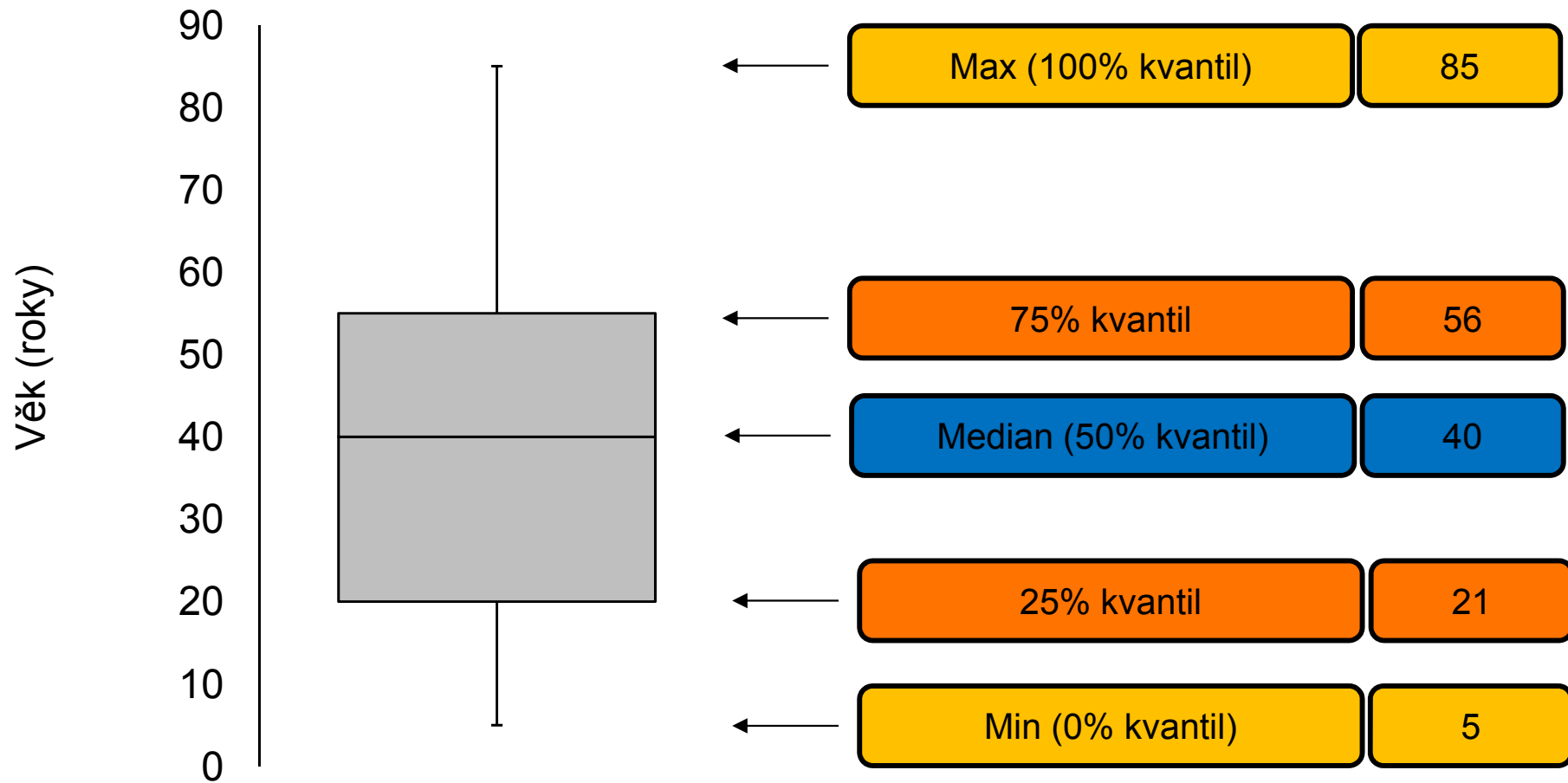


– SE – neplést s SD

- Říká, jak přesný je výpočet průměru:
- velký počet subjektů (n), z nichž počítáme průměr → tím menší je SE (tzn. tím přesnější je průměr)
- malý počet subjektů (n), z nichž počítáme průměr → tím větší je SE (tzn. tím méně přesný je průměr)

$$SE = \frac{SD}{\sqrt{n}}$$

Kvantily



Popis kvantitativních dat

- **Popis kvantitativních dat:** charakteristika středu (průměr, medián aj.), charakteristika variability (rozptyl, rozsah hodnot, interkvartilové rozpětí aj.)

Příklad: Popis výšky pacientů (cm)

Popisné statistiky

Charakteristika	
N	61
Průměr (cm)	161,5
Medián (cm)	161,0
Sm. odchylka (cm)	4,7
Rozptyl (cm ²)	22,2
Min–max (cm)	144–169
Dolní–horní kvartil (cm)	158–164



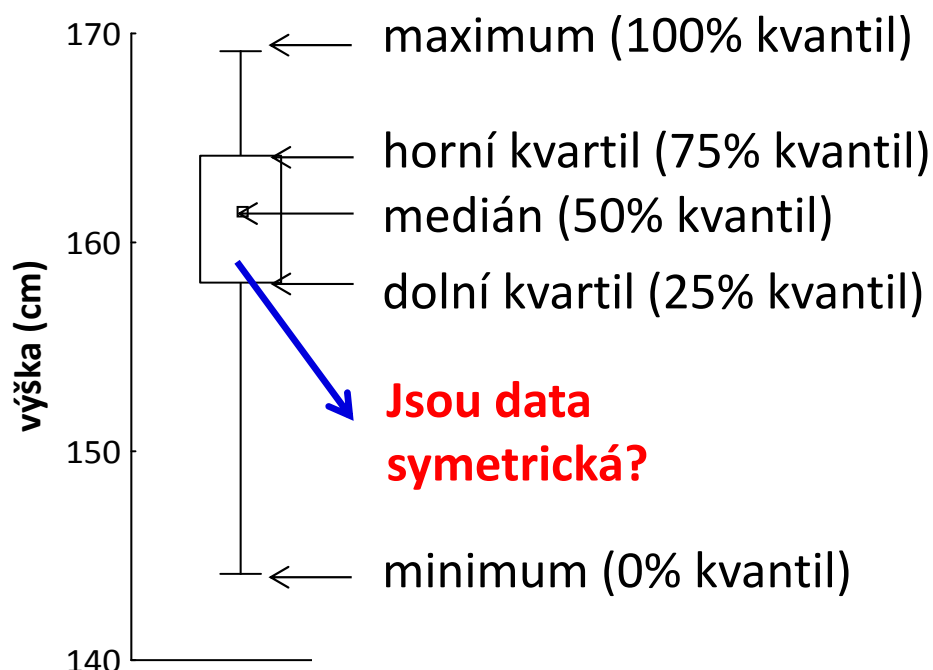
Průměr a medián se téměř shodují. Co nám to říká?

Vizualizace kvantitativních dat

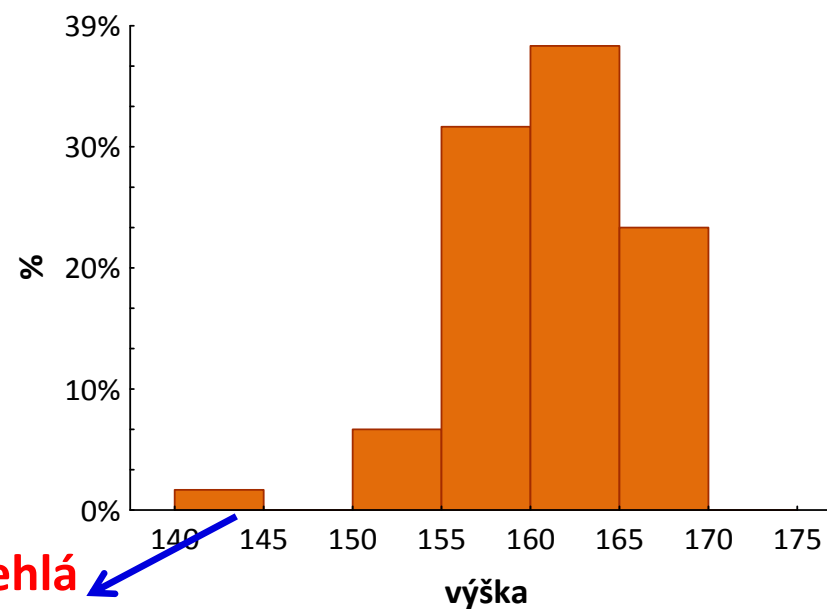
- **Vizualizace kvantitativních dat:** nejčastěji pomocí krabicového grafu nebo histogramu

Příklad: Popis výšky pacientů (cm)

Krabicový graf



Histogram



**M U N I
M E D**

Program Statistica

Představení programu Statistica
Praktické cvičení v programu Statistica

Program Statistica

Jak získat program Statistica:

<https://inet.muni.cz>

Login a heslo: UČO a primární heslo jako do IS-u.

V nabídce zvolit: **Provozní služby – Software – Nabídka softwaru**

Nalézt: **Statistica 14** – kliknout **Získat** a postupovat dle návodu

**M U N I
M E D**

Praktické cvičení v programu Statistica



Datový soubor

Rehabilitace po mozkovém infarktu

Data: 02_Biostatistika_Data02.sta* (24v by 407c)

	Rehabilitace po mozkovém infarktu: data									
	1	2	3	4	5	6	7	8	9	10
	ID	Pohlavi	Vek	Etiologie	Lokalizace	Terapie	Komorbid	Barthel_inc	Kategorie_zavislosti_p	Ukoncen
1	1	muž	82	okluze nek	mozkové tepny	jiná farmakolog	0	25	vysoce závislý	propuště
2	2	žena	81	embolie	mozkové tepny	jiná farmakolog	2	20	vysoce závislý	přeložen
3	3	muž	55	okluze nek	mozkové tepny	jiná farmakolog	0	35	vysoce závislý	propuště
4	4	žena	46	embolie	mozkové tepny	intravenózní trc	0	20	vysoce závislý	propuště
5	5	muž	76	okluze nek	mozkové tepny	jiná farmakolog	0	45	částečně soběstačný	propuště
6	6	muž	72	okluze nek	mozkové tepny	jiná farmakolog	0	25	vysoce závislý	přeložen
7	7	muž	62	trombóza	mozkové tepny	jiná farmakolog	0	40	vysoce závislý	propuště
8	8	muž	64	trombóza	přívodní tepny	jiná farmakolog	0	15	vysoce závislý	propuště
9	9	žena	82	okluze nek	mozkové tepny	jiná farmakolog	0	10	vysoce závislý	přeložen
10	10	muž	58	trombóza	mozkové tepny	jiná farmakolog	0	25	vysoce závislý	propuště
11	11	muž	84	okluze nek	mozkové tepny	jiná farmakolog	0	40	vysoce závislý	propuště
12	12	žena	92	okluze nek	mozkové tepny	jiná farmakolog	0	30	vysoce závislý	propuště
13	13	žena	79	embolie	mozkové tepny	jiná farmakolog	1	40	vysoce závislý	propuště
14	14	muž	69	trombóza	mozkové tepny	jiná farmakolog	3	45	částečně soběstačný	propuště

Rehabilitace po mozkovém infarktu

- Cvičný datový soubor obsahuje záznamy o **celkem 407 pacientech hospitalizovaných pro mozkový infarkt** na neurologickém oddělení akutní péče, kde jim byla poskytnuta terapie pro obnovu krevního oběhu v postižené části mozku.
- Po zvládnutí akutní fáze byl u pacientů vyhodnocen stupeň soběstačnosti v základních denních aktivitách (ADL) pomocí tzv. **indexu Barthelové (BI)** a byli přeloženi na **rehabilitační oddělení**.
- Po dvou týdnech byl opět dle BI vyhodnocen stupeň soběstačnosti a pacienti byli buď propuštěni do ambulantní péče, nebo přeloženi na oddělení následné péče.

Rehabilitace po mozkovém infarktu

Sbírané informace:

- základní demografické údaje (**pohlaví a věk**),
- informace o samotné diagnóze mozkové příhody (**etiologie a lokalizace uzávěru cévy**),
- informace o léčbě (typ indikované **terapie a výskyt komplikací**)
- informace o **způsobu ukončení rehabilitace**.
- Stupeň soběstačnosti před rehabilitací byl dodatečně zjištěn z neurologie a na konci rehabilitace byl vyplněn nový dotazník pro určení výsledného **indexu Barthelové**.

Úkol č. 1 – Popis kategoriálních dat

Zadání: „Proveďte základní popis zastoupení pohlaví u pacientů s mozkovým infarktem. Následně také srovnejte zastoupení pohlaví mezi třemi skupinami pacientů dle etiologie mozkové příhody.“

Postup:

1. Pro popis dat je vhodné určit absolutní a relativní četnosti.
2. Grafické znázornění je ideální pomocí koláčového grafu.
3. V programu Statistica lze získat výsledky pro jakoukoli podskupinu souboru pomocí obecné funkce „By Group“ nebo „Select Cases“.

Úkol č. 1 – Řešení v programu Statistica

- V menu **Statistics** zvolíme **Basic statistics** a vybereme **Frequency tables**.
- Vybereme proměnnou (**Variables**), kterou chceme analyzovat a na záložce **Options** zaškrtneme možnost **Percentages (relative frequencies)**.

The screenshot shows the Statistica 64 interface. A red arrow labeled '1' points to the 'Statistics' menu. A red arrow labeled '2' points to the 'Basic Statistics' icon in the ribbon. A red arrow labeled '3' points to the 'Frequency tables' option in the 'Quick' list of the 'Basic Statistics and Tables' dialog box. The background shows a data table with columns 'ID', 'Pohlavi', 'Vek', and 'Etica'.

The screenshot shows the 'Frequency Tables: 02_Biostatistika_Data02.sta' dialog box. A red arrow labeled '4' points to the 'Variables' field containing 'Pohlavi'. A red arrow labeled '5' points to the 'Options' tab. A red arrow labeled '6' points to the 'Percentages (relative frequencies)' checkbox, which is checked. A blue arrow labeled '7' points to the 'By Group...' button. The 'Display options for frequency tables' section includes: Cumulative frequencies, Percentages (relative frequencies), Cumulative percentages, 100% minus cumulative percentages, Logit transformed proportions, Probit transformed proportions, Normal expected frequencies. At the bottom, Count and report missing data (MD) is selected.

The screenshot shows the 'By Group' dialog box. A blue circle highlights the 'Grouping Variable(s)...' field, which contains 'Etica'. Another blue circle highlights the 'Accumulate tabular results in a single spreadsheet' checkbox, which is checked. The 'Sorting of Groups' section has 'Unsorted' selected.

Chceme-li získat výsledky zvlášť pro podskupiny jiné proměnné, aktivujeme funkci **By Group** (vybrat třídící proměnnou a zaškrtnout **Accumulate tabular results in a single spreadsheet**).

Úkol č. 1 – Výsledky v Statistica

Frekvenční tabulka proměnné pohlaví u pacientů s CMP

Category	Frequency table: Pohlaví	
	Count	Percent
muž	248	60,93366
žena	159	39,06634

Frekvenční tabulka proměnné pohlaví u pacientů s CMP dle etiologie centrální mozkové příhody

Category	Aggregate Results Frequency table: Pohlaví (02_Biostatistika_Dat		
	Etiologie	Count	Percent
muž	okluze nebo stenóza	125	62,18905
žena	okluze nebo stenóza	76	37,81095
žena	embolie	36	46,15385
muž	embolie	42	53,84615
muž	trombóza	81	63,28125
žena	trombóza	47	36,71875

Zastoupení mužů a žen v celém souboru je 61 % oproti 39 %. Při srovnání pacientů dle etiologie mozkového infarktu se nejvíce liší pacienti s embolií, u kterých je podíl mužů jen 54 % oproti 46 % žen.

Úkol č. 1 – Řešení v programu Statistica

- V menu **Graphs** zvolíme nabídku **2D grafů** a vybereme **Pie Charts**.
- Vybereme proměnnou (**Variables**), kterou chceme analyzovat a na záložce **Advanced** vybereme v nastavení legendy možnost **Text and Percent**.
- Chceme-li získat výsledky zvlášť pro podskupiny jiné proměnné, aktivujeme **By Group** (opět vybrat třídící proměnnou).

ID	Pohlavi	Vek	Etiologie	Lokali
1	muž	82	okluze nek	mozke
2	žena	81	embolie	mozke
3	muž	55	okluze nek	mozke
4	žena	46	embolie	mozke
5	muž	76	okluze nek	mozke
6	muž	72	okluze nek	mozke
7	muž	62	trombóza	mozke
8	muž	64	trombóza	privoc
9	žena	82	okluze nek	mozke
10	muž	58	trombóza	mozke
11	muž	84	okluze nek	mozke
12	žena	92	okluze nek	mozke
13	žena	79	embolie	mozke
14	muž	69	trombóza	mozke
15	muž	67	okluze nek	mozke
16	žena	67	okluze nek	mozke
17	žena	67	okluze nek	mozke
18	žena	63	okluze nek	privoc
19	muž	87	embolie	mozke
20	muž	84	trombóza	mozke

Other specific pie layout, location, proportions, and exploding options are available in the Graph Options dialog (accessible from the graph).

Grouping Variable(s)... Etiologie

Enabled

Output to single folder

Label Outputs

Output "All Groups" results

Accumulate tabular results in a single spreadsheet

Sorting of Groups

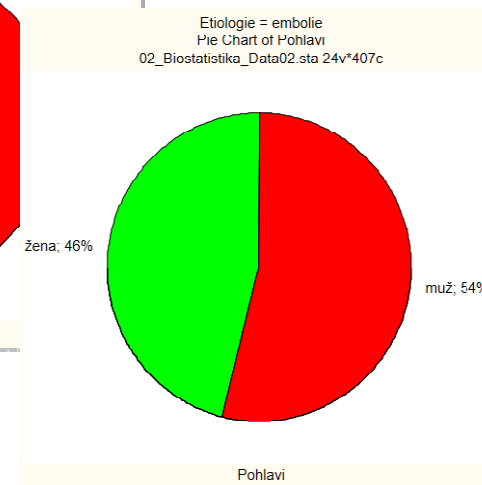
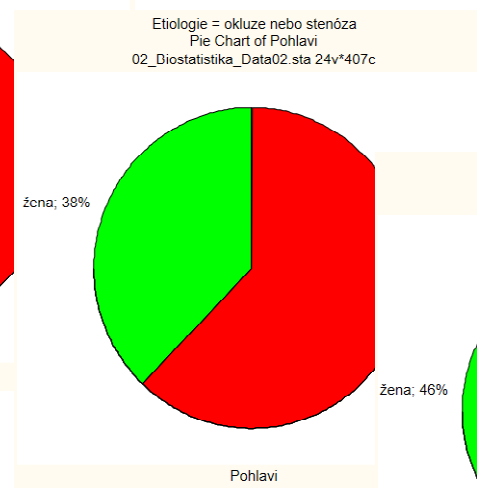
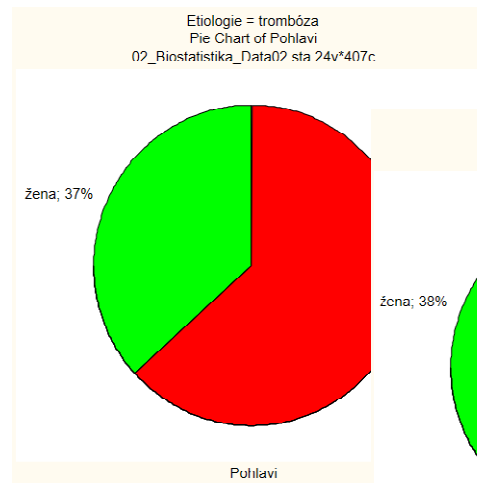
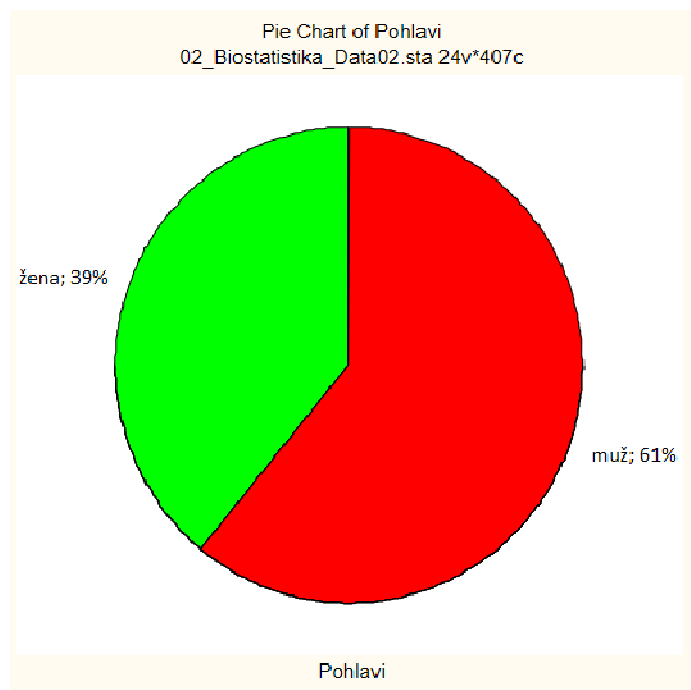
Unsorted

Ascending

Descending

Úkol č. 1 – Výsledky v Statistica

Koláčový graf proměnné pohlaví u pacientů s CMP



Zastoupení mužů v celém souboru je 61 % oproti 39 % žen.

Při srovnání pacientů dle etiologie mozkového infarktu se nejvíce liší pacienti s embolií, u kterých je podíl mužů jen 54 % oproti 46 % žen.

Úkol č. 2 – Popis kvantitativních dat

Zadání: „Provedte základní popis soběstačnosti dle indexu Barthelové na konci rehabilitace po mozkovém infarktu. Následně také tento ukazatel srovnejte podle míry komplikací během léčby.“

Postup:

1. Pro popis dat je vhodné určit průměr, medián, směrodatnou odchylku, případně minimum a maximum.
2. Grafické znázornění je ideální pomocí histogramu. V případě srovnávání různých skupin je vhodný krabicový graf.
3. V programu Statistica lze získat výsledky pro jakoukoli podskupinu pomocí „By Group“ nebo „Select Cases“.

Úkol č. 2 – Řešení v programu Statistica

- V menu **Statistics** zvolíme **Basic statistics** a vybereme **Descriptive statistics**.
- Vybereme proměnnou (**Variables**), kterou chceme analyzovat a na záložce **Advanced** zaškrtneme možnosti výpočtu **Mean, Median, Std. Dev, Min. & Max.**
- Chceme-li získat výsledky zvlášť pro podskupiny jiné proměnné, použijeme **By Group** (vybrat třídící proměnnou a zaškrtnout **Accumulate tabular results in a single spreadsheet**).

1. Click on the **Statistics** menu.

2. Click on **Basic Statistics**.

3. In the **Basic Statistics and Tables** dialog box, select **Descriptive statistics** in the Quick list.

ID	Pohlavi	Vek
1	1 muž	
2		
3		
4	4 žena	
5	5 muž	
6	6 muž	
7	7 muž	
8	8 muž	
9	9 žena	
10	10 muž	
11	11 muž	
12	12 žena	
13	13 žena	
14	14 muž	
15	15 muž	
16	16 žena	
17	17 žena	

4. Click on the **Advanced** tab.

5. Check **Mean**, **Median**, **Standard Deviation**, and **Minimum & maximum**.

7. Click on the **By Group...** button.

Grouping Variable(s): Komorbidity_komplikace

Accumulate tabular results in a single spreadsheet

Úkol č. 2 – Výsledky v Statistica

Popisná statistika indexu Barthelové na konci rehabilitace u pacientů s CMP

All Groups Descriptive Statistics (02_Biostatistika_Data02.sta)						
Variable	Valid N	Mean	Median	Minimum	Maximum	Std.Dev.
Barthel_index_po_rehabilitaci	407	62,01474	70,00000	10,00000	95,00000	19,44095

Popisná statistika indexu Barthelové na konci rehabilitace u pacientů s CMP dle stupně komplikací

Aggregate Results Descriptive Statistics (02_Biostatistika_Data02.sta)							
Variable	Komorbidity_komplikace	Valid N	Mean	Median	Minimum	Maximum	Std.Dev.
Barthel_index_po_rehabilitaci	0	239	66,84100	70,00000	15,00000	95,00000	17,45221
Barthel_index_po_rehabilitaci	1	79	57,15190	60,00000	10,00000	90,00000	19,15998
Barthel_index_po_rehabilitaci	2	50	56,70000	60,00000	10,00000	90,00000	20,01556
Barthel_index_po_rehabilitaci	3	39	49,10256	50,00000	10,00000	85,00000	21,36376

Celková průměrná hodnota indexu Barthelové je 62 bodů s mediánem 70 bodů. V závislosti na skóre komorbidit a komplikací během léčby je zřetelný pokles výsledné soběstačnosti s průměrem 67 bodů u nekomplikovaných případů až k 49 bodům u pacientů se stupněm komplikací 3.

Úkol č. 2 – Řešení v programu Statistica

- V menu **Graphs** zvolíme rovnou **Histogram (a)** nebo **Box (b)**.
- U histogramu pouze vybereme proměnnou (**Variables**), kterou chceme analyzovat.
- U box-plotu vybereme proměnnou (**Variables**), kterou chceme analyzovat (**dependent**), a proměnnou obsahující skupiny, které srovnáváme (**grouping**).

	1	2	3	5	6	7	
	ID	Pohlavi	Vek	Etiologie	Lokalizace	Terapie	Komorbidity
	1	1 muž	82	okluze nebo embolie	mozkové tepny	jiná farmak	
	2	2 žena	81	embolie	mozkové tepny	jiná farmak	
	3	3 muž	55	okluze nebo embolie	mozkové tepny	jiná farmak	
	4	4 žena	46	embolie	mozkové tepny	intravenózní	
	5	5 muž	76	okluze nebo embolie	mozkové tepny	jiná farmak	

2D Histograms

Quick Advanced Appearance Options 1 Options 2

Variables: Barthel_index_po_rehabilitaci

Graph type: Regular

Intervals: Integer mode (checked), Auto (checked)

Fit type: Normal (checked)

2D Box Plots

Quick Advanced Appearance Categorized Options 1 Options 2

Graph Type: Box-Whiskers

Variables: 21

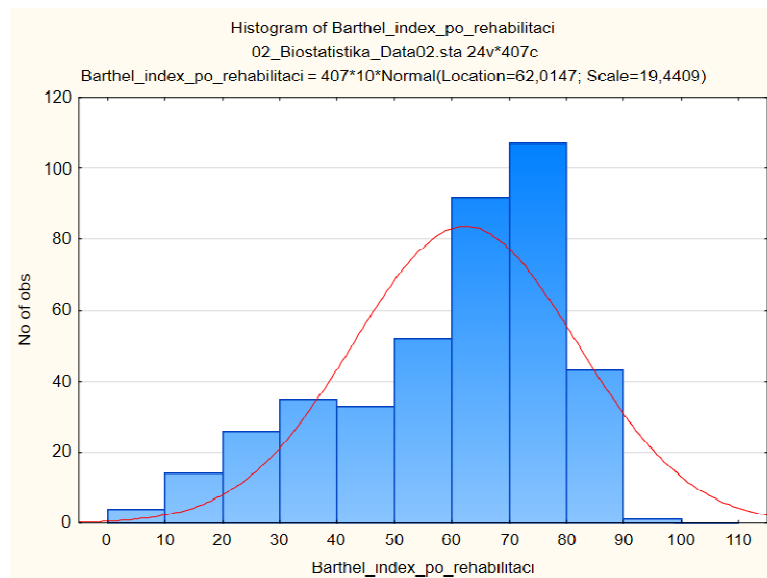
Grouping variable: Komorbidity_komplik

Grouping intervals: Unique values (checked), Auto (checked)

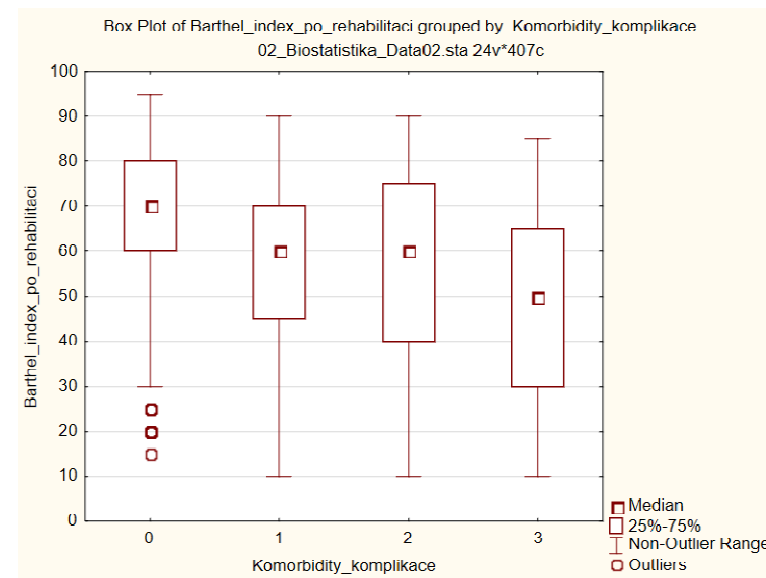
Middle point: Median

Úkol č. 2 – Výsledky v Statistica

Histogram indexu Barthelové na konci rehabilitace u pacientů s CMP



Krabicový graf indexu Barthelové na konci rehabilitace dle stupně komplikací



Celková průměrná hodnota indexu Barthelové je 62 bodů a tvar distribuce je asymetrický s hodnotami vyskytujícími se hlavně v rozmezí cca 50 až 90 bodů. V závislosti na skóre komorbidit a komplikací během léčby je zřetelný pokles výsledné soběstačnosti.