

# Přednáška III.

## Data, jejich popis a vizualizace

- ➡ Náhodný výběr, cílová a výběrová populace
- ➡ Typy dat
- ➡ Vizualizace různých typů dat
- ➡ Popisné statistiky



evropský  
sociální  
fond v ČR



EVROPSKÁ UNIE



MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání  
pro konkurenceschopnost



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

# 1. Jak vznikají data?

# Jak vznikají data?

➡ Záznamem skutečnosti...

... kterou chceme dále studovat → smysluplnost?

... více či méně dokonalým → kvalita?

# Jak vznikají data?

➡ Záznamem skutečnosti...

... kterou chceme dále studovat → smysluplnost?

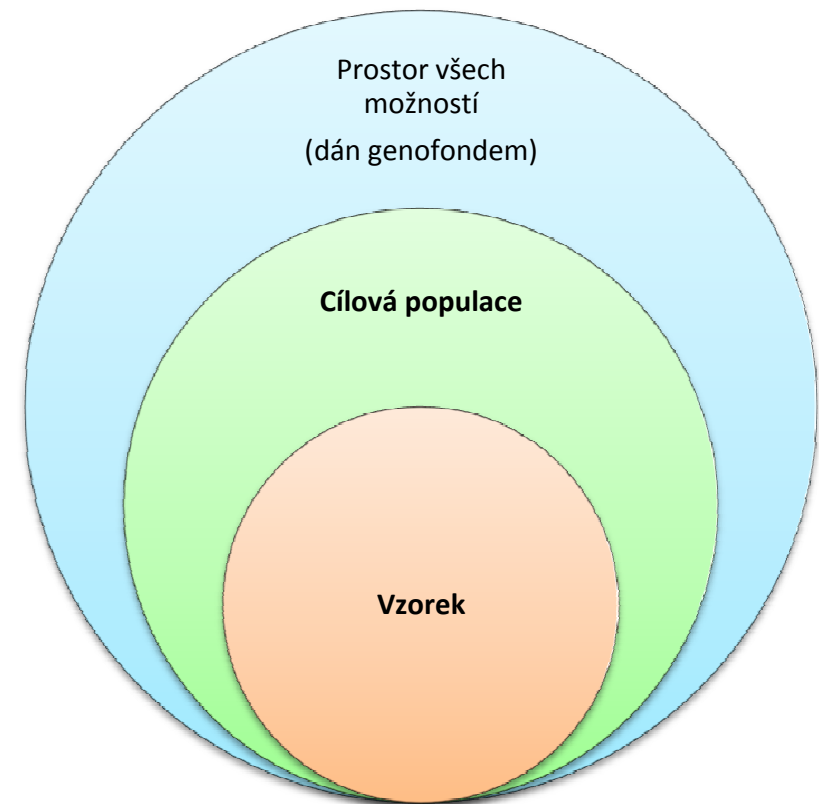
(krevní tlak, glykémie × počet srdcí, počet domů)

... více či méně dokonalým → kvalita?

(variabilita = informace + chyba)

# Cílová populace, výběrová populace

- ➔ **Cílová populace** – skupina subjektů, o které chceme zjistit nějakou informaci. Odpovídá základnímu prostoru  $\Omega$ .
- ➔ **Experimentální vzorek** neboli **výběrová populace** – podskupina cílové populace, kterou pozorujeme, měříme a analyzujeme. Jakékoliv výsledky chceme zobecnit na celou cílovou populaci. **Výběrová populace musí svými charakteristikami odpovídat cílové populaci (reprezentativnost)**. Toho můžeme docílit náhodným, ale i záměrným výběrem.



## 2. Typy dat a jejich vizualizace

# Typy dat

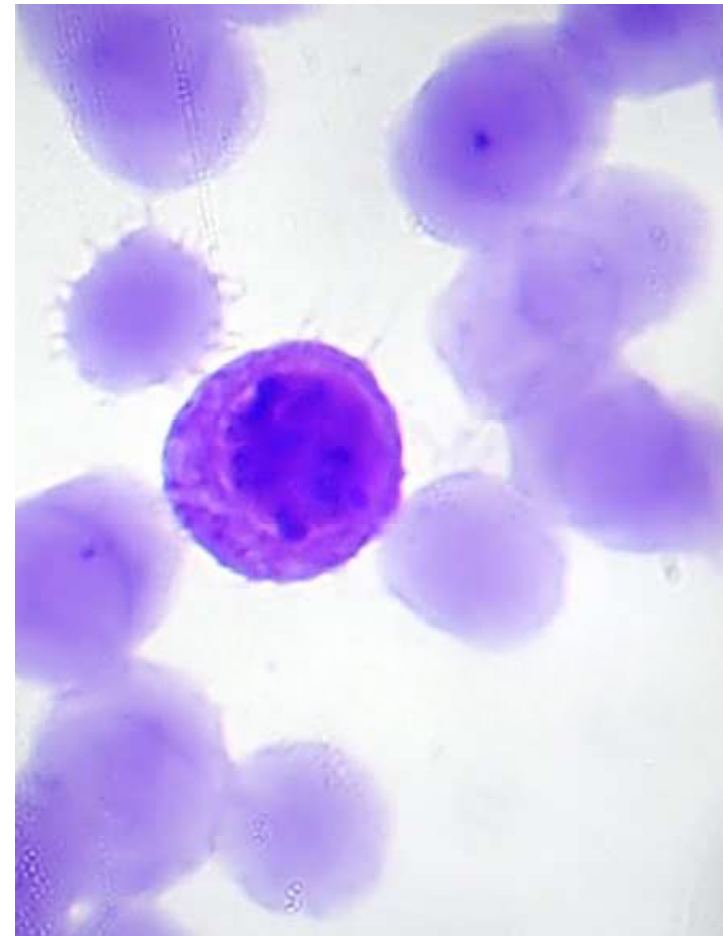
- ➔ **Kvalitativní** proměnná (kategoriální) – lze ji řadit do kategorií, ale nelze ji kvantifikovat, resp. nemá smysl přiřadit jednotlivým kategoriím číselné vyjádření.
  - ➔ Příklady: pohlaví, HIV status, užívání drog, barva vlasů
- ➔ **Kvantitativní** proměnná (numerická) – můžeme jí přiřadit číselnou hodnotu.  
Rozlišujeme dva typy kvantitativních proměnných:
  - ➔ **Spojité**: může nabývat jakýchkoliv hodnot v určitém rozmezí.  
Příklady: výška, váha, vzdálenost, čas, teplota.
  - ➔ **Diskrétní**: může nabývat pouze spočetně mnoha hodnot.  
Příklady: počet krevních buněk, počet hospitalizací, počet krvácivých epizod za rok, počet dětí v rodině.

# Typy dat – příklady

## Kvalitativní proměnná



## Kvantitativní proměnná





# Kvalitativní data lze dělit dále

- ➔ **Binární data** – pouze dvě kategorie typu ano / ne.
- ➔ **Nominální data** – více kategorií, které nelze vzájemně seřadit.  
Nemá smysl ptát se na relaci větší/menší.
- ➔ **Ordinální data** – více kategorií, které lze vzájemně seřadit.  
Má smysl ptát se na relaci větší/menší.

# Kvalitativní data – příklady

## ➔ Binární data

- ➔ diabetes (ano/ne)
- ➔ pohlaví (muž/žena)
- ➔ stav (ženatý/svobodný)

## ➔ Nominální data

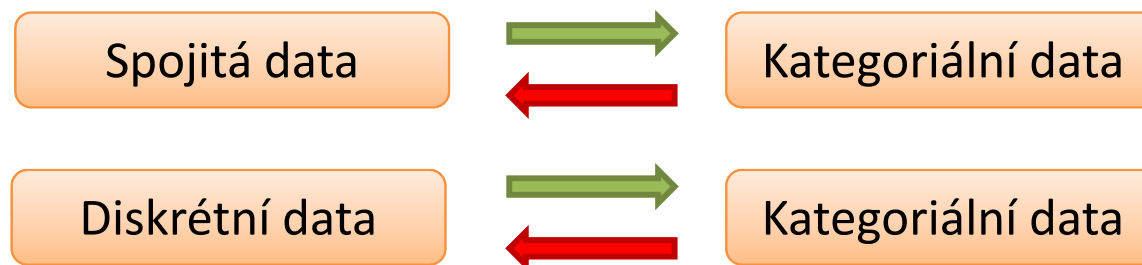
- ➔ krevní skupiny (A/B/AB/0)
- ➔ stát EU (Belgie/.../Česká republika/.../Velká Británie)
- ➔ stav (ženatý/svobodný/rozvedený/vdovec)

## ➔ Ordinální data

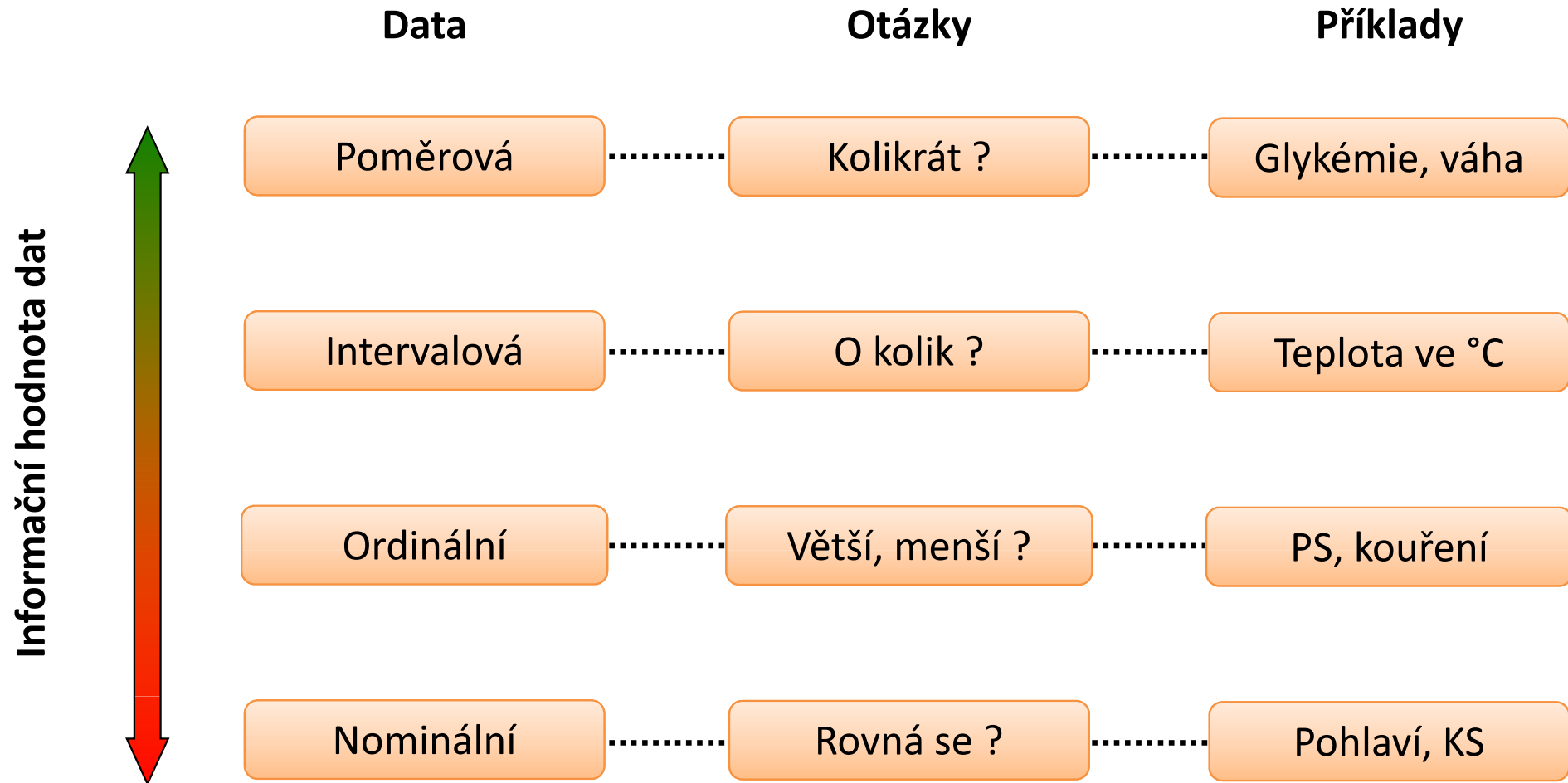
- ➔ stupeň bolesti (mírná/střední/velká/nesnesitelná)
- ➔ spotřeba cigaret (nekuřák/ex-kuřák/občasný kuřák/pravidelný kuřák)
- ➔ stadium maligního onemocnění (I/II/III/IV)

# Kvantitativní data

- ➔ Kvantitativní data poskytují větší **informaci** než data kvalitativní.
- ➔ Spojitá data poskytují větší informaci než data diskrétní.
- ➔ Větší informace znamená, že nám stačí méně pozorování na detekci určitého rozdílu (pokud ten rozdíl samozřejmě existuje).
- ➔ Kvůli interpretaci je někdy výhodné kvantitativní data **agregovat** do kategorií (např. věk) – **tímto krokem však ztrácíme část informace**. Zpětně nejsme schopni data rekonstruovat.



# Typy dat dle škály hodnot



# Další typy dat – odvozená data

- ➔ **Pořadí** (rank) – místo absolutních hodnot známe někdy pouze jejich pořadí. Jedná se sice o ztrátu určitého množství informace, nicméně i pořadí lze v biostatistice využít.
- ➔ **Procento** (percentage) – sledujeme-li např. zlepšení v určitém parametru, je výhodné sledovat procentuální zlepšení. Příklad: ejekční frakce levé srdeční komory.
- ➔ **Podíl** (ratio) – mnoho indexů je odvozeno jako podíl dvou měřených veličin. Příklad: BMI.
- ➔ **Míra pravděpodobnosti** (rate) – týká se výskytu různých onemocnění, kdy počet nových pacientů v daném čase (studii) je vztažen na celkový počet zaznamenaných osobo-roků. Příklad: výskyt nádorového onemocnění u pacientů ve studii.
- ➔ **Skóre** (score) – jedná se o uměle vytvořené hodnoty charakterizující určitý stav, který nelze jednoduše měřit jako číselné hodnoty. Příklad: indexy kvality života.
- ➔ **Vizuální škála** (visual scale) – pacienti často hodnotí svoje obtíže na škále, která má formu úsečky o délce např. 10 cm. Příklad: hodnocení kvality života.

# Další typy dat – odvozená data

-----  
-----  
7 1 1 1 2 3 2 2 SUI.MYŠLENKY

## 10. Suicidální myšlenky

Život nestojí za to žít, myšlenky o vitanosti přirozené smrti, myšlenky na sebevraždu, příprava sebevraždy. Fakticky provedené suicidální pokusy neberte při skórování v úvahu

-----  
0 - má zájem na životě a nebo jej bere tak jak je

1 - potěšení ze života je oproti obvyklému stavu zdraví poněkud sníženo

2 - otrávený životem, občasné úvahy o suicidiu

3 - připouští, že nebýt by bylo momentálně příjemnější než být, o suicidiu jako řešení situace však neuvažuje

4 - raději by nežil, úvahy o suicidiu časté, suicidium by bylo možným řešením situace, plány na suicidium však dosud nejsou konkrétní a promyšlené

5 - představa o způsobu suicidia je již konkrétní, konání však k tomu zatím nesměřovalo

6 - konkrétní plány na suicidium, kdyby byla možnost. Aktivní příprava suicidia

# Absolutní vs. relativní četnost

- ➔ Vyjádření výsledků v relativní formě (procento) má často příjemnou interpretaci, ale může být zavádějící.
- ➔ Relativní vyjádření účinnosti by mělo být vždy doprovázeno absolutním vyjádřením účinnosti.

➔ **Příklad:** Srovnání účinnosti léčiva ve smyslu prevence CMP u kardiaků.

Studie 1: Výskyt CMP ve skupině A je 12 %, ve skupině B je 20 %.

Relativní změna v účinnosti = **40 %**; absolutní změna = **8 %**.

Studie 2: výskyt CMP ve skupině A je 0,9 %, ve skupině B je 1,5 %.

Relativní změna v účinnosti = **40 %**; absolutní změna = **0,6 %**.

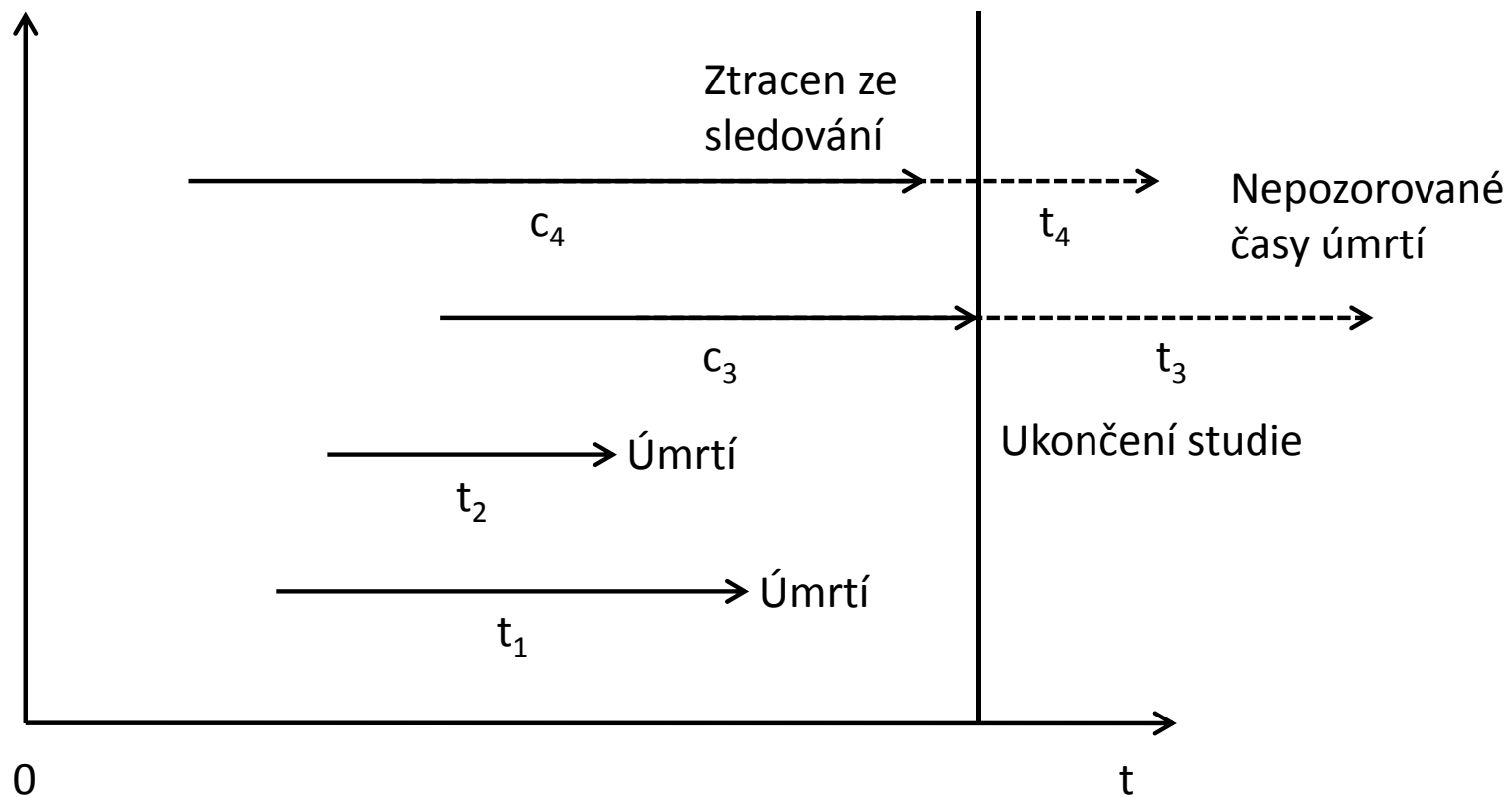
- ➔ Výsledkem je rozdílný přínos léčby při stejné relativní účinnosti.

# Další typy dat – cenzorovaná data

- ➔ **Cenzorovaná data** charakterizují experimenty, kde sledujeme čas do výskytu předem definované události.
- ➔ V průběhu sledování událost nemusí nastat u všech subjektů. Subjekty však nelze vinit z toho, že jsme u nich nebyli schopni danou událost pozorovat a už vůbec je nelze z hodnocení vyloučit.
- ➔ O čase sledování takového subjektu pak mluvíme jako o **cenzorovaném**.
- ➔ Toto označení indikuje, že sledování bylo ukončeno dříve, než u subjektu došlo k definované události. Nevíme tedy, kdy a jestli vůbec daná událost u subjektu nastala, víme pouze, že nenastala před ukončením sledování.



# Další typy dat – cenzorovaná data



# 3. Vizualizace a popis různých typů dat

# Reálná data

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
ID_uniq	INICIALY	Věk	LEKAR	SEX	NHL_STUP	DG_1	DATUM_DG	IPI	LDH	B2M	KS	RT_OD	RT_DO	STAV	ZEMREL
1	MZ	59	Pytlík	F	DLCL	DLCL	28.04.99	0	5.7	1.5	I			KR	
4	JS	64	Pytlík	F	DLCL	DLCL	03.11.99	1	13.3	NA	II			ZTR	
6	VK	66	Pytlík	F	difusní velkobuněčný B-lymfom	DLCL	19.01.00	2	11.1	2.5	III			EX	31.01.01
7	BK	41	Pytlík	F	difusní lymfom z velkých bb	DLCL	27.04.00	0	8.3	2.3	I	12.09.00	13.10.00	KR	
8	ZV	74	Pytlík	M	centroblastický B-lymfom	DLCL	13.11.00	3	12.6	2.6	III			KR	
11	DH	75	Pytlík	M	DLCL	DLCL	15.03.01	0	7.1	3.0	II	25.06.01	18.07.01	KR	
12	JS	60	Jankovská	M	DLCL	DLCL	19.04.01	0	5.6	0.2	I			KR	
13	PF	26	Pytlík	F	DLCL, bude 2. Čtení	DLCL	29.08.01	20	17.9	1.9	II			EX	07.09.02
14	JK	47	Jankovská	F	B-velkobuněčný	DLCL	17.10.01	0	8.6	2.1	III	xx.04.02		KR	
15	JJ	67	Jankovska	M	DLBCL	DLCL	07.02.02	0	8.4	5.6	I			KR	
16	HJ	73	Jankovská	F	DLCL	DLCL	15.02.02	0	6.5	1.4	II	27.05.02	14.06.02	KR	
17	VV	51	Jankovská	Ž	FCL/DLCL	DLCL	20.02.02	0	8.3	1.3	I			EX	18.05.02
22	FŘ	69	Jankovská	M	DLCL	DLCL	07.06.02	0	6.7	NA	I	22.08.03	20.09.03	PR	
23	OH	72	Jankovská	M	difusní velkobuněčný B lymfom	DLCL	25.10.02	1	8.2	2.5	III			KR	
24	JK	30	Jankovská	M	DLBCL	DLCL	31.01.03	1	13.8	1.8	II	plánovaná		KR	
25	EH	72	Jankovská	F	DLBCL	DLCL	06.08.03	2	9.2	1.7	III			KR	
26	MM	50	Jankovská	F	DLBCL	DLCL	05.09.03	1	7.3	1.7	III			KR	
32	MS	75	Kubáčková	F	DLCL	DLCL	03.03.99	1	8.8	1.5	I	20.07.99	16.08.99	KR	
33	RS	31	Kubáčková	M	DLCL	DLCL	17.08.00	1	8.8	2.0	I	27.02.01	26.03.01	KR	
34	JS	60	Kubáčková	M		DLCL	Motol	2	8	2.7	III			KR	
35	ZB	56	Kubáčková	M	DLCL	DLCL	19.02.01	1	9.8	2.4	II			KR	
36	JN	37	Kubáčková	M	DLCL	DLCL	13.03.01	1	16.1	2.0	I	24.10.01	21.11.01	KR	
37	AŠ	58	Kubáčková	F	difuzní B-lymfom, HG	DLCL	15.06.01	0	5.7	3.2	II	26.11.01	21.12.01	KR	
39	MH	56	Kubáčková	F		DLCL		1	11.4	2.0	I			EX	08.01.05
40	KŠ	83	Hrabětová	F	difusní velkobuněčný B lymfom	DLCL	01.07.02	2	32.0	6.0	I	28.01.03	10.02.03	EX	27.6.2003
41	LČ	53	Hrabětová	M		DLCL	Motol	0	5.2	1.9	I	21.1.2003	20.2.2003	KR	
48	MF	52	Kubáčková		DLBCL	DLCL	07.02.03	0	5.9	2.3	I			PR	
49	MČ	31	Kubáčková	F	DLBCL	DLCL		3	10.6	1.25	IV			KR	
50	VP		Papajík	M	DLBCL	DLCL	28.04.99	1	8.4	2.2	II			KR	15.11.02
51	AP		Papajík	M	DLBCL	DLCL	05.05.99	2	23,3	4.1	IV			EX	14.06.00

# Proč je popis a vizualizace dat třeba?

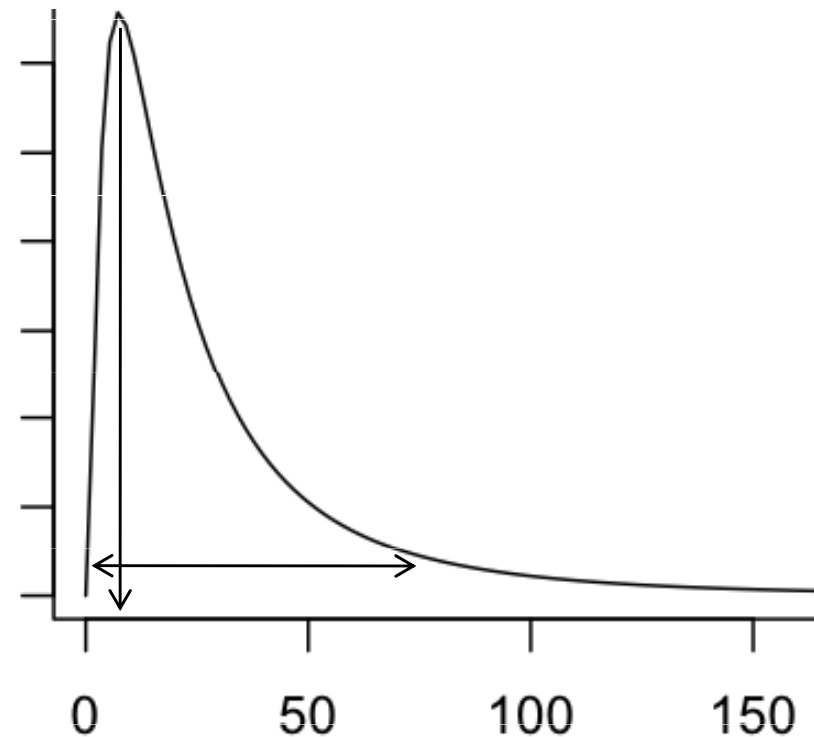
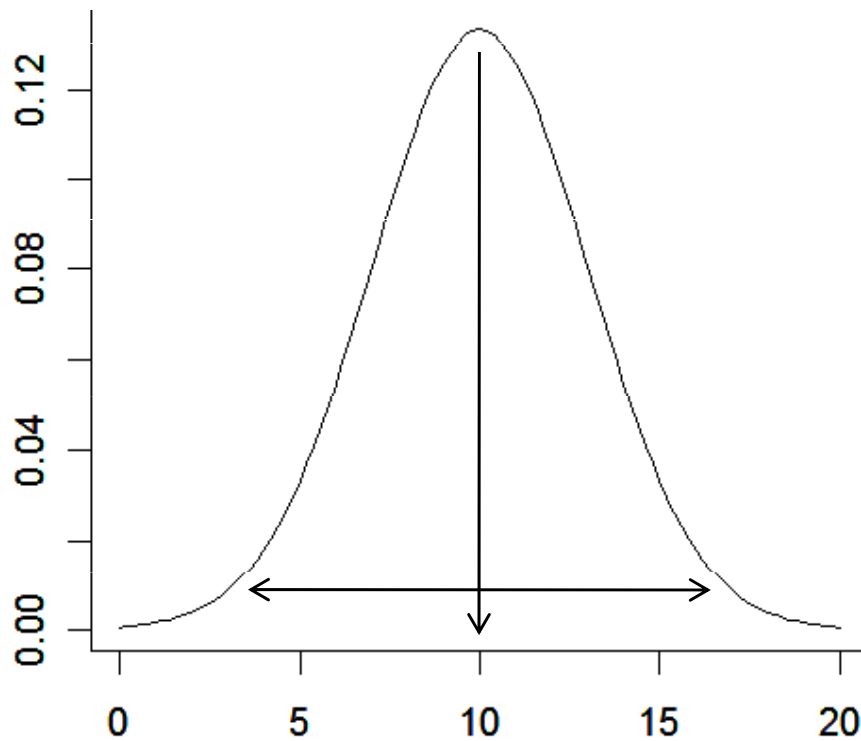
- ➡ Chceme **zpřehlednit** pozorovaná data – ve vhodných grafech.
- ➡ Chceme **zachytit** případné odlehlé a **extrémní** body nebo nečekané, **nelogické** hodnoty.
- ➡ Chceme **popsat** naměřené hodnoty.
- ➡ Chceme vypočítat vhodné sumární statistiky, které budou pozorovaná data dále **zastupovat** při prezentaci, srovnáních apod. Chceme pozorovanou informaci „uložit“ v zástupných statistikách, použití všech pozorovaných dat je nepraktické až nemožné.

# Jaké jsou výstupy popisné analýzy?

- ➔ Obecně neformální, jde o **shrnutí pozorovaného** a ne o formální testování.
- ➔ **Vztahují se pouze na pozorovaná data** (respektive na experimentální vzorek).
- ➔ Mohou sloužit jako **podklad pro stanovení hypotéz**.

# Co chceme u dat popsat?

- ➔ **Kvalitativní data** – četnosti (absolutní i relativní) jednotlivých kategorií.
- ➔ **Kvantitativní data** – těžiště a rozsah pozorovaných hodnot.



# Popis „těžiště“ – míry polohy

➔ Mějme pozorované hodnoty:  $x_1, x_2, \dots, x_n$

➔ Seřadíme je podle velikosti:  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

➔ **Minimum** a **maximum** – nejmenší a největší pozorovaná hodnota nám dávají obraz o tom, kde se na ose  $x$  pohybujeme.

$$x_{\min} = x_{(1)}$$

$$x_{\max} = x_{(n)}$$

➔ **Průměr** – charakterizuje hodnotu, kolem které kolísají ostatní pozorované hodnoty. Je to fyzikální obraz těžiště stejně hmotných bodů ose  $x$ .

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

➔ **Medián** – je to prostřední pozorovaná hodnota. Dělí pozorované hodnoty na dvě půlky, půlka hodnot je menší a půlka hodnot je větší než medián.

$$\tilde{x} = x_{((n+1)/2)} \quad \text{pro } n \text{ liché}$$

$$\tilde{x} = \frac{1}{2} (x_{(n/2)} + x_{(n/2+1)}) \quad \text{pro } n \text{ sudé}$$

# Výpočet mediánu

## ➔ Příklad 1: $N = 8$

$(n + 1) / 2$  pozice je „mezi“ 4. a 5. prvkem po seřazení – uděláme průměr

Data = 6 1 7 4 3 2 7 8

Seřazená data = 1 2 3 4 6 7 7 8

Medián =  $(4 + 6) / 2 = 5$

## ➔ Příklad 2: $N = 9$

$(n + 1) / 2$  pozice znamená 5. pozice po seřazení

Data = 3,0 4,2 1,1 2,5 2,2 3,8 5,6 2,7 1,7

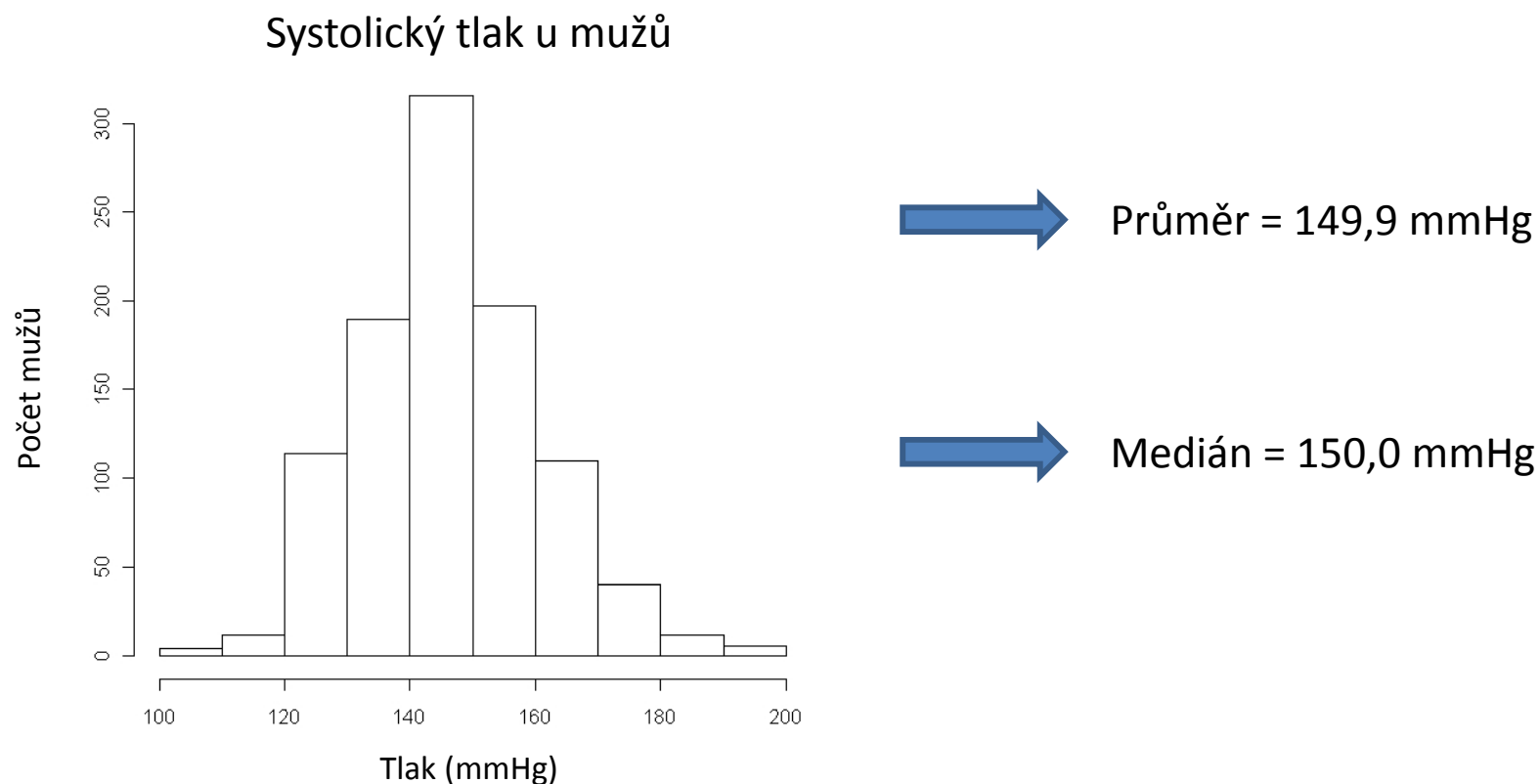
Seřazená data = 1,1 1,7 2,2 2,5 2,7 3,0 3,8 4,2 5,6

Medián = 2,7



# Průměr vs. medián

- ➔ Máme-li symetrická data, je výsledek výpočtu průměru i mediánu podobný.
- ➔ Vše je OK.



# Průměr vs. medián

- ➔ Nemáme-li symetrická data, je výsledek výpočtu průměru i mediánu rozdílný.
- ➔ Není to OK. Výpočet průměru je v tuto chvíli nevhodný!

## ➔ **Příklad 1:** známkování ve škole

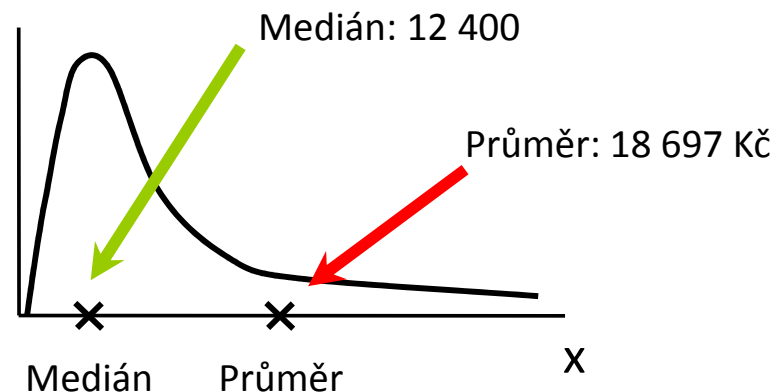
➔ Student A: 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 5

Průměr = 1,35                      Medián = 1,00

➔ Student B: 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2

Průměr = 1,13                      Medián = 1,00

## ➔ **Příklad 2:** plat v ČR v roce 2003



# Pojem kvantil

- ➔ Ve statistice je **kvantil** definován pomocí kvantilové funkce, což je inverzní funkce k distribuční funkci – budeme se jí věnovat příště.
- ➔ Laicky lze kvantil definovat jako číslo na reálné ose, které rozděluje pozorovaná data na dvě části:  $p\%$  kvantil rozděluje data na  $p\%$  hodnot a  $(100-p)\%$  hodnot.

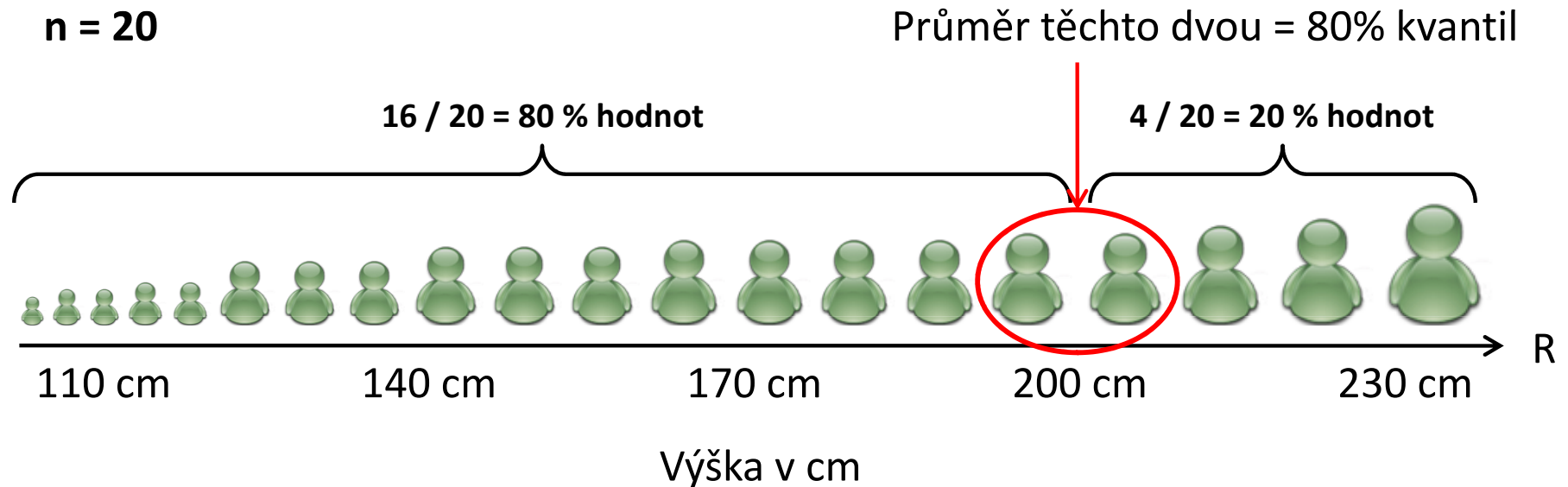
$$x_{p/100} = x_{(k)} \quad \text{pro } np/100 \text{ neceločíselné, pak } k = \lceil np/100 \rceil$$

$$x_{p/100} = \frac{1}{2} (x_{(k)} + x_{(k+1)}) \quad \text{pro } np/100 \text{ celočíselné, pak } k = np/100;$$



# Kvantil - příklad

- ➡ Máme soubor 20 osob, u nichž měříme výšku. Chceme zjistit 80% kvantil souboru pozorovaných dat.



# Významné kvantily

- ➔ Minimum = 0% kvantil
  - ➔ Dolní kvartil = 25% kvantil
  - ➔ **Medián = 50% kvantil**
  - ➔ Horní kvartil = 75% kvantil
  - ➔ Maximum = 100% kvantil
- ➔ **Medián** je významná charakteristika vypovídající o „těžišti“ pozorovaných hodnot. Není to ale jenom popisná charakteristika, na mediánu (a kvantilech obecně) je založeno mnoho **neparametrických statistických metod**.

# Popis „rozsahu“ – míry variability

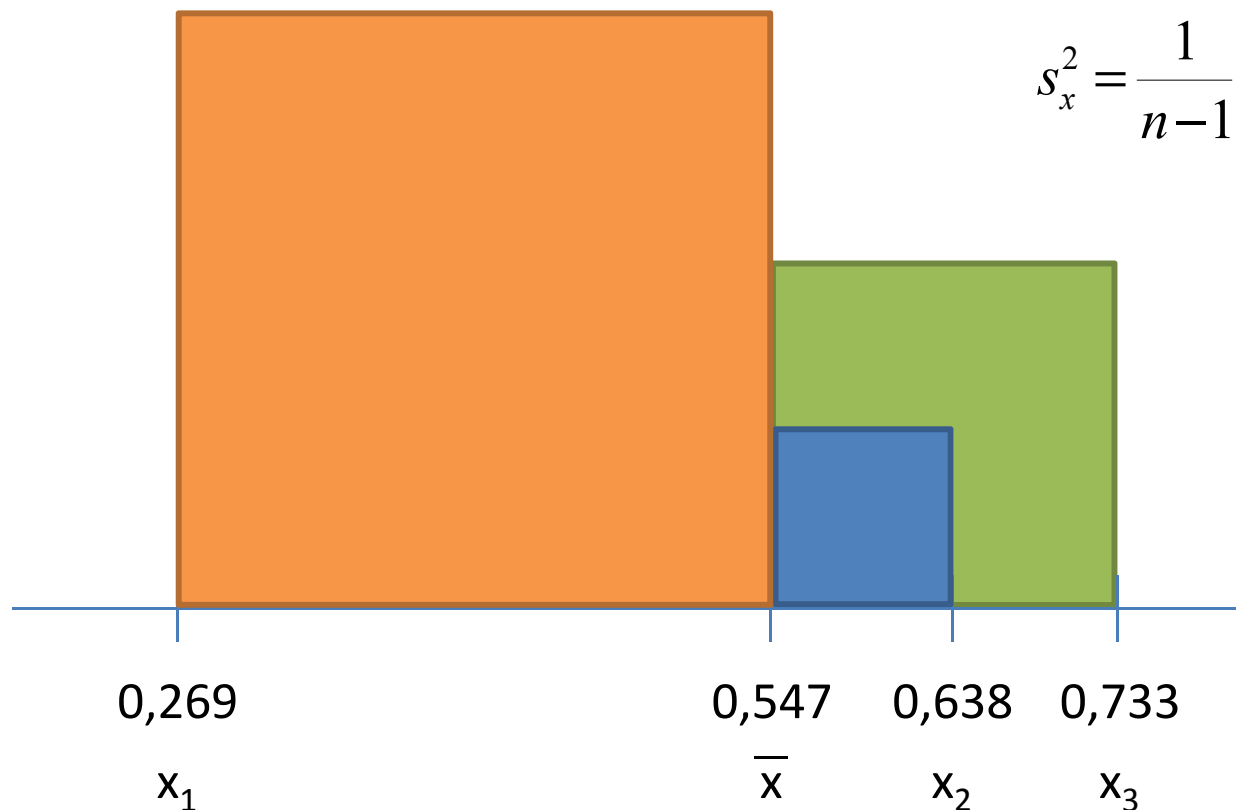
- ➔ Nejjednodušší charakteristikou variability pozorovaných dat je **rozsah hodnot** (rozpětí) = maximum – minimum. Je snadno ovlivnitelný netypickými (odlehlými) hodnotami.
- ➔ **Kvantilové rozpětí** je definováno  $p\%$  kvantilem a  $(100-p)\%$  kvantilem a je méně ovlivněno odlehlými hodnotami. Speciálním případem je **kvartilové rozpětí**, které pokrývá 50 % pozorovaných hodnot.
- ➔ **Výběrový rozptyl** – průměrný čtverec odchylky od průměru. Velmi ovlivnitelný odlehlými hodnotami.

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)$$

- ➔ **Výběrová směrodatná odchylka** – odmocnina z rozptylu. Výhodou směrodatné odchylky je, že má stejné jednotky jako pozorovaná data.

# Popis „rozsahu“ – míry variability

- ➔ Příklad čtverců odchylek od průměru pro  $n = 3$ .
- ➔ Rozptyl je možno značně ovlivnit odlehlými pozorováními.



$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

# 4. Kvalitativní data



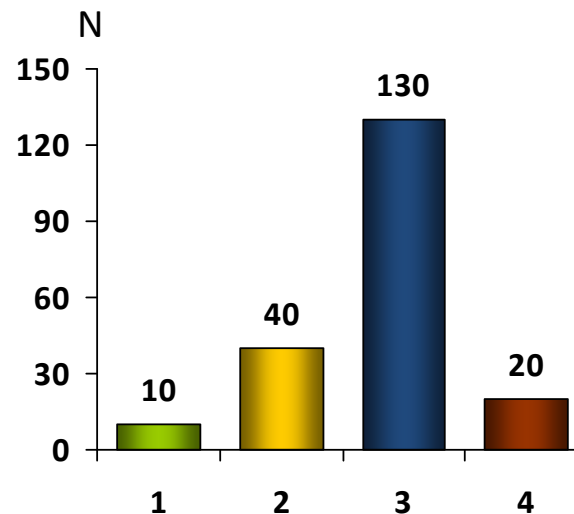
# Vizualizace a popis nominálních dat

- ➔ Vizualizace sloupcovým / koláčovým grafem – **absolutní i relativní četnost**.
- ➔ Sumarizace procentuálním výskytem kategorií v tzv. **frekvenční tabulce**.
- ➔ **Smysluplná agregace** kategorií zjednodušuje interpretaci i validitu výsledků.
- ➔ K popisu může sloužit i tzv. **modus** – nejčetnější pozorovaná hodnota.

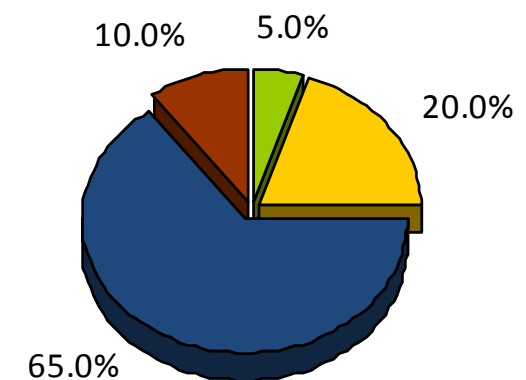
Frekvenční tabulka

Proměnná	n	%
Kategorie 1	10	5.0
Kategorie 2	40	20.0
Kategorie 3	130	65.0
Kategorie 4	20	10.0
<b>Celkem</b>	<b>200</b>	<b>100.0</b>

Sloupcový graf



Koláčový graf



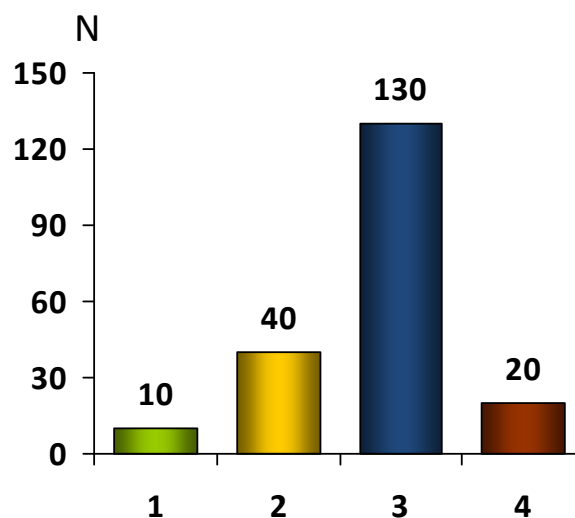
# Vizualizace a popis ordinálních dat

- ➔ Vizualizace sloupcovým / koláčovým grafem – **absolutní i relativní četnost**.
- ➔ Sumarizace procentuálním výskytem kategorií v tzv. **frekvenční tabulce**.
- ➔ **Smysluplná agregace** kategorií zjednodušuje interpretaci i validitu výsledků.
- ➔ K popisu může sloužit i tzv. **modus**, případně **medián** (pouze dává-li to smysl).

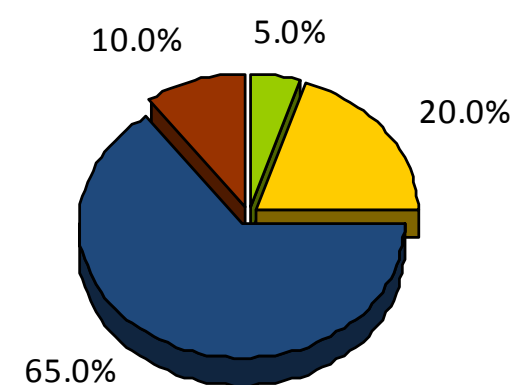
Frekvenční tabulka

Proměnná	n	%
Kategorie 1	10	5.0
Kategorie 2	40	20.0
Kategorie 3	130	65.0
Kategorie 4	20	10.0
<b>Celkem</b>	<b>200</b>	<b>100.0</b>

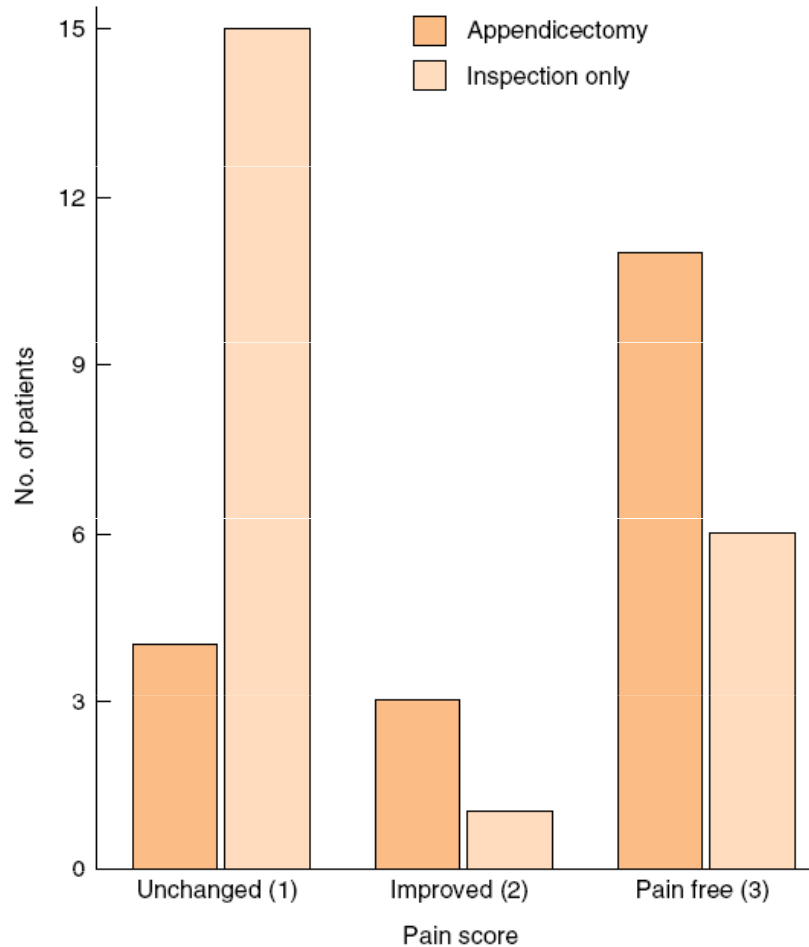
Sloupcový graf



Koláčový graf



# Co je na tom obrázku zavádějící?



A significantly higher proportion of patients in the appendicectomy group than in the inspection-only group had an improvement in pain (14 of 18 *versus* seven of 22;  $P = 0.005$ ). The relative risk was 2.4 (95 per cent c.i. 1.3 to 4.0), indicating that patients who had an appendicec-

Fig. 2 Distribution of pain scores in patients whose appendix was removed during laparoscopy and those in whom it was left *in situ*

# Co je na tom obrázku zavádějící?

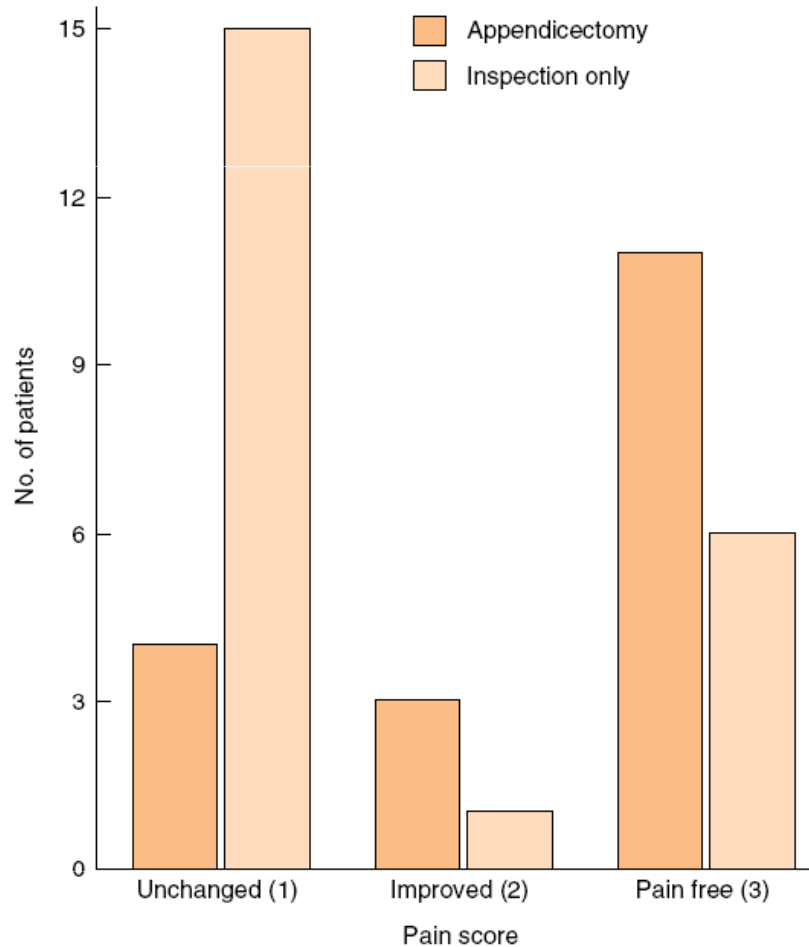


Fig. 2 Distribution of pain scores in patients whose appendix was removed during laparoscopy and those in whom it was left *in situ*

A significantly higher proportion of patients in the appendicectomy group than in the inspection-only group had an improvement in pain (14 of 18 *versus* seven of 22;  $P = 0.005$ ). The relative risk was 2.4 (95 per cent c.i. 1.3 to 4.0), indicating that patients who had an appendic-

➡ Ve chvíli, kdy obě skupiny mají různý počet pacientů, je srovnání absolutních čísel nekorektní.

# 5. Kvantitativní data

# Frekvenční tabulka pro kvantitativní data

Primární data

1,21  
1,48  
1,56  
0,31  
1,21  
1,33  
0,33  
0,21  
1,32  
1,11  
.  
.  
.  
.  
.  
 **$n = 100$**



Frekvenční tabulka

- ➡  $d_i$  – šířka intervalu
- ➡  $n_i$  – absolutní četnost v daném intervalu
- ➡  $n_i / n$  – relativní četnost v daném intervalu

$i$ -tý interval	$d_i$	$n_i$	$n_i / n$	%
$<0 - 0,4)$	0,4	20	0,2	20
$<0,4 - 0,8)$	0,4	10	0,1	10
$<0,8 - 1,2)$	0,4	40	0,4	40
$<1,2 - 1,4)$	0,2	20	0,2	20
$<1,4 - 1,6)$	0,2	10	0,1	10
Celkem	1,6	100	1	100

# Histogram

- ➔ Histogram je grafický nástroj pro vizualizaci **kvantitativních dat** (poměrových, intervalových, spojitých i diskrétních).
- ➔ Každá oblast histogramu odráží **absolutní nebo relativní četnost na jednotku** sledované proměnné na ose x.
- ➔ Histogram není sloupcový graf!

➔ Histogram pro relativní četnost:

$$f(i) = \frac{n_i / n}{d_i}$$

➔ Histogram pro absolutní četnost:

$$f(i) = \frac{n_i}{d_i}$$

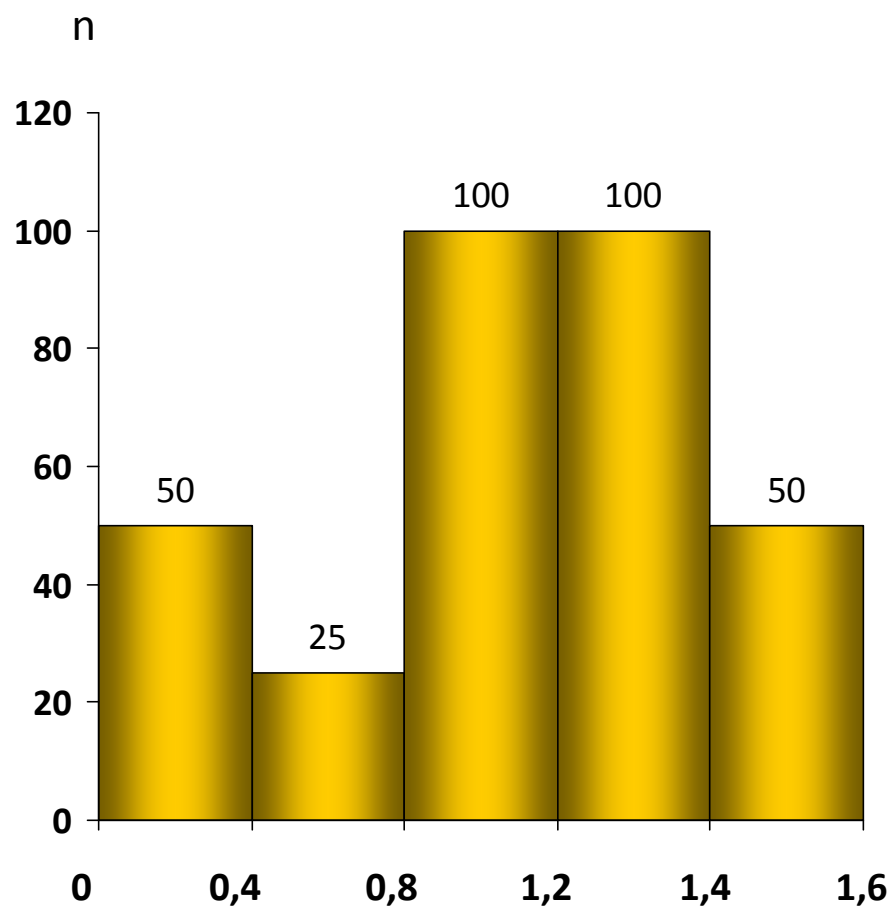
# Sumarizace kvantitativních dat histogramem

- ➔ Pozorovaná data: 1,21; 1,48; 1,56; 0,31; 1,21; 1,33; 0,33; 0,21; 1,32 ... .. n
- ➔ Setřídění dat podle velikosti
- ➔ Vytvoření intervalů na ose x
- ➔ Výpočet relativních nebo absolutních četností  $f(i)$
- ➔ Vykreslení histogramu

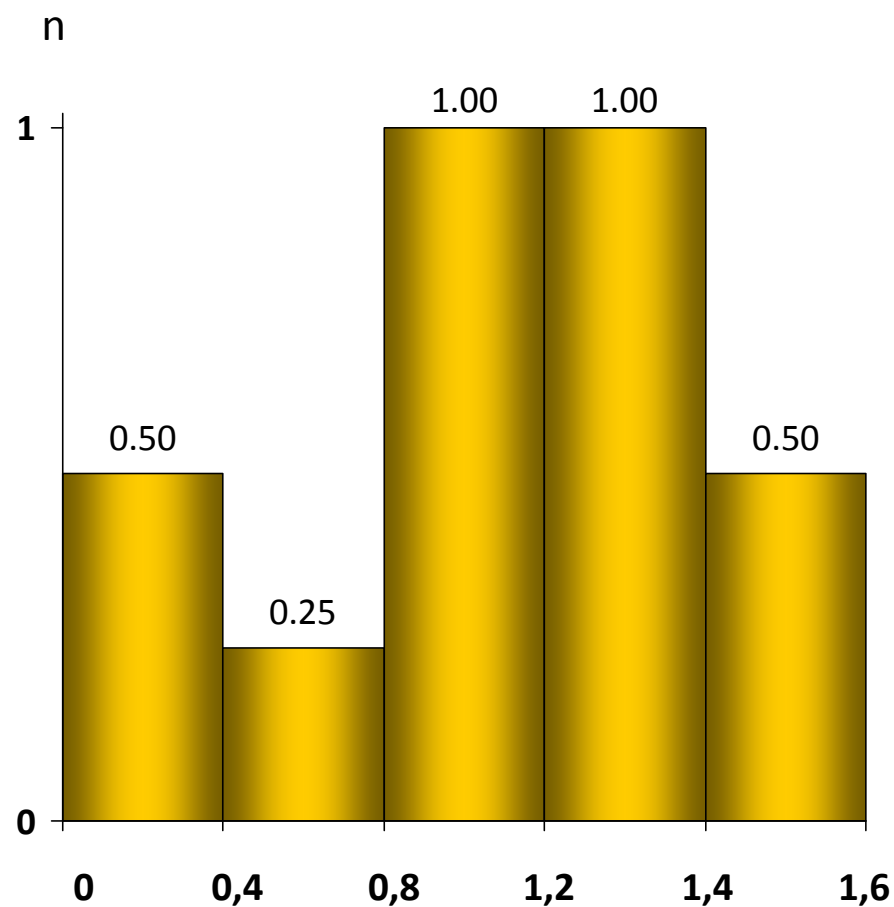


# Histogram – příklad

Histogram pro absolutní četnost



Histogram pro relativní četnost

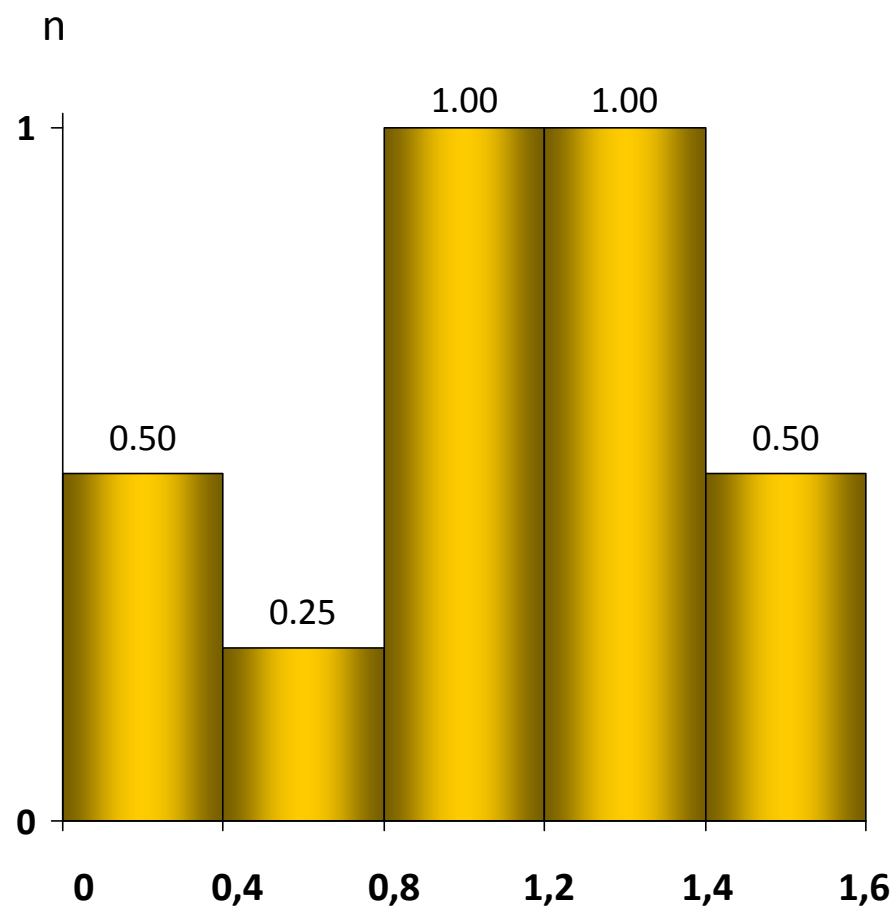


# Histogram – příklad

➡ Jaký obsah má plocha histogramu pro relativní četnost?

➡ A proč?

Histogram pro relativní četnost



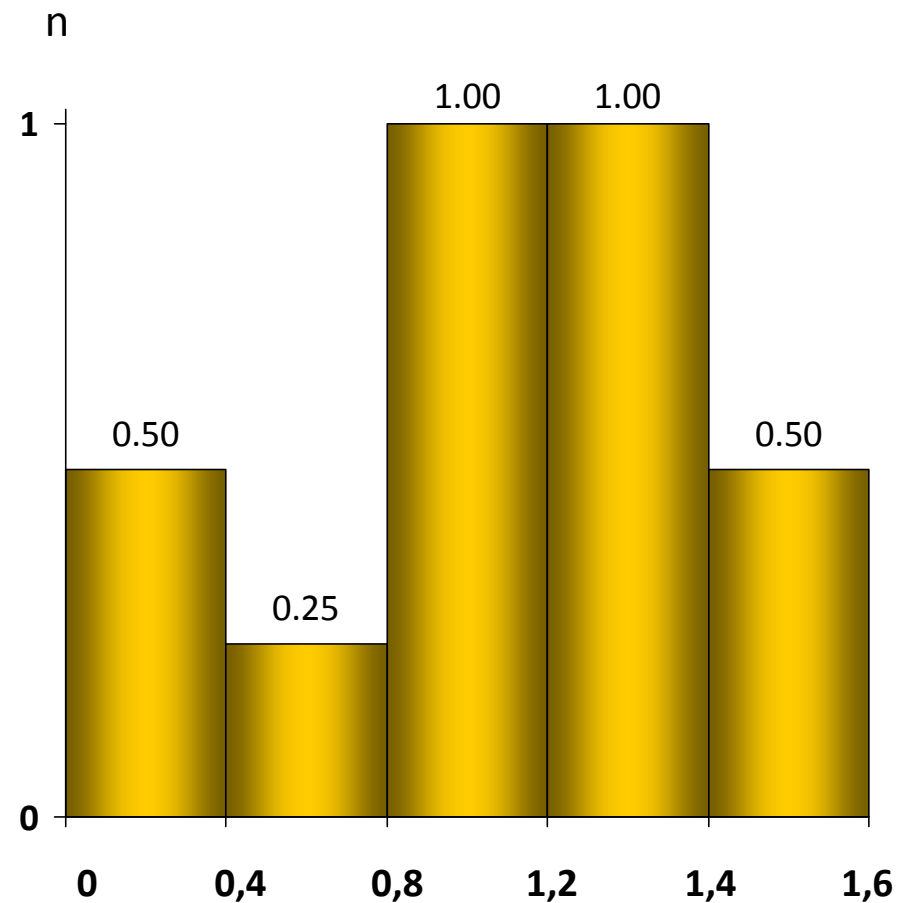
# Histogram – příklad

- ➔ Jaký obsah má plocha histogramu pro relativní četnost?

$$\sum_i f(i) = \sum_i \frac{n_i / n}{d_i} = 1$$

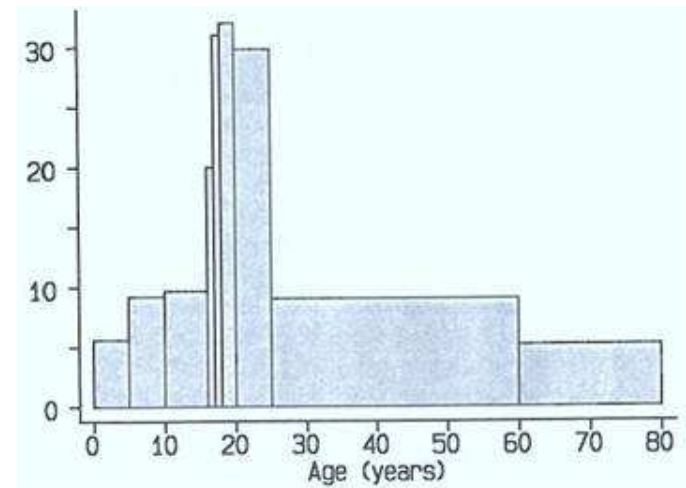
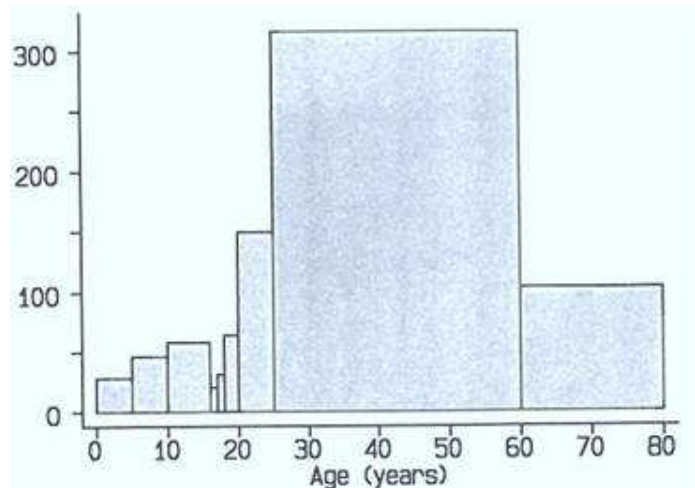
- ➔ A proč?
- ➔ Histogram lze použít pro odhad hustoty pravděpodobnosti. Je to tedy grafická vizualizace rozložení pravděpodobnosti kvantitativních (zejména spojitých) dat.

Histogram pro relativní četnost



# Který histogram je správný a proč?

- ➔ Chceme pomocí histogramu vykreslit počty zraněných při automobilových haváriích na předměstí Londýna v roce 1985. Data máme zadána jako počty v daných věkových kategoriích.



# Histogram ve skutečnosti

➡ Histogram je ve skutečnosti zřídka vyjadřován pomocí výrazů:

$$f(i) = \frac{n_i / n}{d_i} \quad f(i) = \frac{n_i}{d_i}$$

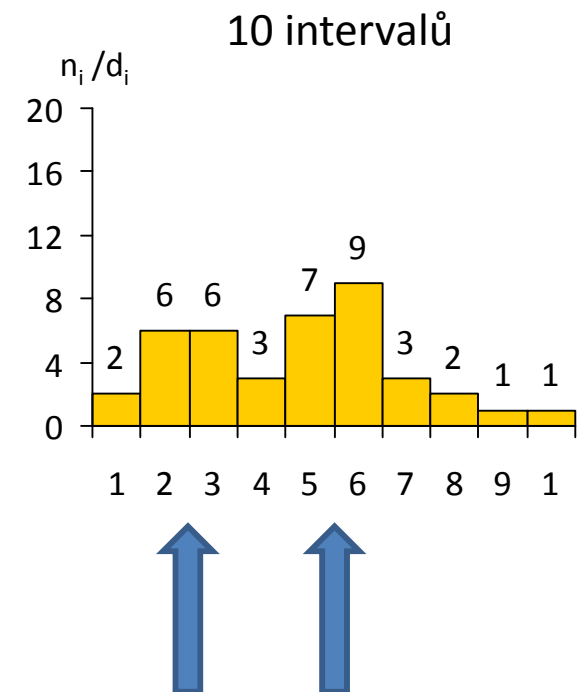
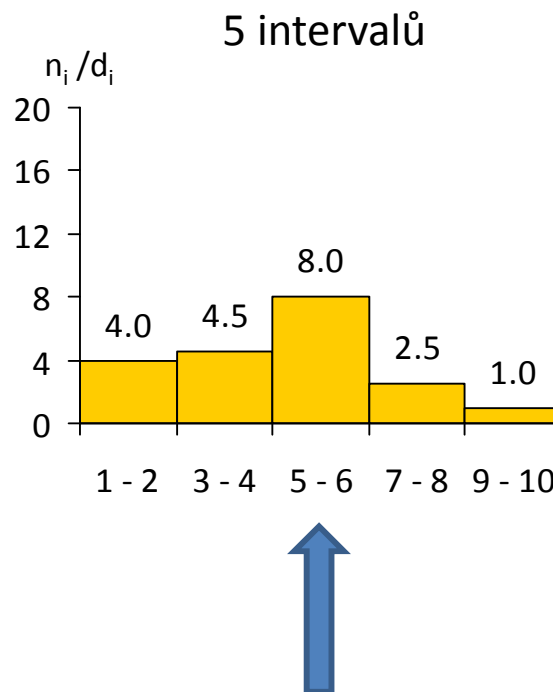
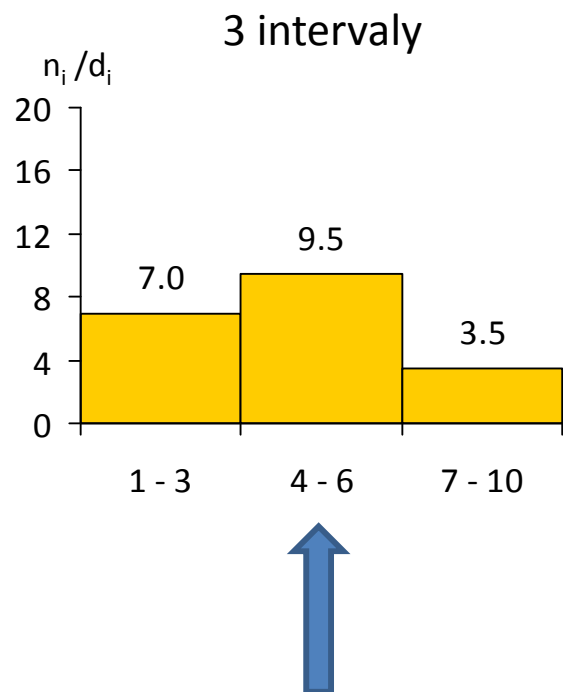
➡ Daleko častěji se jedná o prosté absolutní nebo relativní počty pozorování v daném intervalu (výhodné kvůli snadné čitelnosti a interpretaci):

$$f(i) = n_i / n \quad f(i) = n_i$$

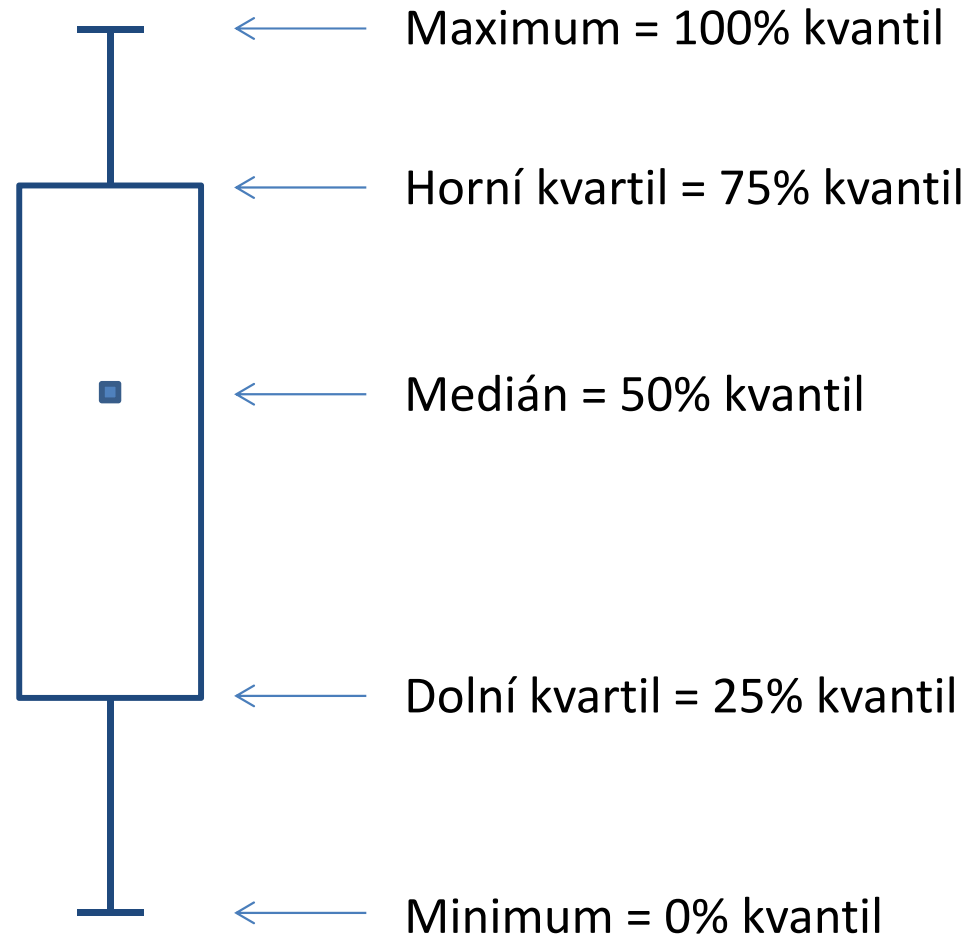
➡ **Důležité však je, aby intervaly měly stejnou šířku, aby výsledky byly srovnatelné!**

# Počet intervalů určuje kvalitu výstupu

➡ Počtem zvolených intervalů v histogramu rozhodujeme o tom, jak bude vypadat. Při malém počtu můžeme přehlédnout důležité prvky v datech, při velkém zase může být informace roztržštěná.



# Krabicový graf – box plot



# Co je extrémní (odlehlá) hodnota?

- ➡ Jednoduše řečeno se jedná o netypické pozorování, které nezapadá do pravděpodobnostního chování souboru dat.
- ➡ Definujeme ji jako hodnotu, která leží několikanásobek (3, 5, 7) směrodatné odchylky , respektive kvartilového rozpětí, od průměru, respektive mediánu.
- ➡ Definice je ale vágní, závisí na naší znalosti dané problematiky, které hodnoty jsou či nejsou možné!



# Vliv odlehlé hodnoty na popisné statistiky

Cílem je určit průměrnou hladinu cholesterolu vybrané populace mužů  
(hodnoty v mmol/l)

Správná data

6.3
7.6
6.3
9.1
4.2
5.8
5.65
6.3
8.6
6
6.2
6.7
4.6
6.25
6.3
4.04
6.3
9.1
6.3
5.2
6.4
5.75

Průměrná hodnota  
6,32

Směrodatná odchylka  
1,34

Která charakteristika se zvýší výrazněji?  
Průměr nebo směrodatná odchylka?

Průměrná hodnota  
?

Směrodatná odchylka  
?

Nesprávná data

6.3
7.6
6.3
9.1
4.2
5.8
5.65
6.3
8.6
6
6.2
6.7
4.6
6.25
6.3
4.04
6.3
9.1
6.3
5.2
64
5.75

# Vliv odlehlé hodnoty na popisné statistiky

Cílem je určit průměrnou hladinu cholesterolu vybrané populace mužů  
(hodnoty v mmol/l)

Správná data

6.3
7.6
6.3
9.1
4.2
5.8
5.65
6.3
8.6
6
6.2
6.7
4.6
6.25
6.3
4.04
6.3
9.1
6.3
5.2
6.4
5.75

Průměrná hodnota  
6,32

**Směrodatná odchylka**  
**1,34**

Průměrná hodnota  
8,94

**Směrodatná odchylka**  
**12,37**

Nesprávná data

6.3
7.6
6.3
9.1
4.2
5.8
5.65
6.3
8.6
6
6.2
6.7
4.6
6.25
6.3
4.04
6.3
9.1
6.3
5.2
64
5.75

# Identifikace odlehlých hodnot

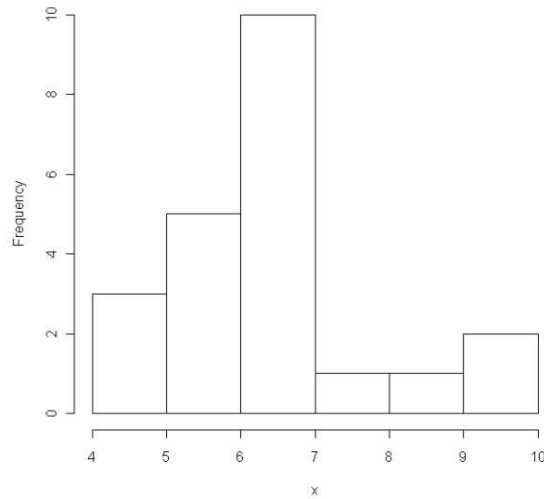
- ➔ Na menších souborech stačí vizualizace.
- ➔ Na větších datových souborech nelze bez vizualizace a popisných statistik.
- ➔ Grafická identifikace: pomocí histogramu a box plotu.
- ➔ Identifikace pomocí popisných statistik: srovnání mediánu a průměru.

# Identifikace odlehlých hodnot – příklad

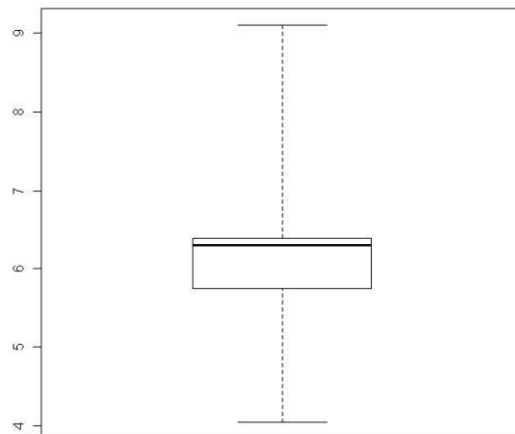
Správná data

- 6.3
- 7.6
- 6.3
- 9.1
- 4.2
- 5.8
- 5.65
- 6.3
- 8.6
- 6
- 6.2
- 6.7
- 4.6
- 6.25
- 6.3
- 4.04
- 6.3
- 9.1
- 6.3
- 5.2
- 6.4
- 5.75

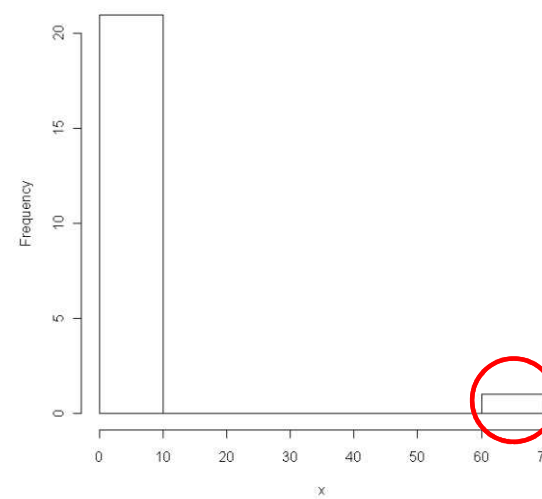
Histogram



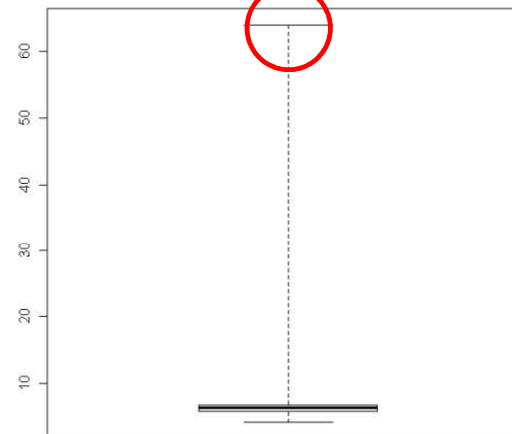
Box plot



Histogram



Box plot



Nesprávná data

- 6.3
- 7.6
- 6.3
- 9.1
- 4.2
- 5.8
- 5.65
- 6.3
- 8.6
- 6
- 6.2
- 6.7
- 4.6
- 6.25
- 6.3
- 4.04
- 6.3
- 9.1
- 6.3
- 5.2
- 64
- 5.75

# Identifikace odlehlých hodnot – příklad

Cílem je určit průměrnou hladinu cholesterolu vybrané populace mužů  
(hodnoty v mmol/l)

Správná data

6.3
7.6
6.3
9.1
4.2
5.8
5.65
6.3
8.6
6
6.2
6.7
4.6
6.25
6.3
4.04
6.3
9.1
6.3
5.2
6.4
5.75

Medián  
**6,30**

Průměrná hodnota  
**6,32**

Směrodatná odchylka  
1,34

Medián  
**6,30**

Průměrná hodnota  
**8,94**

Směrodatná odchylka  
12,37

6.3
7.6
6.3
9.1
4.2
5.8
5.65
6.3
8.6
6
6.2
6.7
4.6
6.25
6.3
4.04
6.3
9.1
6.3
5.2
<b>64</b>
5.75

Nesprávná data

# Reklama na příští týden...

Středem zájmu statistiky a biostatistiky je tzv. náhodná veličina.

