ASTAc/03 Biostatistika
2. cvičení

Kontingenční tabulky v Excelu
Základní popisné statistiky
Představení programu Statistica
Import a základní popis dat ve Statistice

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, L. Dušek, M. Cvanová

I. Kontingenční tabulky v Excelu

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, L. Dušek, M. Cvanová

Kontingenční tabulka

- Frekvenční sumarizace dvou kategoriálních proměnných (binárních, nominálních nebo ordinálních proměnných).
- Obecně: **R x C kontingenční tabulka** (R – počet kategorií jedné proměnné, C – počet kategorií druhé proměnné).
- Speciální případ: 2 x 2 tabulka = čtyřpolní tabulka.
- Kontingenční tabulky: **absolutních četností, celkových procent, řádkových/sloupcových četností**
- Př.: Sumarizace vyšetřených osob podle pohlaví a výsledku diagnostického testu.

Pohlaví	**Výsledek vyšetření**		
	Nemocný	Zdravý	**Celkem**
Muž	45	11	56
Žena	25	6	31
Celkem	70	17	87

Ukázka kontingenční tabulky

- **Vztah pohlaví a výskytu onemocnění** (pozor na hodnocení nesmyslného vztahu)

	Nemocný	Zdravý	Celkem
Muž	**a**	**b**	**a + b**
Žena	**c**	**d**	**c + d**
Celkem	**a + c**	**b + d**	**a + b + c + d = N**

Marginální absolutní četnost

Celkový počet hodnot

Simultánní absolutní četnost

	Nemocný	Zdravý	Celkem
Muž	45	11	56
Žena	25	6	31
Celkem	70	17	87

Jsou více nemocní muži nebo ženy?

Ukázka kontingenční tabulky

Kontingenční tabulka absolutních četností

	Nemocný	Zdravý	Celkem
Muž	45	11	56
Žena	25	6	31
Celkem	70	17	87

Větší počet nemocných mužů, který je dán pouze vyšším zastoupení mužů v celkovém vzorku (56 z 87)

Kontingenční tabulka řádkových procent

	Nemocný	Zdravý	Celkem
Muž	80,4 %	19,6 %	100,0 %
Žena	80,6 %	19,4 %	100,0 %

Po výpočtu relativních četností vidíme, že se muži a ženy neliší ve výskytu onemocnění

Jsou více nemocní muži nebo ženy?

Kontingenční tabulky v Excelu: zdroj dat a příprava dat

Kontingenční tabulka se dá vytvořit:

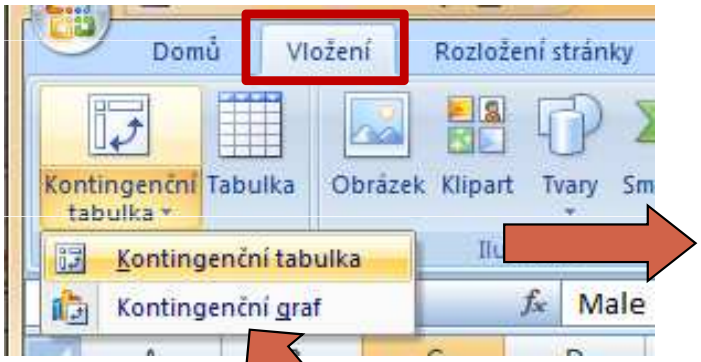
1. z tabulky v daném sešitě
2. z dat z jiného sešitu Excelu
3. z externích dat (např. MS Access)
4. ze sloučených dat z více oblastí - z různých listů nebo různých sešitů
5. z jiné kontingenční tabulky

Data musí být uspořádána formou standardního databázového seznamu:

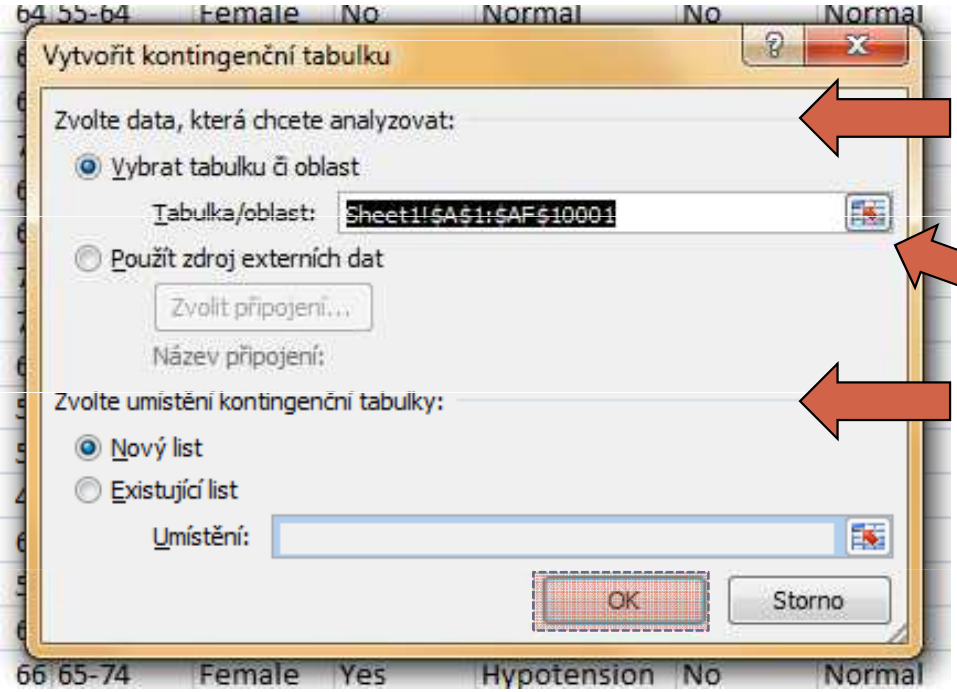
- V prvním řádku: názvy polí
- Další řádky: data

Vzhled tabulky: karta **Domů** → **Formátovat jako tabulku**

Vytvoření kontingenční tabulky v Excelu



Graf nebo tabulka

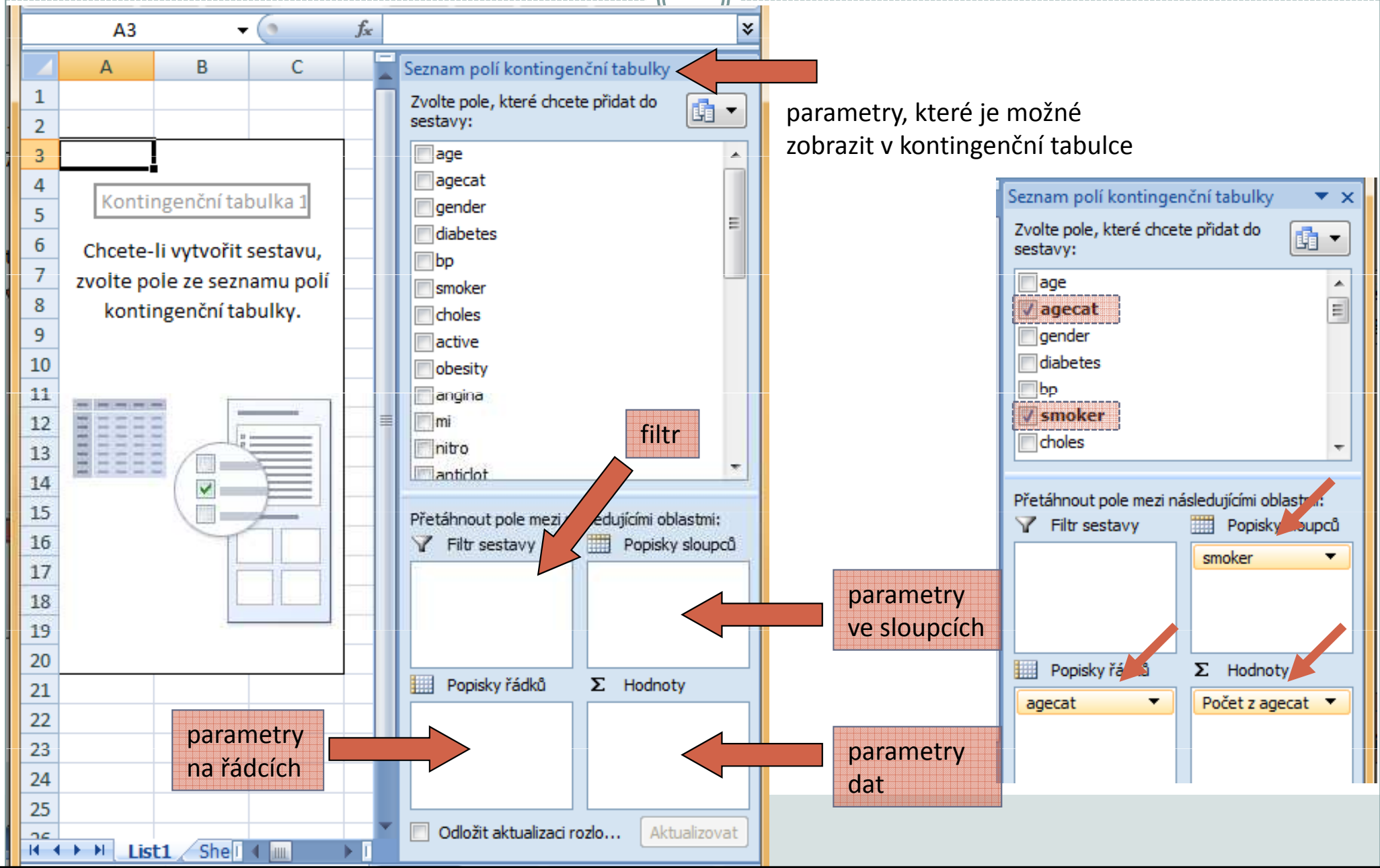


Zdroj dat (kromě Excelu i např. externí databáze)

Zdrojová oblast dat

Umístění tabulky

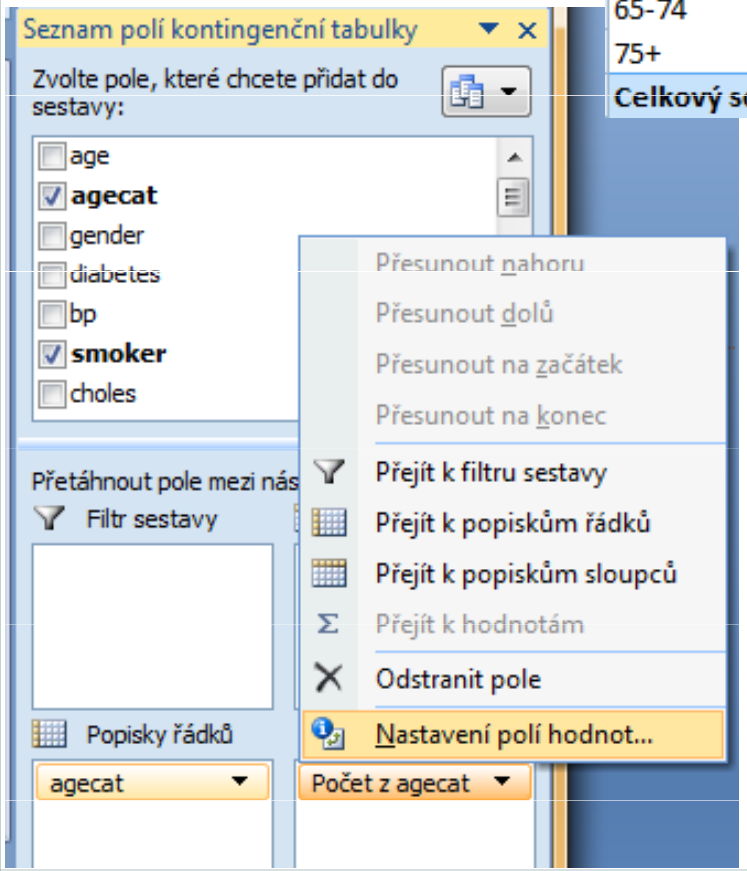
Kontingenční tabulky – rozvržení

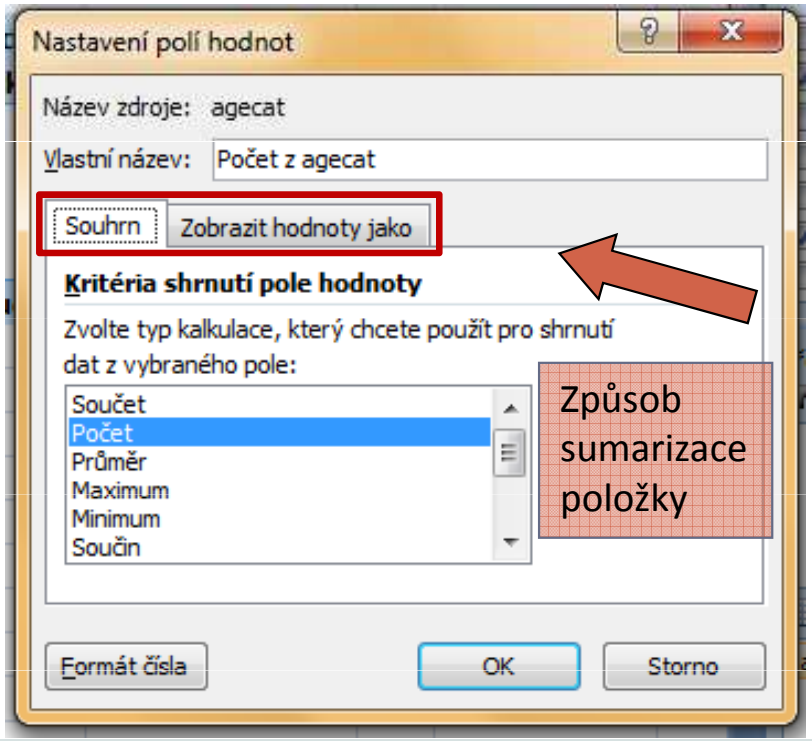


Kontingenční tabulky – nastavení II.

Kontingenční tabulka

Počet z agecat	Popisky sloupců		
Popisky řádků	No	Yes	Celkový součet
45-54	1694	501	2195
55-64	3015	863	3878
65-74	2200	661	2861
75+	816	250	1066
Celkový součet	7725	2275	10000





Způsob sumarizace položky

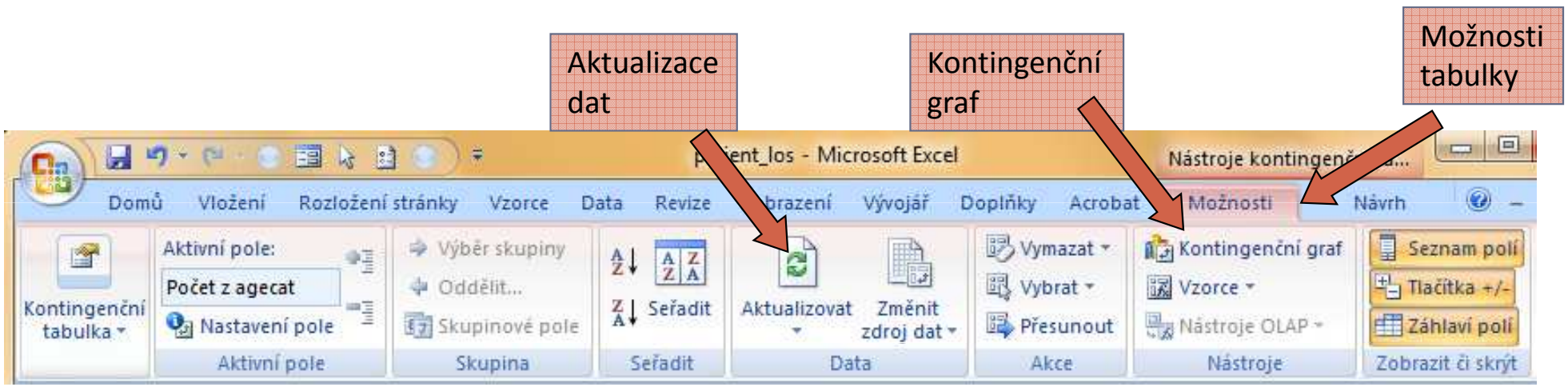
Aktualizace dat v kontingenční tabulce

Při změně dat v tabulce se zdrojovými daty **nedojde** automaticky k aktualizaci dat v kontingenční tabulce.

Musíte provést aktualizaci dat.

1. Stůjte kdekoliv v kontingenční tabulce
2. Na kartě **Možnosti** ve skupině **Data** klikněte na **Aktualizovat** (Alt+F5), nebo na **Aktualizovat vše** (Ctrl+Alt+F5)

Data z kontingenční tabulky lze vizualizovat pomocí **kontingenčního grafu**



Rozložení kontingenční tabulky

Po vytvoření se kontingenční tabulka zobrazí v tzv. **kompaktním formátu**. Lze ji zobrazit ale i ve formě **tabulky**, nebo ve formě **osnovy**.

1. Stůjte kdekoliv v kontingenční tabulce
2. Na kartě **Návrh** vyberte tlačítko **Rozložení sestavy** a volbu **Zobrazit ve formě osnovy nebo zobrazit ve formě tabulky**

Kompaktní formát - uspořádání tabulky aby zabírala co nejméně místa

Forma osnovy - řádková pole nižší úrovně je od vyšších úrovní odsazena, řádky nejsou odděleny čarami

Forma tabulky - klasická forma tabulky, pole nižší úrovně jsou v dalším sloupci

Vyzkoušej!

II. Základy popisné statistiky

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, L. Dušek, M. Cvanová

Jaké úlohy řeší biostatistika?

- **Popis cílové populace** – odhady charakteristik cílové populace
- **Srovnání skupin** – testování hypotéz
- **Regresní analýza** – stochastické modelování pro vysvětlení variability
- **Predikce a klasifikace** – stochastické modelování a klasifikační algoritmy pro předpovídání neznámých hodnot

Motivace

- Realitu můžeme popisovat různými typy dat, každý z nich se specifickými vlastnostmi, výhodami, nevýhodami a vlastní sadou využitelných statistických metod - od binárních přes kategoriální, ordinální až po spojitá data roste míra informace v nich obsažené.

- Základním přístupem k popisné analýze dat je tvorba frekvenčních tabulek a jejich grafických reprezentací – histogramů.

Typy proměnných

Kvalitativní (kategoriální) proměnná

- lze ji řadit do kategorií, ale nelze ji kvantifikovat

  Příklad: ??

Kvantitativní (numerická) proměnná

- můžeme ji přiřadit číselnou hodnotu

  Příklad: ??

Typy proměnných

Kvalitativní (kategoriální) proměnná

- lze ji řadit do kategorií, ale nelze ji kvantifikovat
- Příklady: *pohlaví, HIV status, barva vlasů ...*

Kvantitativní (numerická) proměnná

- můžeme ji přiřadit číselnou hodnotu
- Příklady: *výška, váha, teplota, počet hospitalizací ...*

Kvalitativní znaky

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

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

  Příklady: 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).

  Příklady: stupeň bolesti (mírná/střední/velká), stadium maligního onemocnění (I/II/III/IV).

Kvantitativní znaky

- **Intervalové znaky:** interpretace rozdílu dvou hodnot (stejný interval mezi jednou a druhou dvojicí hodnot vyjadřuje i stejný rozdíl v intenzitě zkoumané vlastnosti). Společný znak intervalových znaků: nula byla stanovena uměle, tedy pouhou konvencí. *Příklad: teplota měřená ve stupních Celsia, letopočet.*

Den	Teplota	Rozdíl [1]	Podíl [1]
1.	2 °C	-	-
2.	4 °C	+2	2x
3.	6 °C	+2	1.5x

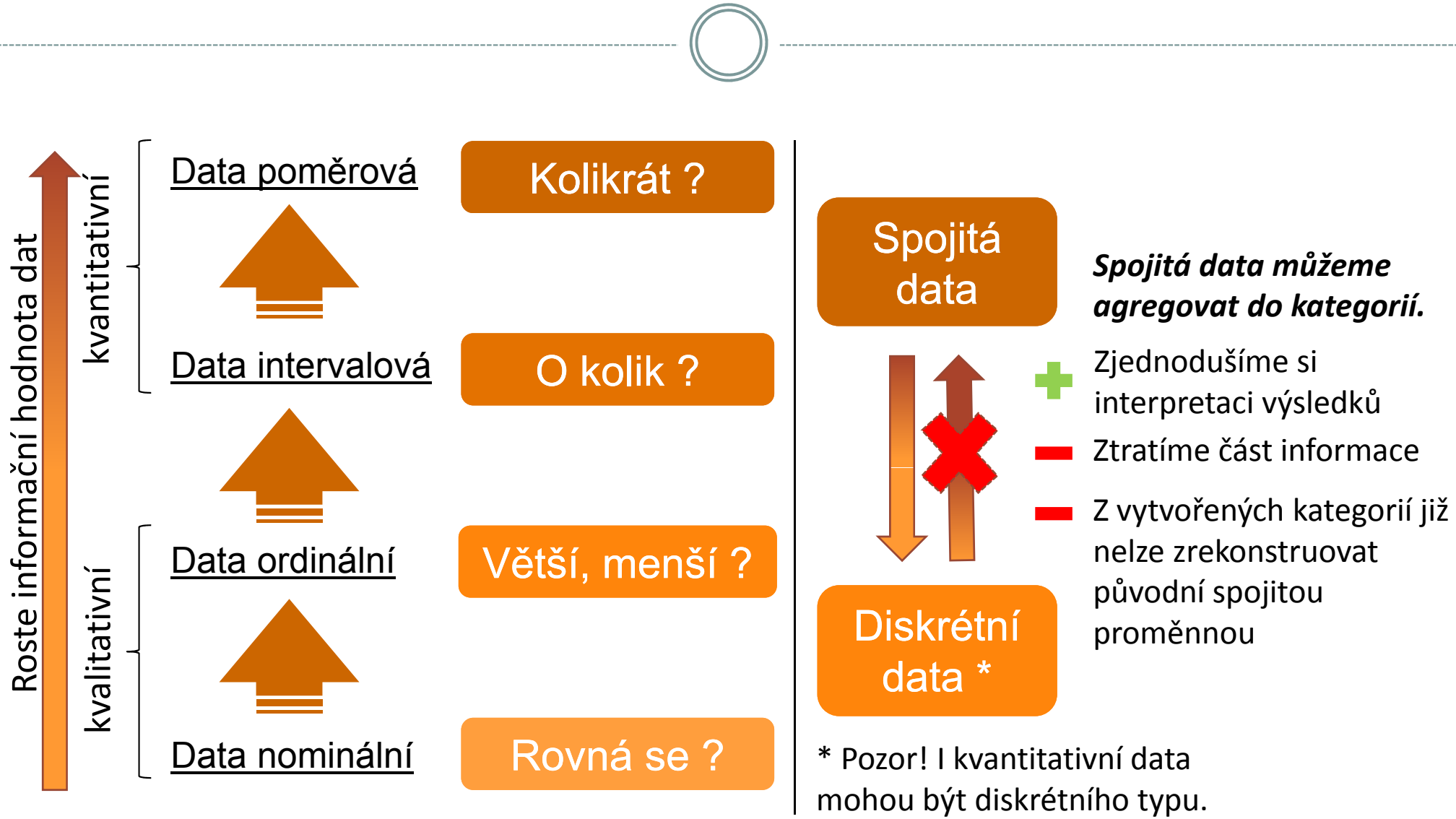
[1] Srovnání s měřením z předchozího dne

← 1.5krát vyšší teplota ve srovnání s 2. dnem, přičemž došlo ke stejnému nárůstu teploty jako při srovnání 2. a 1. dne

- **Poměrové znaky:** kromě rozdílu interpretujeme i podíl dvou hodnot.

  Příklady: výška v cm, váha v kg, ...

Různé typy dat znamenají různou informaci



Jak vznikají informace ?
- frekvenční tabulka jako základní nástroj popisu

DISKRÉTNÍ DATA

Primární data

Počty epizod pro n = 100 hemofiliků

0
0
1
2
1
1
3
1
1
2
.
.
.
.
.
.
n = 100

Frekvenční sumarizace

N: 100 dětí (hemofiliků)
x: znak: počet krvácivých epizod za měsíc

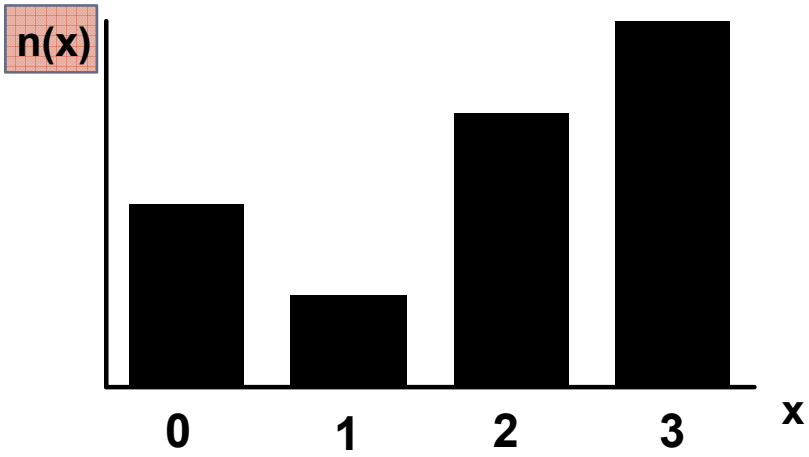
x	n(x)	N(x)	p(x)	F(x)
0	20	20	0,2	0,2
1	10	30	0,1	0,3
2	30	60	0,3	0,6
3	40	100	0,4	1,0

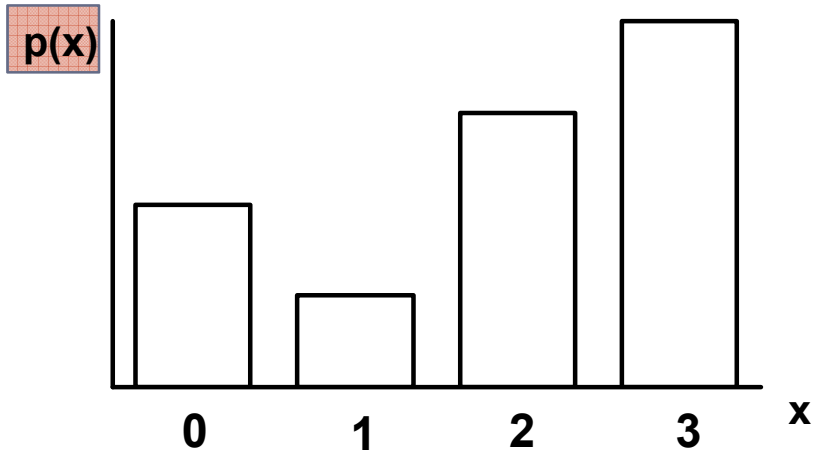
n(x) – absolutní četnost x
N(x) – kumulativní četnost hodnot nepřevyšujících x;

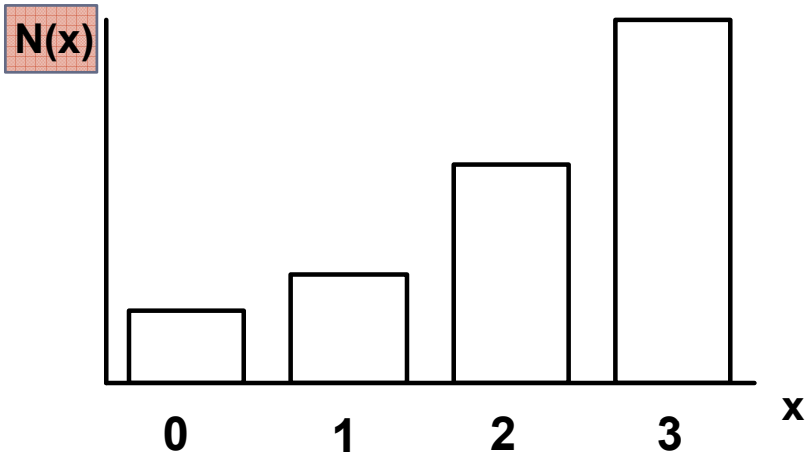
$$N(x) = \sum_{t \le x} n(t)$$

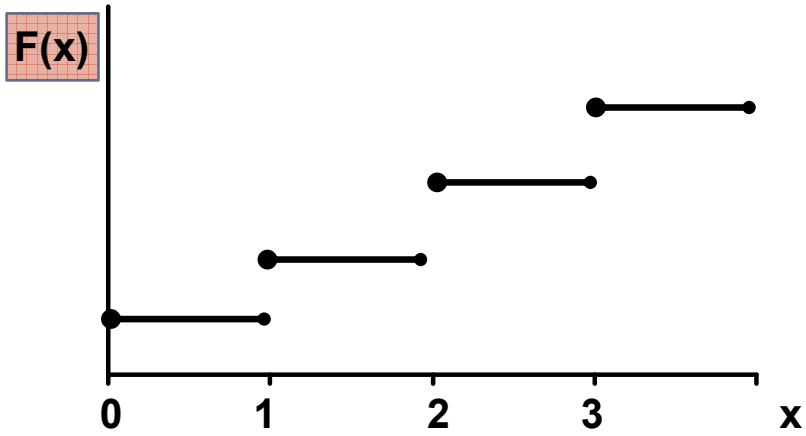
p(x) – relativní četnost; p(x) = n(x) / n
F(x) – kumulativní relativní četnost hodnot nepřevyšujících x; F(x) = N(x) / n

Jak vznikají informace ?
Grafické výstupy z frekvenční tabulky









Jak vznikají informace ?
- frekvenční tabulka jako základní nástroj popisu

SPOJITÁ DATA

Příklad: **x: koncentrace látky v krvi**
n = 100 pacientů

Primární data

Hodnoty pro n = 100 osob

1,21
1,48
1,56
0,31
1,21
1,33
0,33
.
.
.
n = 100

Frekvenční sumarizace

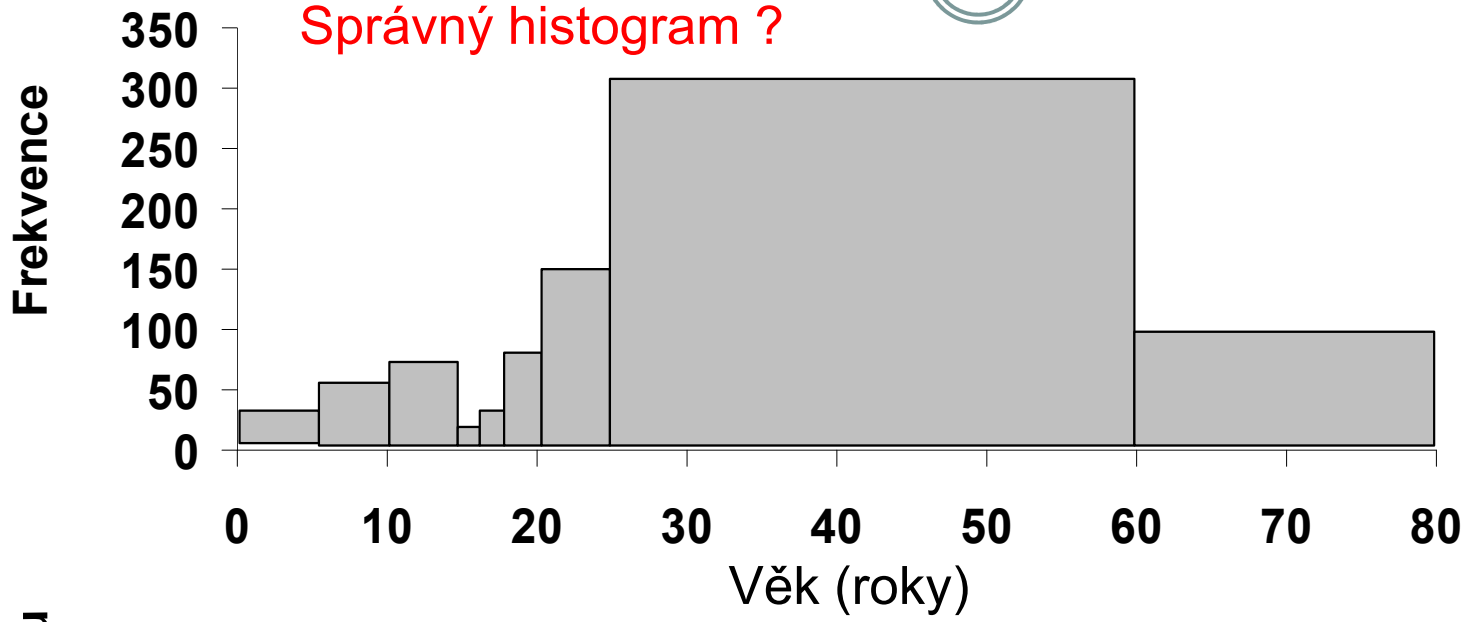
n = 100 opakovaných měření (100 pacientů)
x: koncentrace sledované látky v krvi (20 – 100 jednotek)

Interval*	d(l)	n(l)	n(l)/n	N(x'')	F(x'')
<20, 40)	20	20	0,2	20	0,2
<40, 60)	20	10	0,1	30	0,3
<60, 80)	20	40	0,4	70	0,7
<80, 100)	20	30	0,3	100	1,0

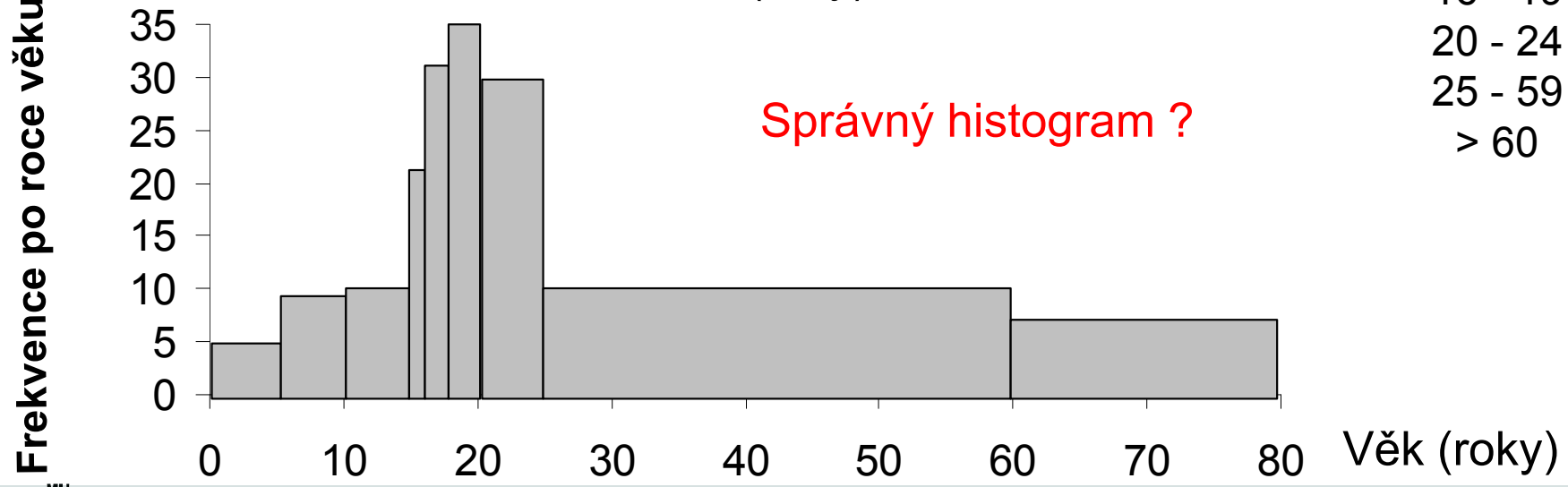
d(l) – šířka intervalu
n(l) – absolutní četnost
n(l) / n – intervalová relativní četnost
N(x") – intervalová kumulativní četnost do horní hranice X"
F(x") – intervalová relativní kumulativní četnost do horní hranice X"

* Třídící interval

Příklad: věk účastníků vážných dopravních nehod

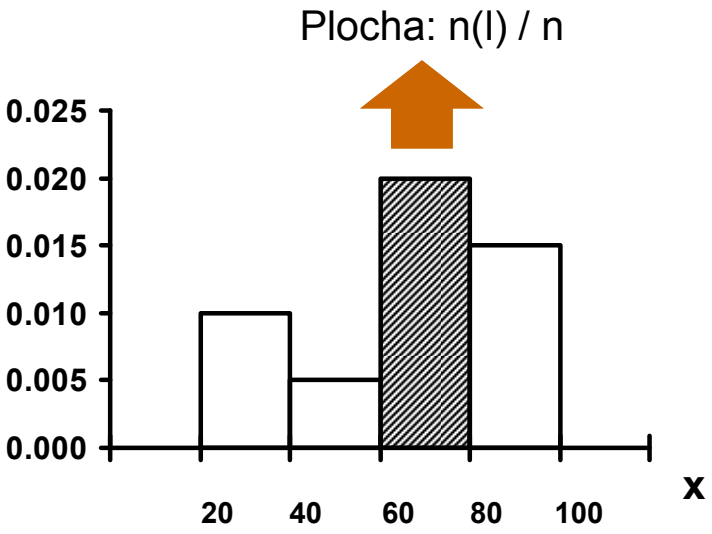


Věk	f
0 - 4	28
5 - 9	46
10 - 15	58
16 - 19	20
20 - 24	114
25 - 59	316
> 60	103



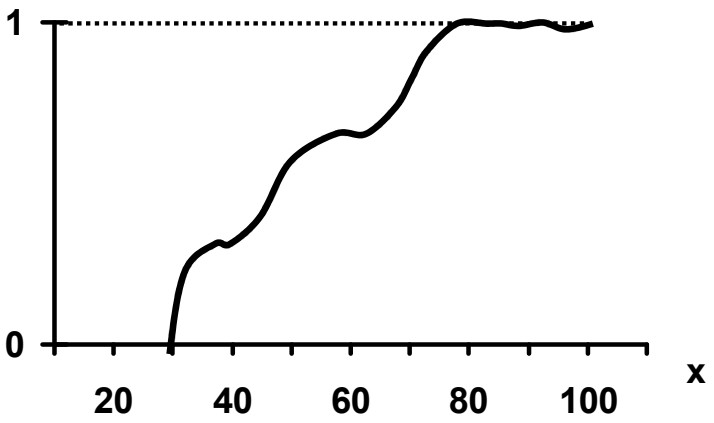
Jak vznikají informace ?
- frekvenční sumarizace spojitých dat

Histogram



$$f(x)=\frac{n(l)/n}{d(l)} \Rightarrow$$ **Intervalová hustota četnosti**

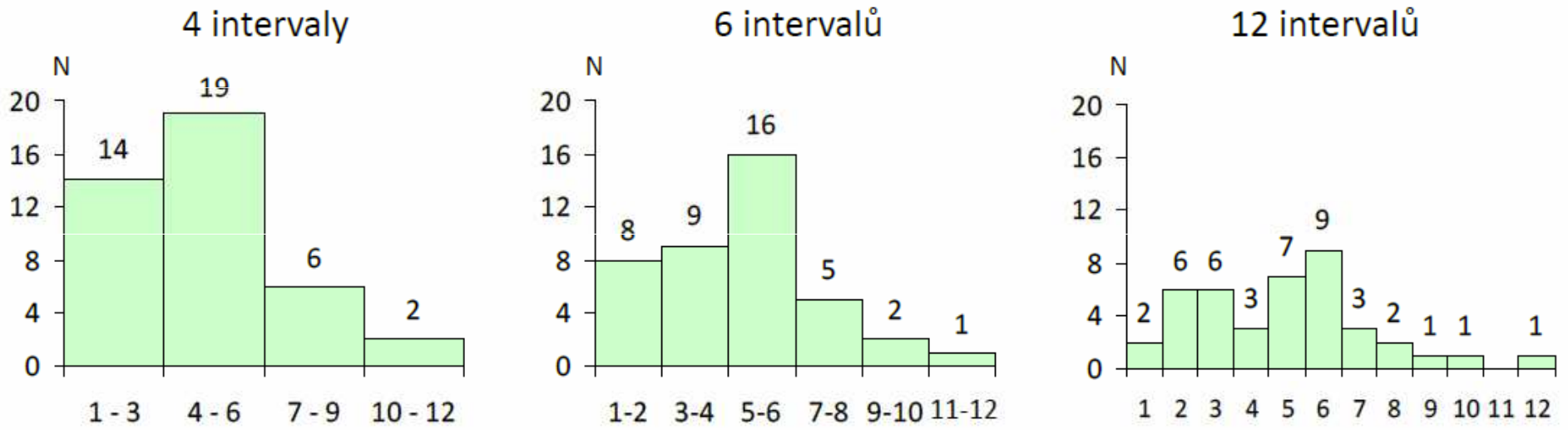
Výběrová distribuční funkce



F(x) ⇒ **Intervalová relativní kumulativní četnost**

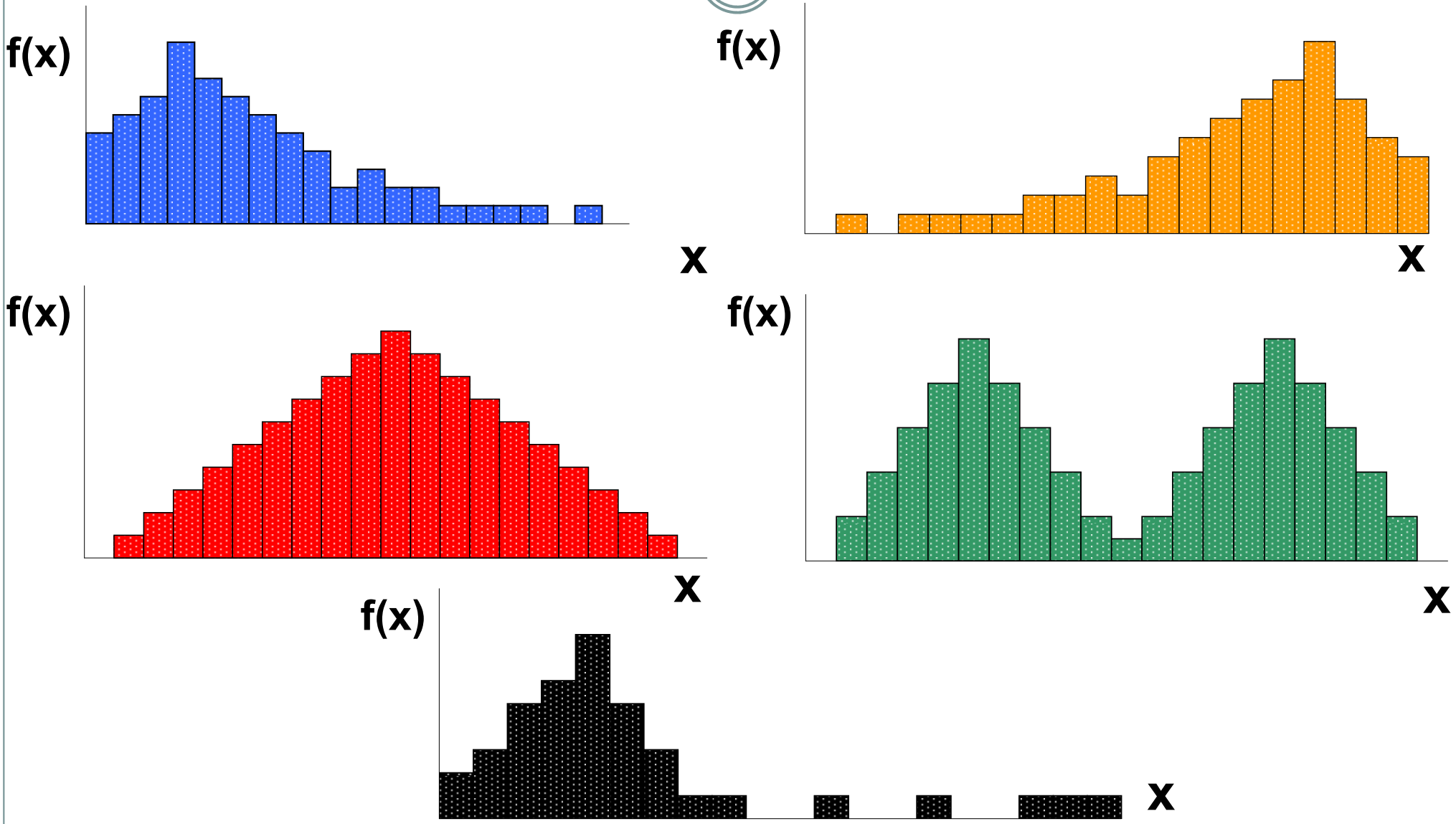
Histogram – počet intervalů

- 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 roztříštěná.

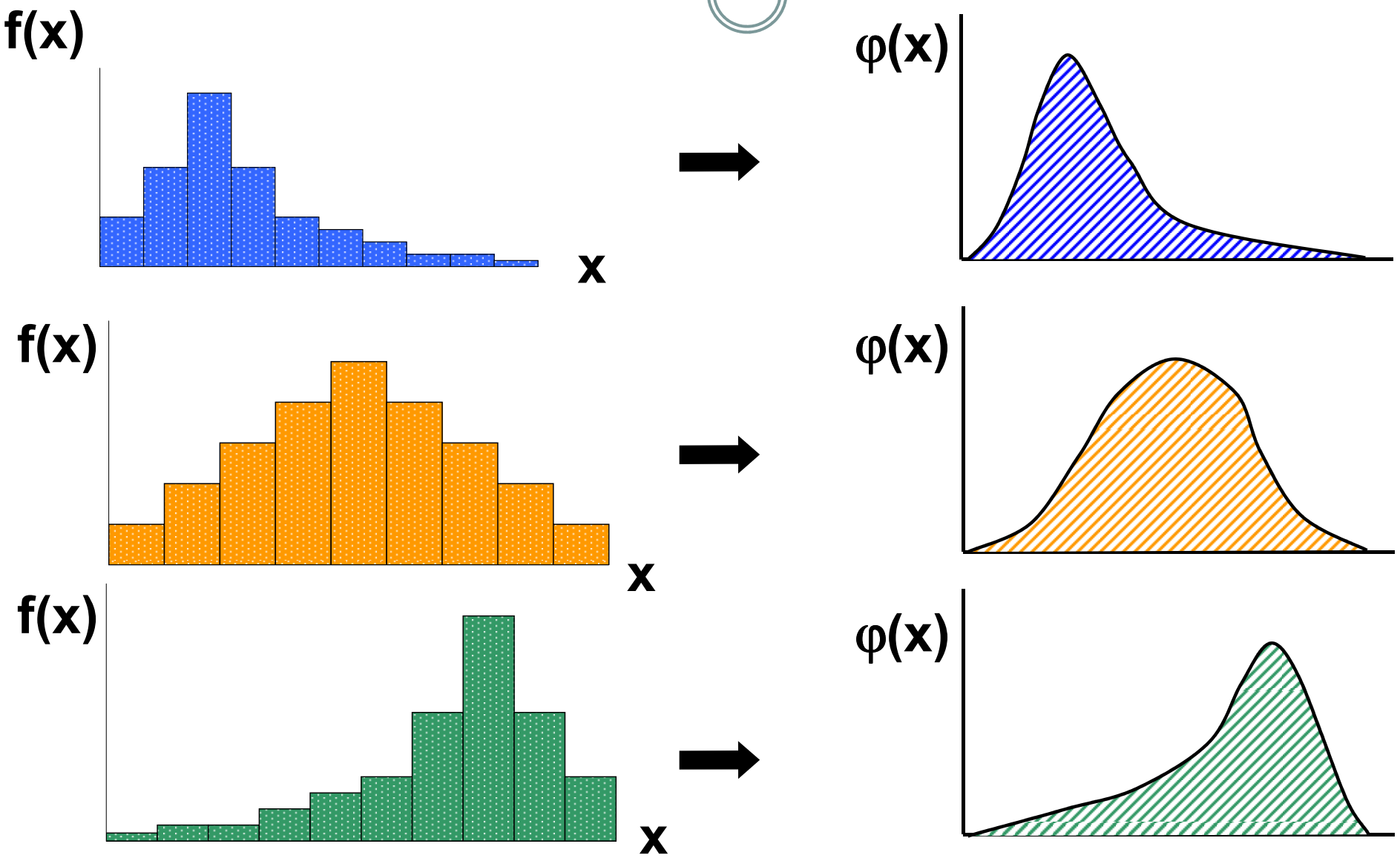


- Dvě základní metody volby počtu intervalů *m*:
  1. **Odmocnina z celkového počtu**: $m = \sqrt{n}$
  2. **Sturgesovo pravidlo**: $m = 1 + \log_2 n$

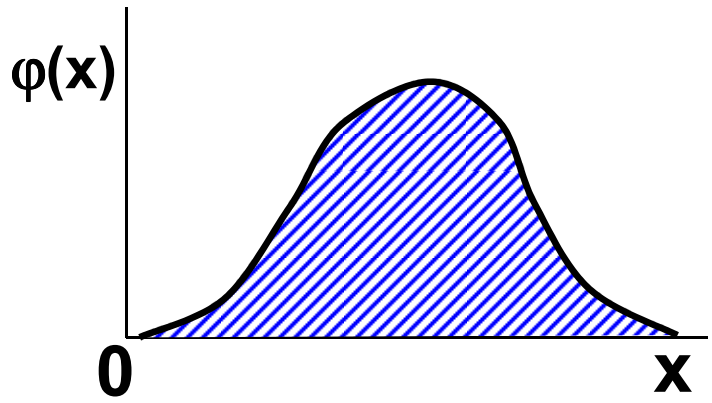
Histogram vyjadřuje tvar výběrového rozložení



Výběrové rozložení hodnot lze modelově popsat a definovat tak pravděpodobnost výskytu X



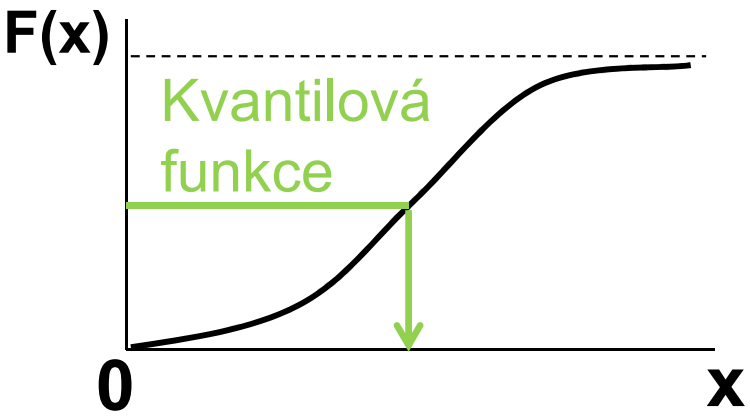
Pojem ROZLOŽENÍ - příklad spojitých dat



Hustota pravděpodobnosti =rozložení

Integrál

Derivace



Distribuční funkce

Je - li dána distribuční funkce, je dáno rozložení

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, geometrický průměr

Charakteristiky variability (proměnlivosti)

- Zachycují rozptýlení hodnot v souboru (proměnlivost dat)
- Variační rozpětí, rozptyl, směrodatná odchylka, variační koeficient, střední chyba průměru

Nominální znaky

Charakteristika polohy

- **Modus**: nejčastěji se vyskytující hodnota proměnné v souboru (hodnota s největší četností). V tabulce rozdělení četností se modus určí jednoduše z hodnoty znaku s největší četností.

Ordinální znaky

Charakteristika polohy

- **α-kvantil**: je-li α ∈ (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-α všech dat.
- Pro speciálně zvolená α užíváme názvů:

  $x_{0,50}$- **medián**, $x_{0,25}$- **dolní kvartil**, $x_{0,75}$-**horní kvartil**, $x_{0,1}$.... $x_{0,9}$-**decily**
- **Medián** znamená hodnotu, jež dělí řadu podle velikosti seřazených výsledků na dvě stejně početné poloviny. Jestliže n je sudé číslo, pak $\tilde{x} = 0{,}5(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$

  Jestliže n je liché číslo, pak $\tilde{x} = x_{(n+1)/2}$

Charakteristika variability

- **Kvartilové rozpětí (odchylka)**: $q=x_{0,75}-x_{0,25}$

Intervalové a poměrové znaky I

Charakteristika polohy

- **Aritmetický průměr**: je definován jako součet všech naměřených údajů vydělený jejich počtem,

$$E(x) = \bar{x} = \sum_{i=1}^{n} x_i / n$$

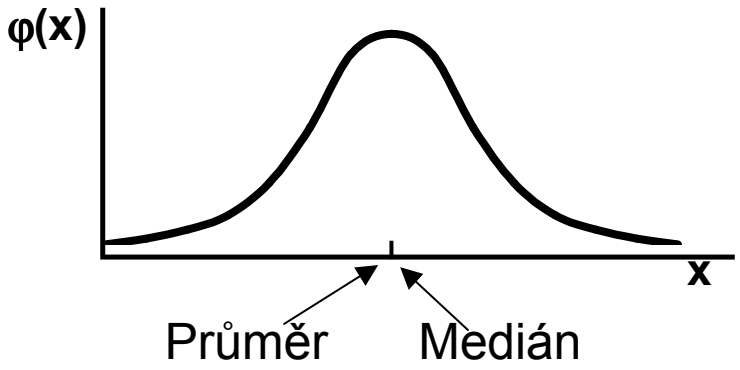
kde x_i jsou jednotlivé hodnoty a n jejich počet

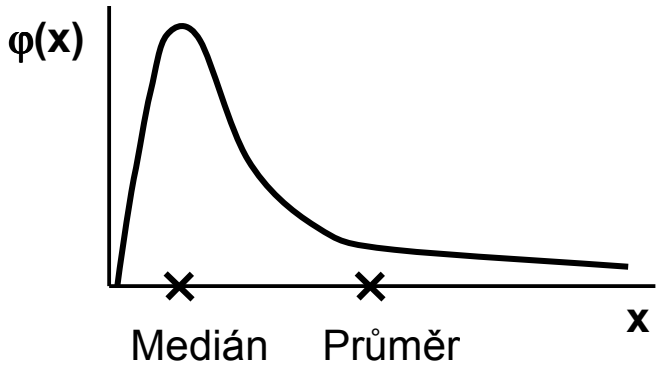
- **Geometrický průměr**: n kladných hodnot x_i, $\sqrt[n]{x_1 * \ldots\ldots * x_n}$, má smysl všude, kde má nějaký informační smysl součin hodnot proměnné. Z praktického hlediska platí, že logaritmus geometrického průměru je roven aritmetickému průměru logaritmovaných hodnot souboru.

Průměr vs medián

PAMATUJ:

- **Průměr je silně ovlivněn extrémními hodnotami (tzv. odlehlá pozorování), medián není ovlivněn vybočujícími pozorováními**
- **Průměr je vhodný ukazatel středu u normálního/symetrického rozložení, medián je vhodnou charakteristikou středu souboru i v případě veličin s neznámým rozdělením**
- V případě symetrického rozložení jsou jejich hodnoty v podstatě shodné, v případě asymetrického rozložení však nikoliv!





Intervalové a poměrové znaky II

Charakteristiky variability

- **Rozptyl (variance)** je ukazatelem šířky rozložení získaný na základě odchylky jednotlivých hodnot od průměru

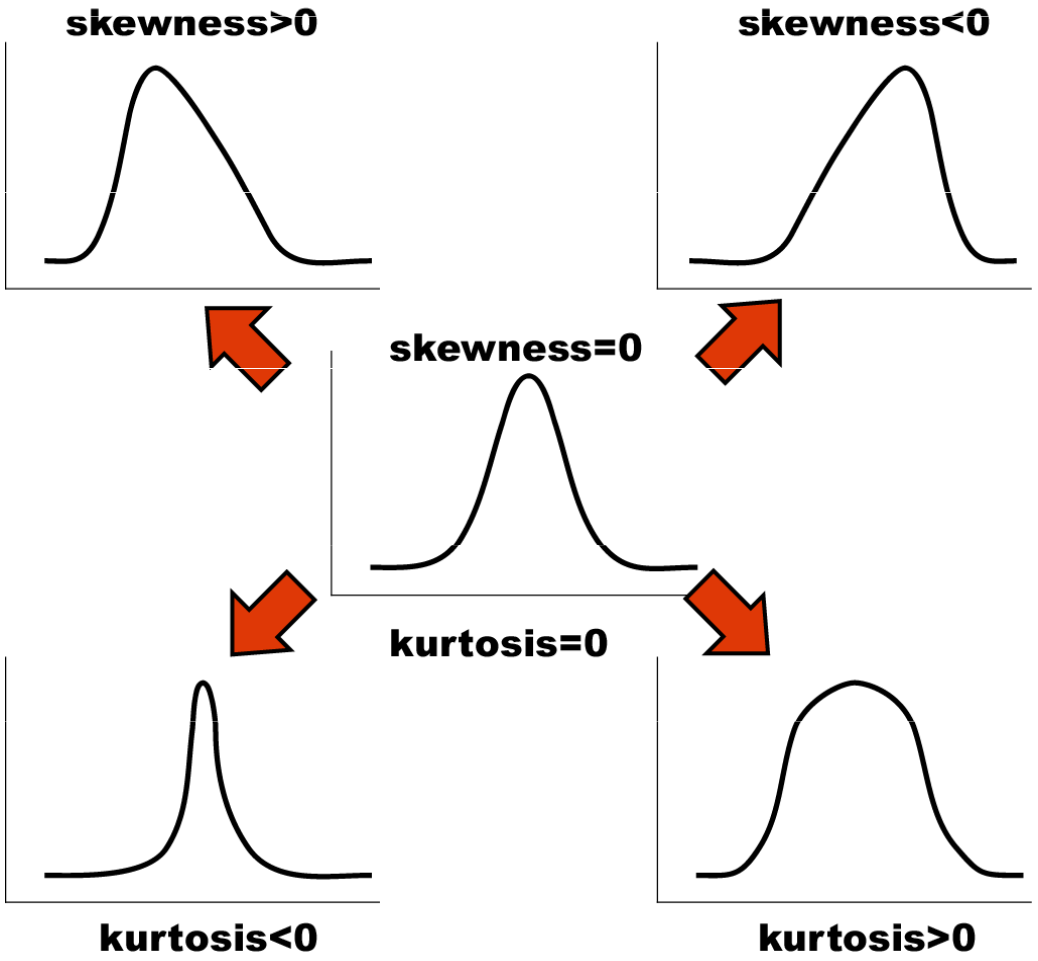
$$s^2 = \frac{\sum (x_i - x)^2}{n - 1}$$

  Obdobně jako u průměru je jeho vypovídací schopnost nejvyšší v případě symetrického/normálního rozložení

- **Směrodatná odchylka (SD-standard deviation)** je druhá odmocnina z rozptylu
- **Koeficient variance** - podíl SD ku průměru, u poměrových znaků, umožňuje porovnat variabilitu několika znaků (často se vyjadřuje v procentech – potom udává, z kolika procent se podílí směrodatná odchylka na aritmetickém průměru)

Ukazatele tvaru rozložení

- **Skewness (šikmost)** – ukazatel „šikmosti“ rozložení, asymetrie rozložení
- **Kurtosis (špičatost)** – ukazatel „špičatosti/plochosti“ rozložení



Další parametry rozložení

- **Počet hodnot –** důležitý ukazatel, znamená jak moc lze na data spoléhat
- **Suma hodnot**
- **Minimum, maximum**
- **Variační rozpětí (rozsah)** – rozdíl mezi největší a nejmenší hodnotou řady
- **Střední chyba průměru (SE)** – měří rozptýlenost vypočítaného aritmetického průměru v různých výběrových souborech vybraných z jednoho základního souboru

Ukázka popisu a vizualizace kvalitativních dat

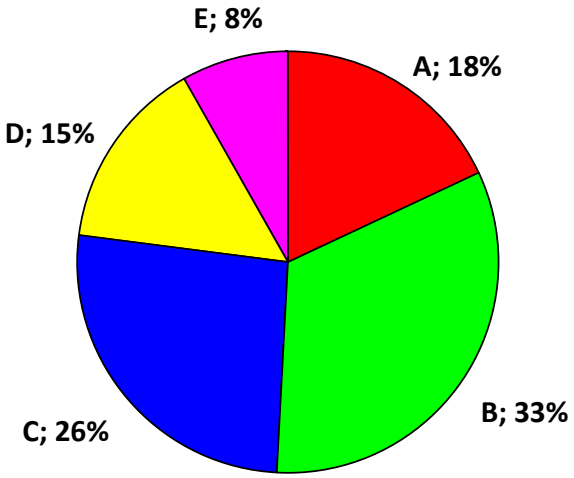
- **Popis kvalitativních dat:** frekvence jednotlivých kategorií
- **Vizualizace kvalitativních dat:** nejčastěji koláčový nebo sloupcový graf

Příklad: Známka z biostatistiky (podzim 2014)

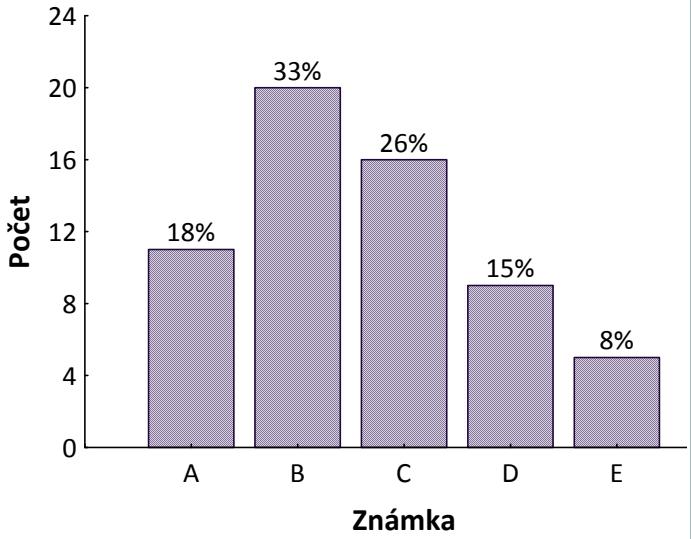
Frekvenční tabulka

Známka	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



Ukázka popisu kvantitativních dat

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

Příklad: Popis výšky (cm) pacientů

Popisné statistiky

Charakteristika	
N	61
Průměr (cm)	161,0
Medián (cm)	161,5
sm. odchylka (cm)	4,7
Rozptyl (cm^2)	22,2
min-max (cm)	144,1 - 169,2
dolní-horní kvartil (cm)	158,1 - 164,2

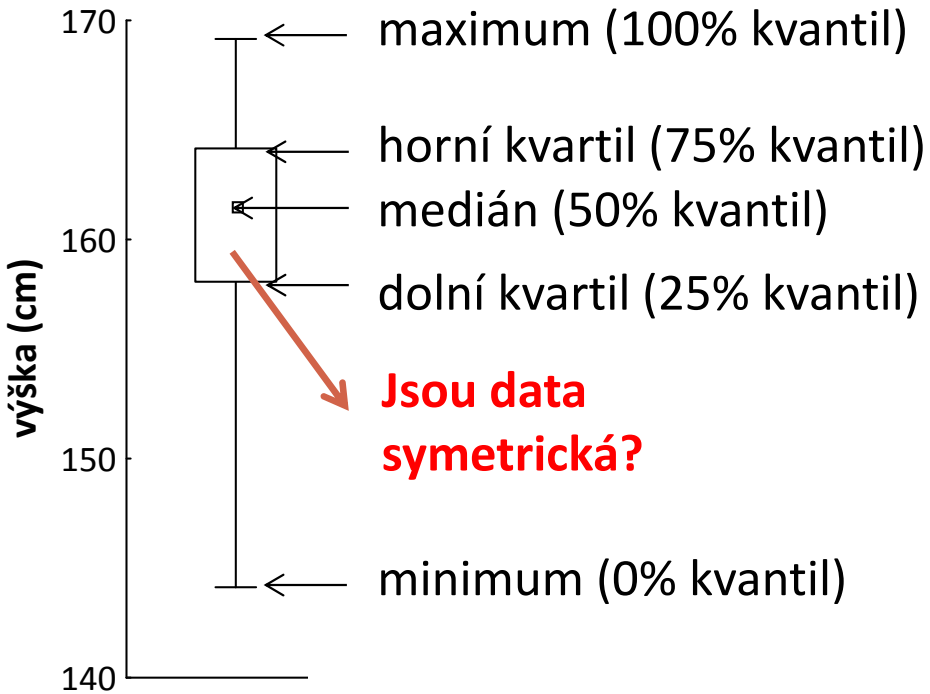
Průměr a medián se téměř shodují. Co nám to říká?

Ukázka vizualizace kvantitativních dat

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

Příklad: Popis výšky (cm) pacientů

Krabicový graf



Histogram