

# Základy popisné statistiky



# Anotace



- 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**

- Ize ji řadit do kategorií, ale nelze ji kvantifikovat
- Příklady: pohlaví, HIV status.....

## **Kvantitativní (numerická) proměnná**

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

# Kvalitativní znaky



- **Binární znaky:** dvě kategorie, obvykle se kódují pomocí čísel 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...*

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

*Příklady: výška v cm, váha v kg..*

- *Někdy je výhodné **kvantitativní data agregovat do kategorií** (např. věk do 10ti -letých věkových skupin)- tímto krokem však ztrácíme část informace.*

# Jak vznikají informace ?

– různé typy dat znamenají různou informaci

Data poměrová

Kolikrát ?



Data intervalová

O kolik ?



Data ordinální

Větší, menší ?



Data nominální

Rovná se ?

Spojité data

Diskrétní data

Kategoriální otázky

Otázky „Ano/Ne“

Podíl hodnot větší/menší než specifikovaná hodnota ?

Procenta odvozené hodnoty

**Samotná znalost typu dat ale na dosažení informace nestačí .....**

# Jak vznikají informace ?

– různé typy dat znamenají různou informaci

## Statistika středu

Data poměrová



**PRŮMĚR**

**Spojitá  
data**

Data intervalová



**MEDIÁN**

Data ordinální

**Diskrétní  
data**



Data nominální

**MODUS**

$Y = f$

X

# 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  
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;

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

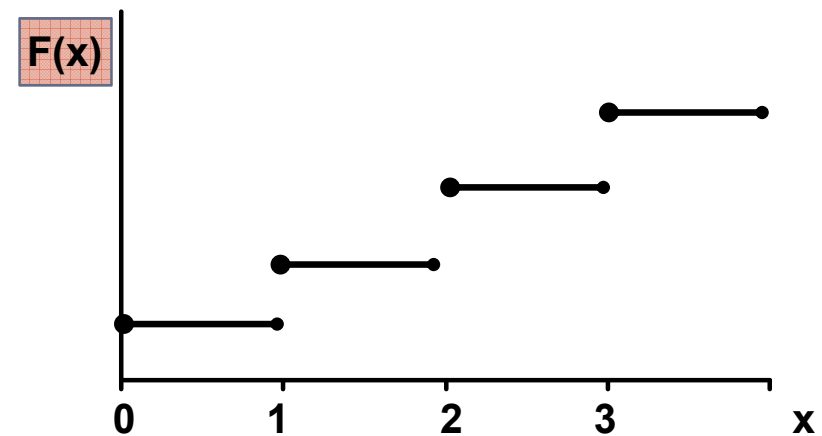
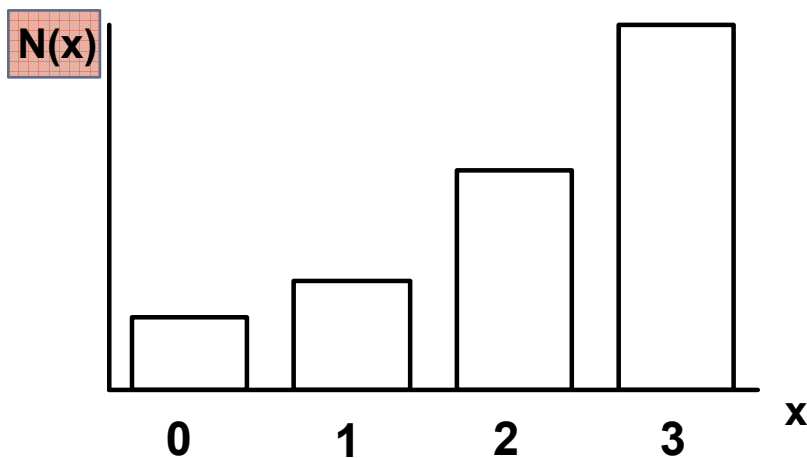
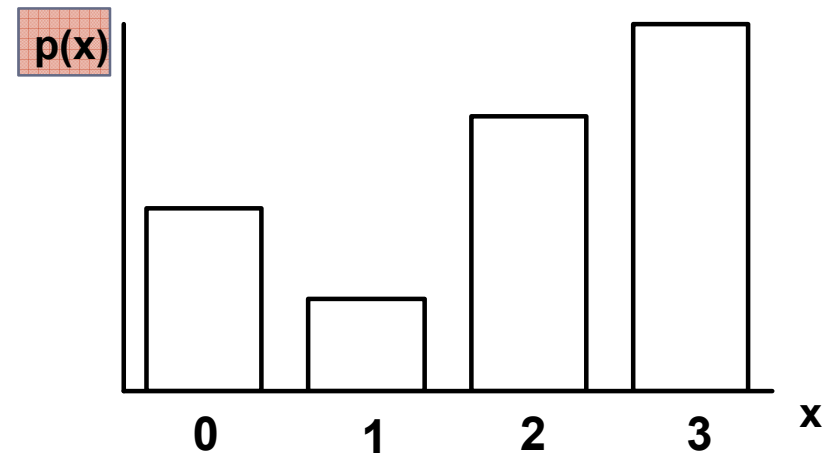
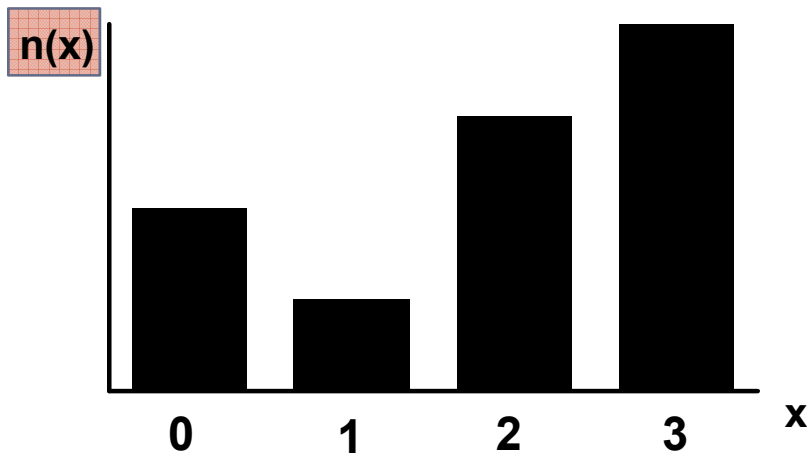
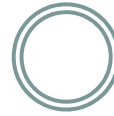
**p(x)** – relativní četnost;  $p(x) = n(x) / n$

**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*	$d(I)$	$n(I)$	$n(I)/n$	$N(x'')$	$F(x'')$
<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

$d(I)$  – šířka intervalu

$n(I)$  – absolutní četnost

$n(I) / n$  – intervalová relativní četnost

$N(x'')$  – intervalová kumulativní četnost do horní hranice  $X''$

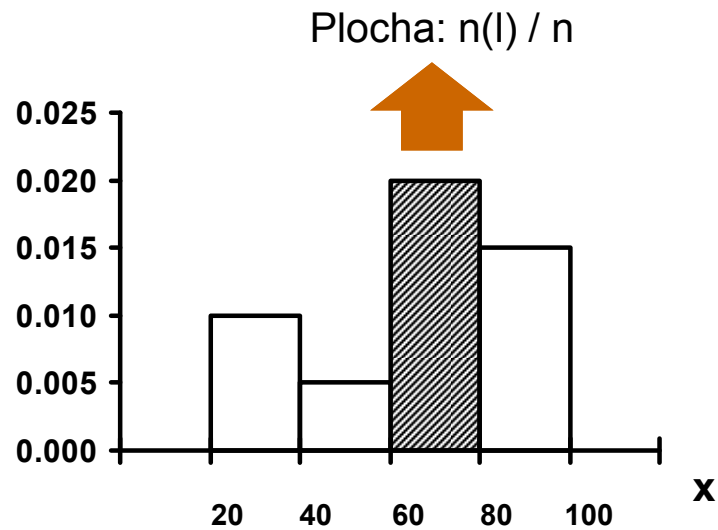
$F(x'')$  – intervalová relativní kumulativní četnost do horní hranice  $X''$

\* Třídící interval

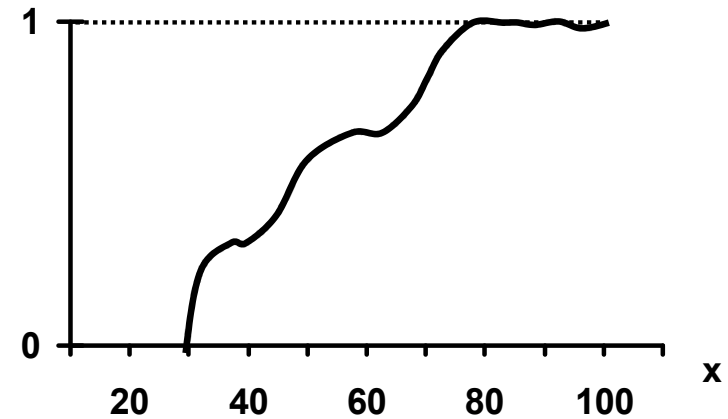
# Jak vznikají informace ?

## - frekvenční sumarizace spojitých dat

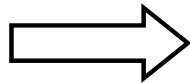
### Histogram



### Výběrová distribuční funkce

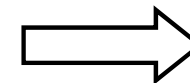


$$f(x) = \frac{n(l) / n}{d(l)}$$



Intervalová  
hustota  
četnosti

$F(x)$

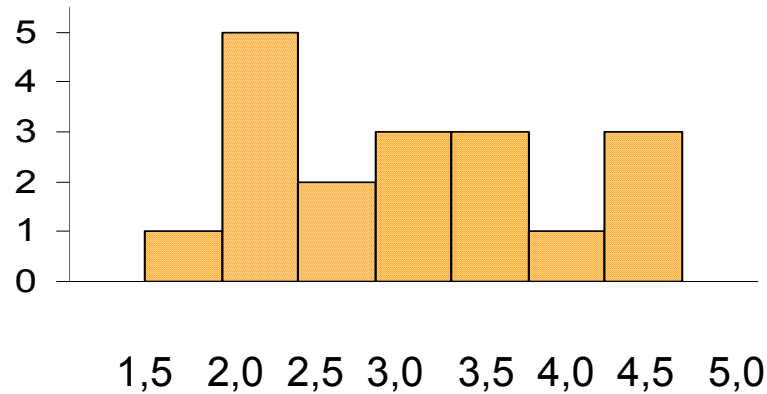


Intervalová  
relativní  
kumulativní  
četnost

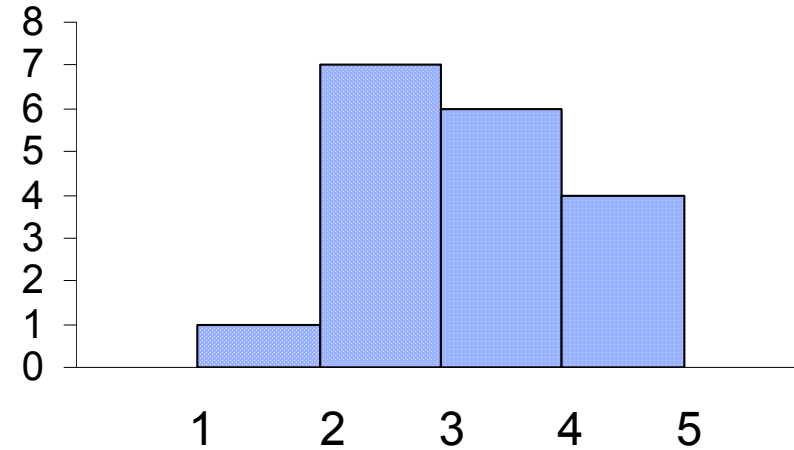
# Počet zvolených tříd a velikost souboru určují kvalitu výstupů



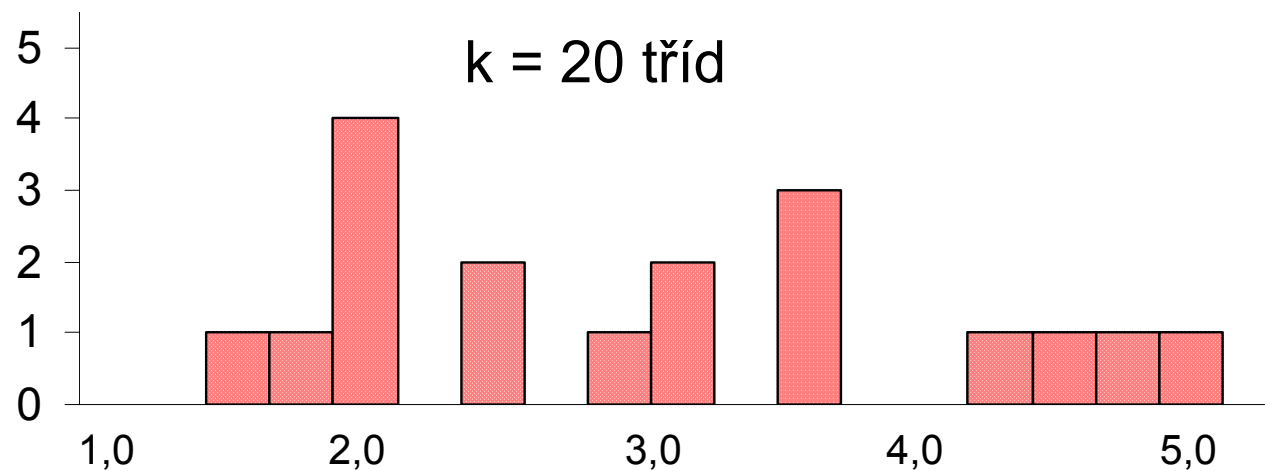
k = 10 tříd



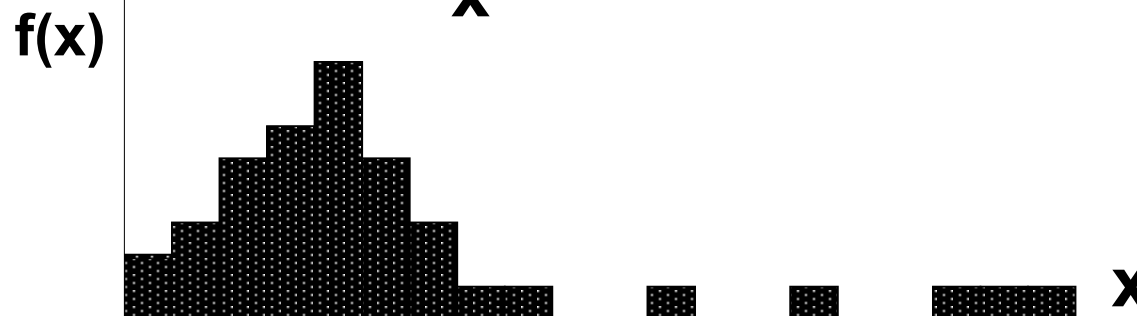
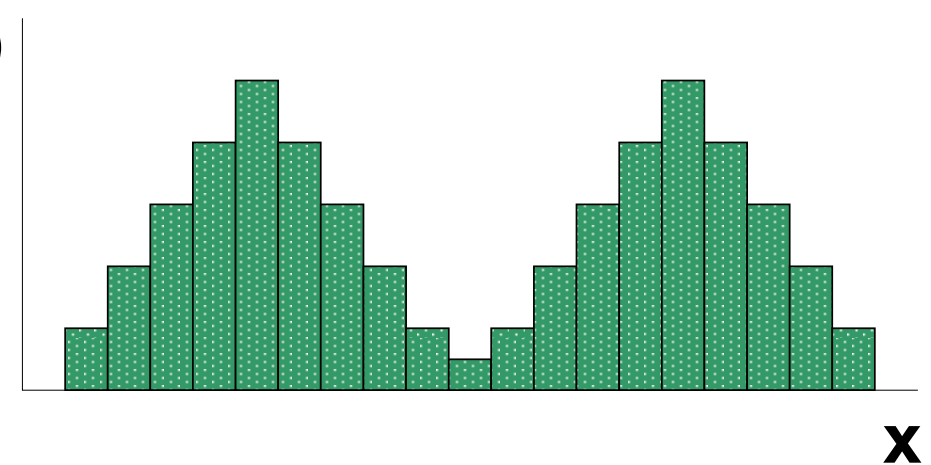
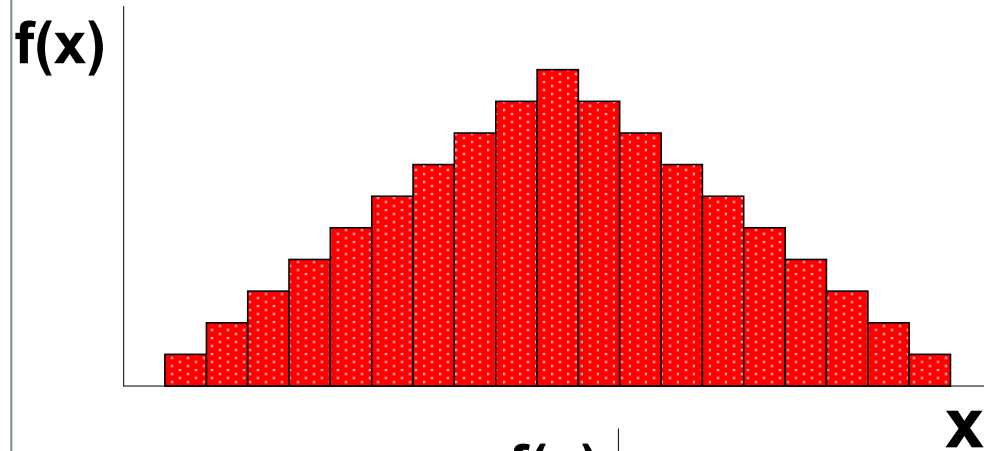
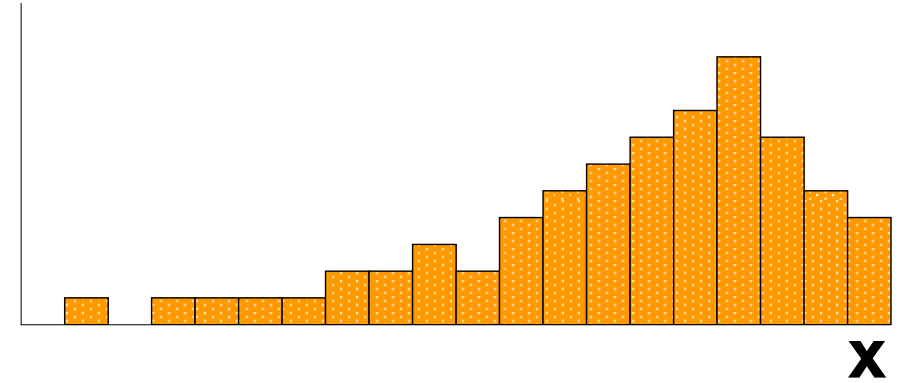
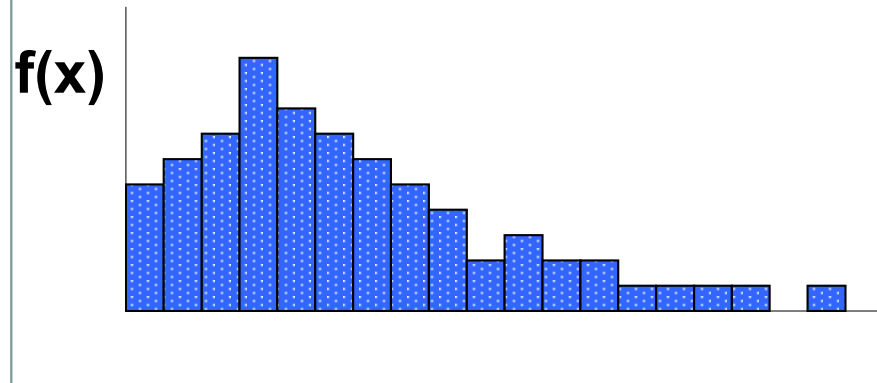
k = 5 tříd



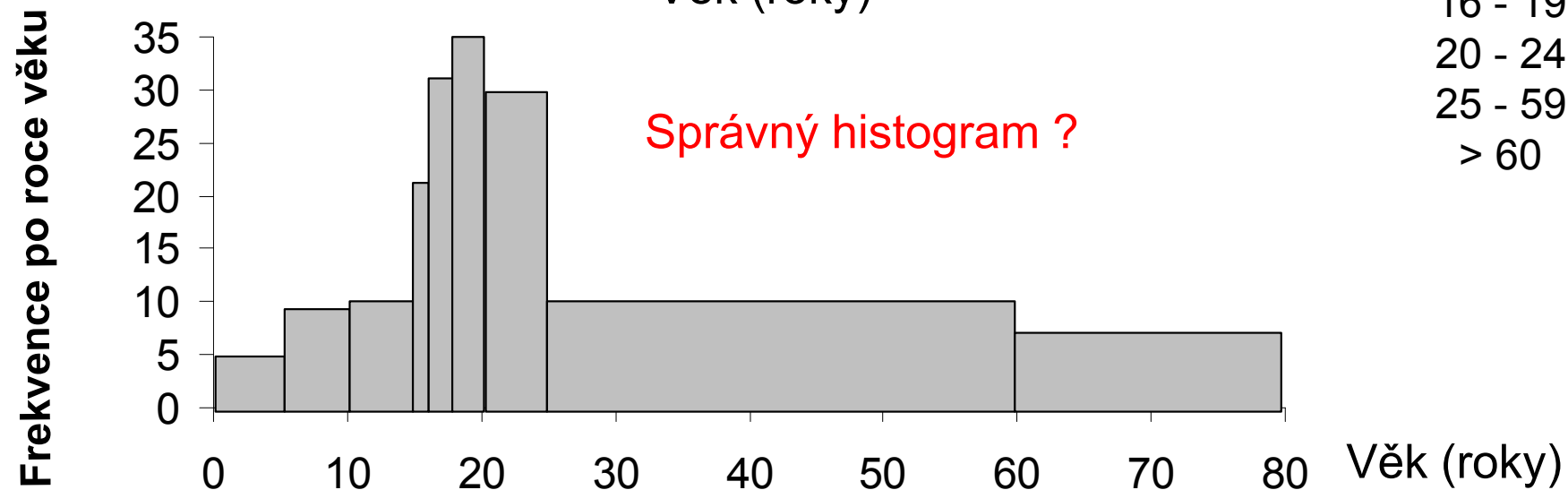
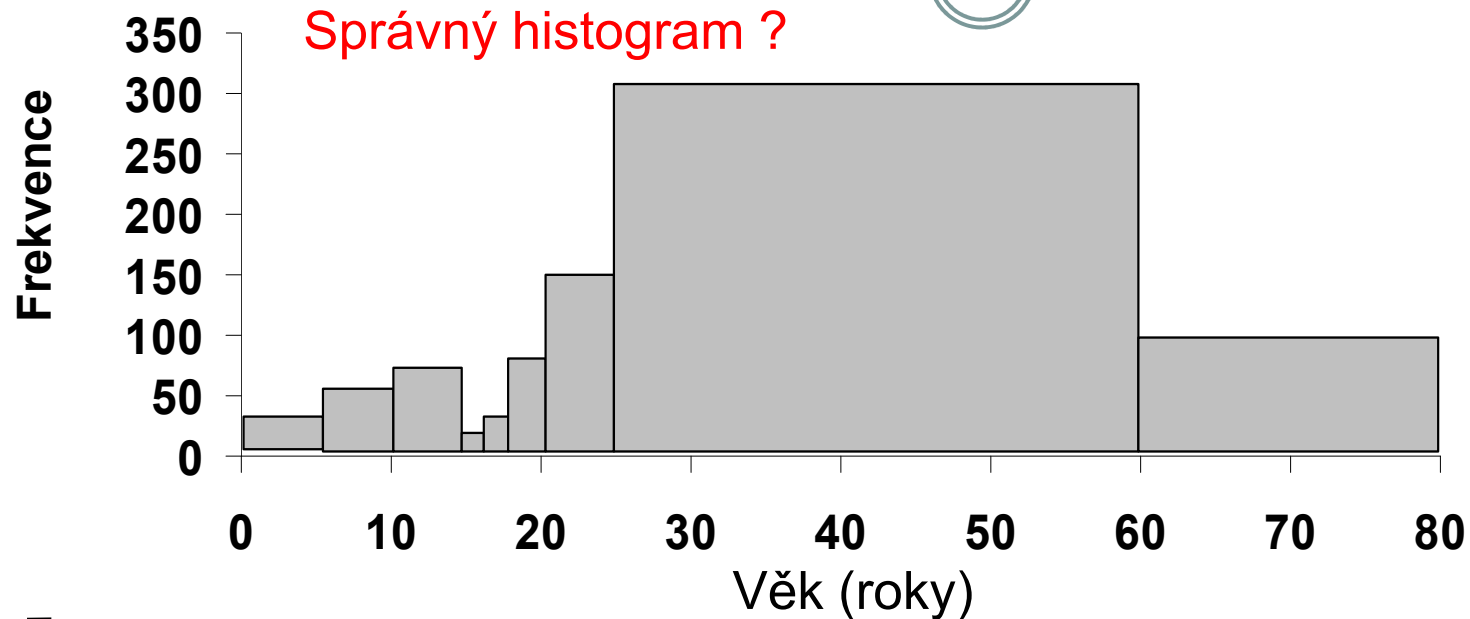
k = 20 tříd



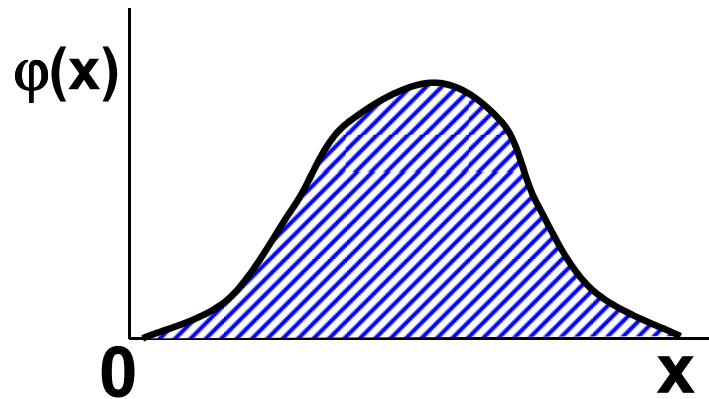
# Histogram vyjadřuje tvar výběrového rozložení



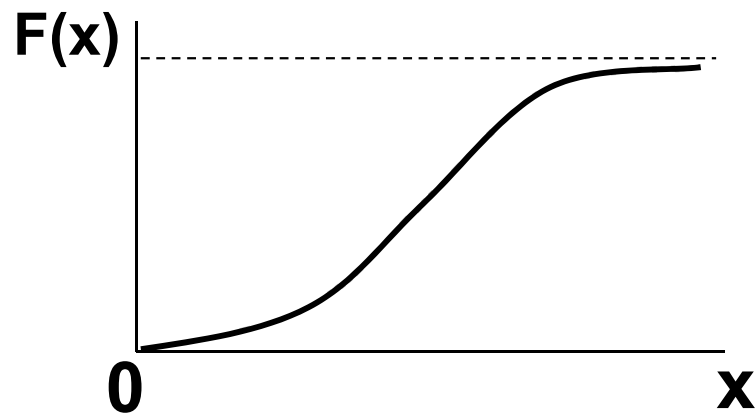
# Příklad: věk účastníků vážných dopravních nehod



# Pojem ROZLOŽENÍ - příklad spojitých dat



Rozložení

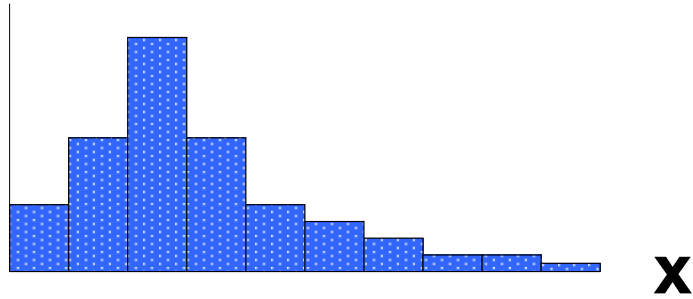


Distribuční funkce

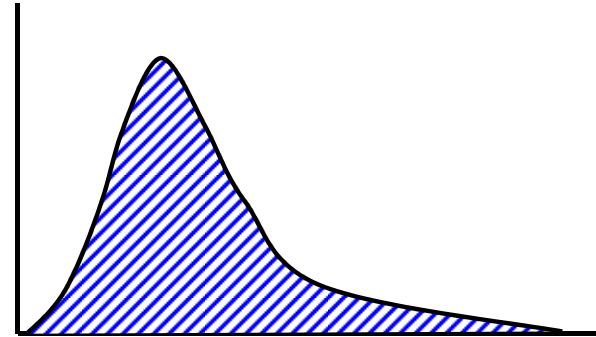
**Je - li dána  
distribuční  
funkce,  
je dáno  
rozložení**

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

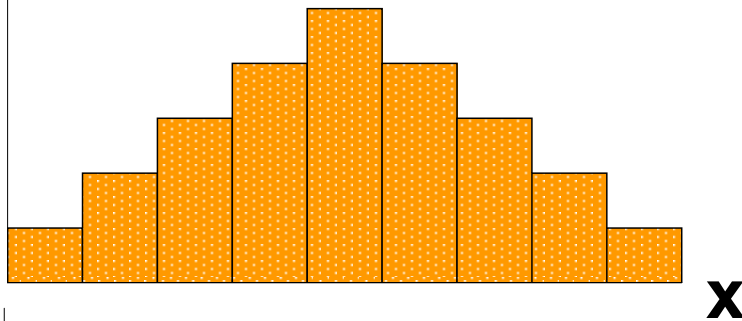
$f(x)$



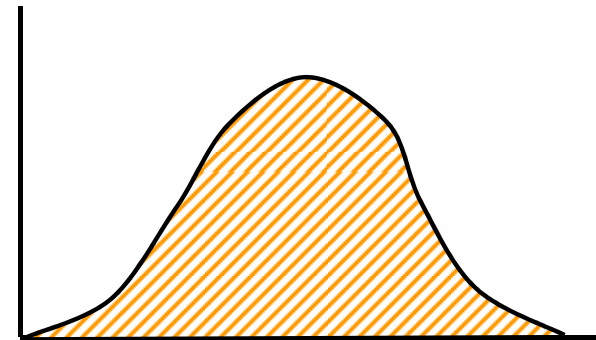
$\varphi(x)$



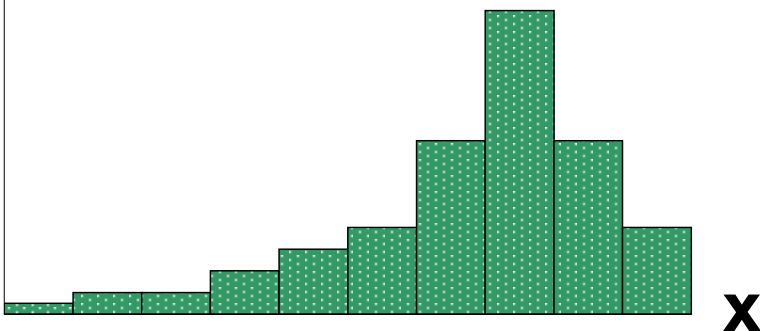
$f(x)$



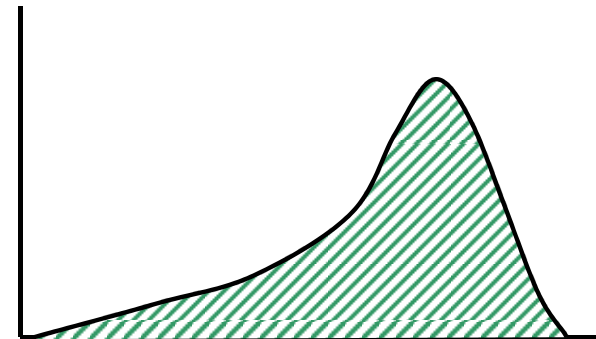
$\varphi(x)$



$f(x)$



$\varphi(x)$





# Popisné statistiky



## Charakteristiky polohy (míry střední hodnoty, míry centrální tendence)

- Udávají, kolem jaké hodnoty se data centrují, resp. které hodnoty jsou nejčastější
- **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

- **$\alpha$ -kvantil**: je-li  $\alpha \in (0,1)$ , pak  $\alpha$ -kvantil  $x_\alpha$  je číslo, které rozděluje uspořádaný datový soubor na dolní úsek, obsahující aspoň podíl  $\alpha$  všech dat a na horní úsek obsahující aspoň podíl  $1-\alpha$  všech dat.
- Pro speciálně zvolená  $\alpha$  užíváme názvů:  
 $x_{0,50}$ - **medián**,  $x_{0,25}$ - **dolní kvartil**,  $x_{0,75}$ - **horní kvartil**,  $x_{0,1}, \dots, x_{0,9}$ - **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,

$$E(x) = \bar{x} = \sum_{i=1}^n \frac{x_i}{n} \quad \text{kde } x_i \text{ jsou jednotlivé hodnoty a } n \text{ jejich počet}$$

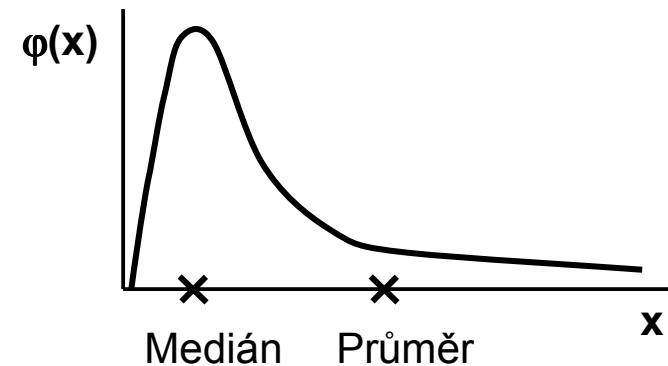
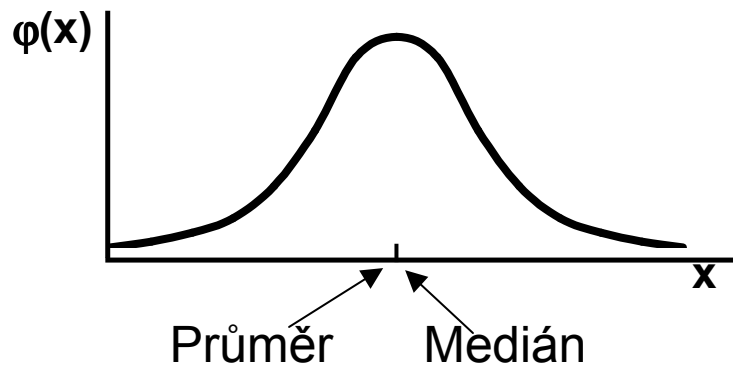
- **Geometrický průměr**: n kladných hodnot  $x_i$ ,  $\sqrt[n]{x_1 * \dots * x_n}$ , má smysl všude, kde má nějaký informační smysl součin hodnot proměnné. Z praktického hlediska platí, že logaritmus geometrického průměru je roven aritmetickému průměru logaritmovaných hodnot souboru.

# 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

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

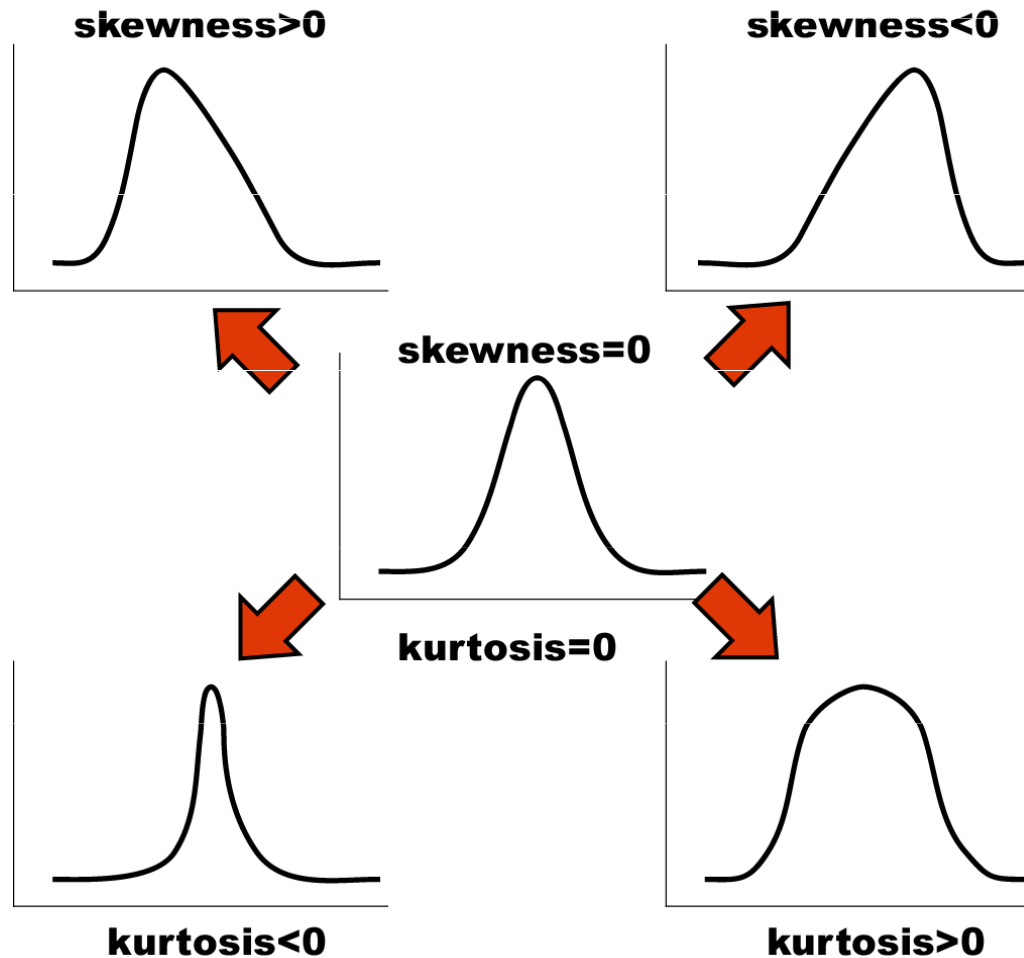
Obdobně jako u průměru je jeho vypovídací schopnost nejvyšší v případě symetrického/normálního rozložení

- **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** – ukazatel „šikmosti“ rozložení, asymetrie rozložení
- **Kurtosis** – ukazatel „špičatosti/plochosti“ rozložení



# Další parametry rozložení



- **Počet hodnot** – důležitý ukazatel, znamená jak moc lze na data spoléhat
- **Suma hodnot**
- **Minimum, maximum**
- **Variační rozpětí** – 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.