# V. Základní typy dat

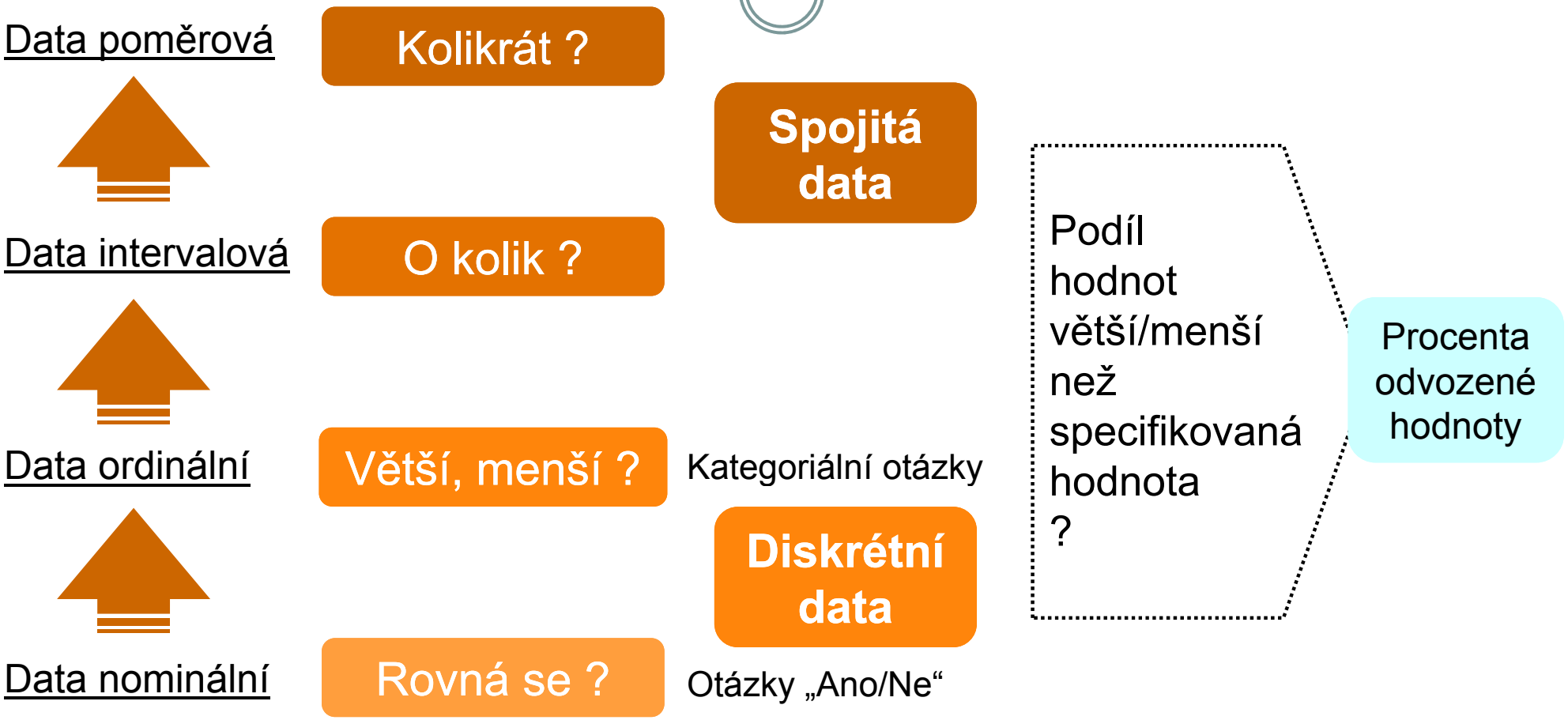
## Spojitá a kategoriální data
## Základní popisné statistiky
## Grafický popis dat

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

# 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.

# Jak vznikají informace ?
# – různé typy dat znamenají různou informaci




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

# Jak vznikají informace ?
# – různé typy dat znamenají různou informaci

## Statistika středu

Data poměrová

Data intervalová

Data ordinální

Data nominální

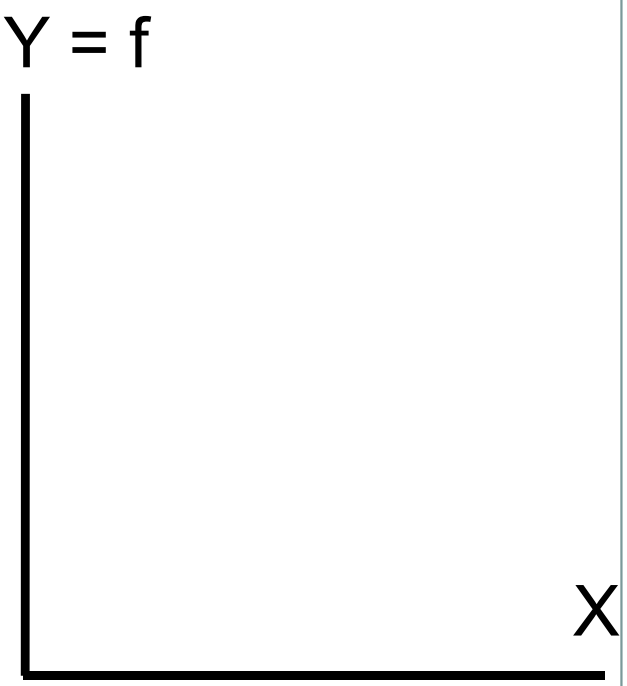
**PRŮMĚR**

**MEDIÁN**

**MODUS**

**Spojitá data**

**Diskrétní data**

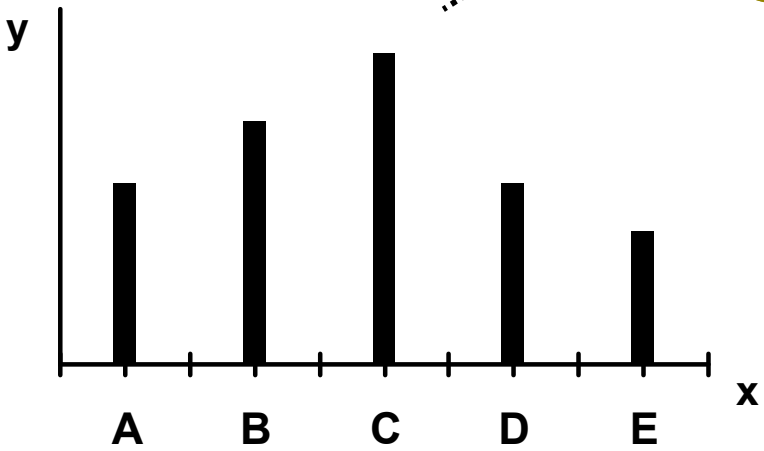


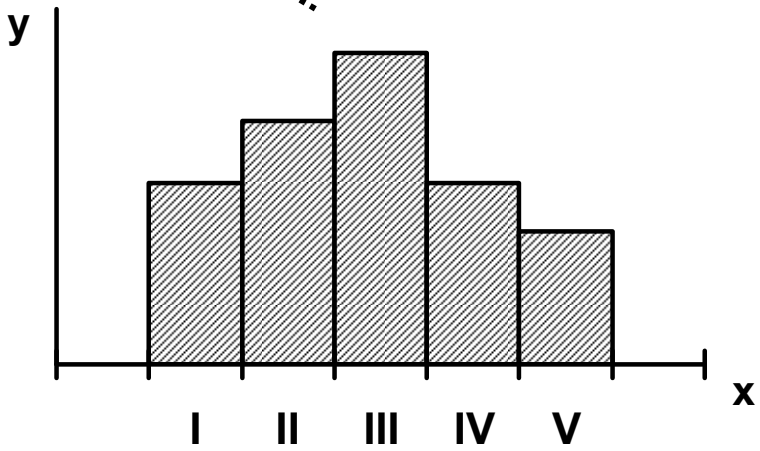
# JAK vznikají informace ?
# - opakovaná měření informují rozložením hodnot

**Y: frekvence**

**- absolutní / relativní**

**KOLIK se naměřilo**




**CO se naměřilo**

**X: měřený znak**

**Diskrétní data** ← → **Spojitá data**

# Odvozená data: Pozor na odvozené indexy

**Příklad I:**

**Znak X**: Hmotnost
**Znak Y**: Plocha

**Příklad II:**

**X**: Průměrný počet výrobků v prodejně
**Y**: Odhad prostoru průměrně nabízeného k vystavení výrobku

průměr : (min - max)

X: 1,2 : (1,15 - 1,24) → **+ / - 3,8 %**

Y: 1,8 : (1,75 - 1,84) → **+ / - 2,5 %**

$$X/Y = 0{,}667 : \left(\frac{1{,}15}{1{,}84} - \frac{1{,}24}{1{,}75}\right)$$ → **+ / - 6,2 %**

**Nová veličina má jinou šířku rozpětí než ty, ze kterých je odvozená**

# 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

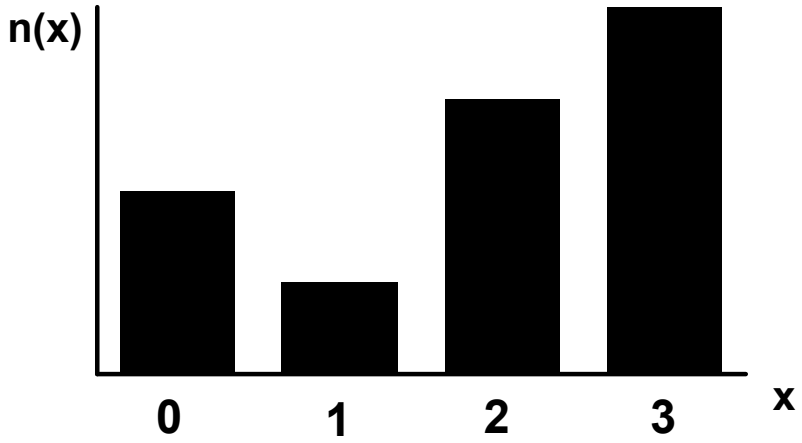
| x | n(x) | p(x) | N(x) | F(x) |
|---|---|---|---|---|
| 0 | 20 | 0,2 | 20 | 0,2 |
| 1 | 10 | 0,1 | 30 | 0,3 |
| 2 | 30 | 0,3 | 60 | 0,6 |
| 3 | 40 | 0,4 | 100 | 1,0 |

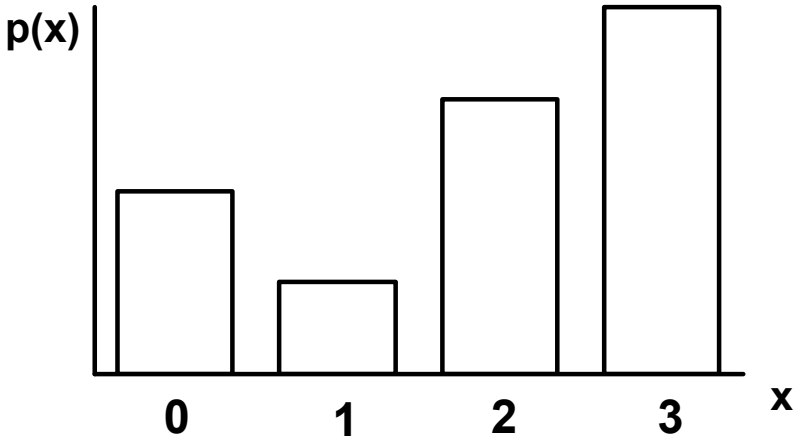
**n(x)** – absolutní četnost x
**p(x)** – relativní četnost; p(x) = n(x) / n
**N(x)** – kumulativní četnost hodnot nepřevyšujících x;

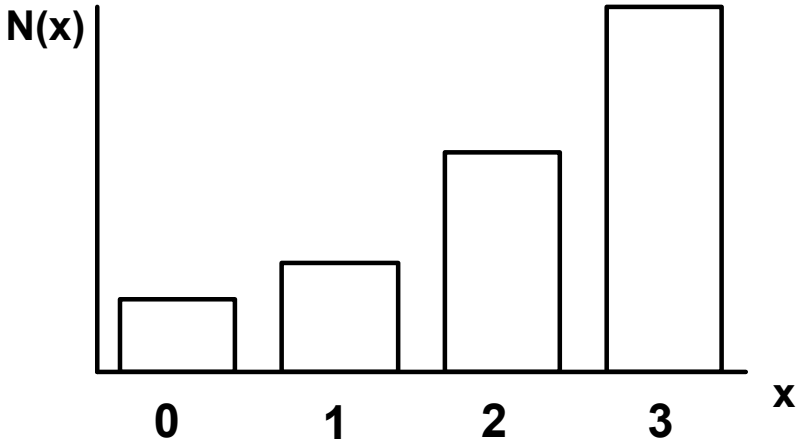
$$N(x) = \sum_{t \le x} n(t)$$

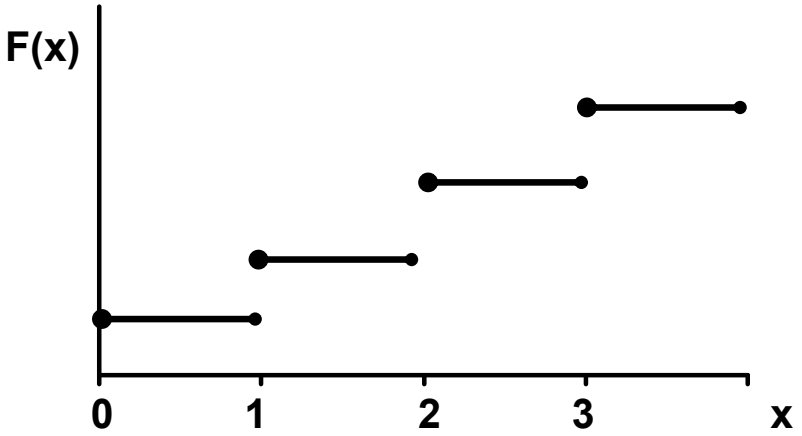
**F(x)** – kumulativní relativní četnost hodnot nepřevyšujících x; F(x) = N(x) / n

# 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**

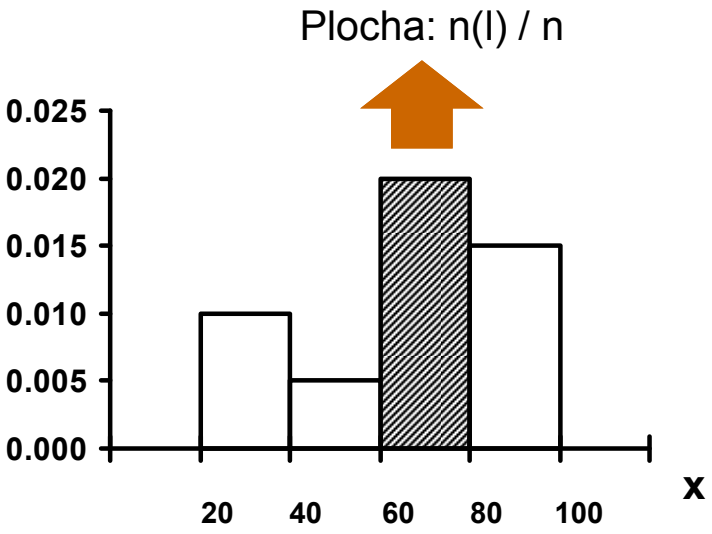
**n** = 100 opakovaných měření (100 pacientů)
**x:** koncentrace sledované látky v krvi (20 – 100 jednotek)

| interv | 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"

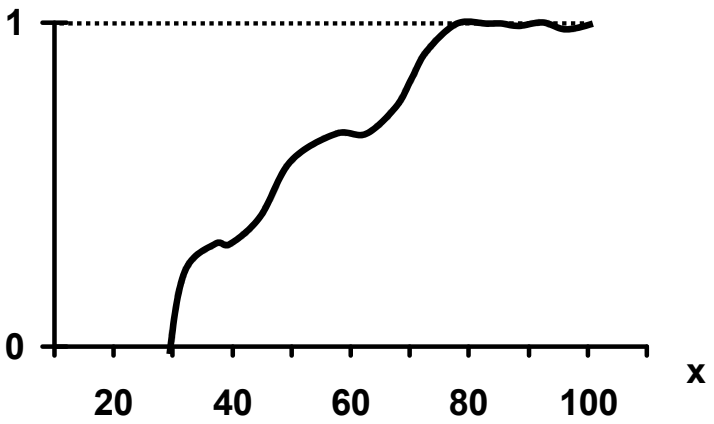
# Jak vznikají informace ?
# - frekvenční sumarizace spojitých dat

**Histogram**



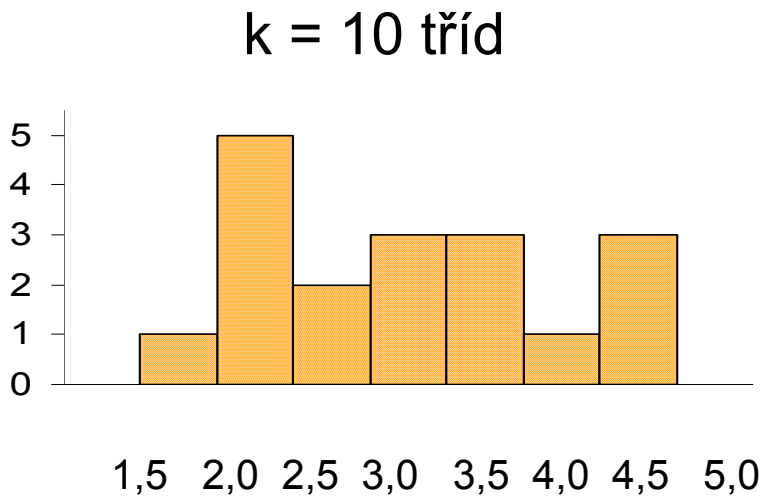

$$f(x) = \frac{n(I) / n}{d(I)} \Rightarrow$$ **Intervalová hustota četnosti**

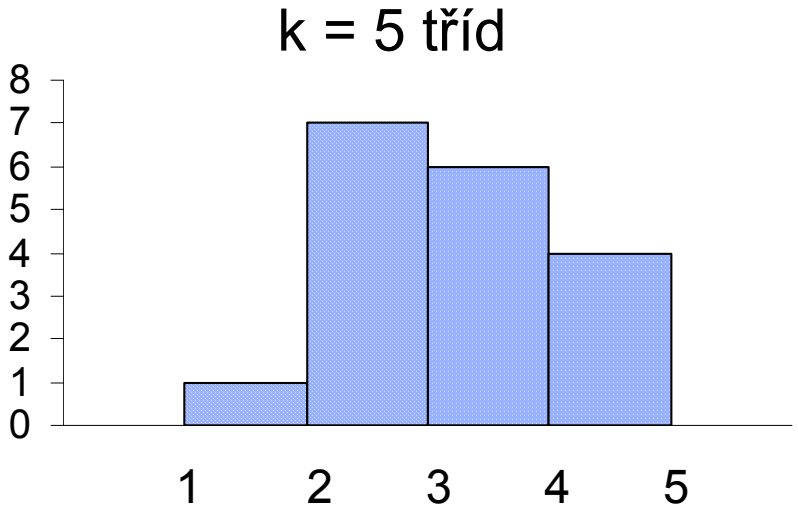
**Výběrová distribuční funkce**

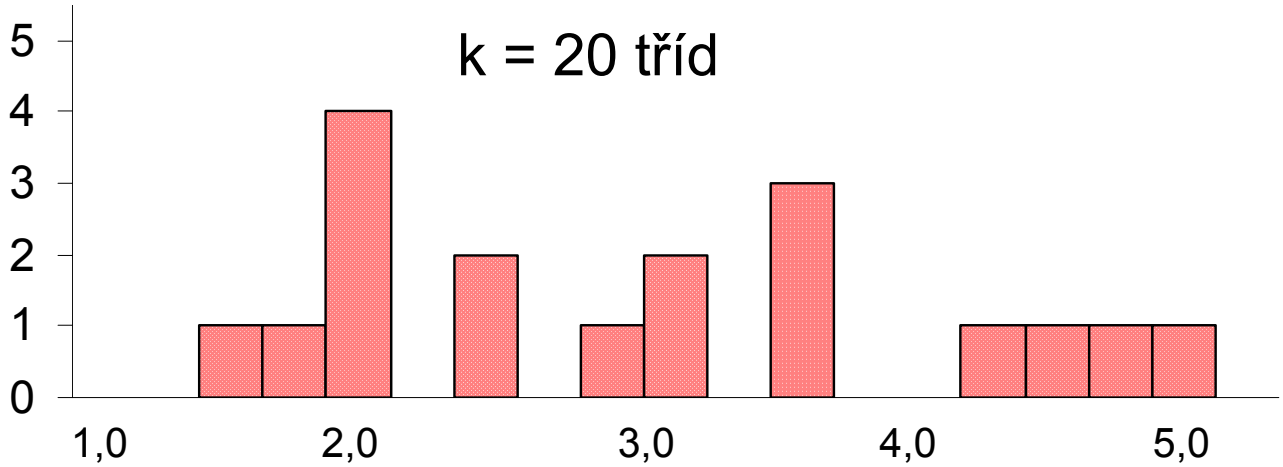



**F(x)** ⇒ **Intervalová relativní kumulativní četnost**

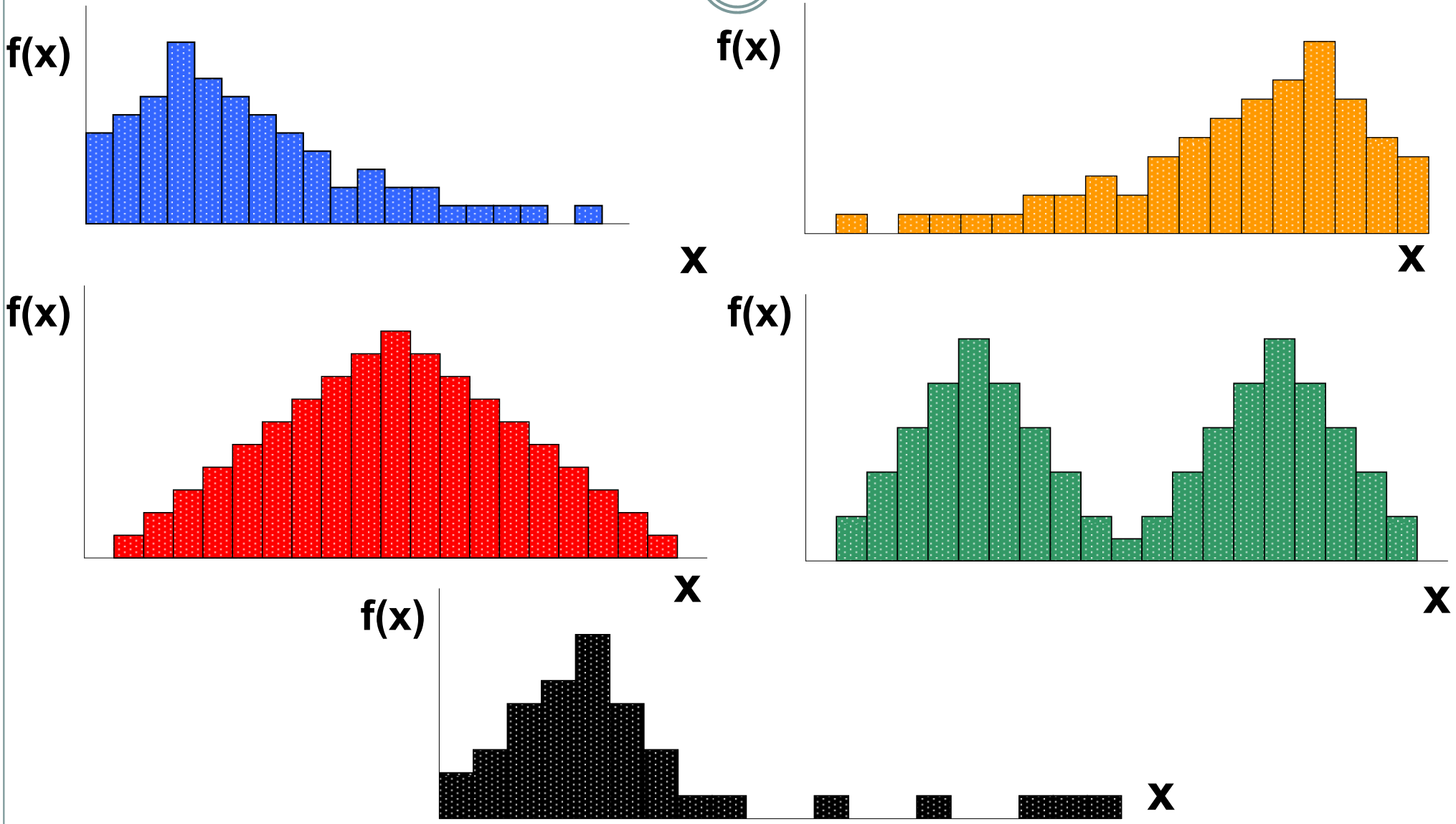
# Počet zvolených tříd a velikost souboru určují kvalitu výstupu



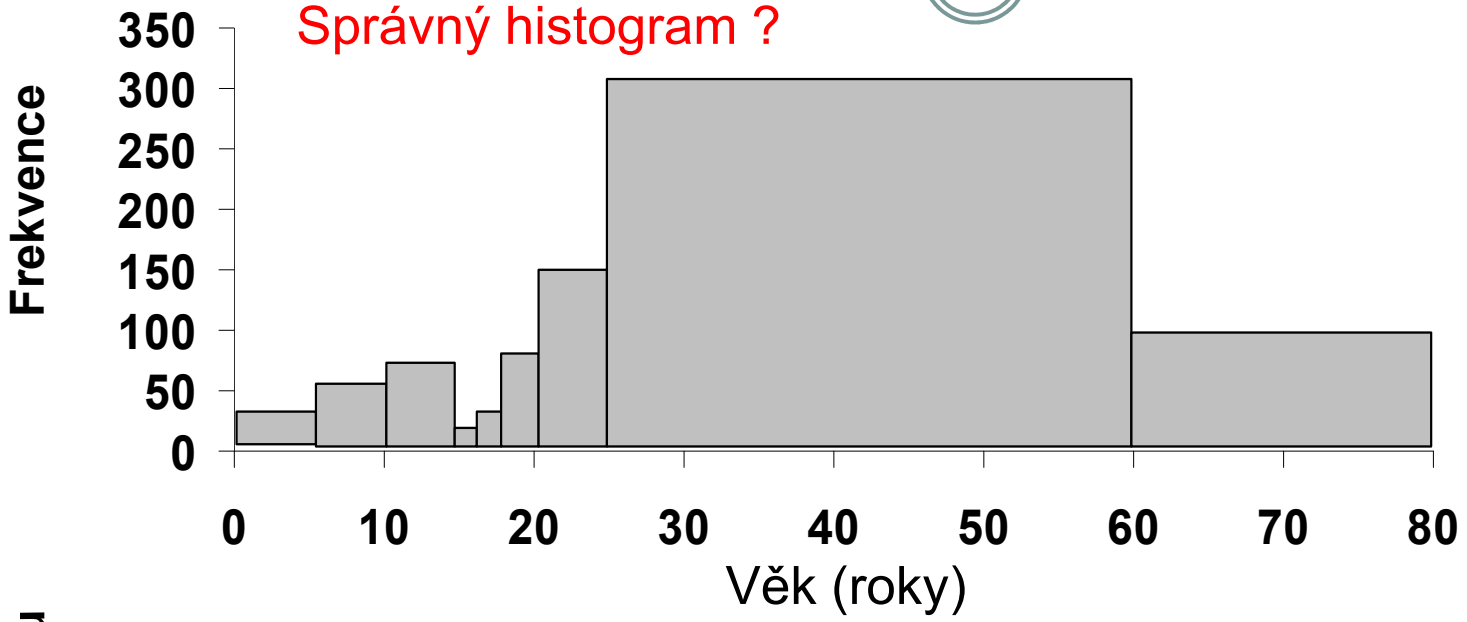



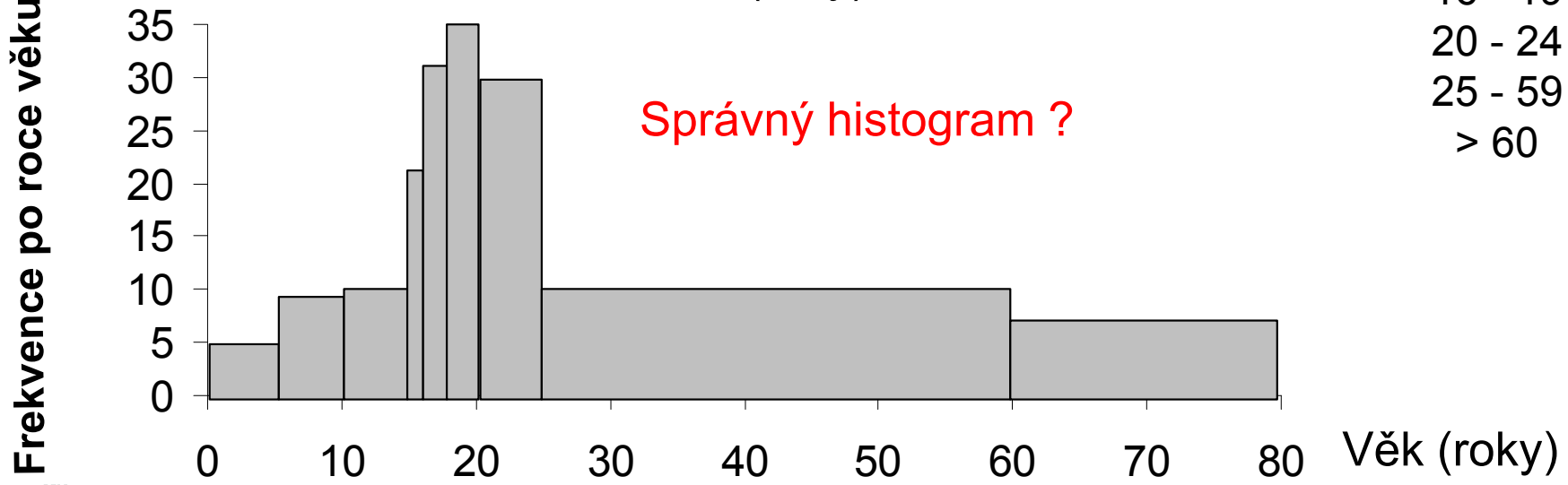
# Histogram vyjadřuje tvar výběrového rozložení



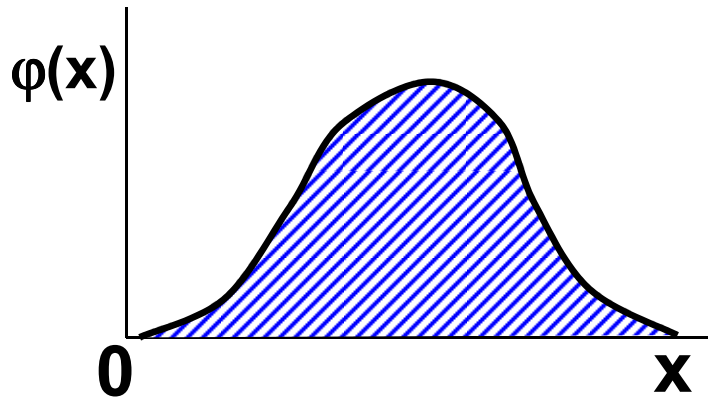
# Příklad: věk účastníků vážných dopravních nehod



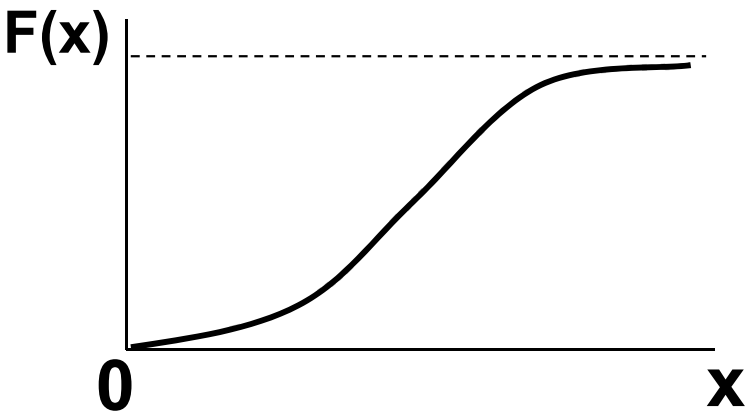

| Věk | f |
|---|---|
| 0 - 4 | 28 |
| 5 - 9 | 46 |
| 10 - 15 | 58 |
| 16 - 19 | 20 |
| 20 - 24 | 114 |
| 25 - 59 | 316 |
| > 60 | 103 |



# Pojem ROZLOŽENÍ - příklad spojitých dat



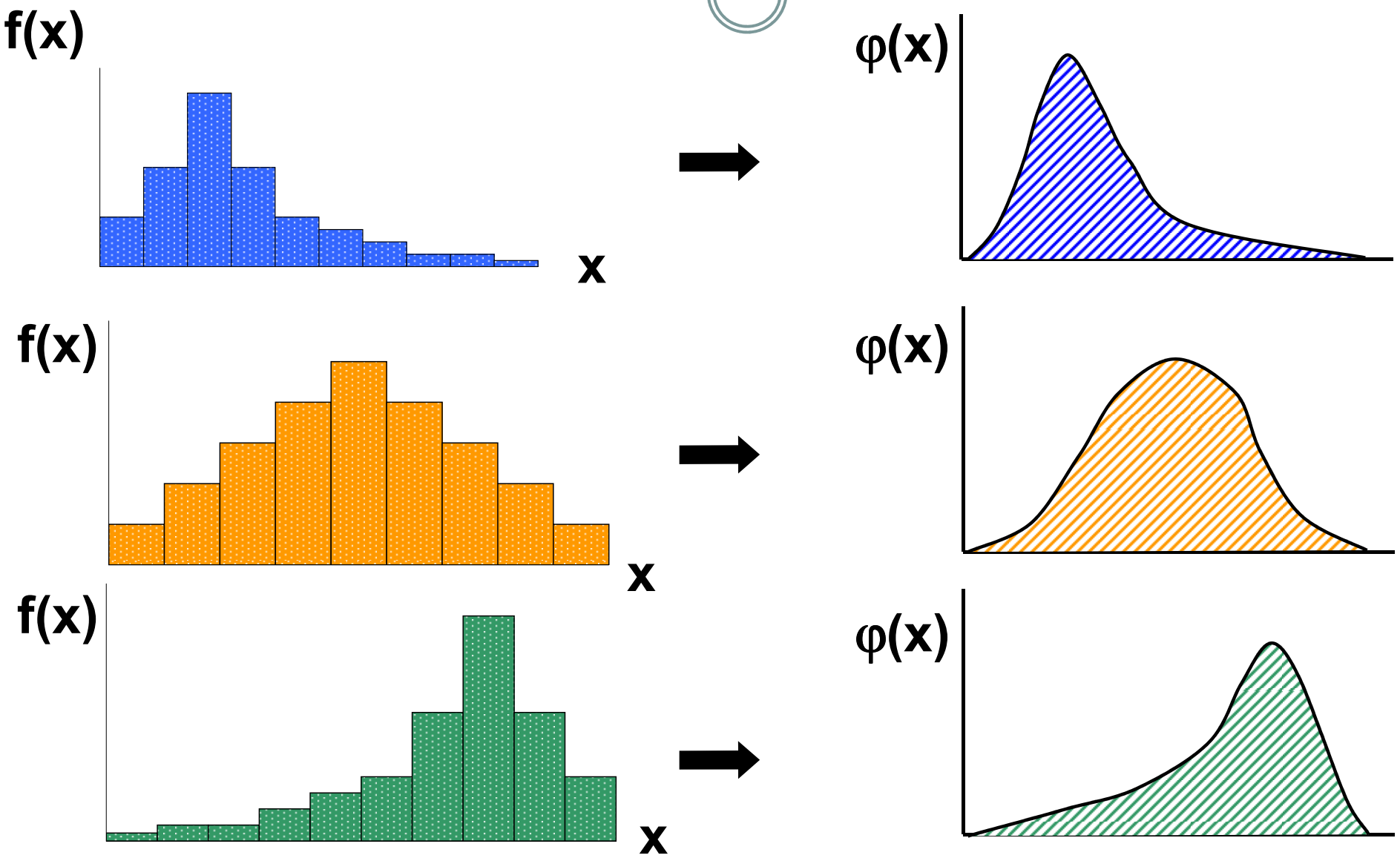

Rozložení




Distribuční funkce

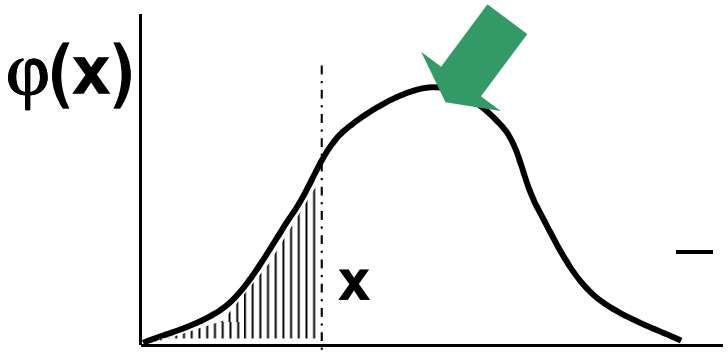
**Je - li dána distribuční funkce, je dáno rozložení**

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



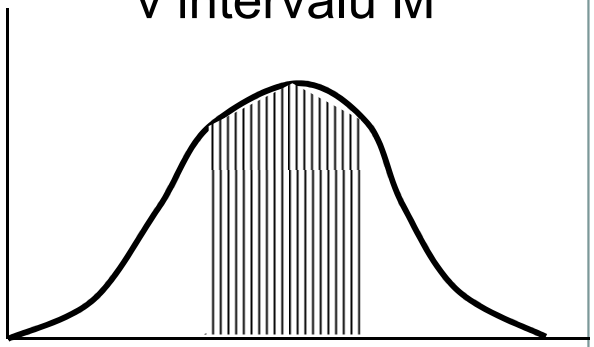
# Distribuční funkce jako užitečný nástroj pro práci s rozložením

Plocha = relativní četnost

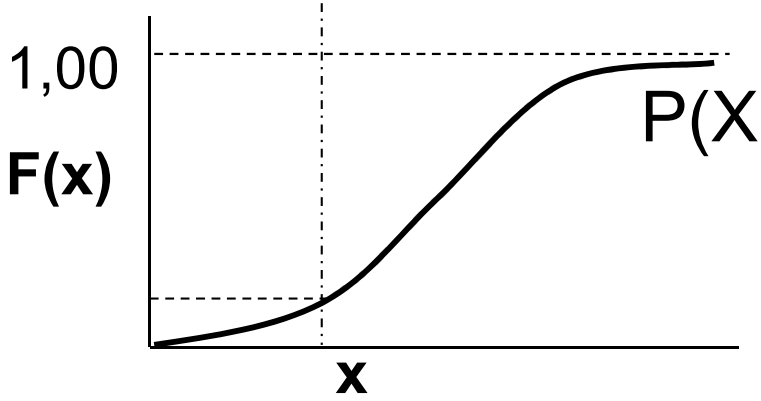



$$\int_{-\infty}^{\infty} \varphi(x)\ d(x) = 1$$

F(x):
Pravděpodobnost, že se X vyskytuje v intervalu M



M




$$P(X \leq x) = \Phi(x) = F(x'')$$

Φ(x) … distribuční funkce

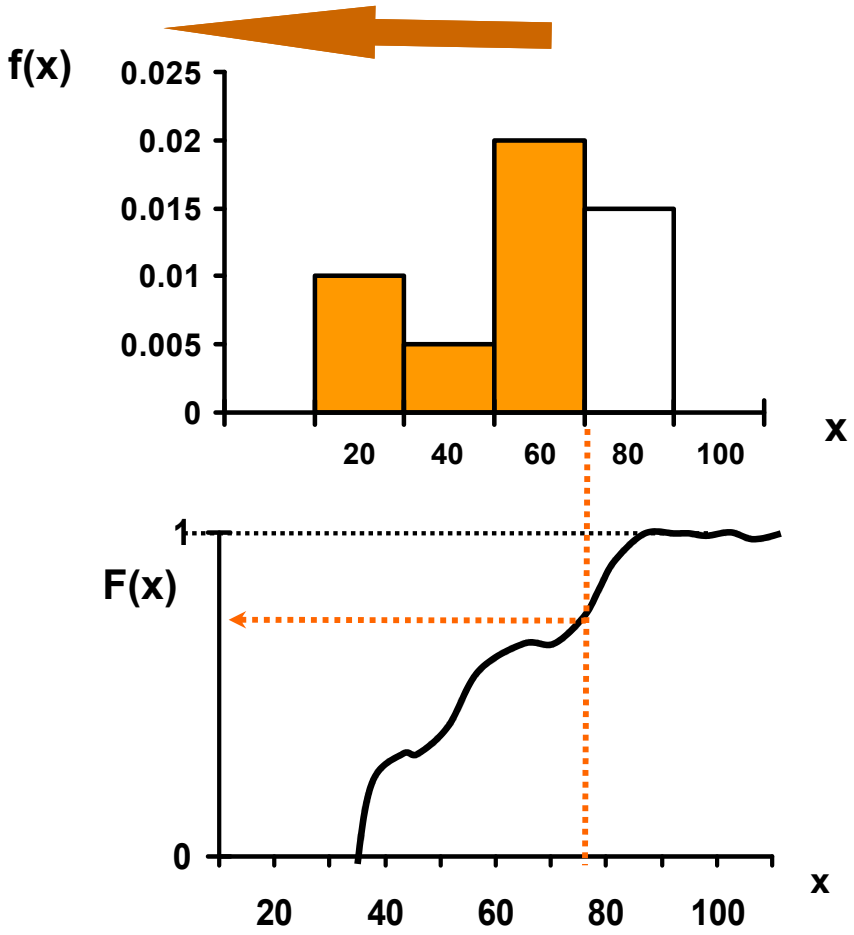
$$P(X \leq x) = \int_{M} \varphi(x)\ d(x)$$

Známe-li distribuční funkci, pak známe rozložení sledované veličiny.

Pro jakoukoli množinu hodnot (**M**) lze určit P, že X do této množiny patří.

# Jak vznikají informace ?
# - frekvenční sumarizace spojitých dat

## Grafické výstupy z frekvenční tabulky – spojitá data




**Uspořádání čísel podle velikosti a konstrukce rozložení umožňuje pravděpodobnostní zařazení každé jednotlivé hodnoty**

**KVANTIL**

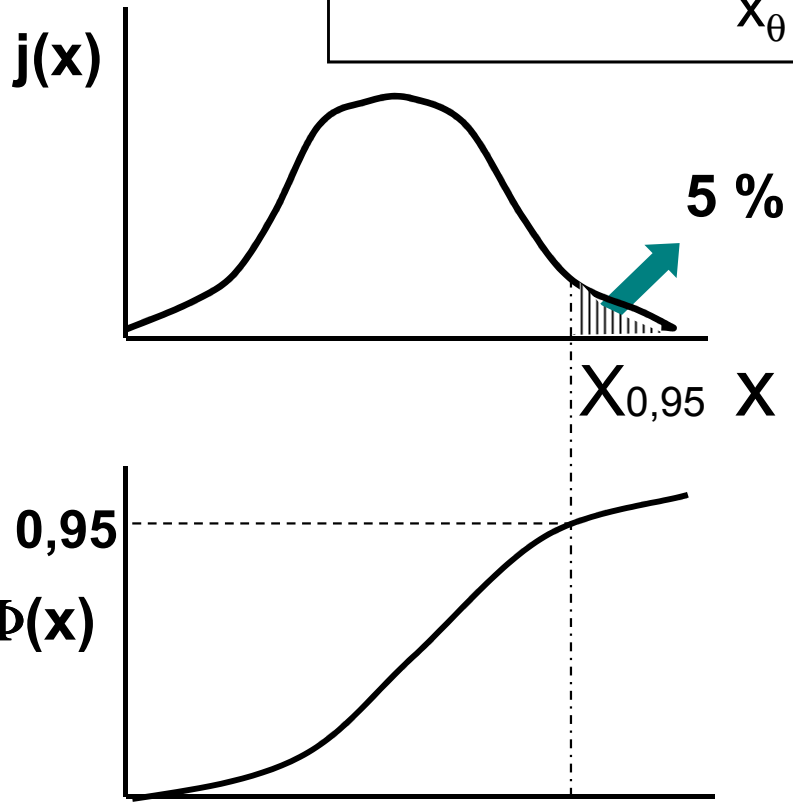
$$X_{0.1};\ X_{0.9};\ X_{0.5};\ X_{\theta}$$

# Otázka: Jak velké musí být X, aby 5 % všech hodnot bylo nad ním?

$\theta$ = 0,95 … Pravděpodobnost

**Hledáme:** $P(X \leq x_\theta) = 0{,}95 = \theta$

$x_\theta = (X_{0,95}) = ?$




$F(x\theta) = \theta$

**Kvantil** je číslo, jehož hodnota distribuční funkce je rovna P, pro kterou je kvantil definován

**Jakékoliv číslo na ose x je kvantilem**

# VI. Modelová rozložení

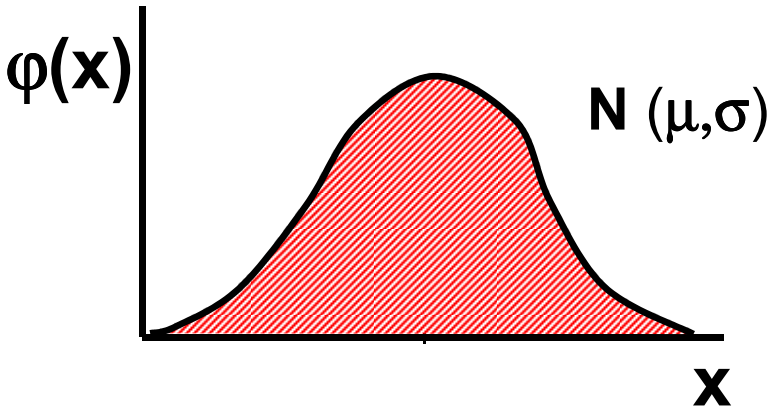
## Normální rozložení jako statistický model
## Aplikace modelových rozložení
## Přehled modelových rozložení

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

# Anotace

- Klasickým postupem statistické analýzy je na základě vzorku cílové populace identifikovat typ a charakteristiky modelového rozložení dat, využít jeho matematického modelu k popisu reality a získané výsledky zobecnit na hodnocenou cílovou populaci.
- Využití tohoto přístupu je možné pouze v případě shody reálných dat s modelovým rozložením, v opačném případě hrozí získání zavádějících výsledků.
- Nejklasičtějším modelovým rozložením, od něhož je odvozena celá řada statistických analýz je tzv. normální rozložení, známé též jako Gaussova křivka.

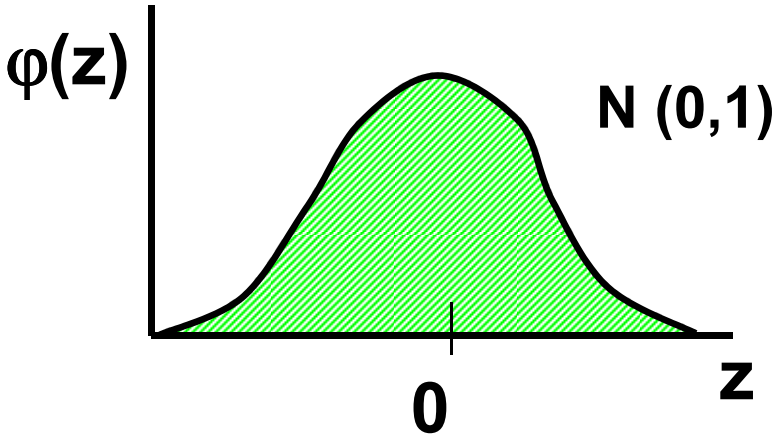
# Rozložení hodnot jako model: Normální rozložení




$$\varphi(x) = \frac{1}{\sigma.\sqrt{2\pi}}.e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$z = \frac{x - \mu}{\sigma}$$

Standardizovaná forma

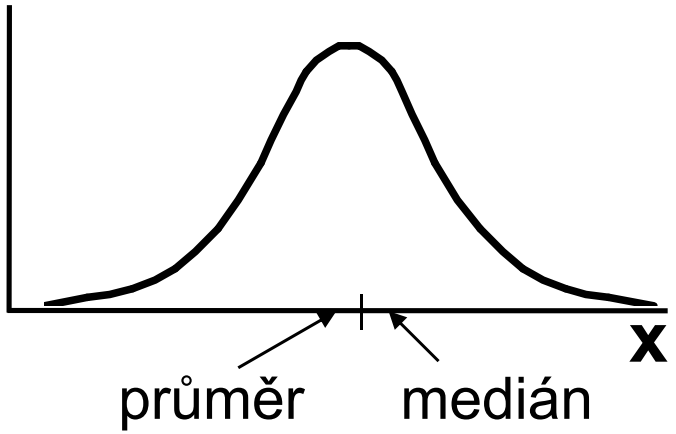



$$\varphi(z) = \frac{1}{\sqrt{2.\pi}}.e^{-\frac{z^2}{2}}$$

Tabelovaná podoba

# Parametry charakterizující normální rozložení a jejich význam

**E (x) ~ $\bar{x}$ ~ μ**
**D (x) ~ $s^2$ ~ $\sigma^2$**



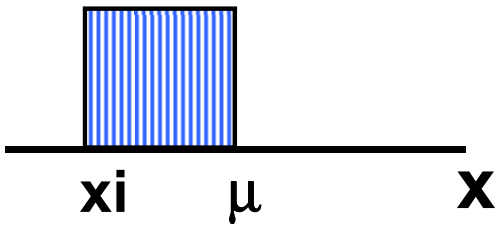

a) $\mu \sim \bar{x}$

**průměr - ukazatel středu**

b) $\sigma^2 \sim s^2$

**rozptyl**

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




c) **σ ~ s**

**směrodatná odchylka**

$$s = \sqrt{s^2}$$

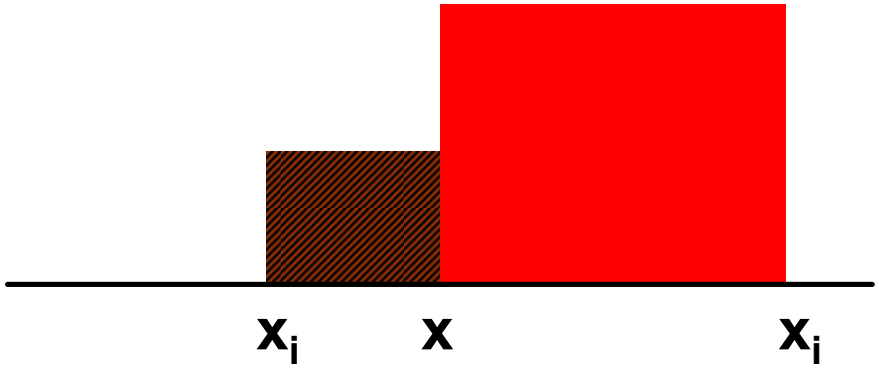
**Pravidlo ± 3s**

d) **koeficient variance**

$$c = s / \bar{x}$$

# Rozptyl není univerzálním ukazatelem variability

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





$\Rightarrow$ neúměrně zvýší $s^2$

# Normální rozložení jako model

## *I. Použitelnost modelu*

### A) X: spojitý znak - hmotnost jedince (myši)

**1,2; 1,4; 1,6; 1,8; 2,0; 2,4; 3.8**

n = 7 opakování

**medián** = 1,8

**průměr** = $\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{7}\sum_{i=1}^{7} x_i = \frac{1}{7}(1{,}2+1{,}4+1{,}6+1{,}8+2{,}0+2{,}4+3{,}8) = \frac{1}{7}14{,}2 = 2{,}03$

**rozptyl ($s^2$)** = $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{7}(x_i-2{,}03)^2}{6} = 0{,}766$

**sm. odchylka (s)** = $\sqrt{s^2} = \sqrt{0{,}766} = 0{,}875$

**?** **Je předpoklad normálního rozložení oprávněný ?**
**Jaký předpokládáte možný rozsah hodnot tohoto znaku ?** **?**

# Normální rozložení jako model

## *I. Použitelnost modelu*

### B) X: spojitý znak - hmotnost jedince (myši)

**1,2; 1,4; 1,6; 1,8; 2,0; 2,2; 2,4; 3,8; 8,9**

n = 9 opakování

medián = 2

průměr = $\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{9}\sum_{i=1}^{9} x_i = \frac{1}{9}(1{,}2+1{,}4+1{,}6+1{,}8+2{,}0+2{,}2+2{,}4+3{,}8+8{,}9) = \frac{1}{9}25{,}3 = 2{,}81$

**rozptyl ($s^2$) =** $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{9}(x_i-2{,}81)^2}{8} = 5{,}79$

**sm. odchylka (s) =** $\sqrt{s^2} = \sqrt{5{,}79} = 2{,}269$

**Jak hodnotíte model u těchto dat ?**

# Stochastické rozložení jako model

1. **Předpoklad: Znak x je rozložen podle daného modelu** ✓

2. **Znak x je naměřen o n hodnotách s modelovými parametry: $\bar{x}$ a s** → **Platnost modelu ?** ?

3. **Znak x je převeden na formu odpovídající tabulkovému standardu:** →

$$Z_i = \frac{x - \mu}{\sigma}$$

4. **Využije se tabelované (modelové) distribuční funkce pro testy o rozložení hodnot x**

# Normální rozložení jako model - příklad

## *Tabulky distribuční funkce*

- **<u>Data z průzkumu jsou publikována jako:</u>**

Kosti prehistorického zvířete:
**n** = 2000
**průměrná délka** = 60 cm
**sm. odchylka (s)** = 10 cm

✓ **Předpokládáme, že je oprávněný model normálního rozložení**

? Jaká je pravděpodobnost, že by velikost dané kosti překročila velikost 66 cm: P (x > 66) ? $Z = \frac{x - \mu}{\sigma}$

$P(x > 66) = 1 - P(x \leq 66)$ a platí, že $P(X \leq x) = F(X)$

tedy $P(x > 66) = 1 - P(x \leq 66) = 1 - P(\frac{x - m}{s} \leq \frac{66 - 60}{10}) = 1 - F(0{,}6) = 0{,}27425$

? Kolik kostí mělo zřejmě délku větší než 66 cm ? $P(x > 66) * n = 0{,}27425 * 2000 = 548$

? Jaký podíl kostí ležel svou délkou v rozsahu x od 60 cm do 66 cm ?

$P(60 < x < 66) = P\left(\frac{60 - 60}{10} < Z < \frac{66 - 60}{10}\right) = F(0{,}6) - F(0) = 0{,}22575$ ➡ 22,6% kostí leží v rozsahu 60-66cm

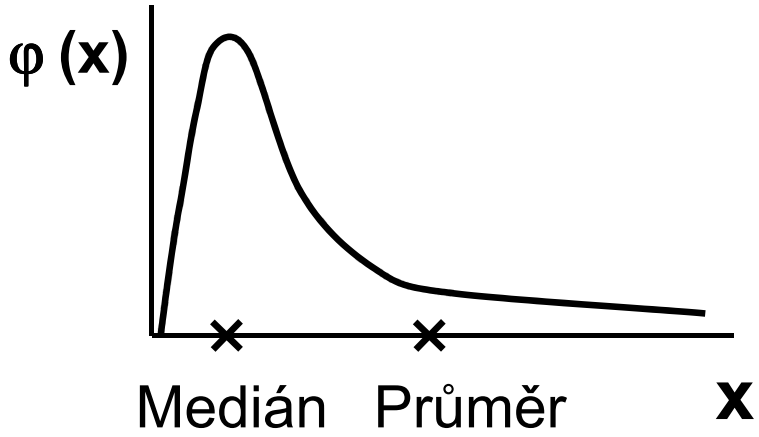
# Stručný přehled modelových rozložení I.

| Rozložení | Parametry | Stručný popis |
|---|---|---|
| **Normální** | **Průměr ($\mu$)**<br>**Rozptyl ($\sigma^2$)** | Symetrická funkce popisující intervalovou hustotu četnosti; nejpravděpodobnější jsou průměrné hodnoty znaku v populaci. |
| **Log-normální** | **Medián**<br>**Geometrický průměr**<br>**Rozptyl ($\sigma^2$)** | Funkce intervalové hustoty četnosti, která po logaritmické transformaci nabude tvaru normálního rozložení. |
| **Weibullovo** | **$\alpha$ - parametr tvaru**<br>**$\beta$ - parametr rozsahu hodnot** | Změnou parametru a lze modelovat distribuci doby přežití, např. stresovaného organismu. Rozložení využívané i jako model k odhahu $LC_{50}$ nebo $EC_{50}$ u testů toxicity. |
| **Rovnoměrné** | **Medián**<br>**Geometrický průměr**<br>**Rozptyl ($\sigma^2$)** | Funkce intervalové hustoty četnosti, která po logaritmické transformaci nabude tvaru normálního rozložení. |
| **Triangulární** | **$f(x) = [b - ABS(x - a)] / b^2$**<br>**$a - b < x < a + b$** | Pravděpodobnostní funkce pro typ rozložení, kdy jsou střední hodnoty výrazně pravděpodobnější než hodnoty okrajové. |
| **Gamma** | **Parametry distribuční funkce:**<br>**$\alpha$ - parametr tvaru**<br>**$\beta$ - parametr rozsahu hodnot** | Umožňuje flexibilně modelování distribučních funkcí nejrůznějších tvarů. Např. $\chi^2$ rozložení je rozložení typu Gamma. Gamma rozložení s a = 1 je známo jako exponenciální rozložení. |

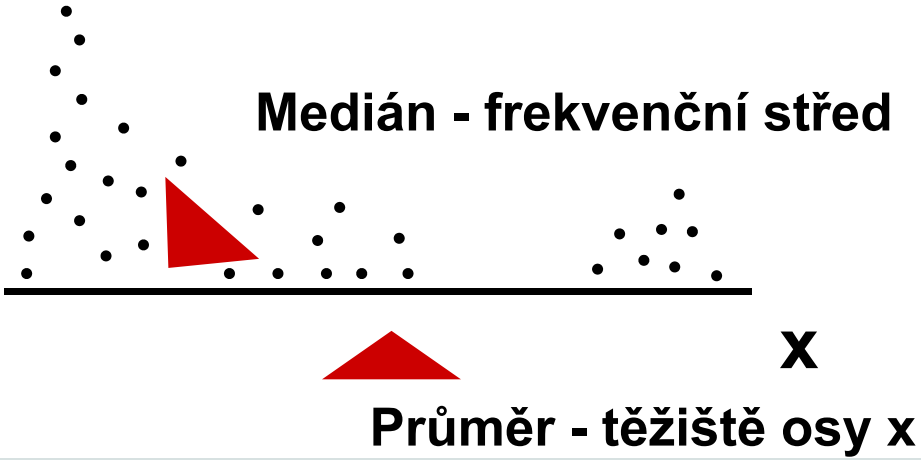
# Stručný přehled modelových rozložení II.

| Rozložení | Parametry | Stručný popis |
|---|---|---|
| **Beta** | **Parametry distribuční funkce:**<br>**$\alpha$ - parametr tvaru**<br>**$\beta$ - parametr rozsahu hodnot** | Pravděpodobnostní funkce pro proměnnou omezenou rozsahem do intervalu [0; 1]. Je matematicky komplikovanější, ale velmi flexibilní při popisu změn hodnot proměnné v ohraničeném intervalu. |
| **Studentovo** | **Stupně volnosti - uvažuje velikost vzorku**<br>**Průměr**<br>**Rozptyl** | Simuluje normální rozložení pro menší vzorky čísel. Pro větší soubory (n > 100) se limitně blíží k normálnímu rozložení. |
| **Pearsonovo** | **Stupně volnosti - uvažuje velikost vzorku** | Slouží především k porovnání četností jevů ve dvou a více kategoriích.<br>Používá se k modelování rozložení odhadu rozptylu normálně rozložených dat. |
| **Fisher-Snedecorovo** | **Dvojí stupně volnosti - uvažuje velikost dvou vzorků** | Používá se k testování hodnot průměrů - F test pro porovnání dvou výběrových rozptylů; F test, ANOVA atd. |

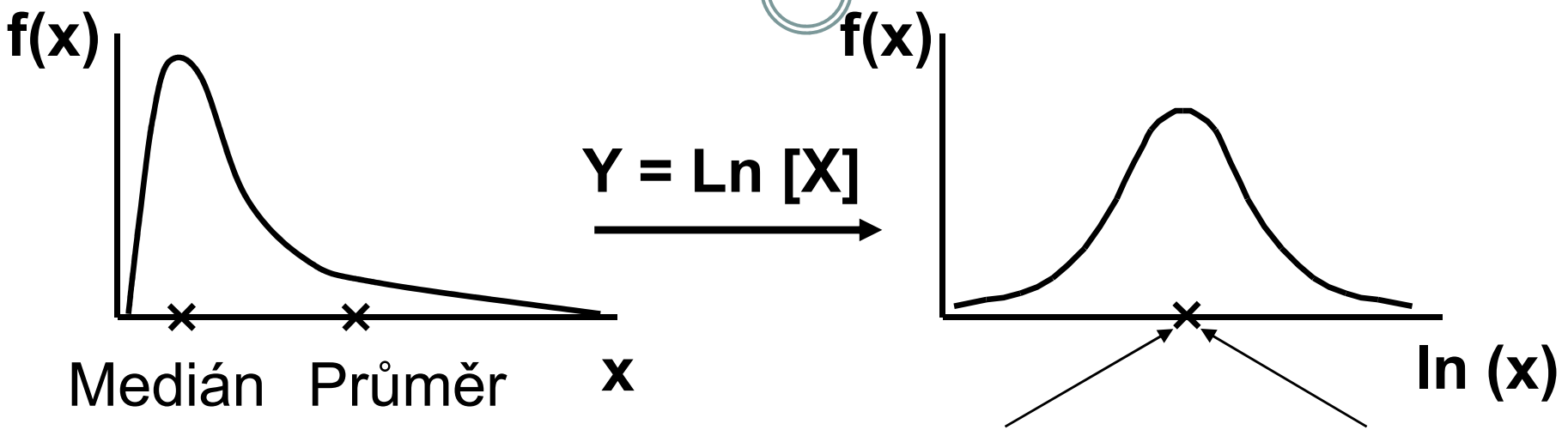
# Log-normální rozložení jako častý model reálných znaků




**U asymetrických rozložení je medián velmi vhodným alternativním ukazatelem středu**



# Log-normální rozložení lze jednoduše transformovat




**EXP (Y) = Geometrický průměr X**

$$\overline{Y} = \sum_{i=1}^{n} \frac{Y_i}{n}$$

**$\overline{Y}$ ± Standardní chyba**

# Transformace dat - legitimní úprava rozložení

✓ **Základní typy transformací vedou k normalitě rozložení nebo k homogenitě rozptylu**

## Logaritmická transformace

Logaritmická transformace je velmi vhodná pro data s odlehlými hodnotami na horní hranici rozsahu. Při porovnání průměrů u více souborů dat je pro tuto transformaci indikující situace, kdy se s rostoucím průměrem mění proporcionálně i směrodatná odchylka, a tedy jednotlivé proměnné mají stejný koeficient variance, ačkoli mají různý průměr.

Za takovéto situace přináší logaritmická transformace nejen zeslabení asymetrie původního rozložení, ale také vyšší homogenitu rozptylu proměnných. Pro transformaci se nejčastěji používá přirozený logaritmus a pokud jsou v původním souboru dat nulové hodnoty, je vhodné použít operaci **Y = ln (X+1)**.

Je-li průměr logaritmovaných dat (tedy průměrný logaritmus) zpětně transformován do původních hodnot, výsledkem není aritmetický, ale geometrický průměr původních dat.

# Transformace dat - legitimní úprava rozložení

✓ **Základní typy transformací vedou k normalitě rozložení nebo k homogenitě rozptylu**

## Odmocninová transformace

Transformace je vhodná pro proměnné mající Poissonovo rozložení, tedy proměnné vyjadřující celkový počet nastání určitého jevu (spíše vzácného) v **n** nezávisle opakovaných pokusech. Obecněji lze tento typ transformace doporučit v případě normalizace dat typu počtu jedinců (buněk, apod.). Jde o transformaci:

$$Y=\sqrt{x} \quad \textbf{nebo} \quad Y=\sqrt{x+1} \quad \textbf{nebo} \quad Y=\sqrt{x}+\sqrt{x+1}$$

Transformace s přičtenou hodnotou 1 jsou efektivní, pokud **X** nabývá velmi malých nebo nulových hodnot. Situace indikující vhodnost odmocninové transformace je také proporcionalita výběrového rozptylu a průměru, tedy obecně jestliže $\mathbf{s^2_x = k}$ (výběrový průměr).

# Transformace dat - legitimní úprava rozložení

## Arcsin transformace

Tzv. **úhlová transformace** - velmi vhodná pro data typu podílů výskytu určitého jevu (znaku) mezi **n** hodnocenými jedinci - tedy pro data mající binomické rozložení. Pokud se určitý znak vyskytuje r-krát mezi **n** možnostmi (jedinci, opakováními), pak lze vyjádřit relativní četnost jeho výskytu jako **p = r/n** s variabilitou **p.(1-p)/n**. Arcsin transformace odstraní ze souborů dat podíly blízké 0 nebo 1, a tak efektivně sníží variabilitu odhadů středu. Transformace však není schopná odstranit variabilitu vyvolanou rozdílným počtem opakování v jednotlivých variantách - v takovém případě lze doporučit provedení vážených transformací dat. Velmi častou formou této transformace je:

$$Y = \arcsin \sqrt{p}$$

- tedy transformace podílů do hodnot, jejichž sinus je roven druhé odmocnině původních hodnot. Pokud celkový počet jedinců (opakování), mezi kterými je výskyt znaku monitorován, je n < 50, pak lze doporučit velmi efektivní empirická opatření pro transformaci podílů blízkých 0 nebo 1. Pro tento případ lze nahrazovat nulové podíly hodnotou 1/4n a 100 % podíly hodnotou (n-1/4)/n. Pokud se mezi hodnotami vyskytuje větší množství krajních hodnot (menší než 0,2 a větší než 0,8), lze doporučit transformaci:

$$Y = \frac{1}{2}\left[\arcsin \sqrt{\frac{x}{n+1}} + \arcsin \sqrt{\frac{x+1}{n+1}}\right]$$