

Modelová rozložení



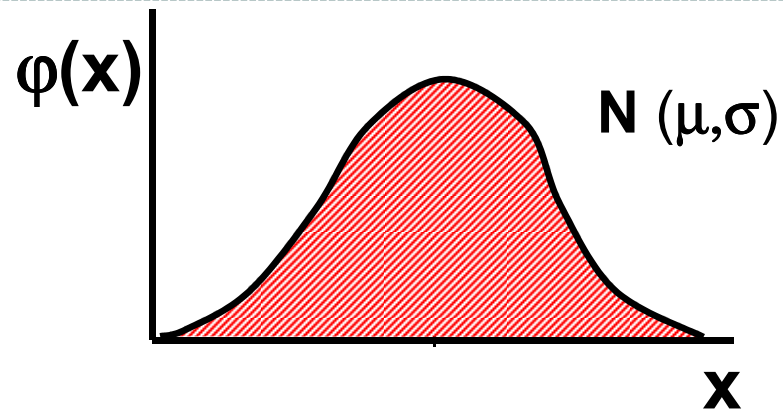
Normální rozložení jako statistický model
Aplikace modelových rozložení
Přehled modelových rozložení

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.

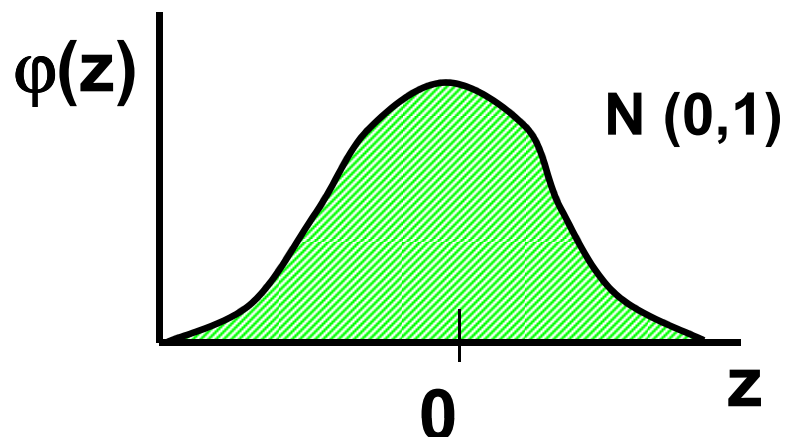
Rozložení hodnot jako model: Normální rozložení



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

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

Standardizovaná forma



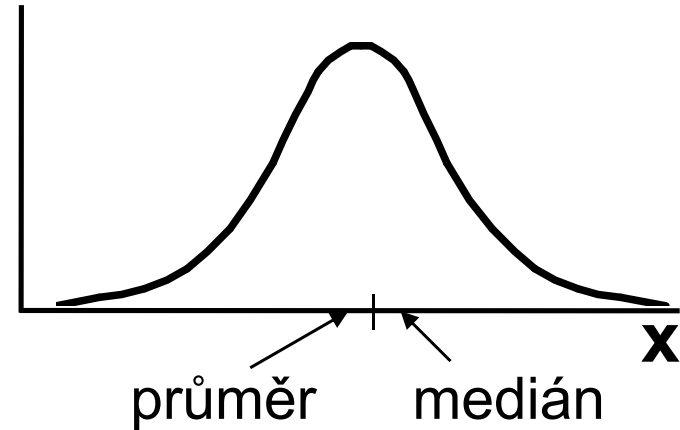
$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}$$

Tabelovaná podoba

Parametry charakterizující normální rozložení a jejich význam

$$E(x) \sim \bar{x} \sim \mu$$
$$D(x) \sim s^2 \sim \sigma^2$$

$\varphi(x)$



a)

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

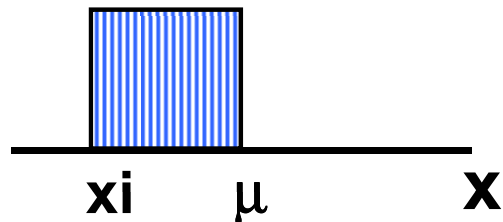
průměr - ukazatel středu

b)

$$\sigma^2 \sim s^2$$

rozptyl

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



c)

$$\sigma \sim s$$

směrodatná odchylka

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

Pravidlo $\pm 3s$

d)

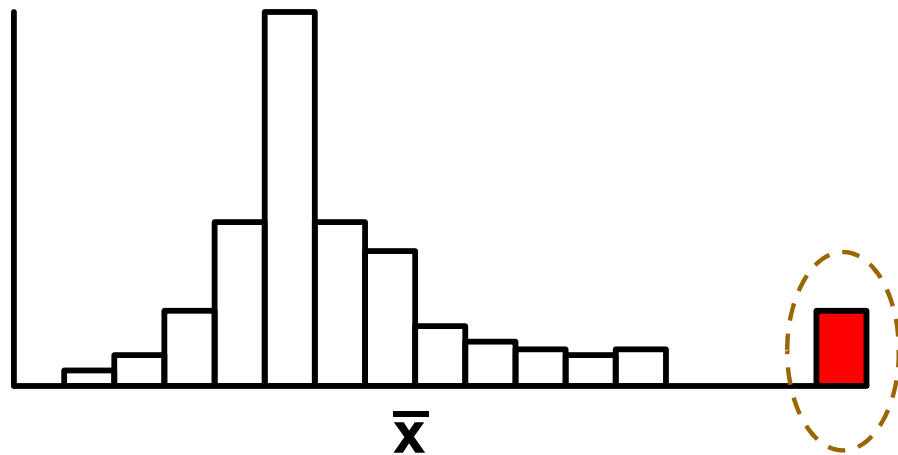
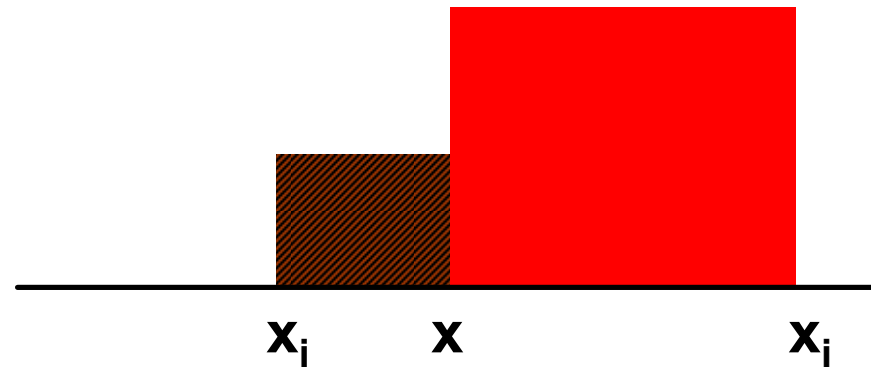
koeficient variance

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

Rozptyl není univerzálním ukazatelem variability



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



⇒ 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

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

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

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



**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

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

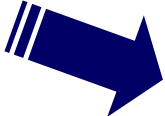

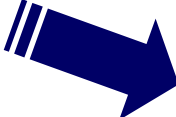
$$\text{rozptyl (s}^2\text{)} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^9 (x_i - 2,81)^2}{8} = 5,79$$

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

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

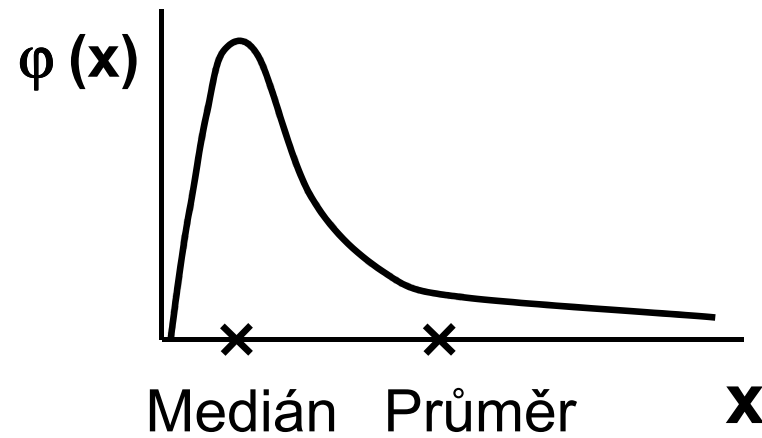
Stručný přehled modelových rozložení I.

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

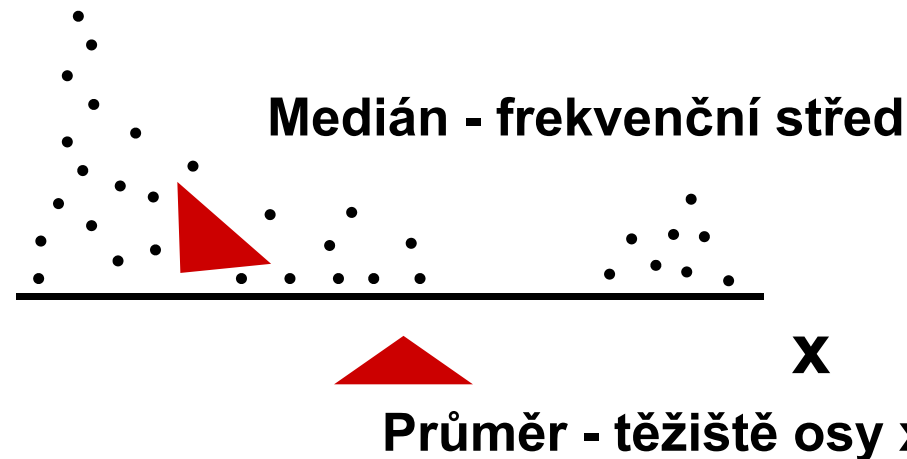
Stručný přehled modelových rozložení II.

Rozložení	Parametry	Stručný popis
Beta	Parametry distribuční funkce: α - parametr tvaru β - 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 Průměr 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. 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.

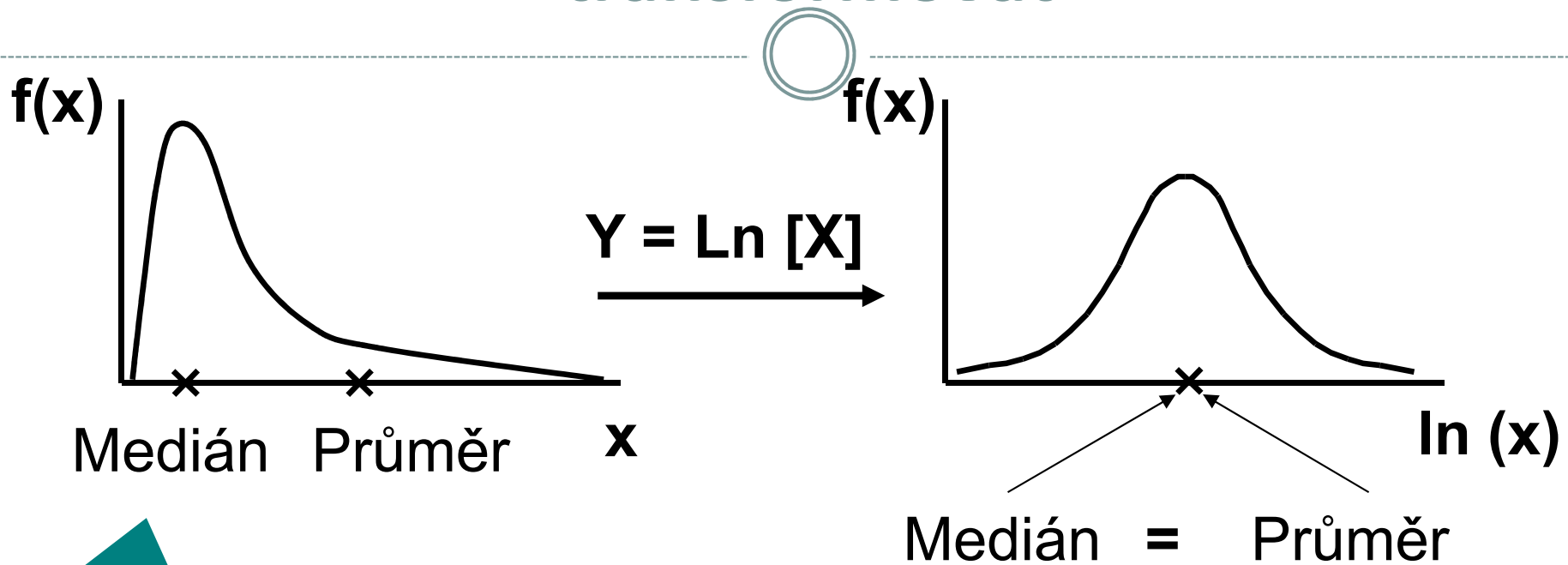
Log-normální rozložení jako častý model reálných znaků



U asymetrických rozložení je medián velmi vhodným alternativním ukazatelem středu



Log-normální rozložení lze jednoduše transformovat



$\text{EXP}(Y) = \text{Geometrický průměr } X$

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

$\bar{Y} \pm \text{Standardní chyba}$

Opakování- Popisná statistika dat



Popisné statistiky dat
Vizualizace dat

Anotace



- Popisná analýza dat je po vizualizaci dat dalším krokem v procesu statistického hodnocení. Poskytuje představu o rozsazích hodnocených dat a umožňuje vyhodnotit, srovnání s literárními údaji nebo dosavadní zkušeností, jejich realističnost.
- Již při výběru vhodné popisné statistiky se uplatňuje znalost rozložení dat. Některé popisné statistiky, odvozené od modelových rozložení, je možné využít pouze v případě, že data mají dané modelové rozložení. Typickým příkladem je průměr a směrodatná odchylka, jejichž předpokladem je přítomnost normálního rozložení.

Typy proměnných



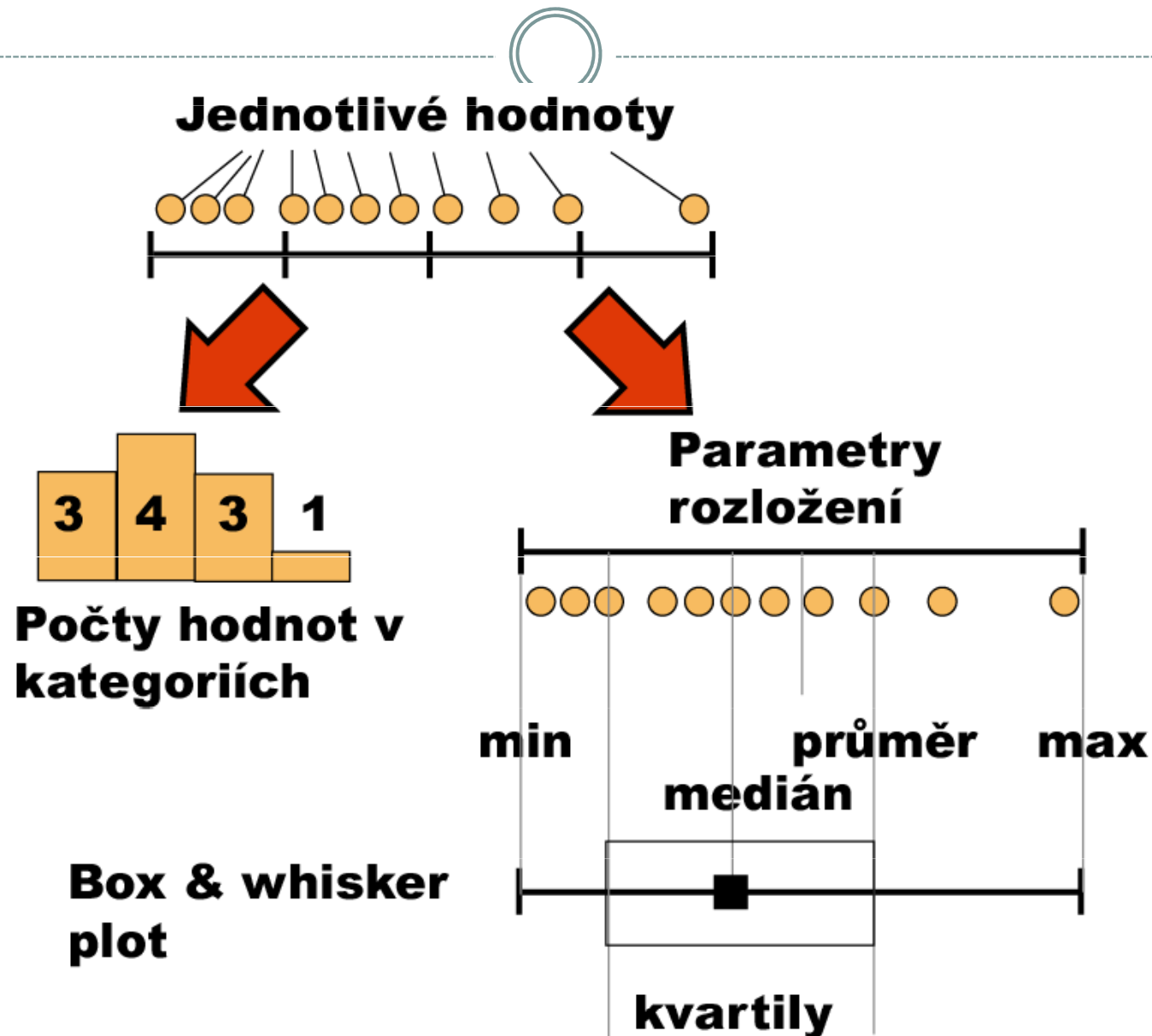
- **Kvalitativní/kategorická**

- binární - ano/ne
- nominální - A,B,C ... několik kategorií
- ordinální- $1 < 2 < 3$...několik kategorií a můžeme se ptát, která je větší

- **Kvantitativní**

- nespojitá – čísla, která však nemohou nabývat všech hodnot (např. počet porodů)
- spojitá – teoreticky jsou možné všechny hodnoty (např. krevní tlak)

Řada dat a její vlastnosti



Frekvenční rozložení



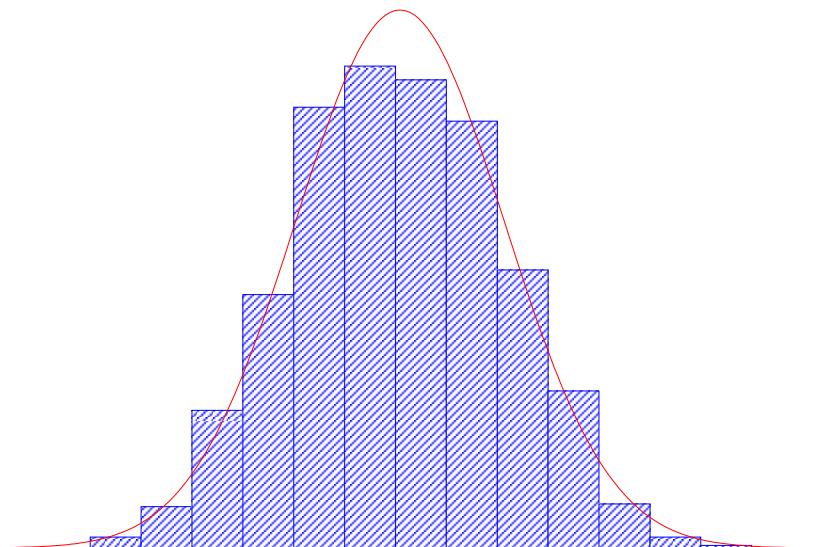
Kategorie	Četnost
B	5
C	8
D	1

Kvalitativní data

Tabulka s četností jednotlivých kategorií.

Kvantitativní data

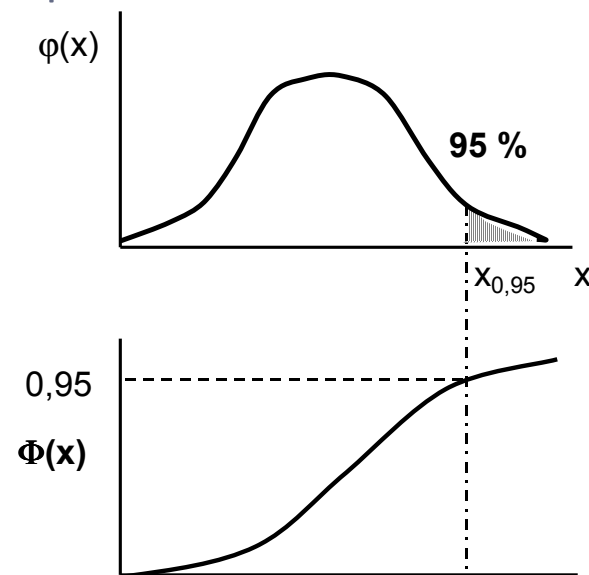
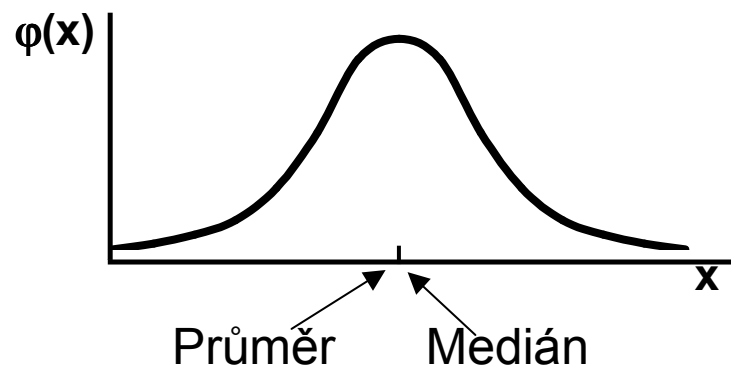
Četnost hodnot rozložení v jednotlivých intervalech.



Parametry rozložení



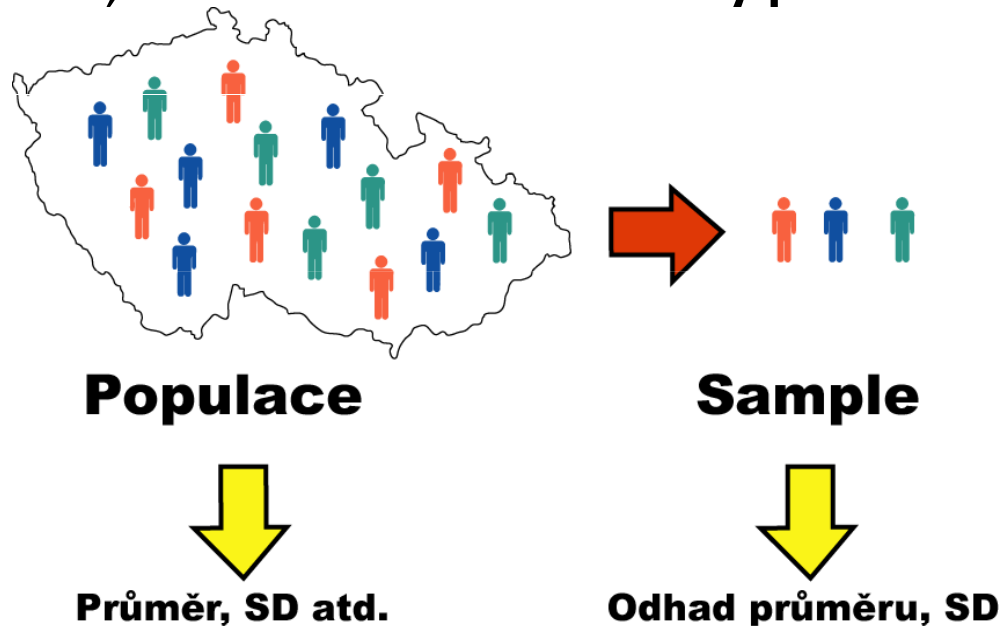
- Soubor dat (řada čísel) můžeme charakterizovat parametry jeho rozložení
- Hlavní skupiny těchto parametrů můžeme charakterizovat jako ukazatele:
 - Středu (medián, průměr, geometrický průměr)
 - Šířky rozložení (rozsah hodnot, rozptyl, směrodatná odchylka)
 - Tvaru rozložení (skewness, kurtosis)
 - Kvantily rozložení – kolik % řady dat leží nad a pod kvantilem



Populace a vzorek



- Populace představuje veškeré možné objekty vzorkování, např. veškeré obyvatelstvo ČR při sledování na úrovni ČR, z populace získáme reálné parametry rozložení
- Z populace je prováděno vzorkování za účelem získání reprezentativního vzorku (**sample**) populace, toto vzorkování by mělo být náhodné, důležitá je také velikost vzorku, ze vzorku získáme **odhady parametrů rozložení**



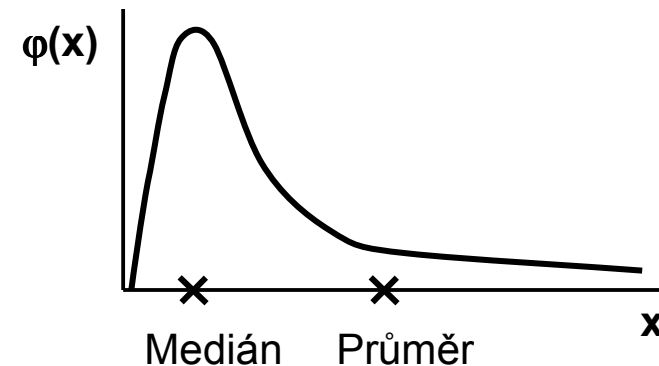
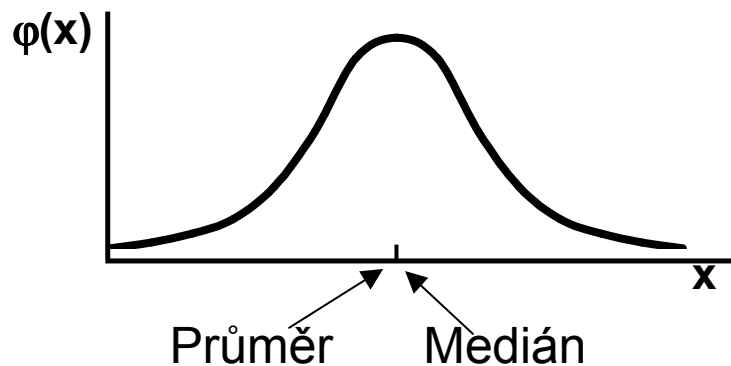
Ukazatele středu rozložení I



- **Průměr** – vhodný ukazatel středu u normálního/symetrického rozložení, kde x_i jsou jednotlivé hodnoty a n jejich počet

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

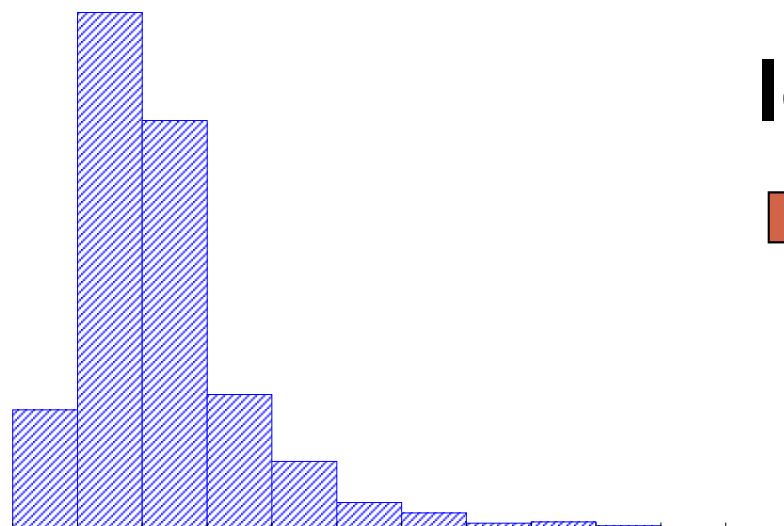
- **Medián** – jde vlastně o 50% kvantil, tj. polovina hodnot leží nad a polovina pod mediánem
- V případě symetrického rozložení jsou jejich hodnoty v podstatě shodné



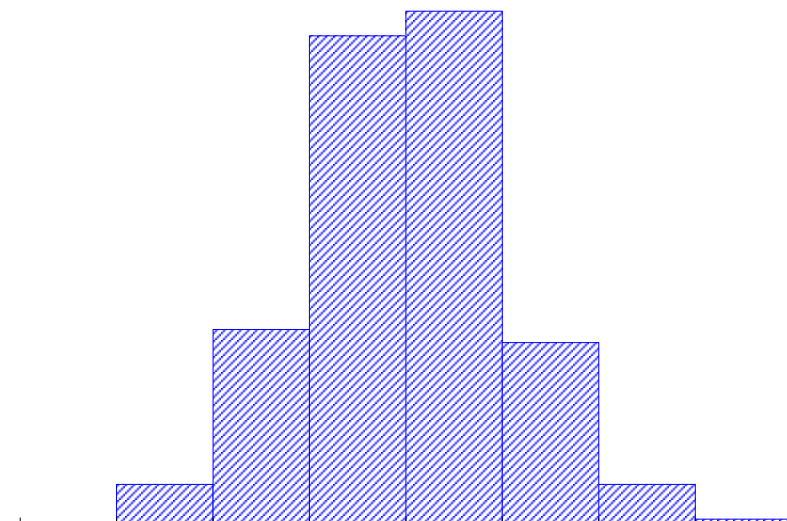
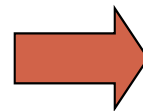
Ukazatele středu rozložení II.



- Geometrický průměr – antilogaritmus průměru logaritmovaných dat, je vhodný pro doleva asymetrická data (lognormální rozložení), která jsou v biologii velmi častá, jeho hodnota v podstatě odpovídá mediánu
- Takto asymetrická data je možné převést logaritmickou transformací na normální rozložení



log



Medián, geometrický průměr

Průměr (logaritmovaných dat)

Ukazatele šířky rozložení



- **Rozptyl** 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** je druhá odmocnina z rozptylu
- **Koeficient variance** - podíl SD ku průměru (u normálního rozložení by se 95% hodnot mělo vejít do průměr ± 3 SD), pokud je SD větší než 1/3 průměru jsou teoreticky pravděpodobné záporné hodnoty v rozložení – ukazatel problémů s normalitou dat