# V. Průzkumová analýza dát

# Motivácia

- Pri spracovaní dát sa často používajú metódy, ktoré sú založené na predpoklade, že dáta pochádzajú z nejakého konkrétneho rozloženia.
- Najčastejšie sa predpokladá normálne rozloženie.
- Prečo to nemusí platiť:
  - Dáta pochádzajú z iného rozloženia.
  - Sú zaťažené chybami.
  - Pochádzajú z niekoľkých rôznych rozložení.

# Základné pojmy

- Dátový súbor – dáta.
- Prípad – pozorovaná  jednotka (napr. pacient), predstavuje jeden riadok v dátovom súbore.
- Znaky = premenné – pozorované vlastnosti prípadu (napr. výška, váha, farba očí).
- Náhodný výber – postupnosť nezávislých rovnako rozložených veličín (prípadov). Keď niekomu dávame dotazník, nevieme vopred ako odpovie.
- Usporiadaný náhodný výber – dátový súbor usporiadaný podľa nejakého znaku.

# Frekvenčná tabuľka alebo tabuľka rozloženia četností I.

- Bodové rozloženie četností:
  - Máme malý počet variant, jednotlivým variantám priraďujeme ich četnosti.
  - n – počet všetkých prípadov

| Varianta | Absolútne četnosti | Relatívna četnosť | Absolútna kumulatívna četnosť | Relatívna kumulatívna četnosť |
|---|---|---|---|---|
| Varianta j $x_j$ | $n_j$ | $p_j$ | $N_j$ | $F_j$ |
| | | $p_j=n_j/n$ | $N_j = n_1+n_2+...+n_j$ | $F_j = N_j /n=$ $p_1+p_2+...+p_j$ |

# Funkcie

- Empirická distribučná funkcia
  - zobrazuje relatívne kumulatívne četnosti
  - končí vždy v 1
- Četnostná funkcia
  - $p(x) = p_j$ ak je x jednou z variant
  - $= 0$ ak x nie je jednou z variant
  - zobrazuje relatívne četnosti

# Grafy

- Graf četností funkcie
  - osa x: možnosti, osa y: četnosti
  - sú zobrazené len body
- Graf empirickej distribučnej funkcie
- Stĺpcový diagram
  - osa x: možnosti, osa y: počet pozorovaní
- Polygon četností
  - osa x: možnosti, osa y: počet pozorovaní
  - spojené čiarou

# Príklad

- U 30 domácností bol zisťovaný počet členov rodiny

| Počet členov | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Počet domácností | 2 | 6 | 4 | 10 | 5 | 3 |

- Vytvorte tabuľku rozloženia četností.
- Nakreslite graf četností, stĺpcový graf a polygon četností.

# Príklad tabuľka rozloženia četností

| $x_j$ | $n_j$ | $p_j$ | $N_j$ | $F_j$ |
|---|---|---|---|---|
| 1 | 2 | 2/30 | 2 | 2/30 |
| 2 | 6 | 6/30 | 8 | 8/30 |
| 3 | 4 | 4/30 | 12 | 12/30 |
| 4 | 10 | 10/30 | 22 | 22/30 |
| 5 | 5 | 5/30 | 27 | 27/30 |
| 6 | 3 | 3/30 | 30 | 30/30=1 |

# Frekvenčná tabuľka alebo tabuľka rozloženia četností II.

- Intervalové rozloženie
  - Veľký počet variant, ktoré rozdelíme do intervalov
  - Určujeme četnosti v jednotlivých intervaloch
  - Určenie počtu intervalov je subjektívne
  - Často sa odporúča ako odmocnina z n (n=počet všetkých prípadov)

# Frekvenčná tabuľka

| Interval | $n_j$ | $p_j$ | $f_j$ | $N_j$ | $F_j$ |
|---|---|---|---|---|---|
| | počet | $n_j / n$ | $p_j / d_j$ | $n_1 + n_2 + \ldots + n_j$ | $p_1 + p_2 + \ldots + p_j$ |
| | | | $d_j$ – šírka intervalu | | |
| | | | intervalová hustota četností | | intervalová empirická distribučná funkcia |

# Grafy

- Histogram
  - osa x: intervaly, osa y: hodnota četnostnej funkcie
  - pomer obsahov  stĺpikov odpovedá pomeru zastúpenia jednotlivých intervalov v dátach
- Intervalová empirická distribučná funkcia
  - osa x: intervaly, osa y: hodnoty intervalovej empirickej funkcie
  - vždy sa vynesú nad koniec intervalu a spoja sa priamkou

# Príklad

- V 70 domácnostiach boli zisťované týždenné výdaje na sladkosti.

| výdaje | (36,65> | (65,95> | (95,125> | (125,155> | (155, 185> | (185, 200> |
|---|---|---|---|---|---|---|
| Počet domácností | 7 | 16 | 27 | 14 | 4 | 2 |

- Napíšte tabuľku rozloženia četností a nakreslite histogram a graf intervalovej empirickej distribučnej funkcie.

# Príklad tabuľka rozloženia četností

| Interval | $n_j$ | $p_j$ | $f_j$ | $N_j$ | $F_j$ |
|---|---|---|---|---|---|
| (35,65> | 7 | 7/70 | 7/2100 | 2 | 7/70 |
| (65,95> | 16 | 16/70 | 16/2100 | 23 | 23/70 |
| (95,125> | 27 | 27/70 | 27/2100 | 50 | 50/70 |
| (125,155> | 14 | 14/70 | 14/2100 | 64 | 64/70 |
| (155,185> | 4 | 4/70 | 4/2100 | 68 | 68/70 |
| 185,215 | 2 | 2/70 | 2/2100 | 70 | 70/70=1 |

# Číselné charakteristiky dátového súboru
# Nominálne znaky

- Modus – najčastejšia varianta

# Číselné charakteristiky dátového súboru
# Ordinálne znaky

- Vieme ich usporiadať
- Alfa – kvantil = $x_{alfa}$ je číslo, ktoré rozdeľuje usporiadaný súbor na dolný úsek, ktorý obsahuje podiel aspoň alfa všetkých dát a na horný úsek, ktorý obsahuje podiel aspoň 1-alfa všetkých dát.
- Alfa- číslo
- Medián: $x_{0,50}$
- $x_{0,25}$ = dolný kvartil, $x_{0,75}$ = horný kvartil
- $x_{0,1}$ ,…, $x_{0,9}$ = decily
- $x_{0,01}$ ,…, $x_{0,99}$ = percentily

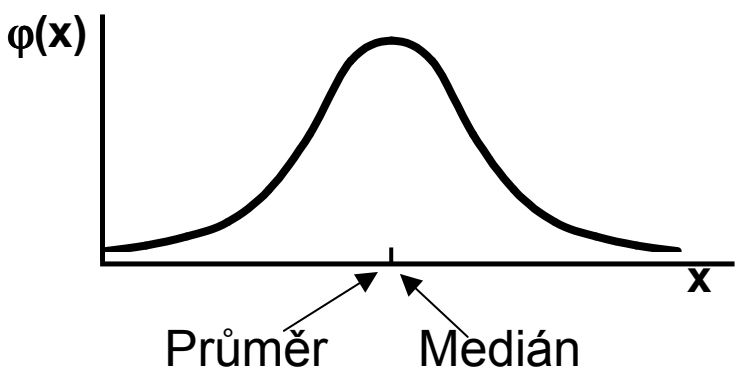
# Číselné charakteristiky dátového súboru
# Intervalové a pomerové znaky-ukazatele stredu

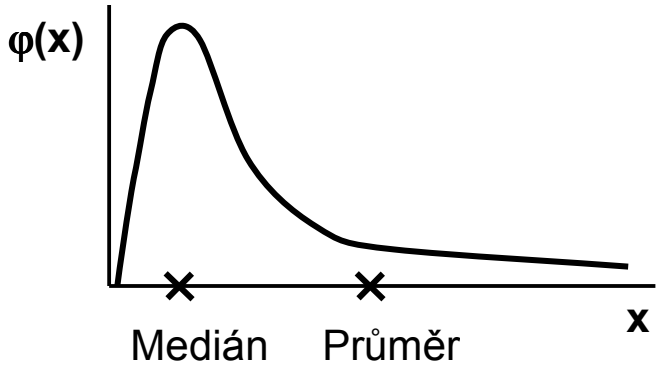
- **Průměr** – vhodný ukazatel středu u normálního/symetrického rozložení, kde $\mathbf{x_i}$ jsou jednotlivé hodnoty a **n** jejich počet

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

- **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



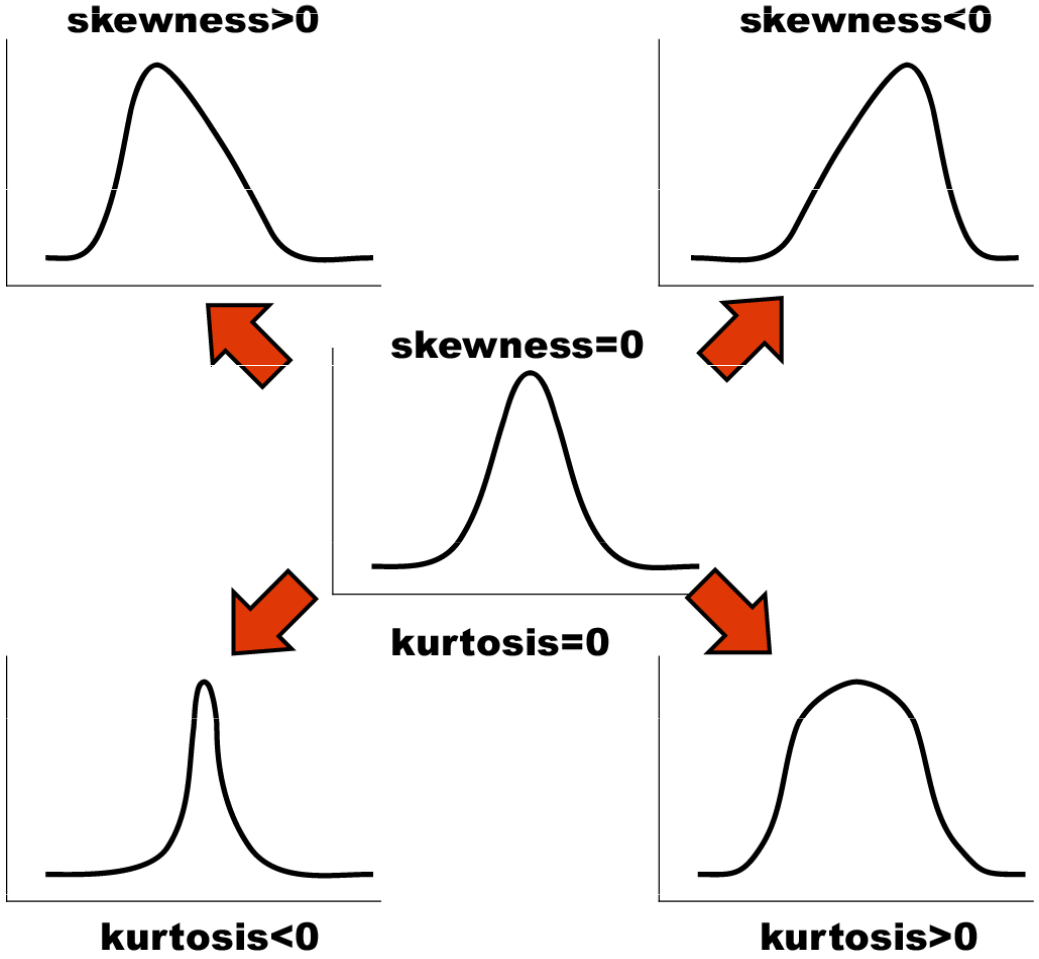



# Číselné charakteristiky dátového súboru Intervalové a pomerové znaky-ukazatele šírky

- **Rozptyl** 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** 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

# 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
- **Střední chyba odhadu průměru** - je založena na směrodatné odchylce rozložení a **počtu hodnot**, vlastně jde o směrodatnou odchylku rozložení průměru. Říká jak přesný je náš výpočet průměru. Čím větší počet hodnot rozložení, tím je náš odhad skutečného průměru přesnější.
- **Suma hodnot**
- **Modus** – nejčastější hodnota, vhodný např. při kategoriálních datech
- **Minimum, maximum**
- **Rozsah hodnot**
- **Harmonický průměr** - převrácená hodnota průměru převrácených hodnot (vždy platí harmonický průměr < geometrický průměr < aritmetický průměr)

# Príklad

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

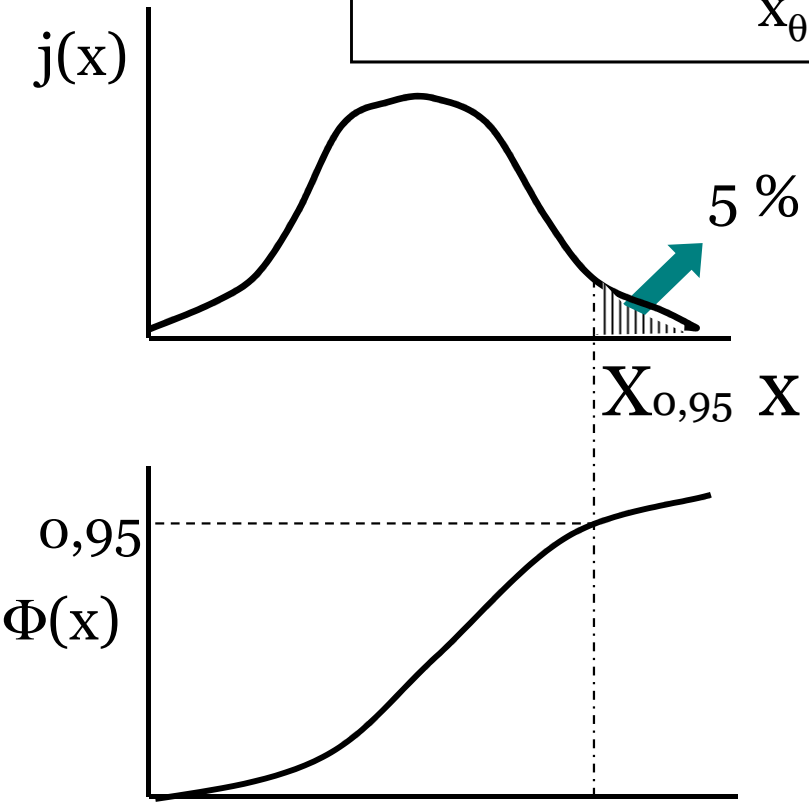
sm. odchylka (s) = $\sqrt{s^2} = \sqrt{0{,}766} = 0{,}875$

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

θ = 0,95 … Pravděpodobnost

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

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

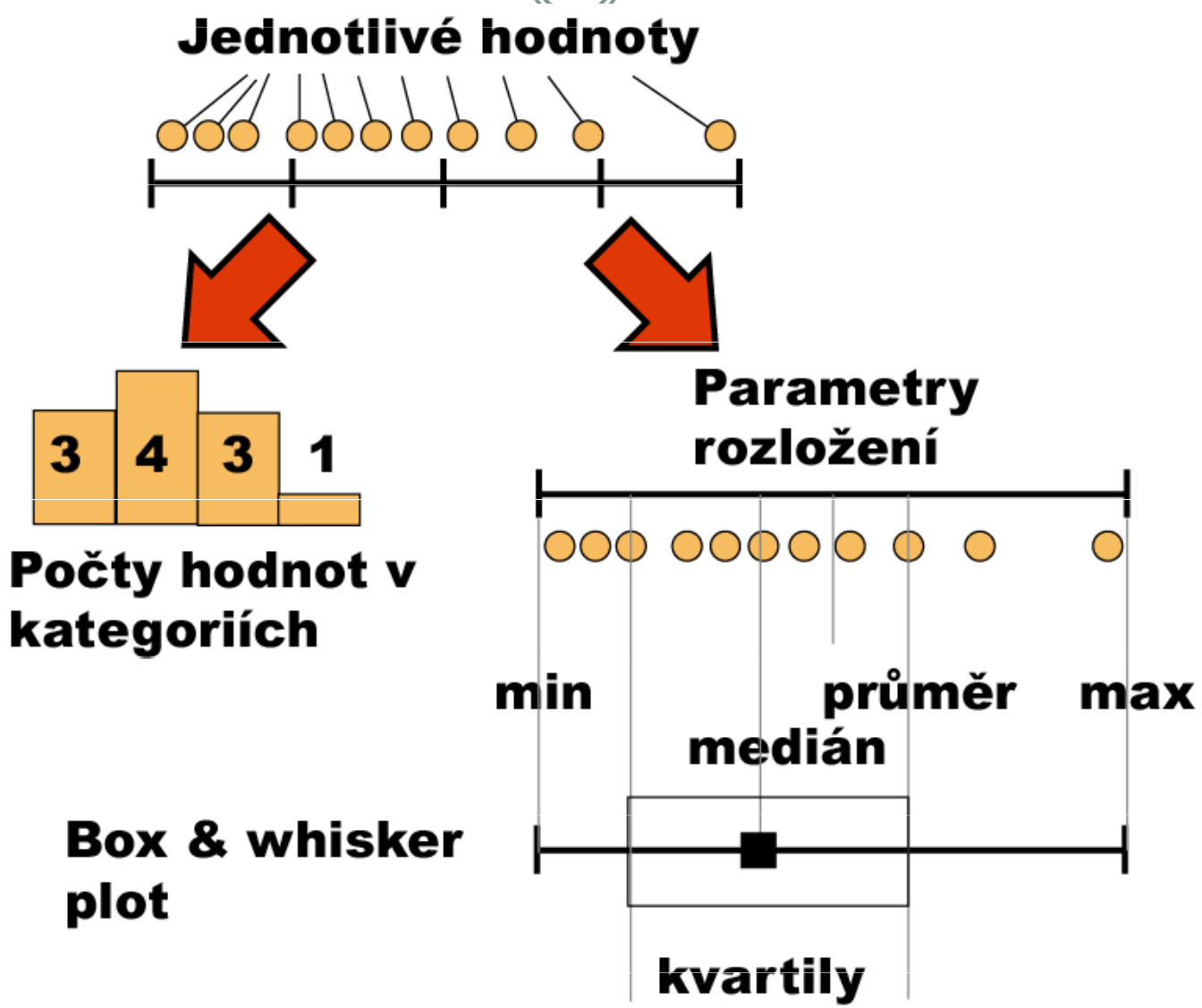



$F(x\theta) = \theta$

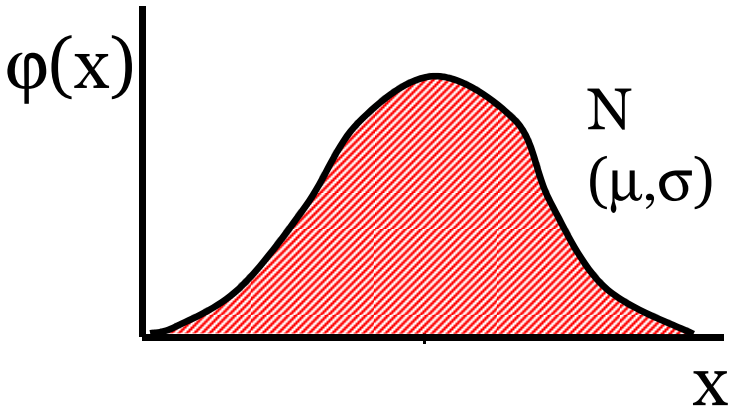
Kvantil je číslo, jehož hodnota distribuční funkce je rovna P, pro kterou je kvantil definován

Jakékoliv číslo na ose x je kvantilem

# Diagnostické grafy-krabicový graf (box plot)



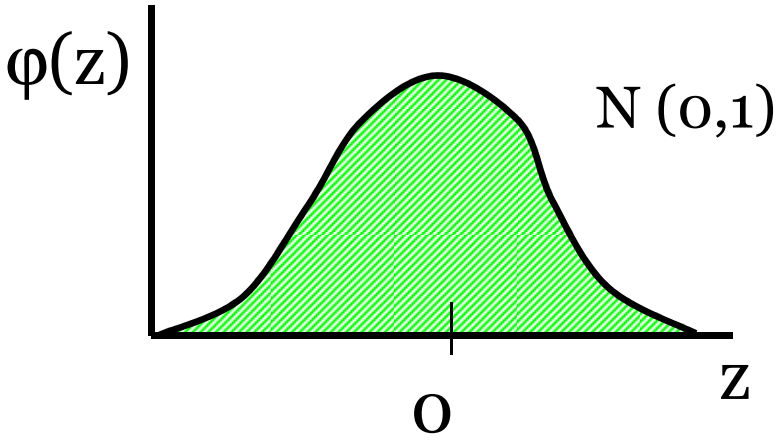
# Normální rozložení




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

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

**Standardizovaná forma**



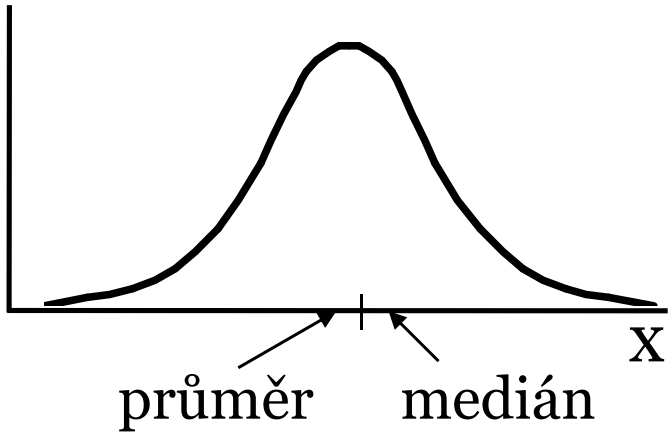

$$\varphi(z) = \frac{1}{\sqrt{2.\pi}}.e^{-\frac{z^2}{2}}$$

**Tabelovaná podoba**

# Parametry charakterizující normální rozložení a jejich význam

$E(x) \sim \bar{x} \sim \mu$

$D(x) \sim s^2 \sim \sigma^2$



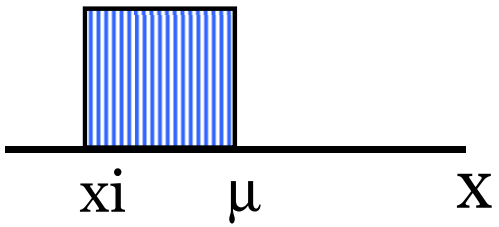

a) $\mu \sim \overline{x}$

průměr - ukazatel středu

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

rozptyl

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




c) $\sigma \sim s$

směrodatná odchylka

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

Pravidlo ± 3s

d) koeficient variance

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

# Normální rozložení – příklad

- Data z průzkumu jsou publikována jako:

  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 dalších 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 dalších rozložení II.

| Rozložení | Parametry | Stručný popis |
|---|---|---|
| Beta | Parametry distribuční funkce:<br>α - parametr tvaru<br>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<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. |