# 8. Modelová rozdělení pravděpodobnosti, popisné statistiky

Rozdělení pravděpodobnosti

Normální rozdělení jako statistický model

Přehled a aplikace modelových rozdělení

Popisné statistiky

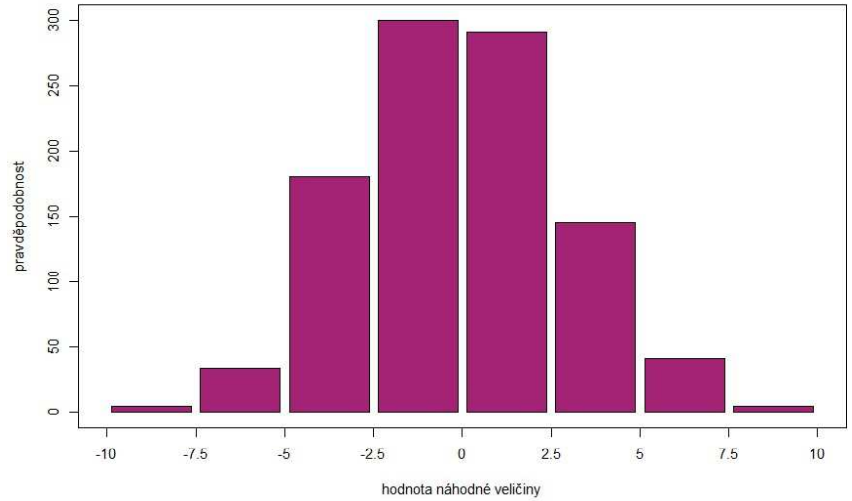
Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
*J. Jarkovský, L. Dušek*

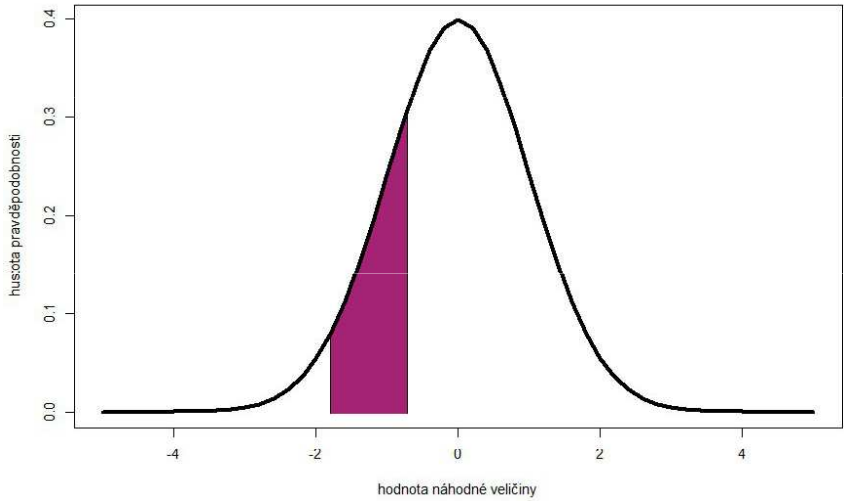
# Anotace

- Klasickým postupem statistické analýzy je na základě vzorku cílové populace identifikovat typ a charakteristiky modelového rozdělení 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 rozdělením, v opačném případě hrozí získání zavádějících výsledků (neparametrické statistiky).
- Nejklasičtějším modelovým rozdělením, od něhož je odvozena celá řada statistických analýz je tzv. normální rozdělení, známé též jako Gaussova křivka.

# Rozdělení (rozložení, distribuce) pravděpodobnosti (dat)

- Funkce přiřazující intervalu hodnot náhodné veličiny pravděpodobnost (obecně), resp. přiřazující hodnotě náhodné veličiny určitou hustotu pravděpodobnosti (derivace pravděpodobnosti podle náhodné veličiny).
- V případě diskrétní náhodné veličiny lze ztotožnit intervaly s konkrétními hodnotami a tvrdit, že rozdělení pravděpodobnosti přiřazuje jednotlivým hodnotám přímo pravděpodobnost.

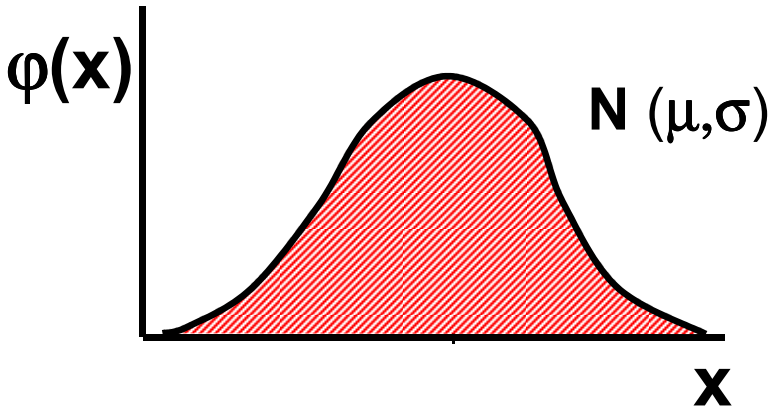


# Rozdělení (rozdělení, distribuce) pravděpodobnosti (dat)

- Rozdělení pravděpodobnosti pro spojité a diskrétní náhodné veličiny se liší (páry podobných rozdělení).
- Každá náhodná veličina má určité rozdělení, které může a nemusí být známé (plyne z definice náhodné veličiny).
- Rozdělení je určeno charakteristickými parametry. Jejich typ a počet se liší na základě komplexity rozdělení:
  - průměr,
  - rozptyl,
  - špičatost,
  - šikmost aj.
- Při analýze určujeme výběrové parametry, které nejsou totožné s reálnými parametry rozdělení.

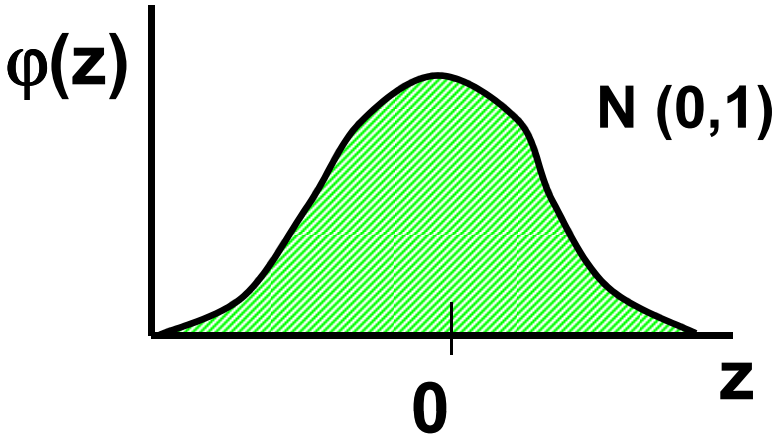
# Rozdělení hodnot jako model: Normální rozdělení




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

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

Standardizovaná forma



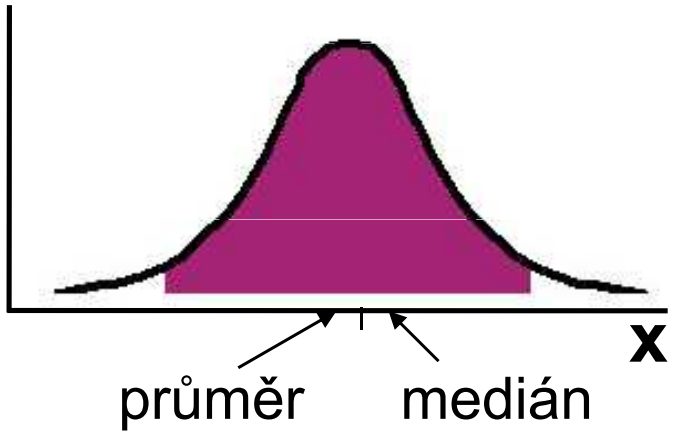

$$\varphi(z) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{z^2}{2}}$$

Tabelovaná podoba

# Parametry charakterizující normální rozdělení a jejich význam

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

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



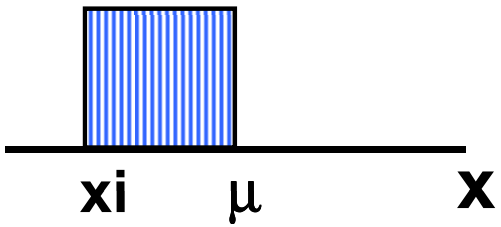

a) $\mu \sim \bar{x}$

**průměr - ukazatel středu**

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

**rozptyl**

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




c) $\sigma \sim s$

**směrodatná odchylka**

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

**Pravidlo ± 3s**

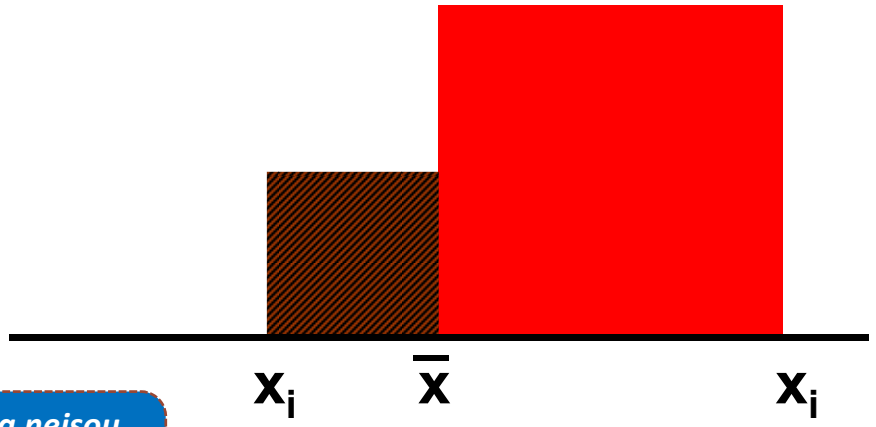
d)

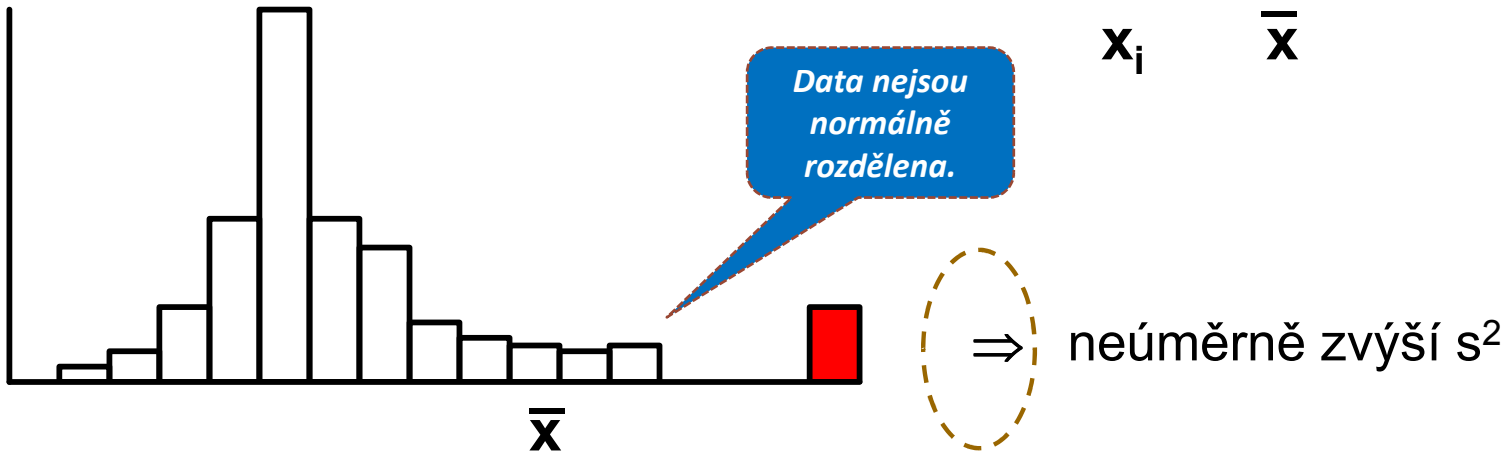
**koeficient variance**

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

# Rozptyl není univerzálním ukazatelem variability

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




$\Rightarrow$ neúměrně zvýší $s^2$

- Rozptyl a směrodatná odchylka jsou citlivé na odlehlé hodnoty (nemají vhodný význam pro jiné než normální rozdělení).

# Normální rozdělení 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

*Kolmogorov-Smirnov: p=n. s.*
*Liliefors: p<1,000*
*Shapiro-Wilks: p=0,1307*

**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 rozdělení oprávněný ?**
**Jaký předpokládáte možný rozsah hodnot tohoto znaku ?** **?**

# Normální rozdělení 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 = 2,0

*Kolmogorov-Smirnov: p<0,200*
*Liliefors: p<0,010*
*Shapiro-Wilks: p<0,001*

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 ?**

# Normální rozdělení 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 rozdělení hodnot x**

# Normální rozdělení jako model - příklad

## *Tabulky distribuční funkce*

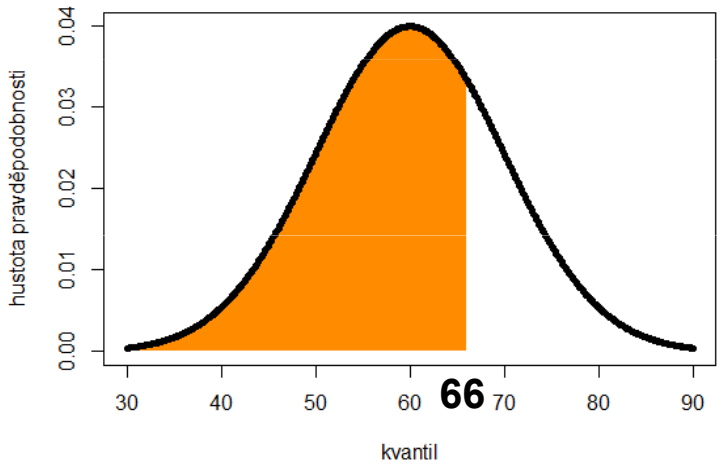
- **<u>Data z průzkumu jsou publikována jako:</u>**

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

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

? Jaká je pravděpodobnost, že by velikost dané nepřekročí velikost 66 cm: P (x < 66) ?

$$Z = \frac{x - \mu}{\sigma} = \frac{66 - 60}{10} = 0,6$$



# Normální rozdělení jako model - příklad

## *Tabulky distribuční funkce*

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

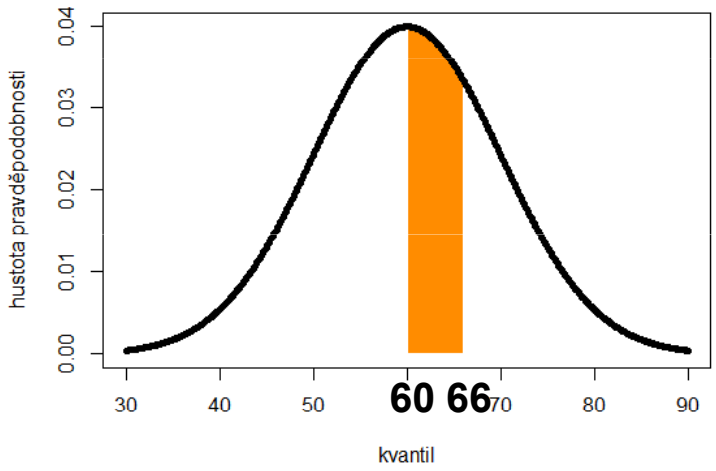
? Kolik kostí mělo zřejmě délku větší než 66 cm ? $P(x > 66) * n = 0{,}27425 * 2000 = 548$

? 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

***Hodnoty distribuční funkce F lze nalézt v tabulkách () nebo zjistit pomocí fce NORMDIST v Excelu.***

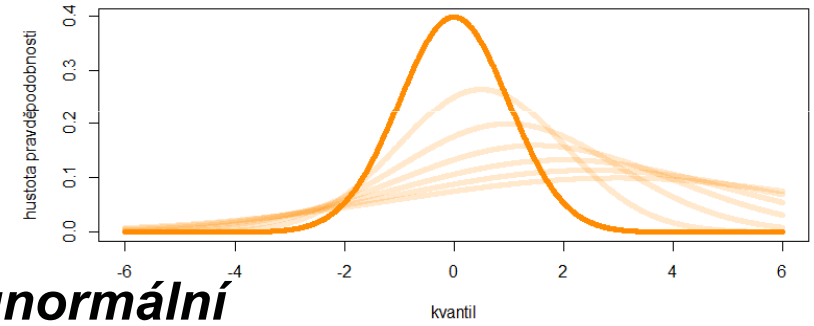


# Stručný přehled modelových rozdělení I.

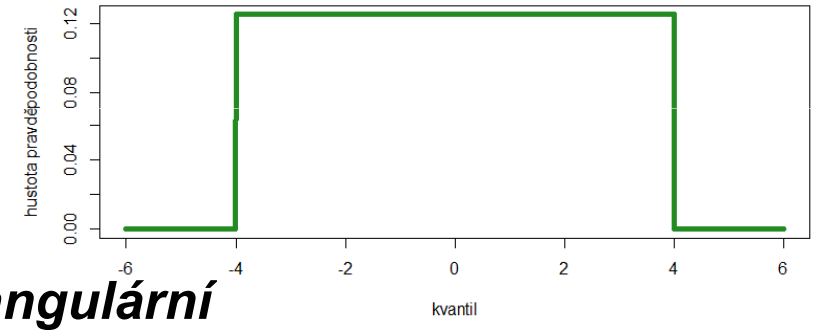
| rozdělení | Parametry | Stručný popis |
|---|---|---|
| **Normální** | **Průměr ($\mu$)**<br>**Rozptyl ($\sigma^2$)** | Symetrická funkce popisující intervalovou hustotu četnosti; nejpravděpodobnější jsou průměrné hodnoty znaku v populaci. |
| **Log-normální** | **Medián**<br>**Geometrický průměr**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

# Stručný přehled modelových rozdělení I.

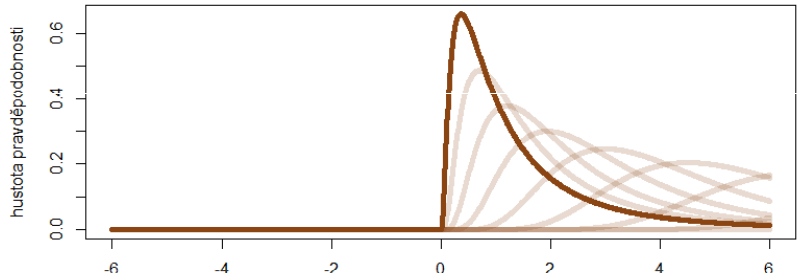
## *Normální*



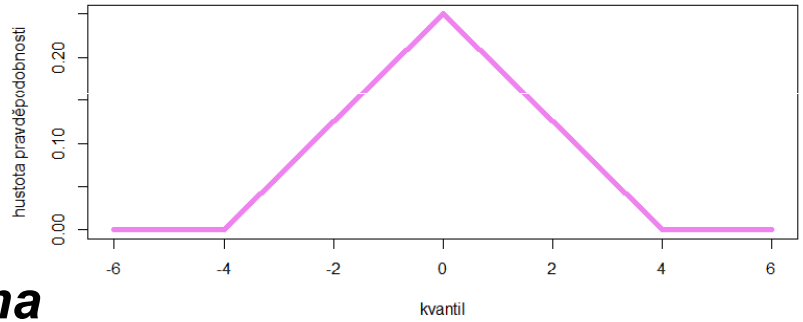

## *Rovnoměrné*



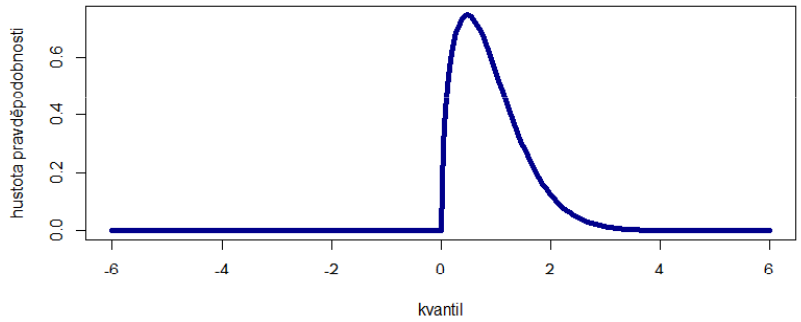

## *Lognormální*



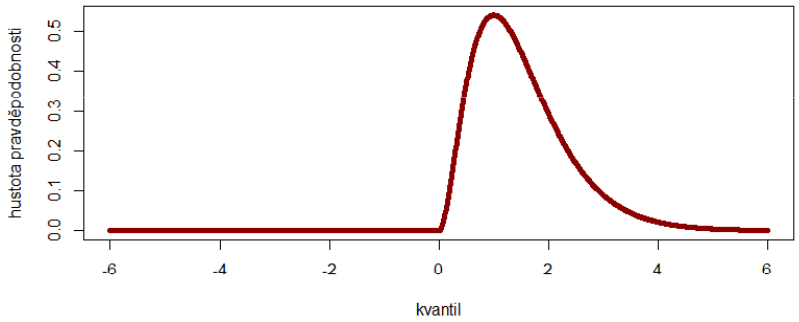

## *Triangulární*




## *Weibullovo*




## *Gama*



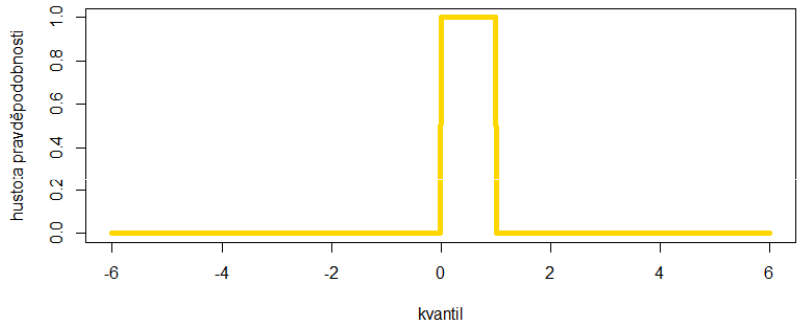

MU IBA

# Stručný přehled modelových rozdělení II.

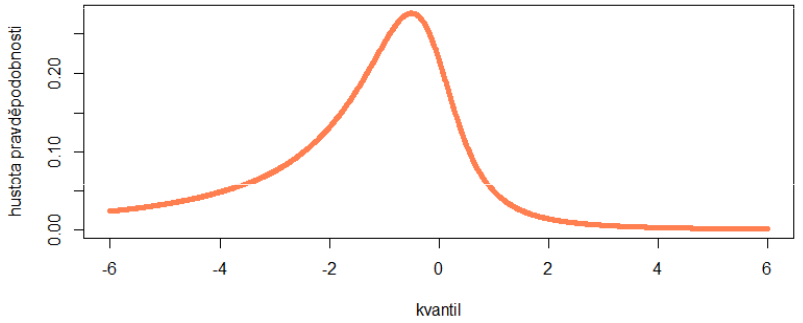
| rozdělení | Parametry | Stručný popis |
|---|---|---|
| **Beta** | **Parametry distribuční funkce:**<br>**$\alpha$ - parametr tvaru**<br>**$\beta$ - parametr rozsahu hodnot** | Pravděpodobnostní funkce pro proměnnou omezenou rozsahem do intervalu [0; 1]. Je matematicky komplikovanější, ale velmi flexibilní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

# Stručný přehled modelových rozdělení II.

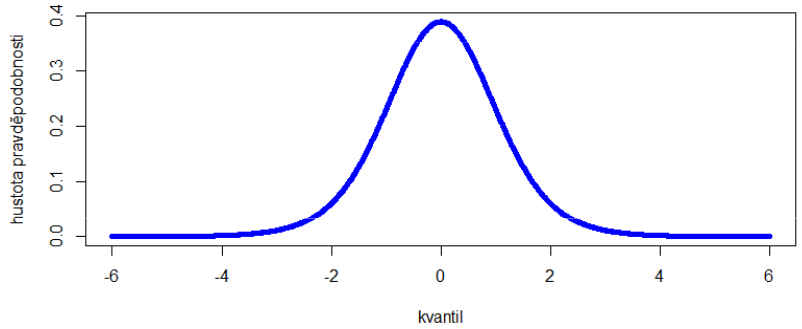
## *Beta*



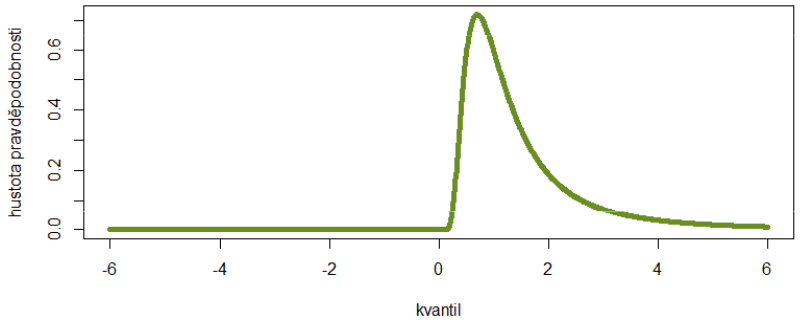

## *Pearsonovo*




## *Studentovo*




## *Fisher-Snedecorovo*

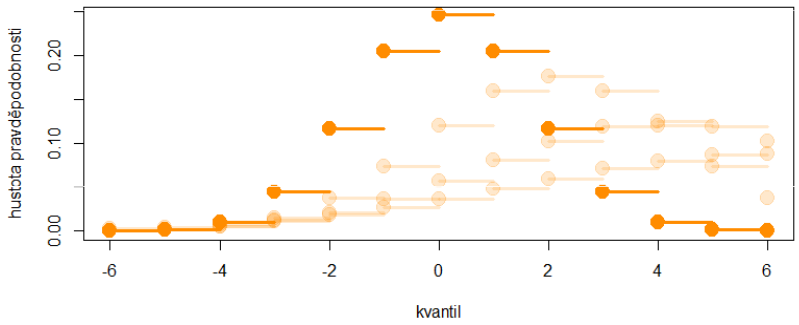


# Stručný přehled modelových rozdělení III.

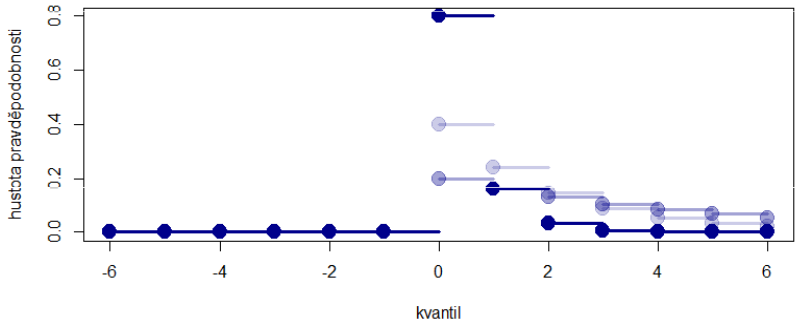
| rozdělení | Parametry | Stručný popis |
|---|---|---|
| **Binomické** | **Průměr ($\mu$)**<br>**Rozptyl ($\sigma^2$)** | Diskrétní obdoba normálního rozdělení - symetrická funkce popisující intervalovou četnost výskytu jevu v nezávislých pokusech; nejpravděpodobnější jsou průměrné hodnoty znaku. |
| **Poissonovo** | **Lambda** | Rozdělení řídkých (málo pravděpodobných) jevů. Pro n > 30 se používá k aproximaci binomického rozdělení (jednoduchá matematická forma funkce). |
| **Geometrické** | **Lambda** | Diskrétní podoba exponenciálního rozdělení. Udává počet opakování experimentu do prvního úspěchu při konstantní pravděpodobnosti úspěchu. |
| **Bernoulliho** | **Pravděpodobnost jevu p** | Binární rozdělení pravděpodobnosti, kdy jev nastane s pravděpodobností p a nenastane s pravděpodobností 1-p. |

# Stručný přehled modelových rozdělení III.

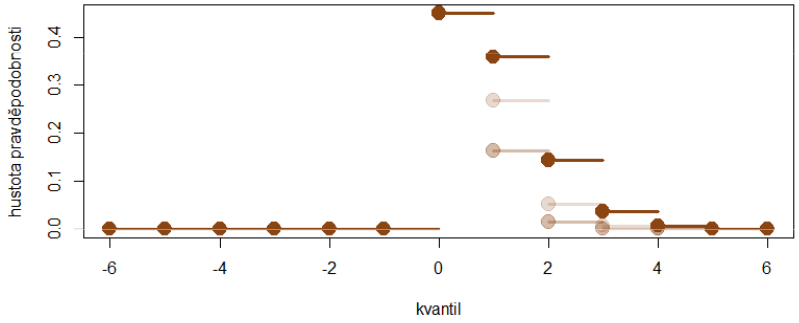
## *Binomické*



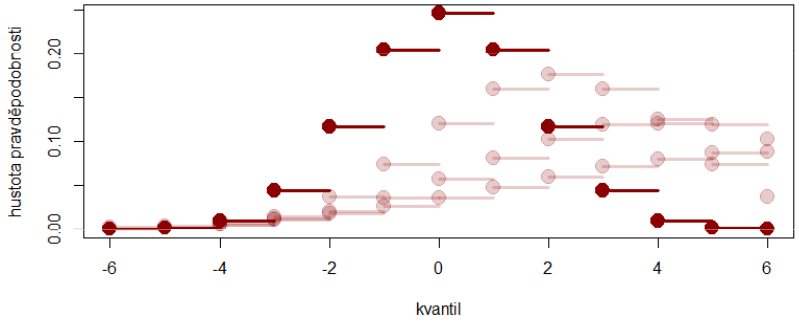

## *Geometrické*



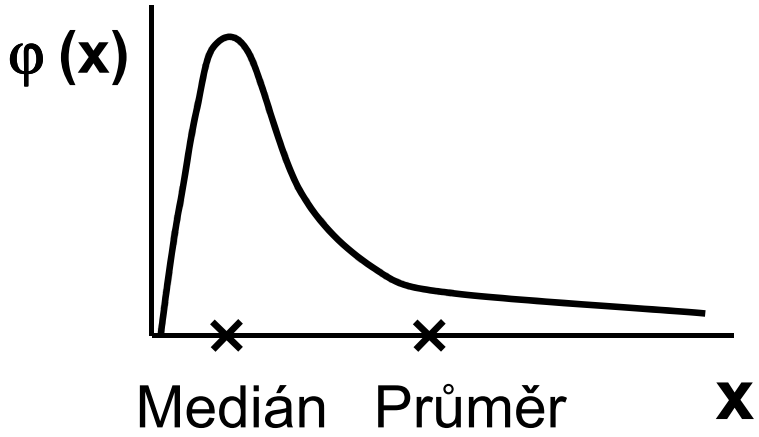

## *Poissonovo*



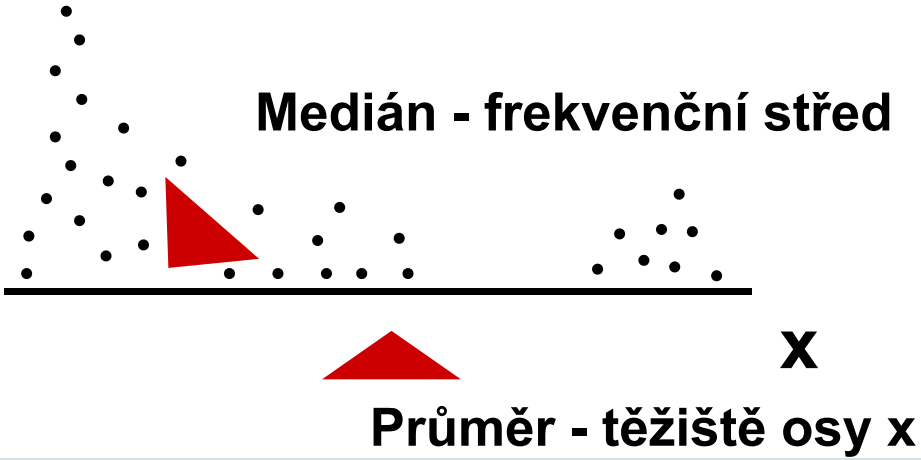

## *Bernoulliho*



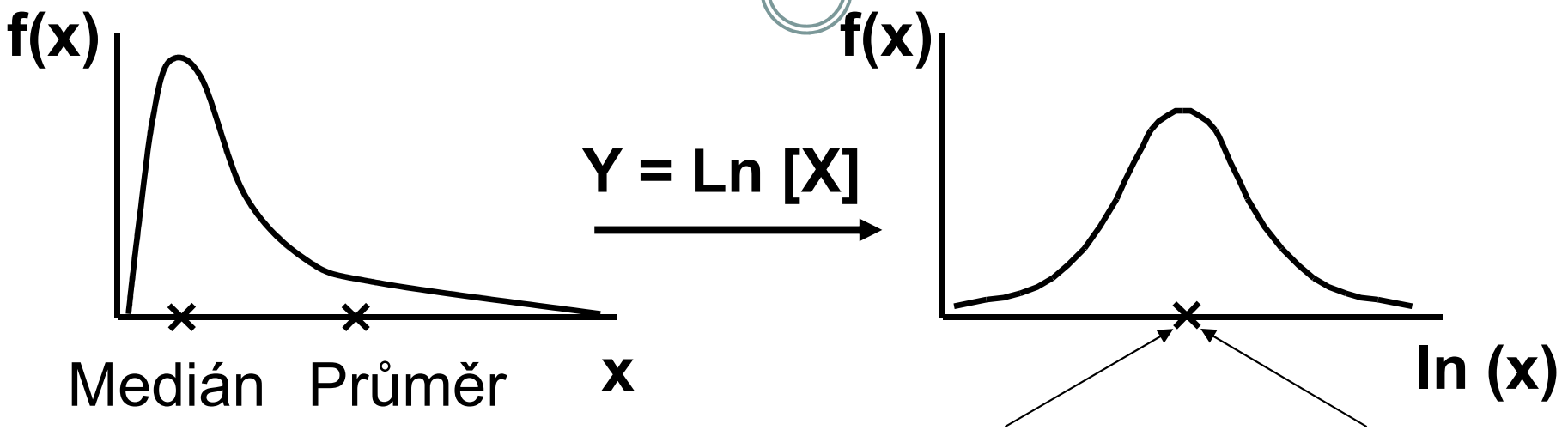
# Log-normální rozdělení jako častý model reálných znaků




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



# Log-normální rozdělení lze jednoduše transformovat




**EXP (Y) = Geometrický průměr X**

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

$\overline{Y}$ ± **Standardní chyba**

# Ukazatele tvaru rozdělení
# Koeficienty šikmosti a špičatosti

- **Skewness** – koeficient šikmosti rozdělení, míra asymetrie rozdělení

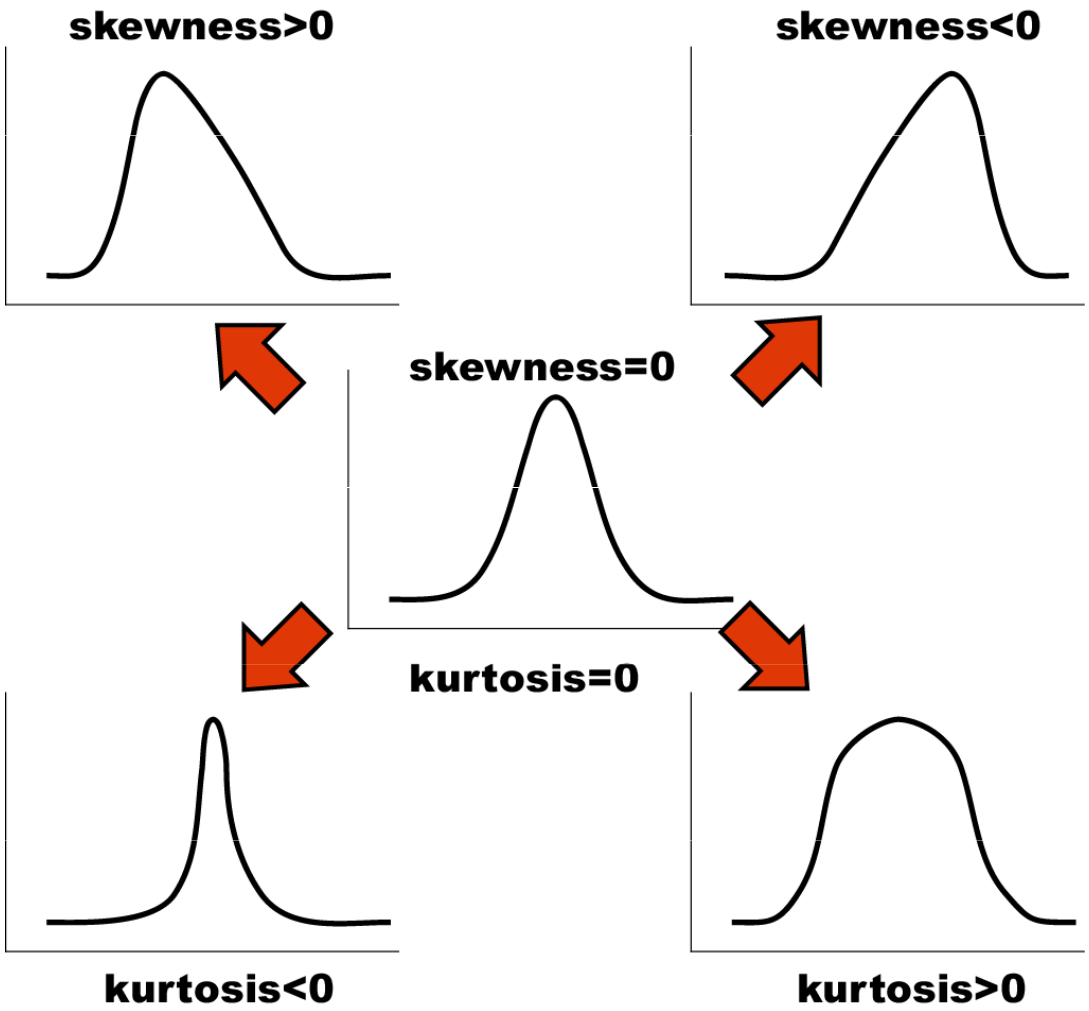
$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}[X - \mathrm{E}(X)]^3}{(\mathrm{var}\ X)^{3/2}}$$

kladná hodnota znamená odlehlé body vpravo, záporná vlevo od střední hodnoty.

- **Kurtosis** – koeficient špičatosti rozdělení,

$$\gamma_2 - \frac{\mu_4}{\sigma^4} - 3 - \frac{\mathrm{E}[X - \mathrm{E}(X)]^4}{(\mathrm{var}\ X)^2} - 3$$

kladná hodnota znamená větší hustotu pravděpodobnosti blíže střední hodnotě rozdělení.



# Transformace dat - legitimní úprava rozdělení

✓ **Základní typy transformací vedou k normalitě rozdělení 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 rozdělení, ale také vyšší homogenitu rozptylu proměnných. Pro transformaci se nejčastěji používá přirozený logaritmus a pokud jsou v původním souboru dat nulové či záporné hodnoty, je vhodné použít operaci **Y = ln(X+i)**, kde **i** je velmi malý pozitivní přírůstek.

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 rozdělení

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

## Odmocninová transformace

Transformace je vhodná pro proměnné mající Poissonovo rozdělení, 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 rozdělení

## Arcsin transformace

Tzv. **úhlová transformace** - velmi vhodná pro data typu podílů výskytu určitého jevu (znaku) mezi **n** hodnocenými jedinci - tedy pro data mající binomické rozdělení. 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:

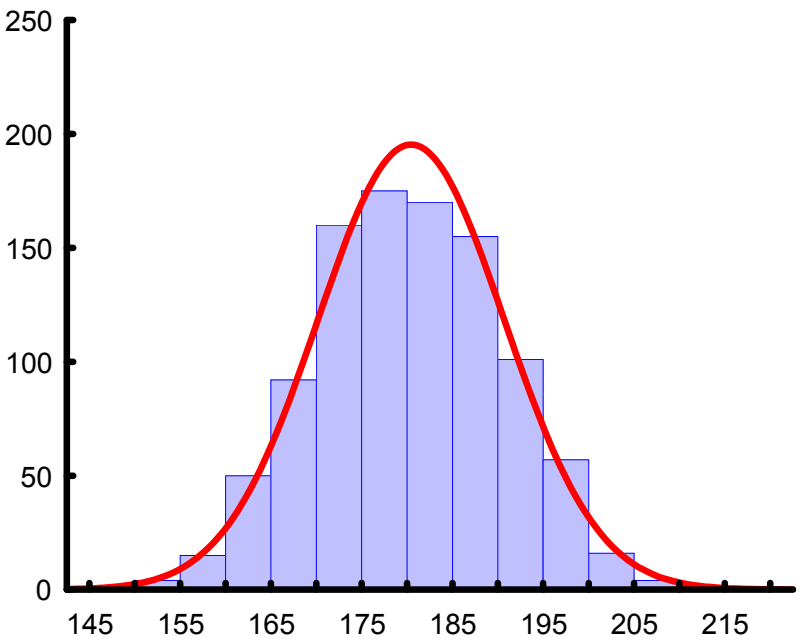
$$Y = \frac{1}{2}\left[\arcsin \sqrt{\frac{x}{n+1}} + \arcsin \sqrt{\frac{x+1}{n+1}}\right]$$

# Popisná statistika

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

# Testy normality

- Testy normality pracují s nulovou hypotézou, že není rozdíl mezi zpracovávaným rozložením a normálním rozložením. Vždy je ovšem dobré prohlédnout si i histogram, protože některé odchylky od normality, např. bimodalitu některé testy neodhalí.



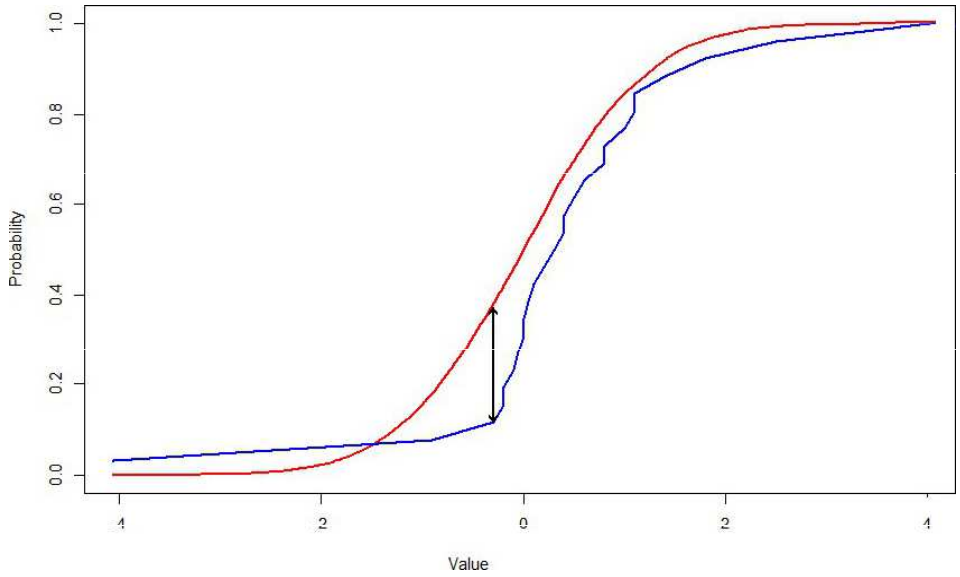

## Test dobré shody

V testu dobré shody jsou data rozdělena do kategorií (obdobně jako při tvorbě histogramu), tyto intervaly jsou normalizovány (převedeny na normální rozložení) a podle obecných vzorců normálního rozložení jsou k nim dopočítány očekávané hodnoty v intervalech, pokud by rozložení bylo normální. Pozorované normalizované četnosti jsou poté srovnány s očekávanými četnostmi pomocí $\chi 2$ testu dobré shody. Test dává dobré výsledky, ale je náročný na n, tedy množství dat, aby bylo možné vytvořit dostatečný počet tříd hodnot.

# Testy normality

## Kolgomorovův-Smirnovův test

Tento test je často používán, dokáže dobře najít odlehlé hodnoty, ale počítá spíše se symetrií hodnot než přímo s normalitou. Jde o neparametrický test pro srovnání rozdílu dvou rozložení. Je založen na zjištění rozdílu mezi reálným kumulativním rozložením (vzorek) a teoretickým kumulativním rozložením. Měl by být počítán pouze v případě, že známe průměr a směrodatnou odchylku hypotetického rozložení, pokud tyto hodnoty neznáme, měla by být použita jeho modifikace – Lilieforsův test.




## *Shapiro-Wilkův test*

***Jde o neparametrický test použitelný i při velmi malých n (10) s dobrou sílou testu, zvláště ve srovnání s alternativními typy testů, je zaměřen na testování symetrie.***

# P-hodnota

Významnost hypotézy hodnotíme dle získané tzv. p-hodnoty, která vyjadřuje pravděpodobnost, s jakou číselné realizace výběru podporují $H_0$, je-li pravdivá.

P-hodnotu porovnáme s α (hladina významnosti, stanovujeme ji na **0,05**, tzn., že připouštíme 5 % chybu testu, tedy, že zamítneme $H_0$, ačkoliv ve skutečnosti platí).

P-hodnotu získáme při testování hypotéz ve statistickém softwaru.

- Je-li **p-hodnota ≤ α**, pak **$H_0$ zamítáme** na hladině významnosti α a **přijímáme $H_A$**
- Je-li **p-hodnota > α**, pak **$H_0$ nezamítáme** na hladině významnosti α

P-hodnota vyjadřuje pravděpodobnost za platnosti $H_0$, s níž bychom získali stejnou nebo extrémnější hodnotu testové statistiky.