



Univerzita Palackého
v Olomouci

Rozdělení pravděpodobnosti, bodové a intervalové odhady

Chemometrie I (ACH/CHEX1)

(c) David MILDE, 2024

1



Univerzita Palackého
v Olomouci

Náplň předmětu a doporučená literatura

- **Náplň předmětu – statistická analýza jednorozměrných dat:**
 - bodové a intervalové odhady,
 - průzkumová analýza jednorozměrných dat,
 - testování hypotéz,
 - analýza rozptylu,
 - lineární regrese, kalibrace,
 - korelace.
- **LITERATURA:**
 - Meloun M., Militký J.: Kompendium statistického zpracování dat. Academia, Praha 2002 (+ novější vydání).
 - Studijní materiály na webových stránkách katedry.
 - Meloun M., Militký J.: Statistické zpracování experimentálních dat. Plus, Praha 1994 (+ novější vydání).

2



Univerzita Palackého
v Olomouci

Rozdělení pravděpodobnosti

3



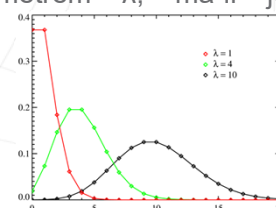
Univerzita Palackého
v Olomouci

Diskrétní rozdělení pravděpodobnosti

- **Alternativní rozdělení:**
 - máme 1 pokus, kdy zkoumaný jev A nastane nebo nenastane,
 - Př. hod mincí, úspěch či neúspěch u zkoušky.
- **Binomické rozdělení:**
 - U 1 pokusu může sledovaný jev A nastat s pravděpodobností P nebo nenastat s pravděpodobností $Q = 1 - P$.
- **Poissonovo rozdělení:**
 - Náhodná veličina x má Poissonovo rozdělení s parametrem λ , má-li její pravděpodobnostní funkce tvar:

$$P(x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}$$

Graf hustoty
pravděpodobnosti
Poissonova rozdělení



4



Univerzita Palackého
v Olomouci

Spojitá rozdělení pravděpodobnosti

– Normální (Gaussovo) rozdělení:

- Poprvé popsáno při studiu chování chyb měření. Později se ukázalo, že mnoho náhodných veličin má toto rozdělení, např.: výška a váha živých organismů, rozměry výrobků atp.

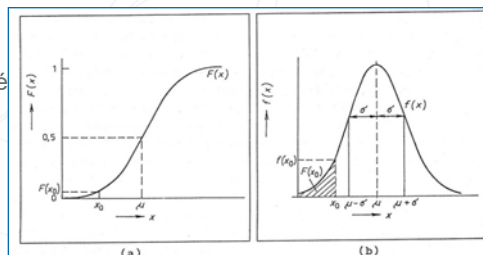
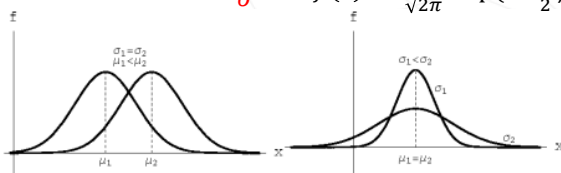
- Náhodná veličina x má normální rozdělení s parametry μ a σ^2 – $N(\mu, \sigma^2)$, má-li hustotu pravděpodobnosti $f(x)$:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{pro } -\infty < x < \infty, \sigma > 0$$

- Normování – transformace z $N(\mu, \sigma^2)$ na $N(0, 1)$, tj. tabelované rozdělení:

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

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$



(a) graf distribuční funkce
(b) graf hustoty pravděpodobnosti

5

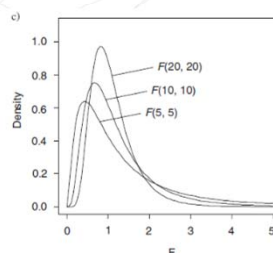
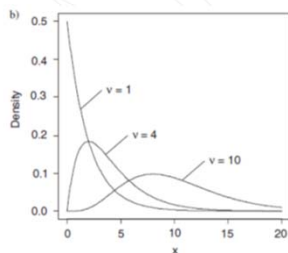
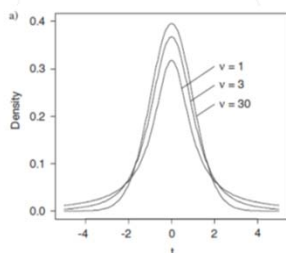


Univerzita Palackého
v Olomouci

Spojitá rozdělení pravděpodobnosti

– Rozdělení odvozená od normálního rozdělení („teoretická“ rozdělení):

- **Studentovo t rozdělení** – aproximuje normální rozdělení pro malý počet dat, využívá se u výpočtů intervalů spolehlivosti či řady statistických testů. Je to symetrické rozdělení s počtem stupňů volnosti $v = n - 1$. Pro $n > 100$ se limitně blíží normálnímu rozdělení.
- **Chí-kvadrát rozdělení** (χ^2) – popisuje rozdělení rozptylu, využívá se u některých neparametrických testů. Jde o asymetrické rozdělení s počtem stupňů volnosti $v = n - 1$.
- **F rozdělení** (Fisherovo-Snedecorovo rozdělení) – popisuje rozdělení podílu dvou rozptylů, využívá se v analýze rozptylu. Je asymetrické a popisuje se dvěma stupni volnosti (v_1, v_2).



Hustoty
pravděpodobnosti pro:
(a) Studentovo t roz.
(b) Chí-kvadrát roz.
(c) F rozdělení

6



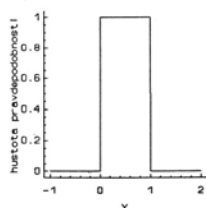
Univerzita Palackého
v Olomouci

Spojité rozdělení pravděpodobnosti

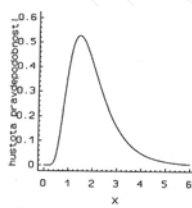
– „Experimentální“ rozdělení:

- Rovnoměrné rozdělení
- Logaritmicko-normální rozdělení (LN)
- Exponenciální rozdělení
- Laplaceovo rozdělení (dvojitě exponenciální)

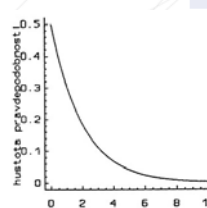
Rozdělení	g_1	g_2
rovnoměrné	0	< 3
log-normální	> 0	> 3
exponenciální	2	9
Laplaceovo	0	6



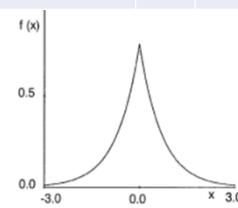
rovnoměrné roz.



logaritmicko-normální roz.



exponenciální roz.



Laplaceovo roz.

Hustoty pravděpodobnosti pro:

7

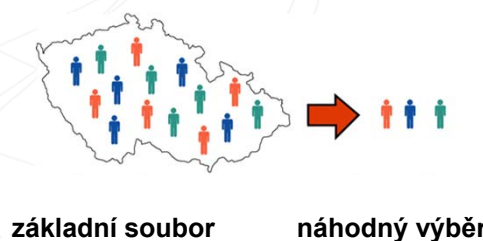
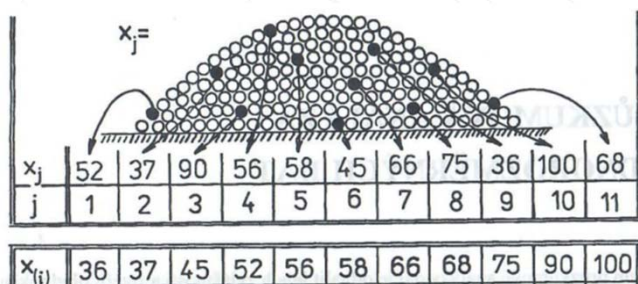


Univerzita Palackého
v Olomouci

Základní soubor a náhodný výběr

– Ve statistice se předpokládá, že sada zpracovávaných hodnot je tzv. náhodným výběrem z tzv. základního souboru.

- Základní soubor (aj. population): teoreticky všechny možné výsledky daného pozorování s parametry: střední (pravdivá) hodnota μ a rozptylem σ^2 .
- Náhodný výběr (aj. sample): náhodně vybraná data ze základního souboru s určitými vlastnostmi a parametry: výběrový odhad střední hodnoty (obvykle) $\bar{x}$ a výběrový odhad rozptylu s^2 .



základní soubor

náhodný výběr

8



Univerzita Palackého
v Olomouci

Náhodný výběr, momentové charakteristiky

– Vlastnosti reprezentativního náhodného výběru:

1. Prvky výběru x_i jsou vzájemně nezávislé.
2. Výběr je homogenní, tj. všechna x_i jsou ze stejného rozdělení.
3. Předpokládá se, že jde o normální rozdělení.
4. Všechny x_i mají stejnou pravděpodobnost, že budou zařazeny do výběru.

– **Momenty** = číselné charakteristiky rozdělení podávající informace o vlastnostech rozdělení:

- Obecný moment r -tého stupně $M_r = \int x^r p(x) dx$
- Centrální moment r -tého stupně $M(\mu)_r = \int (x - \mu)^r p(x) dx$
- M_1 – střední hodnota – charakterizuje polohu
- $M(\mu)_2$ – rozptyl – charakterizuje preciznost
- $M(\mu)_3$ – koeficient šikmosti – charakterizuje tvar – g_1
- $M(\mu)_4$ – koeficient špičatosti – charakterizuje tvar – g_2

9



Univerzita Palackého
v Olomouci

Bodové odhady (polohy, rozptýlení a tvaru)

10



Univerzita Palackého
v Olomouci

Vlastnosti a dělení bodových odhadů

– Vlastnosti bodových odhadů:

- KONZISTENTNOST odhadu – s rostoucí četností n se zmenšuje rozdíl mezi odhadem ($\bar{x}$) a skutečnou hodnotou μ .
- NESTRANOST odhadu – pro n blížící se k nekonečnu $\bar{x} = \mu$
- VYDATNOST odhadu – rozptyl odhadu okolo skutečné hodnoty μ se s rostoucí četností n zmenšuje. Nestranný odhad ani při malém n soustavně nepodhodnocuje ani nenadhodnocuje odhadovaný parametr.
- ROBUSTNOST odhadu – necitlivost na odchylky od předpokládaného rozdělení.

– Dělení bodových odhadů (zejména polohy):

- odhady založené na kvantilech (kvantilové charakteristiky, např. **medián**),
- odhady založené na momentech (momentové charakteristiky, např. **aritmetický průměr**),
- pro diskrétní proměnné se používá **modus** (x_M), tj. nejčastější prvek v datech:

Př.: 1; 3; 3; 3; 3; 4; 5; 5; 6; 8; 9; 9; 9 $x_M = 3$

11



Univerzita Palackého
v Olomouci

Kvantilové odhady polohy

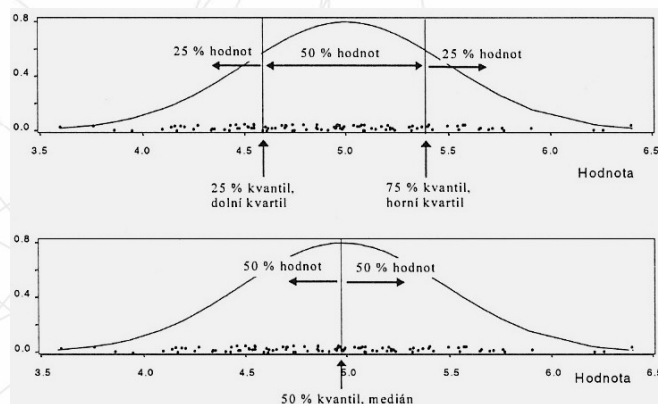
- KVANTILY – hodnoty znaku, které dělí data na určitý počet skupin o stejném počtu prvků.

- **Medián** je kvantil, který rozděluje data na 2 části o 50 % rozsahu souboru.

$$\tilde{x}_{0,5} = \begin{cases} n \text{ je liché } x_{(n+1)/2} \\ n \text{ je sudé } \frac{x_{n/2} + x_{(n+2)/2}}{2} \end{cases}$$

- Kvartily – rozdělují uspořádanou řadu hodnot na 4 skupiny se stejnou četností, prostřední kvartil = medián dolní kvartil se značí $\tilde{x}_{0,25}$ a horní kvartil $\tilde{x}_{0,75}$.

- Decily – rozdělují uspořádanou řadu hodnot na 10 skupin o stejně velké četnosti.



12



Univerzita Palackého
v Olomouci

Momentové odhady polohy

- Aritmetický průměr ($\bar{x}$)

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

- Vážený průměr

$$\bar{x}_w = \frac{\sum w_i \cdot x_i}{\sum w_i}, \text{ kde } w_i \text{ je statistická váha}$$

- Geometrický (harmonizovaný) průměr

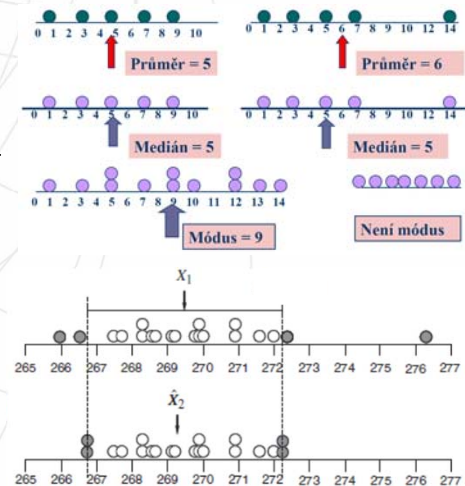
$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

- Uřezaný průměr

$$\bar{x}(v) = \frac{1}{n - 2M} \sum_{i=M+1}^{n-M} x_i \quad M = \text{int}\left(\frac{v \cdot n}{100}\right)$$

$\bar{x}(v)$ 10% uřezání $\bar{x}(10)$

- Winsorizovaný průměr: winsorizace = nahrazení odlehlých výsledků sousedními výsledky uspořádaného souboru, které již nejsou odlehlé; nezmenšuje se četnost souboru a zachovává se charakter dat.



Princip winsorizace

13



Univerzita Palackého
v Olomouci

Bodové odhady rozptýlení

- Rozpětí: rozdíl největší a nejmenší hodnoty statistického výběru $R = x_{\max} - x_{\min}$
- Interkvartilové rozpětí: rozdíl horního ($\tilde{x}_{0,75}$) a dolního ($\tilde{x}_{0,25}$) kvartilu
 - Obsahuje 50 % prostředních hodnot

- Rozptyl σ^2 , směrodatná odchylka σ

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

- Výběrový odhad rozptylu s^2 , výběrový odhad směrodatné odchylky s

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

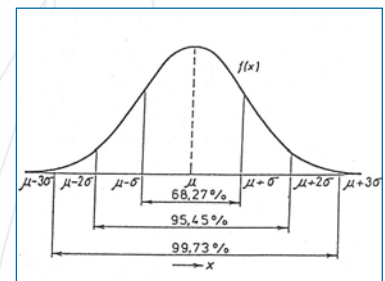
- Relativní směrodatná odchylka – RSD

- Totožná s variačním koeficientem – CV

$$RSD = \frac{s}{\bar{x}} \cdot 100 \quad [\%]$$

- Směrodatná odchylka průměru

$$s(\bar{x}) = \frac{s}{\sqrt{n}}$$



14

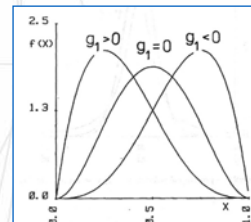


Univerzita Palackého
v Olomouci

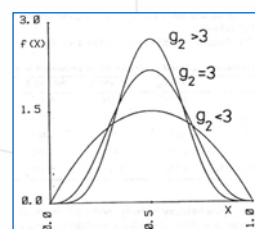
Bodové odhady tvaru

- Koeficient šikmosti g_1 – číslo, které charakterizuje nesouměrnost rozdělení, dává informace o tvaru rozdělení co do zešikmení resp. nesouměrnosti.
 - Pro symetrická rozdělení platí, že $g_1 = 0$
 - Akceptační interval pro $g_1 < -0,3; +0,3$
- Koeficient špičatosti g_2 – číslo, charakterizující zkoncentrování (protažení) prvků souboru v blízkosti určité hodnoty znaku.
 - Akceptační interval pro $g_2 < 2,2; 4,0$
 - Pro normální rozdělení $g_2 = 3$
 - Pozn: některé SW (např. MS Excel) od vzorce pro g_2 odečítají 3

$$g_1 = \frac{\sqrt{n} \cdot \sum (x_i - \bar{x})^3}{[\sum (x_i - \bar{x})^2]^{3/2}} \quad g_2 = \frac{n \cdot \sum (x_i - \bar{x})^4}{[\sum (x_i - \bar{x})^2]^2}$$



$g_1 < 0$ – negativní zešikmení
 $g_1 > 0$ – pozitivní zešikmení



15



Univerzita Palackého
v Olomouci

Intervalové odhady (polohy a rozptýlení)

16



Univerzita Palackého
v Olomouci

Intervalové odhady polohy

- Chceme-li vedle bodového odhadu parametru vyjádřit i preciznost odhadu, užijeme intervalový odhad. Parametr pak odhadujeme nikoli jednou hodnotou, nýbrž dvěma číselnými hodnotami L_1 a L_2 , které tvoří meze intervalu spolehlivosti (IS). Ten pokrývá neznámý parametr souboru, např. μ s předem zvolenou a dostatečně velkou pravděpodobností $1-\alpha$, kterou nazýváme hladinou spolehlivosti $P(L_1 \leq \mu \leq L_2) = 1-\alpha$, kde α je hladina významnosti.
- IS udává rozmezí hodnot střední hodnoty μ , který je v souladu s průměrem náhodného výběru $\bar{x}$ na dané hladině spolehlivosti. Pro základní soubor se střední hodnotou μ a směrodatnou odchylkou σ bude 95 % průměrů z náhodných výběrů ležet v rozmezí dané IS:

$$\mu - 1,96 \cdot \left(\frac{\sigma}{\sqrt{n}}\right) < \bar{x} < \mu + 1,96 \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$$

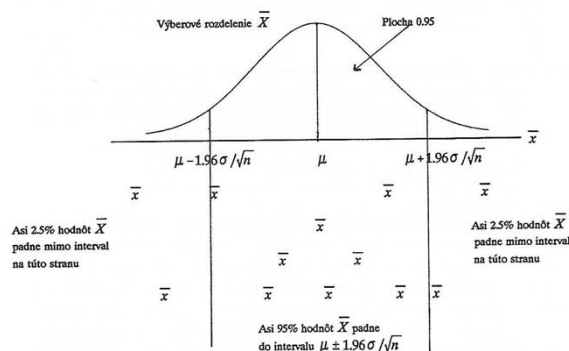
- hodnota 1,96 je hodnota dvoustranného Studentova t pro $\alpha = 0,05$ a $v = \infty$.

17



Univerzita Palackého
v Olomouci

Intervalové odhady polohy



- V praxi budeme mít k dispozici $\bar{x}$ (z experimentálních dat) a potřebujeme znát interval spolehlivosti střední hodnoty základního souboru. Rovnici se upraví:

$$\bar{x} - 1,96 \cdot \left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{x} + 1,96 \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$$

18



Univerzita Palackého
v Olomouci

Interval spolehlivosti střední hodnoty

- **Známe směrodatnou odchylku σ , nebo je-li odhad s určen z výběru o $n > 30$**

$$L_{1,2} = \bar{x} \pm z_{(1-\frac{\alpha}{2})} \frac{\sigma}{\sqrt{n}}$$

- Ve vzorci $z_{(1-\alpha/2)}$ představuje kvantil normovaného normálního rozdělení, v přírodních vědách se nejčastěji užívá $\alpha = 0,05$.

- **Neznáme σ , vypočte se její odhad s , používá se pro $n < 30$**

$$L_{1,2} = \bar{x} \pm t_{(1-\frac{\alpha}{2}; n-1)} \frac{s}{\sqrt{n}}$$

- Ve vzorci použité $t_{(1-\alpha/2, n-1)}$ představuje kvantil Studentova t-rozdělení s $n-1$ stupni volnosti.

19



Univerzita Palackého
v Olomouci

Interval spolehlivosti „malých“ souborů

- **$n = 2$:**

- průměr a rozpětí (směrodatná odchylka není vhodná)
- dovolená diference D_r

$$D_r(\%) = \frac{x_1 - x_2}{(x_1 + x_2)/2} \times 100$$

- **$n = 3$:**

- Vhodnější je použít aritmetický průměr ze dvou bližších hodnot ($\bar{x}_D$) než aritmetický průměr ze všech 3 hodnot a interval vypočítat $L_{1,2} = \bar{x}_D \pm 4,30 \frac{s}{\sqrt{3}}$

- **$n = 2-3$:**

- starší postup podle Deana a Dixona, ve kterém R je rozpětí a K_n je tabelovaný koeficient

$$L_{1,2} = \bar{x} \pm K_n \cdot R$$

K_n :

n	$\alpha = 0,05$	$\alpha = 0,01$
2	6,4	31,8
3	1,3	3,0

20



Univerzita Palackého
v Olomouci

Interval spolehlivosti „malých“ souborů

– Hornův postup analýzy malých výběrů, $n = 4 - 20$ (pořádkové statistiky)

- Pořádkové statistiky (seřadit data podle velikosti),
- Vyčíslení hloubky pivotu:

$$H = \text{int} \frac{(n+1)/2}{2} \quad \text{pro } n \text{ liché}; \quad H = \text{int} \frac{\frac{n+1}{2} + 1}{2} \quad \text{pro } n \text{ sudé}$$

- Určení dolního pivotu $x_{\text{Dol}} = x_{(H)}$ a horního pivotu $x_{\text{Hor}} = x_{(n+1-H)}$
- Parametr polohy = **pivotová polosuma P_L** :

$$P_L = \frac{x_{\text{Dol}} + x_{\text{Hor}}}{2}$$

- Parametr rozptýlení = **pivotové rozpětí $R_L = x_{\text{Hor}} - x_{\text{Dol}}$**

$$L_{1,2} = P_L \pm R_L \cdot T_{H(1-\alpha/2;n)}$$

21



Univerzita Palackého
v Olomouci

Interval spolehlivosti mediánu a rozptylu

– Interval spolehlivosti mediánu:

- Je možné ho vypočítat několika způsoby, zde uveden neparametrický odhad, který je v QC Expertu a využívá mediánovou směrodatnou odchylku $s_{\tilde{x}}$:

$$L_{1,2} = \tilde{x}_{0,5} \pm s_{\tilde{x}} \cdot t_{(1-\frac{\alpha}{2}, n-1)}$$

$$\text{kde } s_{\tilde{x}} = \frac{x_{(n-k+1)} - x_k}{2 \cdot z_{\alpha/2}} \quad \text{a } k = \frac{n+1}{2} - (z_{\alpha/2} \cdot \sqrt{\frac{n}{4}})$$

– Interval spolehlivosti rozptylu:

- Meze oboustranného IS rozptylu σ^2 jsou dány pro dolní mez L_1 a pro horní mez L_2 , kde ve jmenovateli vzorců jsou kritické hodnoty rozdělení χ^2 (chí kvadrát):

$$L_1 = \frac{s^2 \cdot (n-1)}{\chi^2_{(1-\frac{\alpha}{2}, n-1)}} \quad L_2 = \frac{s^2 \cdot (n-1)}{\chi^2_{(\frac{\alpha}{2}, n-1)}}$$

22



Univerzita Palackého
v Olomouci

Intervalové odhady

- Intervaly spolehlivosti lze konstruovat jako:
 - oboustranné $\langle L_1, L_2 \rangle$, kdy se používá kvantil $(1 - \alpha/2)$,
 - jednostranné $\langle L_1, +\infty \rangle$ nebo $(-\infty, L_2 \rangle$, kdy se používá kvantil $(1 - \alpha)$.
- Pravostranný $(1 - \alpha)\%$ IS pro μ $-\infty \leq \mu \leq \bar{x} + z_{(1-\alpha)} \cdot \frac{\sigma}{\sqrt{n}}$
- Levostranný $(1 - \alpha)\%$ IS pro μ $\bar{x} - z_{(1-\alpha)} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq +\infty$
- **Př.:** výrobní podnik potřebuje s 95% spolehlivostí odhadnout horní hranici průměrné denní produkce. Pro náhodný výběr 36 dnů byl průměrný objem výroby 1150 a $\sigma = 312$ kusů.
- **Pozn.:** terminologie použitá v překladu normy ISO 3534: **confidence interval** = konfidenční interval (tedy ne interval spolehlivosti).

23



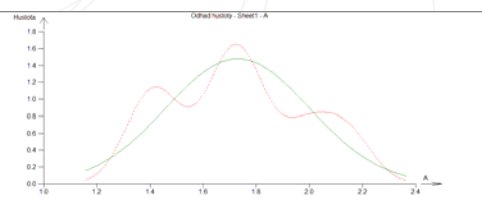
Univerzita Palackého
v Olomouci

Příklad: bodové a intervalové odhady polohy

- **DATA** ($n = 7$):

1,38 1,45 1,66 1,74 1,76 1,98 2,14

Parametr	Bodový odhad	Intervalový odhad (95 %)
průměr	1,73	1,48-1,98
medián	1,74	1,27-2,21
pivotová polosuma (Hornův postup)	1,72	1,33-2,01
10% uřezaný průměr	1,72	1,39-2,05



24