



**Národní informační středisko
pro podporu kvality**

Využití metody bootstrapping při analýze dat

II.část

Doc. Ing. Olga TŮMOVÁ, CSc.

- **Obsah**

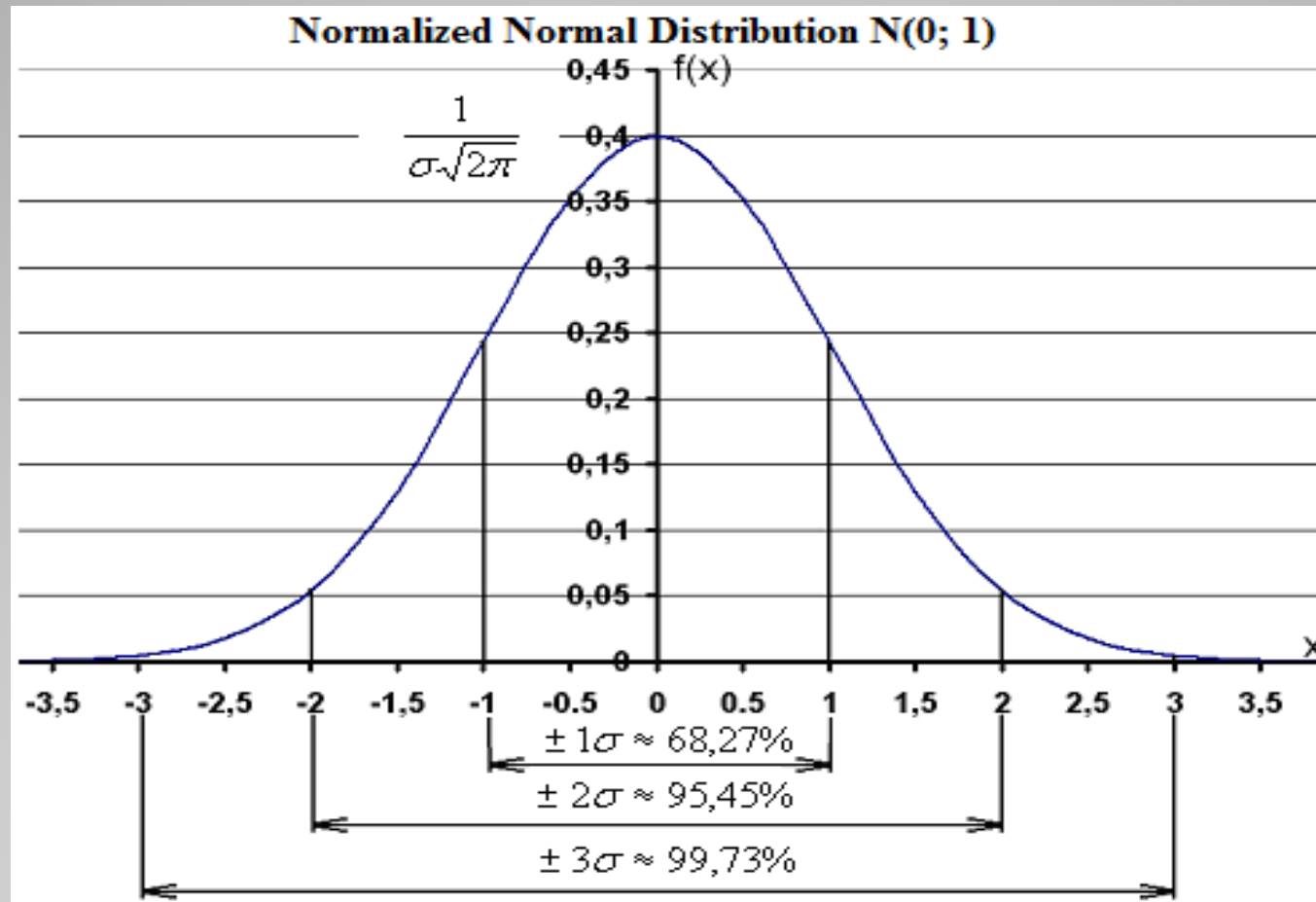
**Klasické procedury a statistické SW - metody výpočtů
konfidenčních intervalů:**

- **popisných statistik,**
- **indexů způsobilosti a výkonnosti.**

**Bootstrap - metoda alternativního odhadu
konfidenčních intervalů:**

- **popisných charakteristik,**
- **indexů výkonnosti.**

- Normované normální rozdělení $N(0,1)$



- Odhady statistik včetně konfidenčních intervalů
- klasické metody

$$\bar{x} - s \frac{t_{1-\alpha/2;(n-1)}}{\sqrt{n}} \leq \mu \leq \bar{x} + s \frac{t_{1-\alpha/2;(n-1)}}{\sqrt{n}}$$

medián \tilde{x}

$$\tilde{x} - s_{\tilde{x}} \cdot t_{1-\alpha/2,(n-1)} \leq \tilde{x} \leq \tilde{x} + s_{\tilde{x}} \cdot t_{1-\alpha/2,(n-1)}$$

- Odhady statistik včetně konfidenčních intervalů
- klasické metody

výběrová směrodatná odchylka s

$$s_{\tilde{x}} = \frac{\tilde{x}_{(n-k+1)} - \tilde{x}_{(k)}}{2 \cdot u_{\alpha/2}}$$

$$k = \frac{n+1}{2} - |u_{\alpha/2}| \cdot \sqrt{\frac{n}{4}}$$

u_p kvantil normovaného normálního rozdělení

- Odhady statistik včetně konfidenčních intervalů
- klasické metody

směrodatná odchylka

σ

$$\left[\frac{(n-1) s^2}{\chi^2_{\alpha/2; n-1}} \right]^{1/2} \leq \sigma \leq \left[\frac{(n-1) s^2}{\chi^2_{(1-\alpha/2); n-1}} \right]^{1/2}$$

- Indexy způsobilosti a výkonnosti

Při statistické regulaci technologického, výrobního nebo měřicího procesu je proces charakterizován souborem naměřených hodnot.

S ohledem na náklady při produkci, neshody apod., je proces nutno udržet ve stabilním (predikovatelném) stavu.

- **Indexy způsobilosti a výkonnosti**

Hodnoty naměřené v regulovaném procesu:

- **mají být centrovány mezi horní a dolní specifikační (toleranční) mezí (USL, LSL),**
- **v datech se nemá vyskytovat žádné seskupení dat podle vzoru systematických (vymezitelných) příčin (viz ČSN ISO 8258).**

Jak dobře hodnoty odpovídají specifikacím, je vyjádřeno indexy způsobilosti a výkonnosti.

Pro vyšší hodnoty indexů je nutná redukce variability.

- **Indexy způsobilosti a výkonnosti**

S ohledem na způsob získání/záznamu dat rozlišujeme:

- Krátkodobá variabilita (inherentní, vnitroskupinová) – odhad z krátkého časového období, obsahuje minimální množství rušivých vlivů a procesních posuvů.

Dlouhodobá variabilita (celková, meziskupinová) - může obsahovat celý proces měření s možnými změnami a posuny.

- Indexy způsobilosti a výkonnosti

Volba správného typu variability je zásadní pro správné určení ukazatele:

$$C_P = \frac{USL - LSL}{6s}$$

kde s je výběrová směrodatná odchylka uvnitř podskupin,

$$P_P = \frac{USL - LSL}{6s_{TOT}}$$

kde s_{TOT} je výběrová směrodatná odchylka mezi podskupinami.

- Indexy způsobilosti a výkonnosti

Informace o centrování procesu - kritické indexy.

Soubor naměřených hodnot představuje pouze výběr z celé populace, proto výpočty všech indexů musí být doplněny konfidenčními intervaly.

- Index výkonnosti

Příklad odhadu intervalů

$$\hat{P}p \sqrt{\frac{\chi^2_{1-\alpha/2;n-1}}{n-1}} \leq Pp \leq \hat{P}p \sqrt{\frac{\chi^2_{\alpha/2;n-1}}{n-1}}$$

- Kritický index výkonnosti

$$P_{pk} = \min \left[\frac{USL - \mu}{3\sigma} ; \frac{\mu - LSL}{3\sigma} \right]$$

- **Příklad použití bootstrap**

Pro ověření této metody byl zvolen:

Základní soubor - cca 10 mil.hodnot (generátor náhodných čísel – SW Statistica),

Výběr - 30 000 generovaných simulačních podskupin

programovací jazyk C++

- **Příklad použití bootstrap**

Ze zdrojového souboru byly generovány podskupiny metodou náhodných výběrů.

Požadovaná statistika:

- **průměr,**
- **medián,**
- **směrodatná odchylka,**
- **index výkonnosti**

je určena z každé podskupiny a tyto výsledky jsou

vzestupně seříděny.

- Ilustrativní příklad řazení náhodných vzorků v podskupinách ($n = 5$)

Vstupní data	1	2	3	4	5	Průměr	Setřídění průměrů
1.Podskupina	1	3	2	1	5	2,4	2,2
2.podskupina	5	2	2	4	3	3,2	2,4
3.podskupina	4	1	3	1	2	2,2	3,2
atd.

- Počet dat pro hodnocení procesu - bootstrap

$$B = 40 \cdot n$$

n počet vstupních dat v jednotlivých
podskupinách

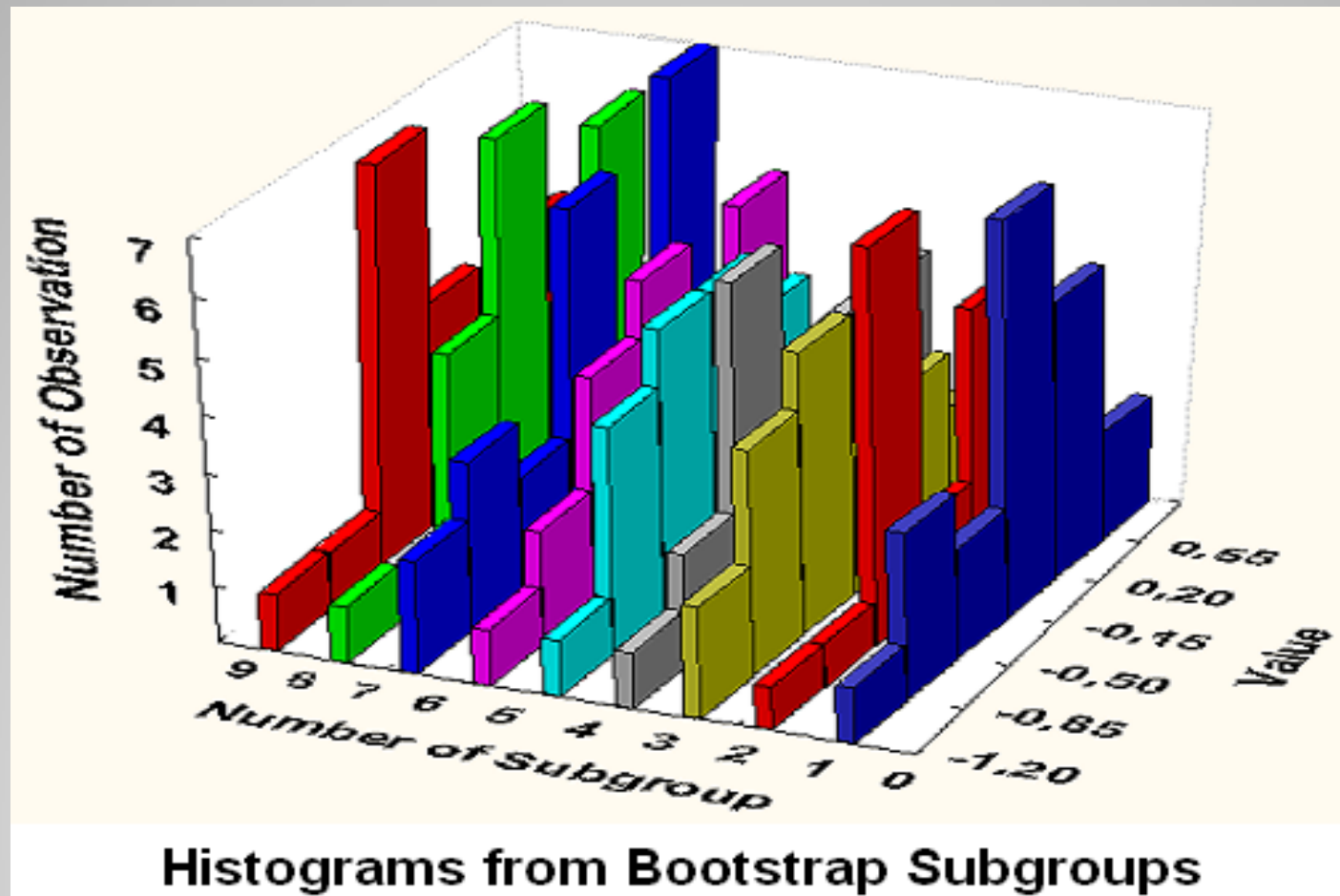
B počet simulačních podskupin
(bootstrap)

- Příklad použití bootstrap

**Grafická ilustrace generovaných podskupin -
následující obr. odhad hustoty pravděpodobnosti
pomocí histogramů z prvních devíti generovaných
podskupin.**

**Pozn.: Z důvodu náhodného převzorkování je každý
histogram poněkud odlišný od ostatních.**

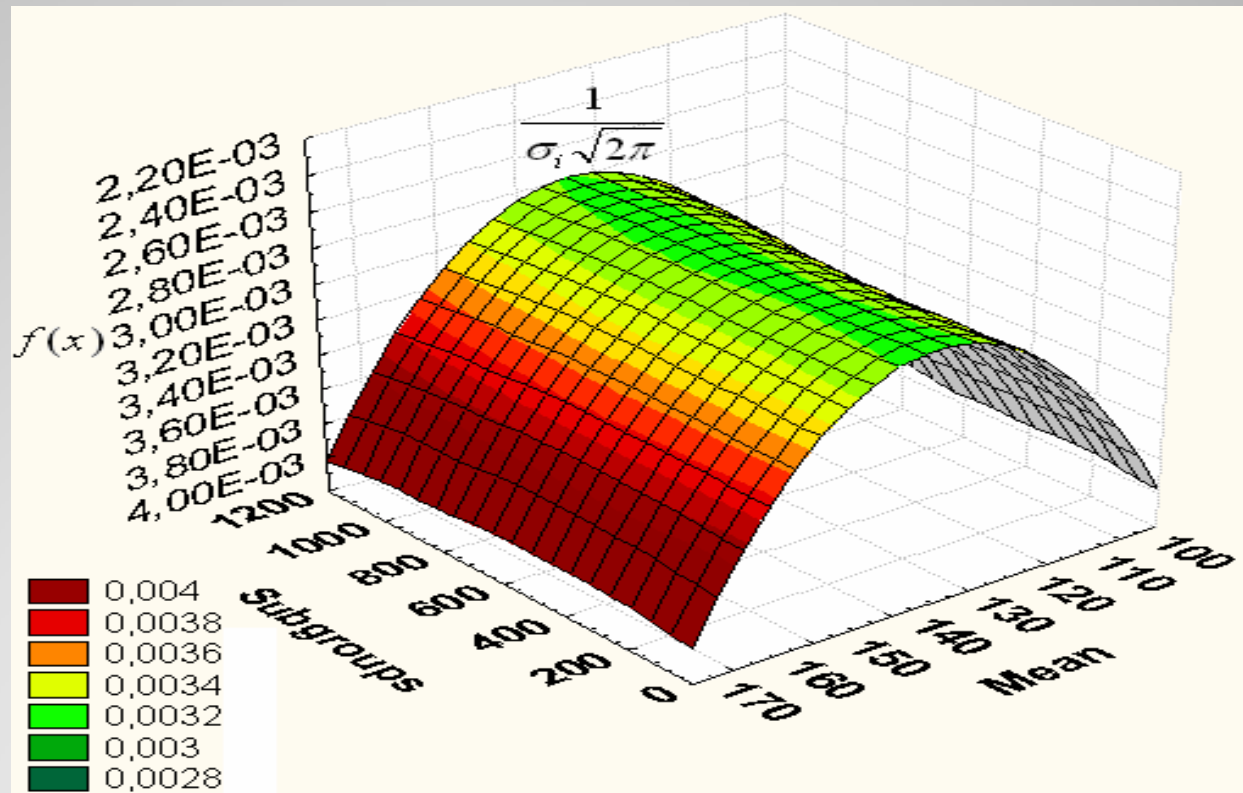
- Příklad použití bootstrap
 $n = 20, B = 800$



- Příklad použití bootstrap

$n = 100, B = 1200$

- Grafické hodnocení - SW STATISTICA 6.0



- **Příklad použití bootstrap**

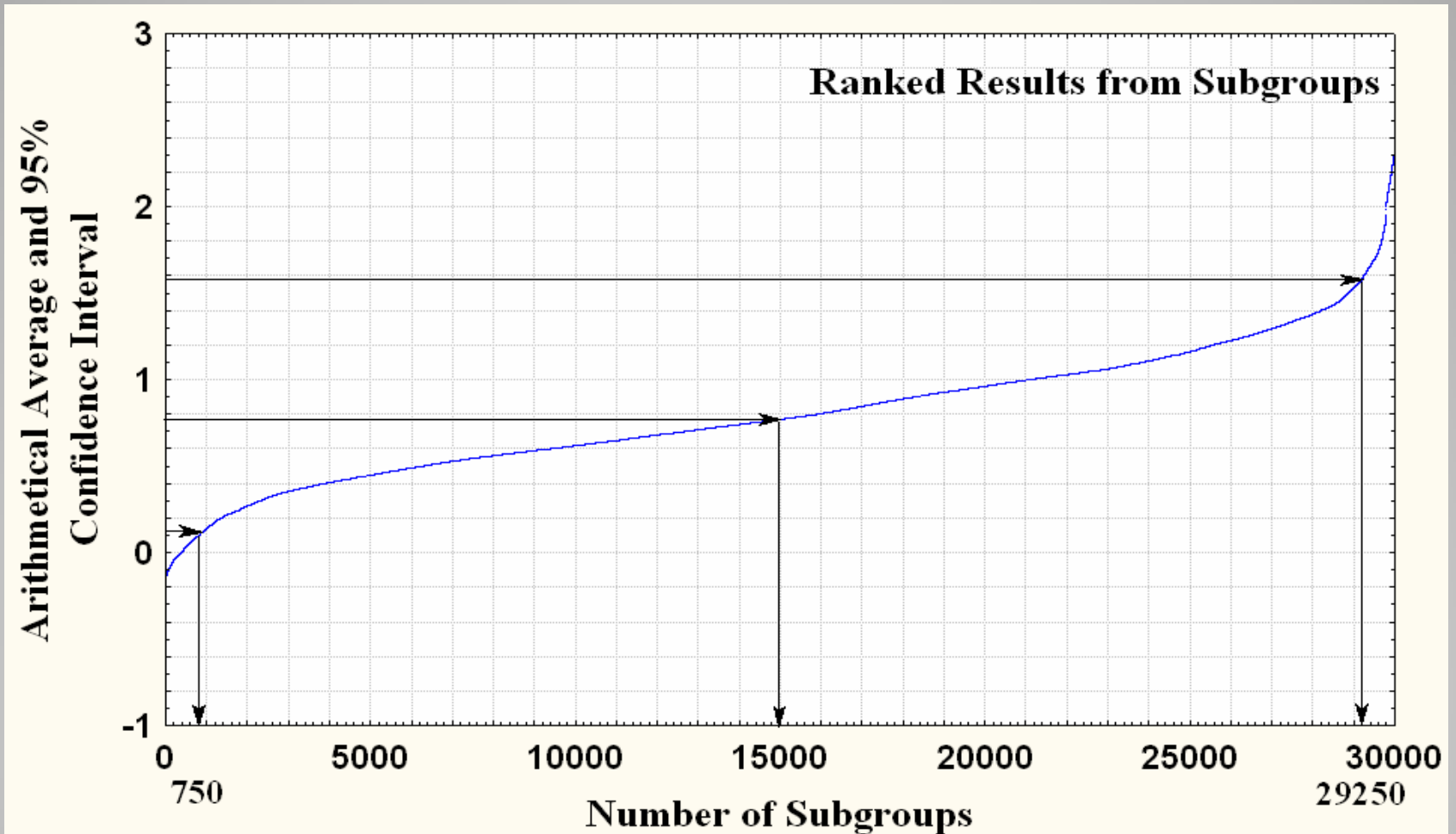
Vzestupné setřídění je uvedeno na obr., které ilustrují odhad aritmetického průměru (mediánu, směrodatné odchylky) a 95% konfidenčního intervalu.

Použito 30 000 generovaných podskupin.

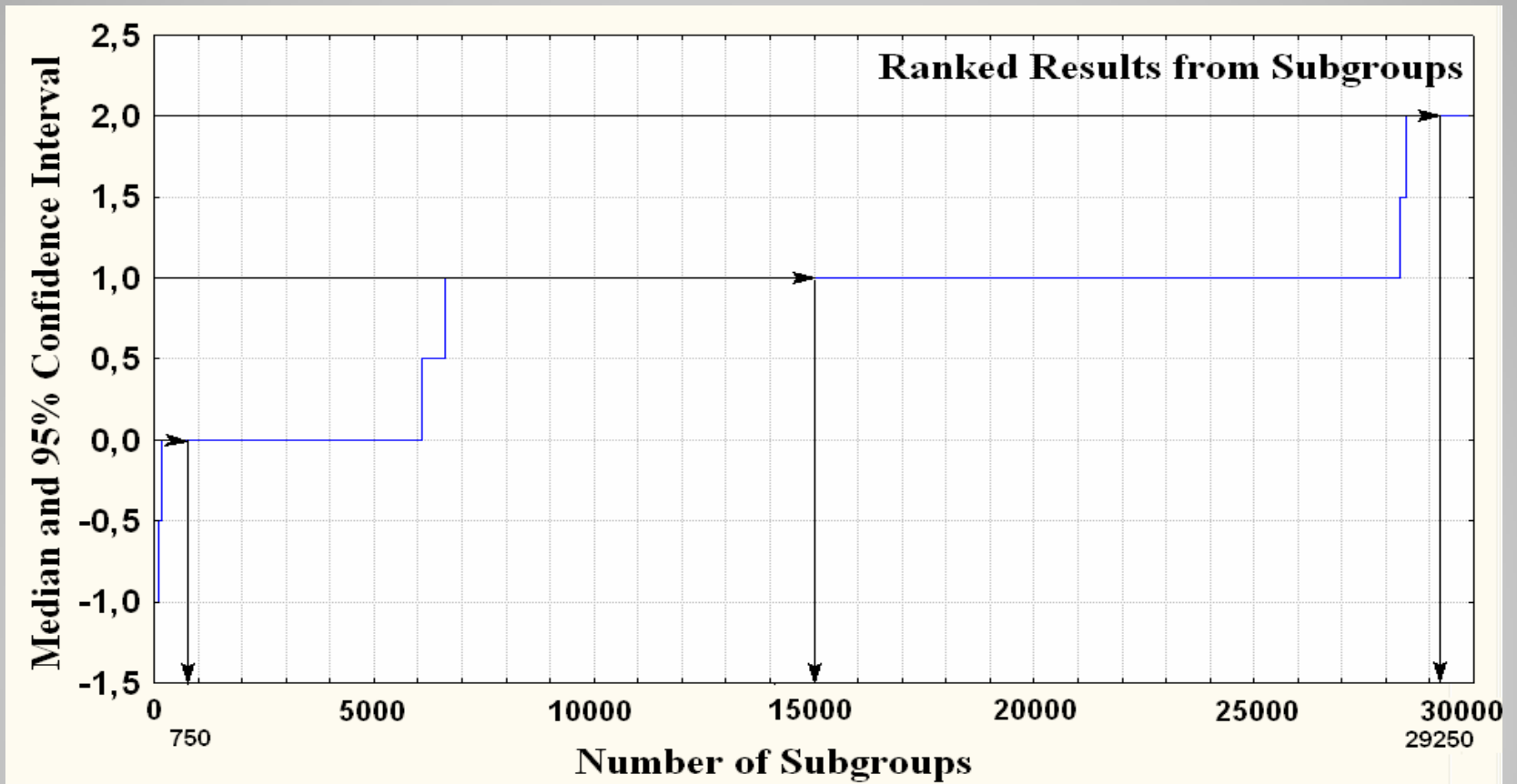
95% konfidenční interval - leží v rozmezí setříděných hodnot od 750. podskupiny do 29250. podskupiny.

Centrální hodnota průměru leží pak uprostřed tohoto intervalu.

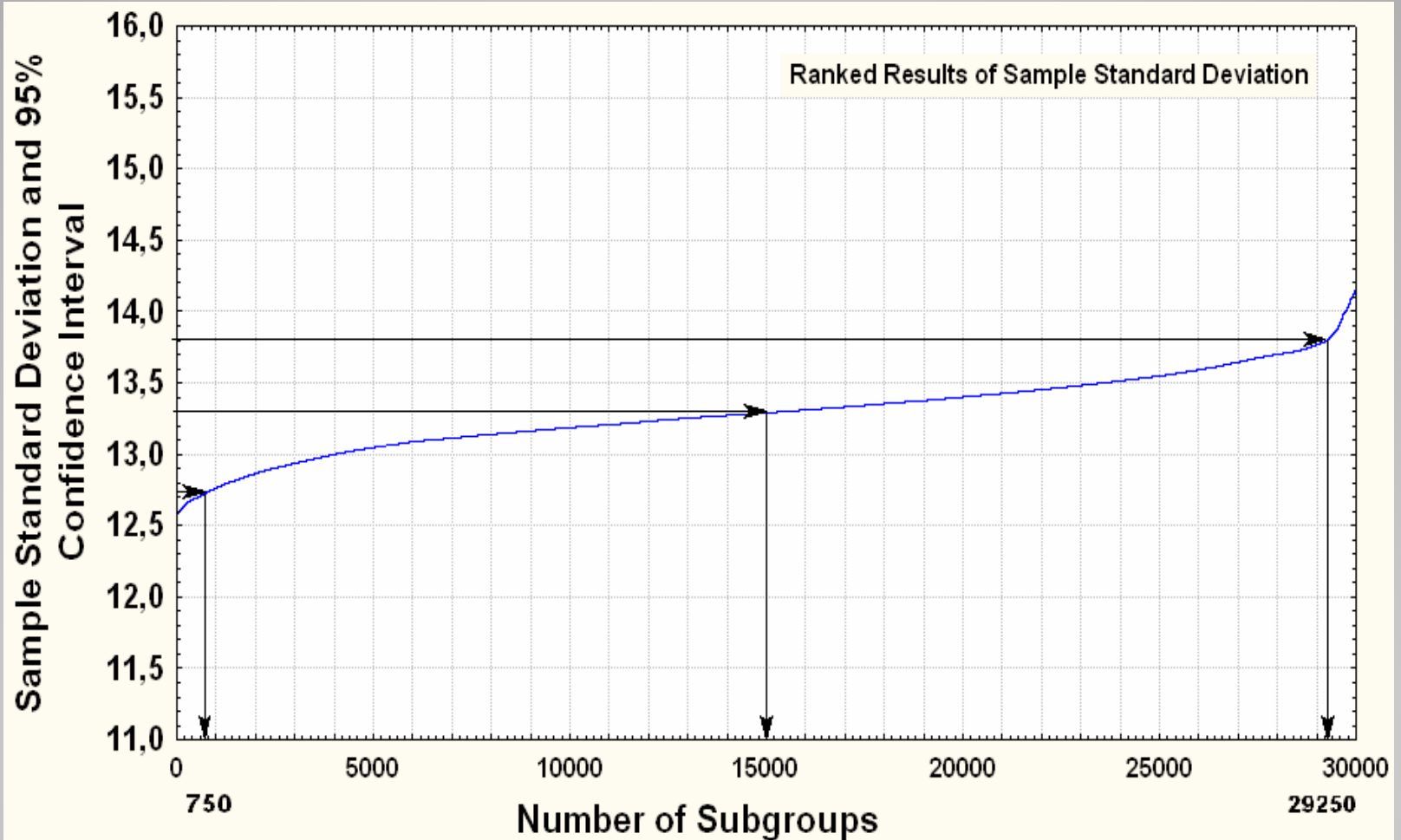
- Příklad použití bootstrap



- Příklad použití bootstrap



- Příklad použití bootstrap



- Příklad použití bootstrap

Ukázka shody výsledků mezi metodou bootstrappingu a tradičním odhadem pro vybrané statistické charakteristiky:

- průměr
- medián
- směrodatná odchylka

- Příklad použití bootstrap

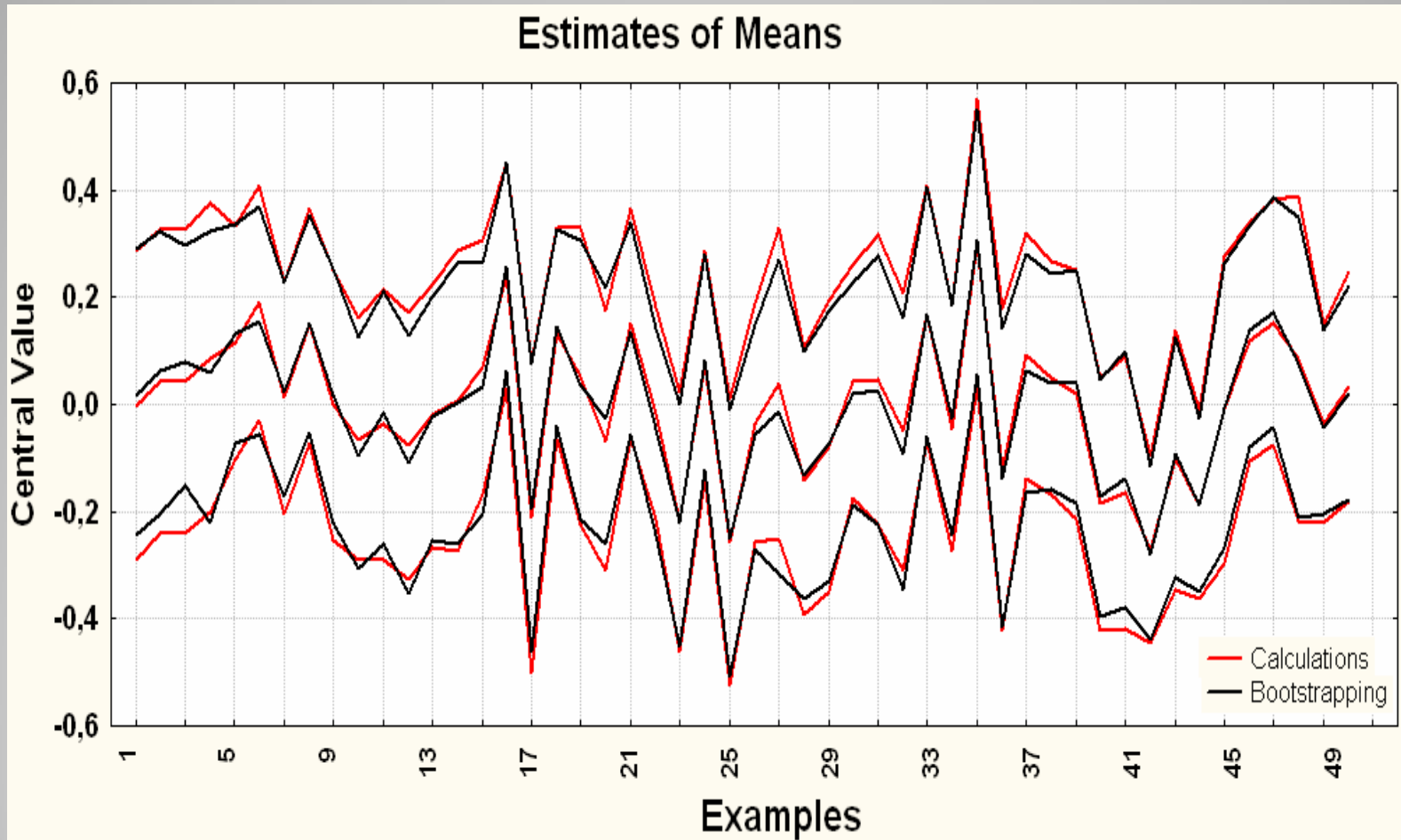
Ověření správné funkce programu v C++ na jednoduchých příkladech

	SW Calculations			Bootstrapping		
Parameter	Value	Lower	Upper	Value	Lower	Upper
Mean	0,799	0,026	1,624	0,798	0,073	1,602
Median	1,000	-0,001	2,001	1,000	0,000	2,000
St. Dev.	13,290	12,732	13,900	13,298	12,725	13,812

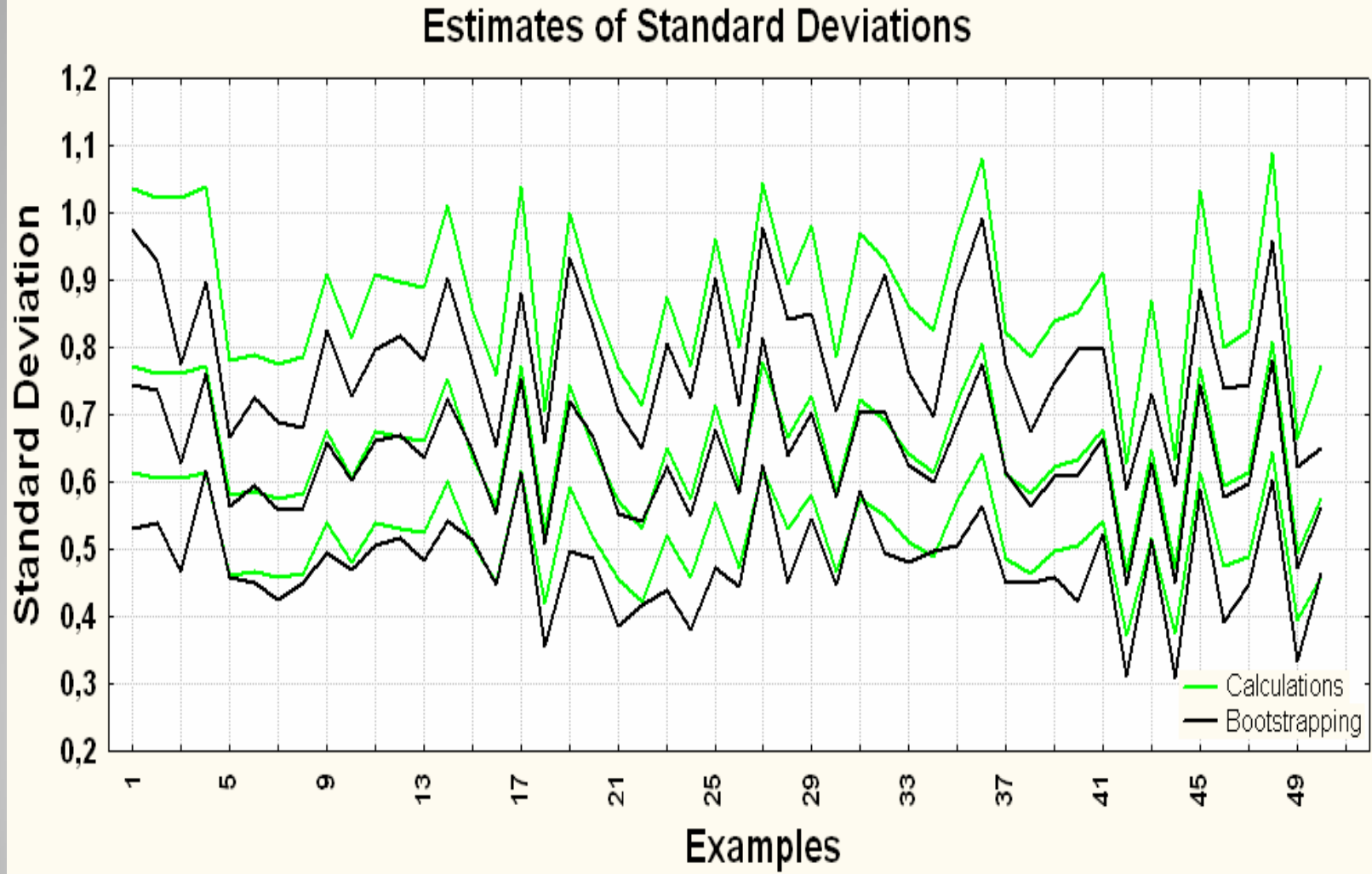
- Příklady použití bootstrap

- Grafické porovnání odhadů parametrů a konfidenčních intervalů
- metoda bootstrap – černé značení
- Klasická metoda – barevné značení

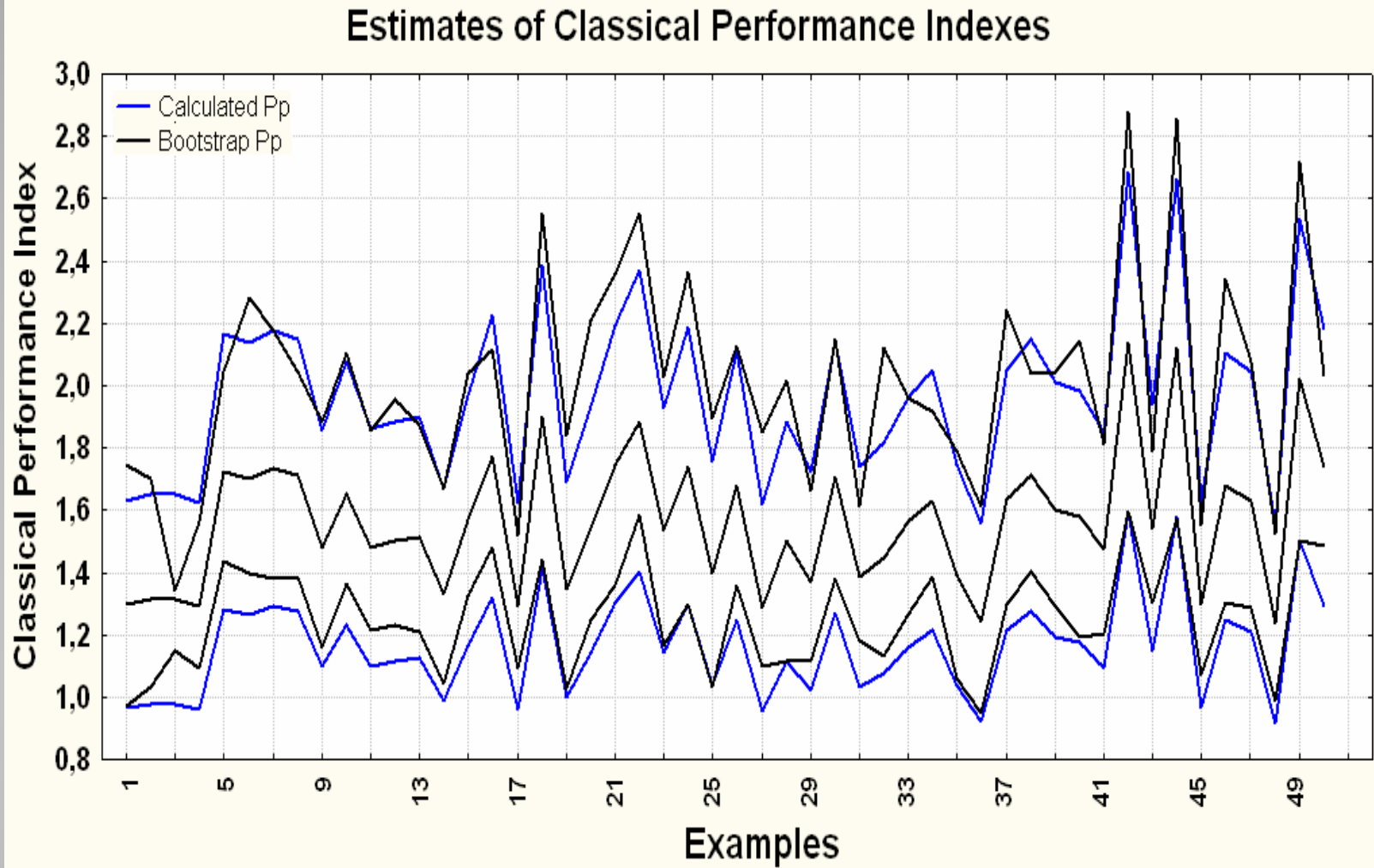
- Odhad průměrů a konfidenčních intervalů



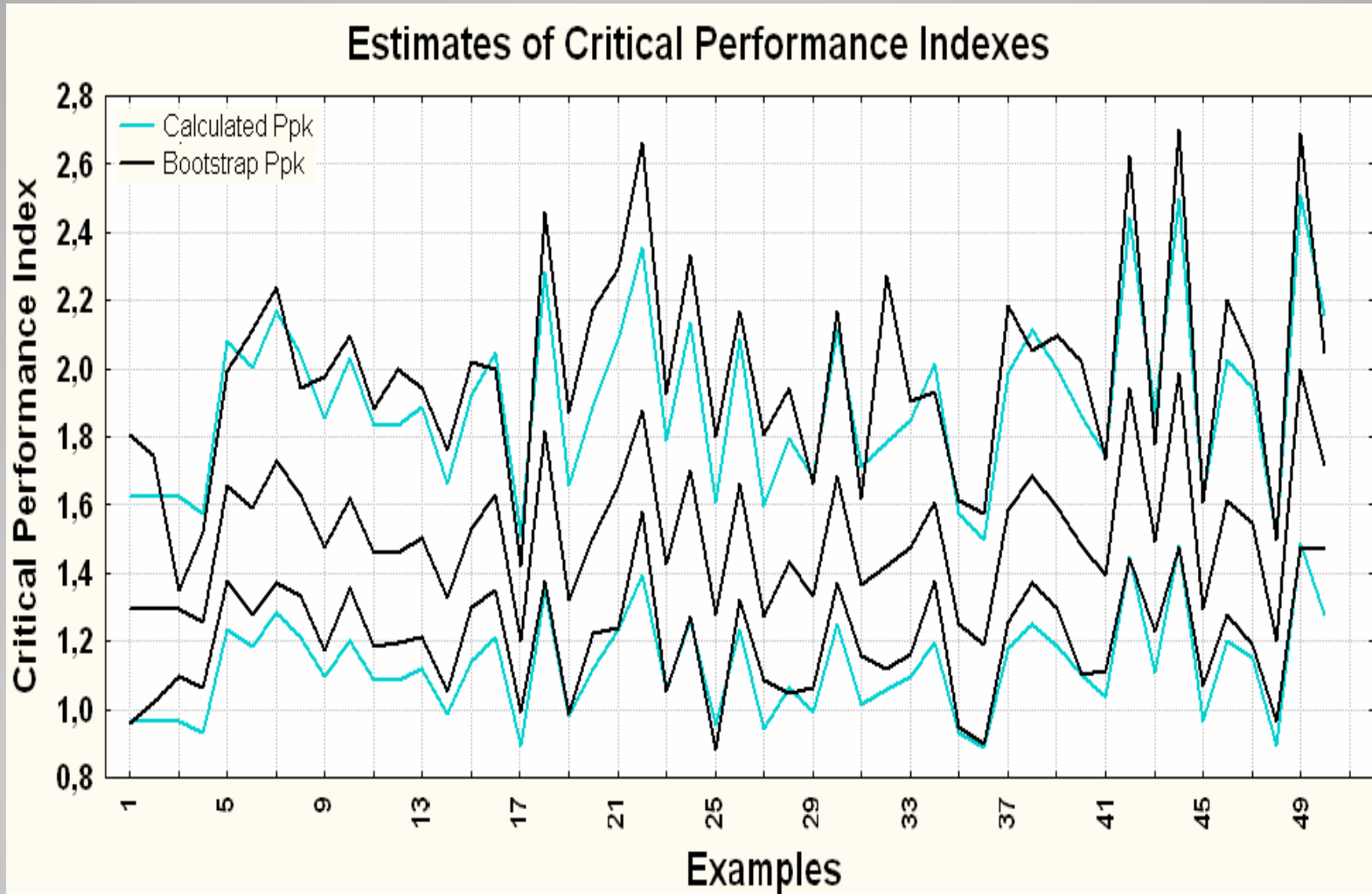
- Odhad směrodatné odchylky



- Odhad indexu výkonnosti

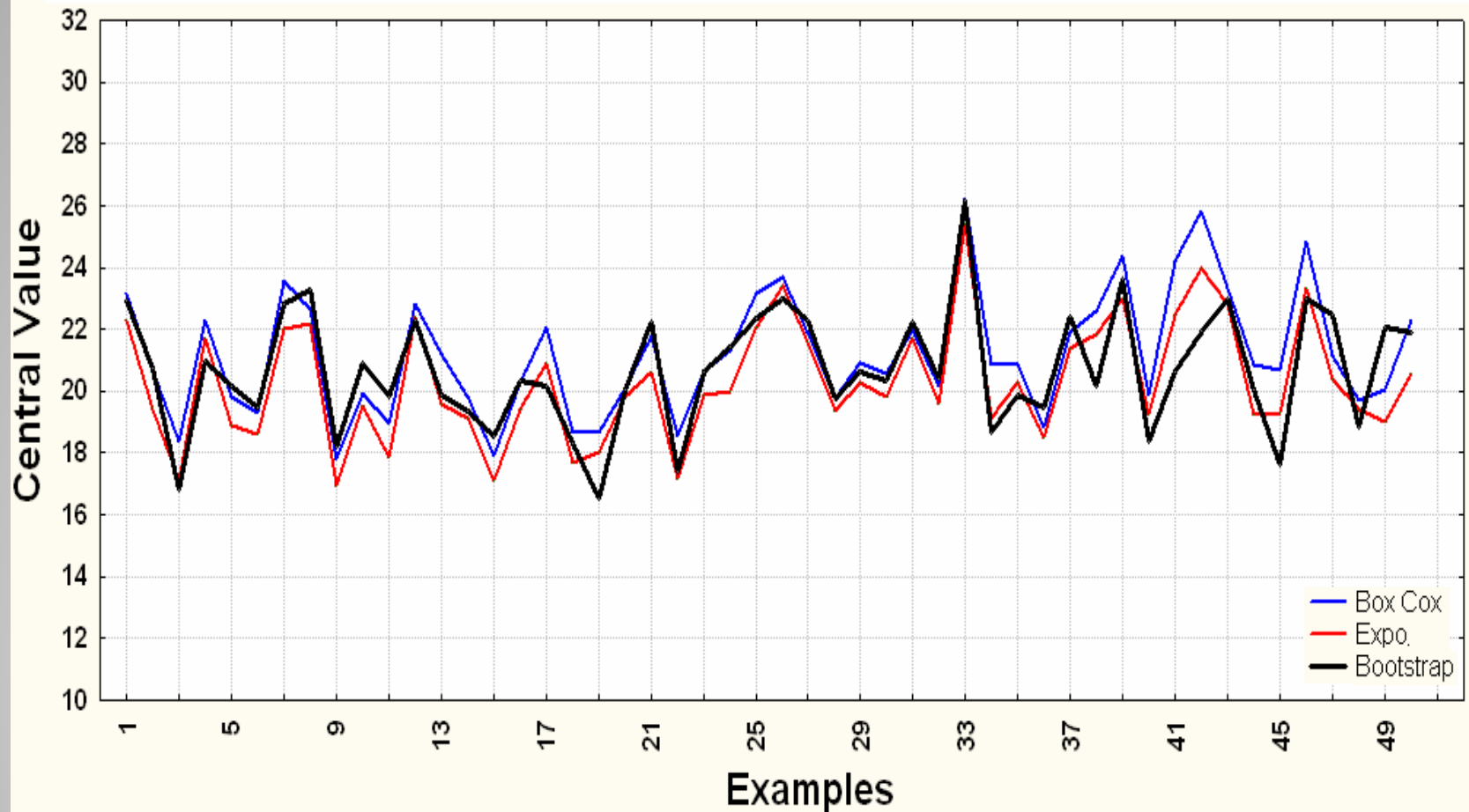


- Odhad kritického indexu výkonnosti

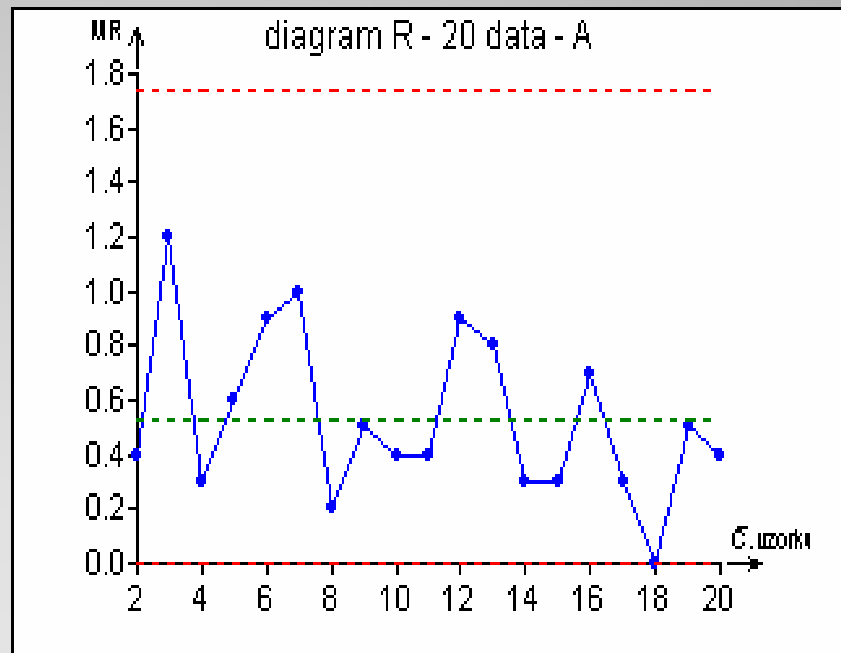
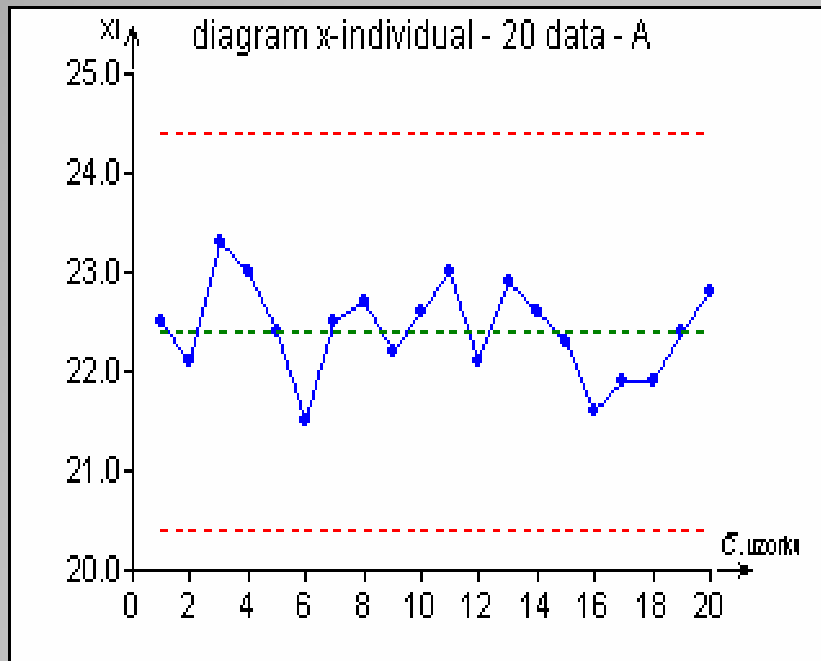


- Odhad střední hodnoty

Central Values of the Bootstrap Simulations, Box-Cox and Exponential Transformations



- Příklad - Regulační diagram měřicího procesu



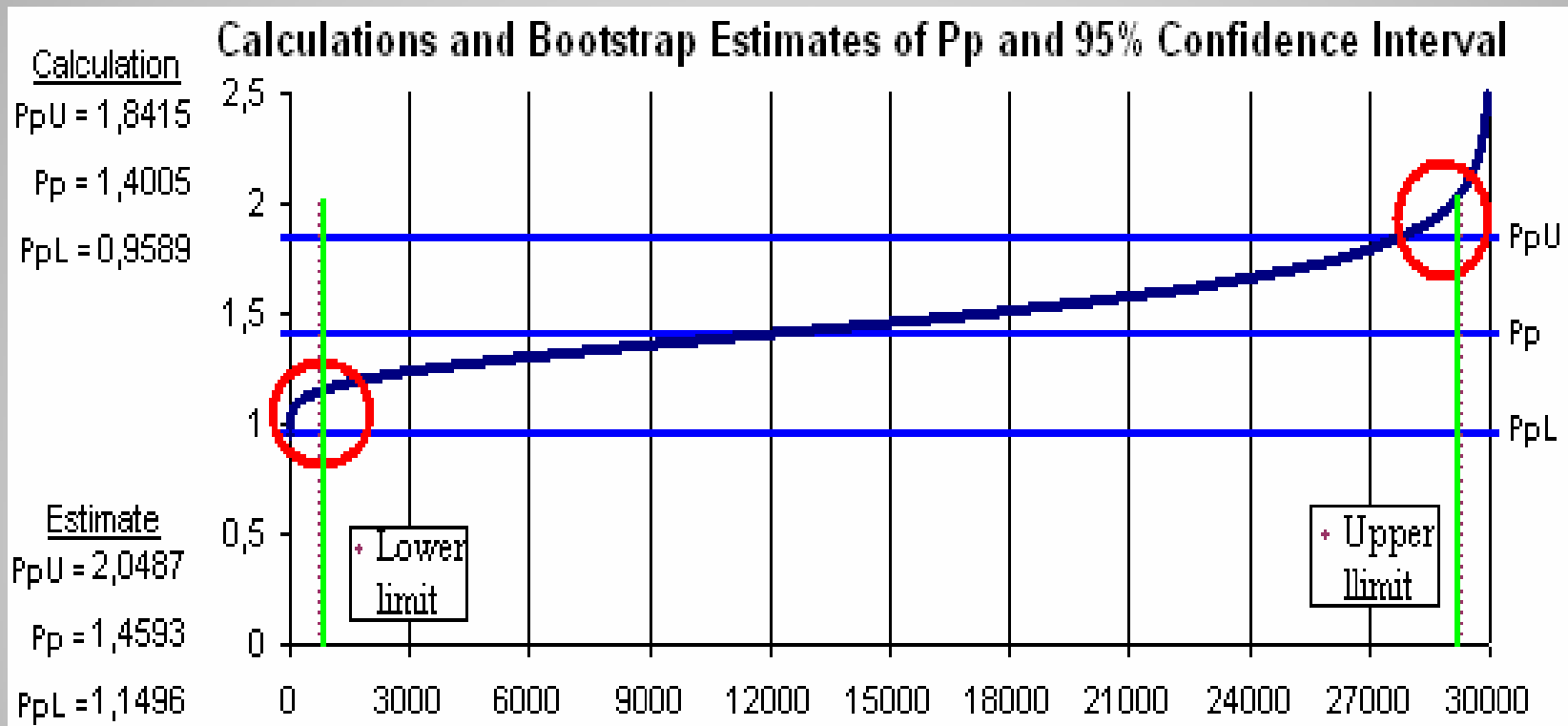
- Porovnání výsledků obou metod

Calc. of Descriptive Statistics			
Char.	Value	Lower	Upper
Mean	22,4150	22,1922	22,6378
Median	22,4500	22,1955	22,7045
St.dev	0,4760	0,3620	0,6953
Pp	1,4005	0,9589	1,8415
Ppk	1,3900	0,9480	1,8319

Estimates from Simulations			
Char.	Value	Lower	Upper
Mean	22,4350	22,2300	22,6300
Median	22,5000	22,2000	22,7000
St.dev	0,4568	0,3253	0,5799
Pp	1,4005	1,0897	1,8191
Ppk	1,3900	1,0907	1,9148

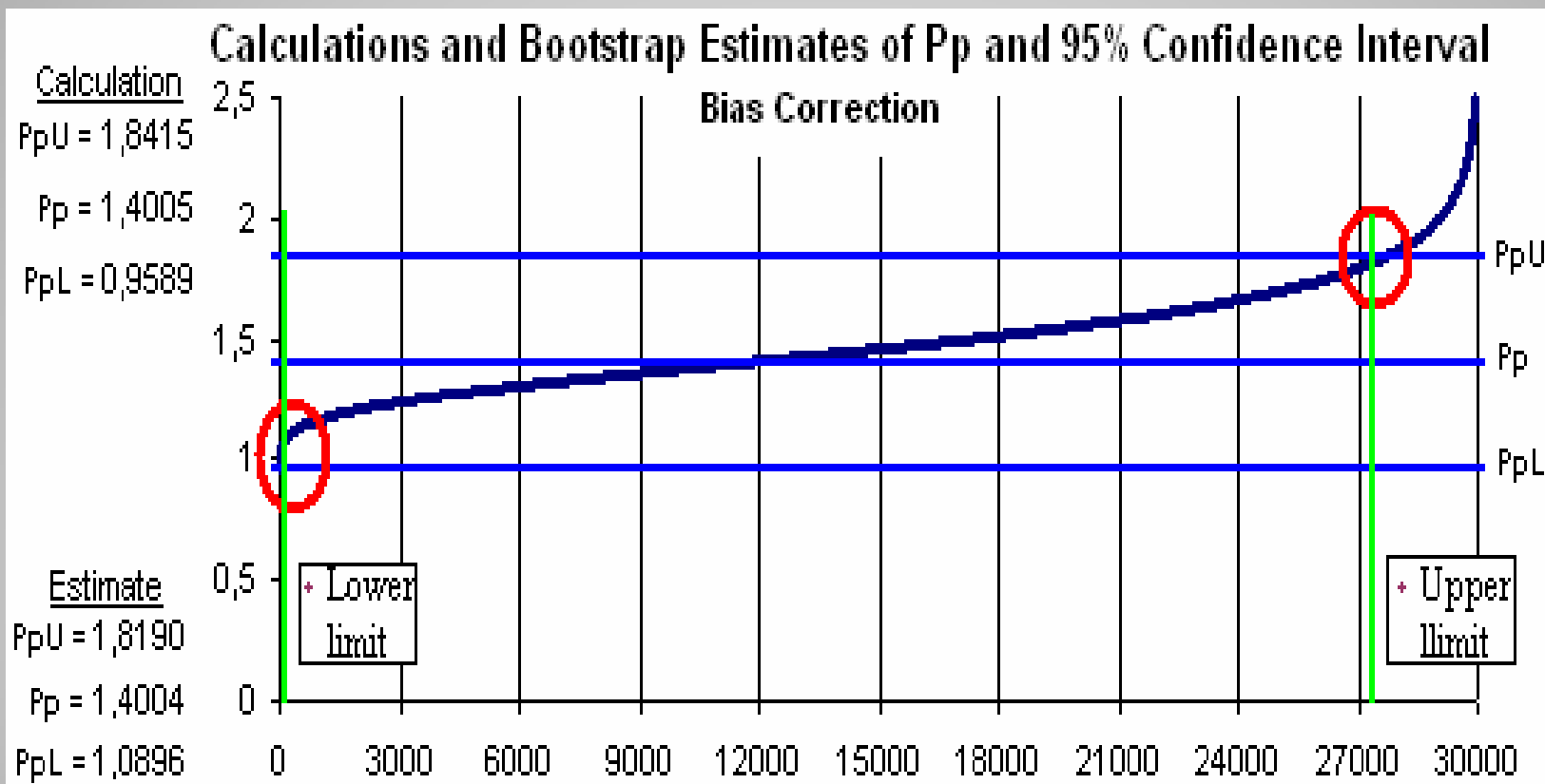
- Index Pp a konfidenční intervaly

(bez korekce)



- Index Pp a kondidenční intervaly

(po Efronově korekci)



- Hlavní faktory odlišnosti výsledků

Rozdíl mezi klasickým výpočtem a metodou bootstrap:

- způsobí nepřesnost výsledků v odhadu parametrů souboru,
- je ovlivněn špičatostí základního souboru dat (viz Tab. Korelace mezi rozdílem výsledků a špičatostí souboru dat),
- odlehlými hodnotami,
- vyšší počet vstupních dat ovlivní lepší shodu výsledků.

- Hlavní faktory odlišnosti výsledků

Difference between simulations and computations	$r=f(\text{Kurt.})$
PpkU (Bootstrap) - PpkU (Computation)	0,86
PpkU (Corrected Bootstrap) - PpkU (Computation)	0,72
PpkL (Corrected Bootstrap) - PpkL (Computation)	-0,85
PpkL (Bootstrap) - PpkL (Computation)	-0,67

- **Výhody simulačních metod:**

- **metoda bootstrap je jednoduchá a intuitivní,**
- **poskytuje výsledky, které jsou velmi blízké tradičním odhadům,**
- **poskytuje relevantní odhady popisných statistik, indexů výkonnosti a konfidencích intervalů i pro malé soubory dat,**
- **je velmi vhodná v případech, kdy nahrazuje komplikované výpočty.**

- **Nevýhody (1):**

-velký počet (tisíce) simulací mohou být časově náročné,

- v případě odhadů klasických a kritických indexů výkonnosti je nutná korekce strannosti,

- chyba metody bootstrap je rozdíl mezi skutečnou a simulovanou distribucí hodnot.

- **Nevýhody (2):**

Chyba metody sestává ze dvou základních složek:

- **statistická chyba (závisí na počtu vstupních dat a na jejich správnosti), nelze ji touto metodou eliminovat,**
- **chyba simulací (nedostatečná náhodnost, nedostatek podskupin), lze redukovat zvýšením počtu generovaných podskupin,**
- **Bootstrapping není rezistentní vůči odlehlým hodnotám.**

• Literatura:

- [1] Breyfogle F.: *Implementing Six Sigm.*, New Jersey 2003.
- [2] Burget D., Tůmová O., Čtvrtník V.: *Bootstrapping Technique and Confidence Intervals*. In Proceedings of the 12th IMEKO TC1-TC7 Joint Symposium on Man, Science & Measurement. Annecy: Université de Savoie, 2008. s. 125-132. ISBN 2-9516453-8-4.
- [3] Herout P.: *Učebnice jazyka C*, III. upravené vydání, České Budějovice 2001.
- [4] Kupka K.: *Statistické řízení jakosti*. TriloByte Ltd., Pardubice 1997.
- [5] Matyáš V.: *Měření, analýza a vytváření náhodných procesů*. SNTL Praha 1976.
- [6] Michálek J.: *Vyhodnocování způsobilosti a výkonnosti výrobního procesu*. CQR, Praha 2009.
- [7] Reif J.: *Metody matematické statistiky*. Vydavatelství ZČU Plzeň 2000.
- [8] Tůmová O., Pirich D.: *Nástroje řízení jakosti a základy technické diagnostiky*. Vydavatelství ZČU Plzeň 2003.
- [9] Tůmová O., Veverková M., Burget D.: *Měřicí procesy a analýza dat při sledování interaktivních dějů v elektrotechnice*, dílčí zpráva pro MSM 4977751310, FEL ZČU Plzeň 2008.
- [10] ČSN ISO 8258 *Shewhartovy regulační diagramy*.

- Závěr
-

Děkuji

za pozornost

O. Tůmová