Národní informační středisko pro podporu kvality

Využití metody bootstrapping při analýze dat

Eva Jarošová

18. listopadu 2010

Použití

- Určení přesnosti odhadu neznámých charakteristik
- Výpočet konfidenčních mezí pro neznámou charakteristiku
- Testování hypotéz

Využití naměřených dat a počítače k simulaci neznámého výběrového rozdělení

- Nesplňují-li data předpoklad normálního rozdělení
- Neznáme-li výběrové rozdělení odhadu
- Nemůžeme-li využít centrální limitní věty

Příklad simulace

Schéma výběru s opakováním

1	1	1	4
2	2	5	3
3	4	3	2
4	4	5	1
5	2	1	1

Naměřené hodnoty	Simulované výběry			
	1	2	3	…
6,1	6,1	6,1	6,6	
6,2	6,2	6,9	6,5	
6,5	6,6	6,5	6,2	
6,6	6,6	6,9	6,1	
6,9	6,2	6,1	6,1	
6,46	6,34	6,50	6,30	…

simulované rozdělení průměrů

Konfidenční intervaly

Metody

- Podle výpočtu konfidenčních mezí
  - percentilový, s korekcí na zkreslení a s akcelerací, studentizovaný, základní; standardní (normální)
- Podle způsobu simulace neznámého rozdělení
  - neparametrický či parametrický bootstrap
- Podle výběru vzorků
  - obyčejný, vyvážený, s klouzavými bloky

Značení

θ odhadovaná charakteristika

$\hat{\theta}$ odhad charakteristiky na základě n naměřených hodnot

$\hat{\theta}_i^*$ odhad charakteristiky na základě n hodnot v i-tém simulovaném výběru

$\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \ldots, \hat{\theta}^*_{(B)}$ simulované rozdělení (hodnoty seřazené vzestupně)

$\left(\hat{\theta}^*_{(k_L)}; \hat{\theta}^*_{(k_U)}\right)$ konfidenční interval pro θ

Příklad – odhad střední hodnoty

μ odhadovaná charakteristika

$\hat{\mu} = \bar{x}$ odhad střední hodnoty, výběrový průměr (z n naměřených hodnot)

Rozdělení hodnot X není normální, malý rozsah výběru, nemůžeme uplatnit centrální limitní větu

$\hat{\mu}_i^* = \bar{x}_i^*$ průměr v i-tém simulovaném výběru

$\bar{x}_{(1)}^*, \bar{x}_{(2)}^*, \ldots, \bar{x}_{(B)}^*$ simulované rozdělení

$\left(\bar{x}_{(k_L)}^*; \bar{x}_{(k_U)}^*\right)$ konfidenční interval pro μ

Vlastnosti odhadu

- Zkreslení odhadu $E(\hat{\theta}) - \theta$

odhad $\frac{1}{B}\sum_{i=1}^{B}(\hat{\theta}_i^* - \hat{\theta})$

- Směrodatná chyba odhadu $\sqrt{E[\hat{\theta} - E(\hat{\theta})]^2}$

odhad $s(\hat{\theta}^*) = \sqrt{\frac{\sum_{i=1}^{B}(\hat{\theta}_i^* - \overline{\theta^*})^2}{B-1}}$ $\quad \overline{\theta}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$

Percentilový interval

$$\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \ldots, \hat{\theta}^*_{(B)}$$

Zvolená konfidence 1 - α

$$k_L = \left\lfloor \frac{\alpha}{2}(B+1) \right\rfloor \qquad k_U = (B+1) - k_L$$

$\lfloor x \rfloor$ Největší celé číslo menší nebo rovné x

Např. pro konfidenci 1 - α = 0,95 a B = 999 $\left(\hat{\theta}^*_{(25)}; \hat{\theta}^*_{(975)}\right)$

Jednoduchý, dobře funguje u symetrických rozdělení
U nesymetrických rozdělení pokrytí neodpovídá deklarované konfidenci

Základní interval

$$\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \ldots, \hat{\theta}^*_{(B)}$$

$$\hat{\theta}^*_{(k_L)}; \hat{\theta}^*_{(k_U)}$$

$$\hat{\theta}^*_{(k_L)} - \hat{\theta} < \hat{\theta} - \theta < \hat{\theta}^*_{(k_U)} - \hat{\theta}$$

$$2\hat{\theta} - \hat{\theta}^*_{(k_U)} < \theta < 2\hat{\theta} - \hat{\theta}^*_{(k_L)}$$

$$\left(2\hat{\theta} - \hat{\theta}^*_{(k_U)};\quad 2\hat{\theta} - \hat{\theta}^*_{(k_L)}\right)$$

BCa interval

$$\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \ldots, \hat{\theta}^*_{(B)}$$

p_0 empirická distribuční funkce v bodě $\hat{\theta}$

podíl hodnot $\hat{\theta}^*_i$ menších než $\hat{\theta}$

$\hat{z}_0 = \Phi^{-1}(p_0)$ korekce zkreslení

$\hat{a}$ akcelerace, korekce nekonstantní směrodatné chyby, odhad viz Efron, Tibshirani (1993)

$$p_L = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{\alpha/2})} \right)$$

$$p_U = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{1-\alpha/2})} \right)$$

$$k_L = \lfloor p_L(B+1) \rfloor$$

$$k_U = (B+1) - k_L$$

$$\left(\hat{\theta}^*_{(k_L)}; \hat{\theta}^*_{(k_U)} \right)$$

$z_{\alpha/2}$ kvantil normovaného normálního rozdělení

- Odhad koeficientu zkreslení z_0 a koeficientu akcelerace a neparametricky či parametricky (za předpokladu určitého rozdělení
- Úprava znamená změnu pořadí k_L a k_U pro určení percentilů
- Zachovává obor hodnot parametrů
- S rostoucím n se pokrytí blíží stanovené konfidenci rychleji než u předešlých (je však třeba minimálně B = 1000)

Studentizovaný (t)

Založen na jiném simulovaném rozdělení

$\hat{\theta}_i^*$

$$t^* = \frac{\hat{\theta}_i^* - \hat{\theta}}{s(\hat{\theta}_i^*)}$$

simulované rozdělení $t^*_{(1)}, t^*_{(2)}, \ldots, t^*_{(B)} \longrightarrow t^*_{(k_L)} \quad t^*_{(k_U)}$

$$\left(\hat{\theta} - t^*_{(k_L)} s(\hat{\theta}^*);\ \hat{\theta} + t^*_{(k_U)} s(\hat{\theta}^*)\right)$$

$$s(\hat{\theta}^*) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^* - \overline{\theta^*})^2}$$

Nevýhody

- Problém při odhadu směrodatné chyby v každém simulovaném vzorku

  Není-li k dispozici vzorec pro odhad směrodatné chyby, použije se znovu bootstrap na každý ze simulovaných vzorků

  $$s(\hat{\theta}_i^*) = \sqrt{\frac{\sum_{k=1}^{b} (\hat{\theta}_{ik}^* - \overline{\theta_i^*})^2}{b-1}}$$

  Např. pro každý z 1000 vzorků 25 vzorků pro odhad směrodatné chyby – celkem 25000
- Může být příliš široký a obsahovat i nepřípustné hodnoty charakteristiky
- Nerespektuje transformaci (záleží na stupnici měření)

Standardní (normální) interval

$$\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \ldots, \hat{\theta}^*_{(B)}$$

$$\overline{\theta^*} = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}^*_i$$

$$s(\hat{\theta}^*) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}^*_i - \overline{\theta^*})^2}$$

$$\left(\overline{\theta^*} - z_{1-\alpha/2}s(\hat{\theta}^*); \overline{\theta^*} + z_{1-\alpha/2}s(\hat{\theta}^*)\right)$$

Nezaručuje dodržení oboru hodnot

Příklady

Interval pro střední hodnotu

Pro velký výběr

$$\left(\bar{x} - z_{1-\alpha/2} \frac{s}{\sqrt{n}}; \bar{x} + z_{1-\alpha/2} \frac{s}{\sqrt{n}} \right)$$

$$P\left(\bar{x} - z_{1-\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2} \frac{s}{\sqrt{n}} \right) = 1 - \alpha$$

Platí přibližně, nemají–li průměry normální rozdělení.

Percentilový interval

$$\bar{x}^*_{(1)}, \bar{x}^*_{(2)}, \ldots, \bar{x}^*_{(B)}$$

$$k_L = \left\lfloor \frac{\alpha}{2}(B+1) \right\rfloor \qquad k_U = (B+1) - k_L$$

$\lfloor x \rfloor$ Největší celé číslo menší nebo rovné x

$$\left(\bar{x}^*_{(k_L)}; \bar{x}^*_{(k_U)} \right)$$

Studentizovaný interval

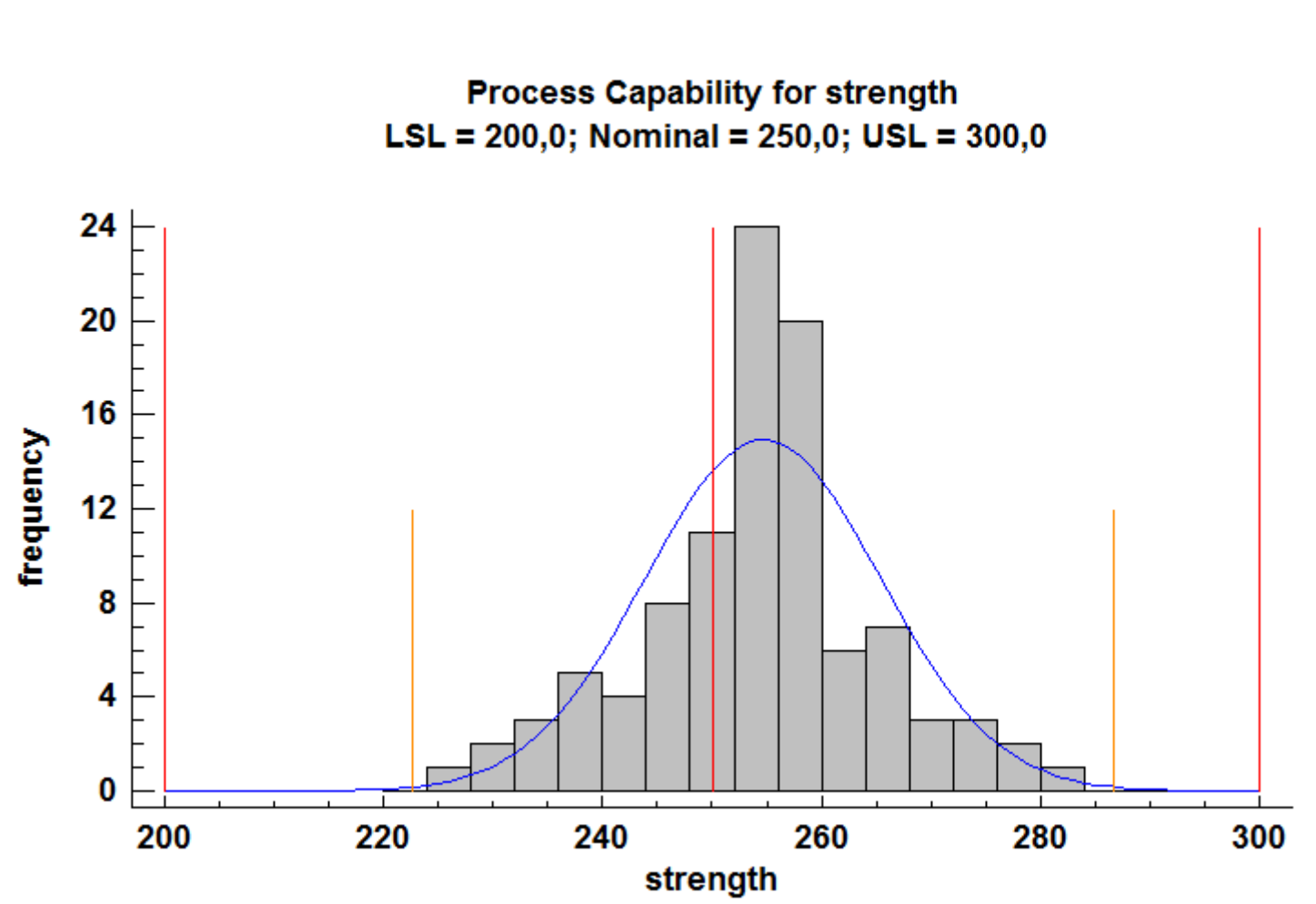
Založen na jiném bootstrapovém rozdělení než předcházející

$$t^* = \frac{\bar{x}_i^* - \bar{x}}{s(\bar{x}_i^*)} \qquad s(\bar{x}_i^*) = \sqrt{\frac{\sum_{j=1}^{n}(x_{ij}^* - \bar{x}_i^*)^2}{n^2}}$$

$$t_{(1)}^*, t_{(2)}^*, \ldots, t_{(B)}^*$$

$$\left(\bar{x} - t_{(k_L)}^* s(\bar{x});\ \bar{x} + t_{(k_U)}^* s(\bar{x})\right) \qquad s(\bar{x}) = \sqrt{\frac{\sum_{j=1}^{n}(x_j - \bar{x})^2}{n^2}}$$

Interval pro index výkonnosti



Odhad indexů za předpokladu normality

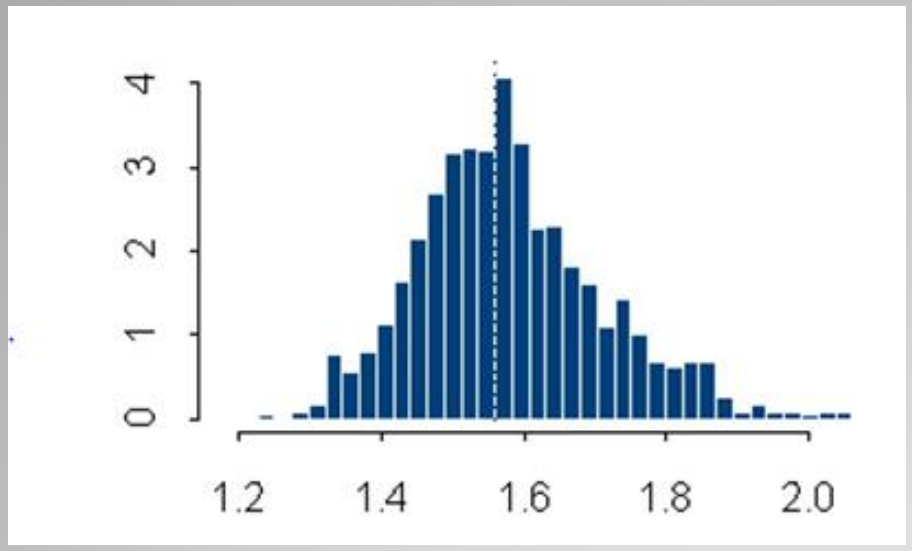
$$P_p = \frac{USL - LSL}{6\sigma}$$

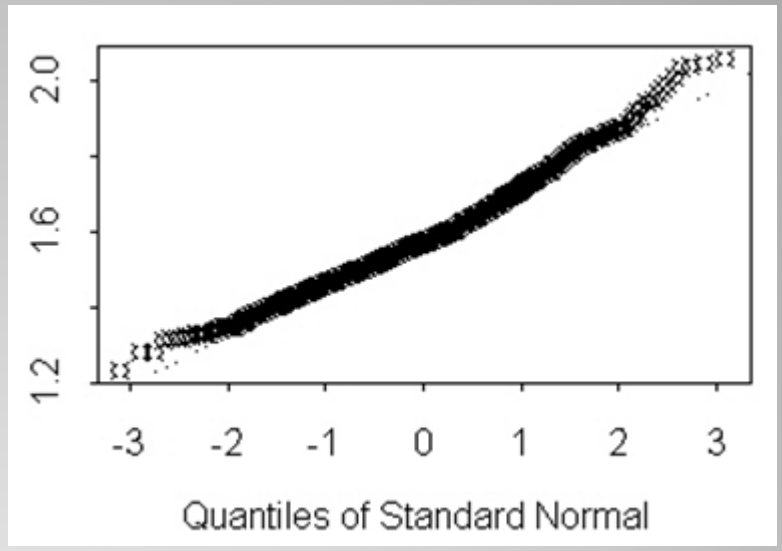
$$\hat{P}_p \sqrt{\frac{\chi^2_{1-\alpha/2}(n-1)}{n-1}} < P_p < \hat{P}_p \sqrt{\frac{\chi^2_{\alpha/2}(n-1)}{n-1}}$$

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

$$\hat{P}_{pk}\left[1 - z_{\alpha/2}\sqrt{\frac{1}{9n\hat{P}^2_{pk}} + \frac{1}{2(n-1)}}\right] < P_{pk} < \hat{P}_{pk}\left[1 + z_{\alpha/2}\sqrt{\frac{1}{9n\hat{P}^2_{pk}} + \frac{1}{2(n-1)}}\right]$$

Index výkonnosti P_p

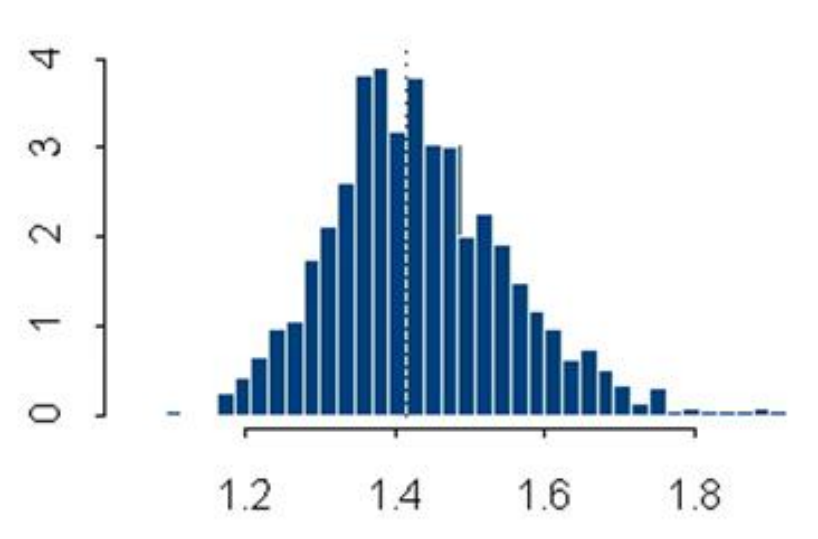


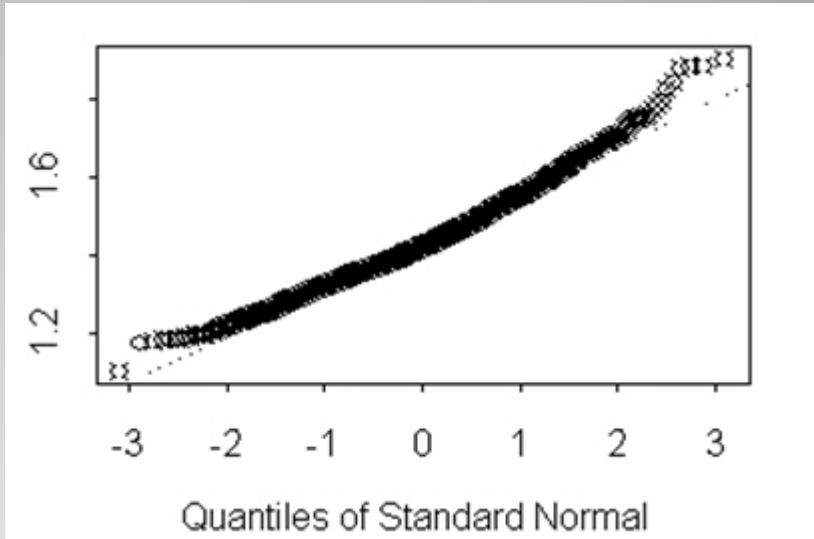



```
                odhad 1.560213
        Normal                  Basic
( 1.284,  1.795 )    ( 1.257,  1.774 )
     Percentile                 BCa
( 1.346,  1.863 )    ( 1.324,  1.829 )
```

Normalita 1,56021 (1,34307, 1,77699)
Clements 1,41825

Index výkonnosti P_{pk}






```
                odhad  1.415426
        Normal                  Basic
( 1.161,  1.633 )    ( 1.131,  1.608 )
     Percentile                 BCa
( 1.223,  1.699 )    ( 1.197,  1.666 )
```

Normalita 1,41543 (1,20773 1,62312)
Clements 1,36902

Metoda vzorkování

- Výkonnost (dlouhodobá způsobilost)

  Pp, Ppk

  odhad na základě variability všech hodnot ve výběru

  vzorkování z celého výběru

- Způsobilost (krátkodobá)

  Cp, Cpk

  odhad na základě variability v podskupinách (případně pomocí klouzavých rozpětí)

  vzorkování z jednotlivých podskupin

Literatura

- Choi, K.C., Nam, K.H., Park, D.H.: Estimation of capability index based on bootstrap method. *Microelectronics and Reliability*, vol.36, no.9, pp. 1141-1153, 1996
- Collins, A.J.: Bootstrap konfidence limits on process capability indices. *The Statistician*, vol.44, no.3, pp. 373-378, 1995
- Dixon, P.M.: The bootstrap and the jackknife: describing the precision of ecological studies, in Design and Analysis of Ecological Experiments, **...** 2001 *www.wiley.com/legacy/wileychi/eoenv/pdf/Vab028-.pdf*
- Efron, B., Tibshirani, R.J.: Bootstrap Methods for Standard Errors, Confidence Intervals, and Other measures of Statistical Accuracy. Statistical Science, vol.1, no.1, pp. 54-77, 1986

- Efron, B., Tibshirani, R.J.: *An Introduction to the Bootstrap*. Chapman&Hall/CRC 1993
- Franklin, L.A., Wasserman, G.: Standard bootstrap confidence interval estimates of C_{PK} *Computers & Industrial Engineering*, vol.21, no.1-4, pp. 129-133, 1991
- Franklin, L.A., Wasserman, G.S.: Bootstrap Lower Confidence Limits for Capability Indices. JQT, vol.24, no.4, 1992
- Yang J.: A Bootstrap Confidence Limit for Process Capability Indices

  http://www.bmtfi.net/upload/product/200910/2007glhy14a3.pdf