

VLASTNOSTI ODHADŮ UKAZATELŮ ZPŮSOBILOSTI II

Jiří Michálek

CQR při Ústavu teorie informace a automatizace AV ČR v Praze

Abstrakt: Autor¹ se zabývá asymptotickým chováním maximálně věrohodného odhadu ukazatele C_p a je dokázána jeho silná konzistence. Tento příspěvek je přímým pokračováním autorova článku publikovaného na konferenci Request'08.

Klíčová slova: ukazatel způsobilosti C_p , maximálně věrohodný odhad, silná konzistence

Úvod

Ukazatele způsobilosti a výkonnosti jsou v průmyslové praxi, hlavně v automobilovém průmyslu, využívány velice často pro jednoduché vyjádření požadavků zákazníka či konstruktéra na vybrané znaky kvality spojité povahy. Na jedné straně je vyjádřený požadavek na znak kvality vyjádřený číselnou hodnotou ukazatele, což klade požadavky na očekávaný počet neshodných kusů mimo toleranční pole a současně tím i požadavek na úroveň variability, kterou má sledovaný znak kvality jako náhodná veličina vykazovat. Požadovaná úroveň variability je nejčastěji vyjádřena hodnotou směrodatné odchylky za předpokladu, že chování znaku kvality se dá popsat normálním rozdělením. Na druhou stranu je reálný výrobní proces produkující výrobek se sledovaným znakem kvality a je otázkou, zdali je proces požadavku na znak kvality kladené splnit. Abychom se o tom přesvědčili, je nutno jít do procesu, odebrat data a ta vyhodnotit pomocí metod matematické statistiky.

Velice často se lze v praxi setkat s následujícím postupem. Z procesu jsou odebírána data, která jsou automaticky vkládána do softwaru, který vede regulační diagramy hlídající stabilitu procesu a současně jsou tatáž data použita pro odhady ukazatelů způsobilosti a výkonnosti. Výsledek je pak bodový odhad ukazatele způsobilosti či výkonnosti, který je získán za období, pro které byl veden regulační diagram. Toto období může být relativně krátké, např. jedna směna či celý den (24 hodin) a získáváme tak informaci spíše o krátkodobé způsobilosti procesu. Máme tedy k dispozici posloupnost konečnou bodových odhadů ukazatelů způsobilosti či výkonnosti a chtěli bychom ji využít pro konstrukci bodového odhadu pro dlouhodobější způsobilosti či výkonnost, protože mnohdy originální data odebraná přímo z procesů již nemusejí být k dispozici. Je tedy otázkou, jak zkonstruovat odhad ukazatele způsobilosti či výkonnosti na základě konečné posloupnosti bodových odhadů získaných z jednotlivých regulačních diagramů.

Máme tedy znak kvality X , o kterém budeme předpokládat, že má normální rozdělení $N(\mu, \sigma^2)$, data z procesu při jeho měření jsou získávána ve formě podskupin rozsahu n ,

¹e-mail: michalek@utia.cas.cz, telefon: 266052241

na každém regulačním diagramu je k podskupin a těchto regulačních karet je celkem N . Je předpokládána vzájemná stochastická nezávislost dat. K dispozici je celkem tedy N bodových odhadů $\hat{C}_p^{(1)}, \dots, \hat{C}_p^{(N)}$, resp. $\hat{P}_p^{(1)}, \dots, \hat{P}_p^{(N)}$.

V praxi se pro odhad směrodatné odchylky σ inherentní variability, která vystupuje v definici ukazatele C_p , používá téměř výhradně následujících tří odhadů pro parametr σ :

$$\text{a) } \hat{\sigma}_R = \frac{\bar{R}}{d_2(n)}, \quad \bar{R} = \frac{1}{k} \sum_{i=1}^k R_i$$

$$\text{b) } \hat{\sigma}_s = \frac{\bar{s}}{C_4(n)}, \quad \bar{s} = \frac{1}{k} \sum_{j=1}^k s_j$$

$$\text{c) } \hat{\sigma}_I = \left(\frac{1}{k} \sum_{i=1}^k s_i^2 \right)^{1/2}.$$

V případě a) se jedná o odhad na základě průměrného výběrového rozpětí \bar{R} , v případě b) je odhad založen na průměrné výběrové směrodatné odchylce \bar{s} . Třetí případ používá tzv. “pooled standard deviation”. Konstanty $d_2(n)$ a $C_4(n)$ závislé na velikosti podskupiny, jsou tabelovány např. v normě ISO ČSN 8258.

Je zřejmé, že chování odhadů ukazatele C_p bude záviset na volbě odhadu parametru σ , což se bohužel v praxi neděje a ani softwary tuto možnost neuvažují a mezi příslušnými odhady nerozlišují.

Pro případ odhadu \hat{P}_p se používá rovněž výhradně celková (totální) výběrová směrodatná odchylka pro odhad tzv. totální směrodatné odchylky, a to

$$\hat{\sigma}_{\text{TOT}} = \left(\frac{1}{k_{n-1}} \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{\bar{x}})^2 \right)^{1/2},$$

tento odhad má pouze smysl, když data lze bez ohledu na sběr do podskupin považovat za normálně rozdělené, neboť obecně u aplikace ukazatele P_p se nepředpokládá stabilní stav procesu. Zde x_{ij} je j -té pozorování z i -té podskupiny, $\bar{\bar{x}}$ je celkový aritmetický průměr ze všech dat.

1. Hustoty pro odhady \hat{C}_p a \hat{P}_p

Abychom mohli zkonstruovat MLE ukazatelů C_p či P_p , musíme nejdříve najít hustoty rozdělení pravděpodobností. Samozřejmě, že tyto hustoty budou závislé na volbě odhadu směrodatné odchylky σ . Jestliže budeme uvažovat odhad $\hat{\sigma}_R$, pak lze najít aproximaci pro hustotu, protože lze využít pouze asymptotické vyjádření pro chování výběrového rozpětí R z normální populace. Platí, viz. např. [1], že

$$P \left\{ \frac{R_n - \alpha_n \sigma}{\beta_n \sigma} < \lambda \right\} \xrightarrow{n \rightarrow \infty} N(0, 1).$$

Z toho ihned plyne pro průměrné výběrové rozpětí \bar{R} , že

$$\frac{\bar{R}_k - \alpha_n \sigma}{\beta_n \sigma} \sqrt{k} \sim N(0, 1).$$

Parametry α_n, β_n jsou tabelovány v [1] a závisí na velikosti podskupiny n . Jestliže tedy odhad ukazatele C_p bude

$$\hat{C}_p = \frac{\text{USL} - \text{LSL}}{6\hat{\sigma}_R},$$

pak lze snadno odvodit, že

$$P \left\{ \frac{\hat{C}_p}{C_p} < x \right\} \doteq 1 - \Phi \left(\frac{\alpha_n \sqrt{k}}{\beta_n} \left(\frac{1}{x} - 1 \right) \right)$$

pro $x > 0$, a jinak nula. ($\Phi(\cdot)$ je distribuce rozdělení $N(0, 1)$.) Tím se dostáváme k přibližnému vyjádření hustoty pro odhad \hat{C}_p , totiž

$$f_R(x) \doteq \frac{\alpha_n \sqrt{k}}{\sqrt{2\pi} \beta_n x^2} C_p \exp \left\{ -\frac{k \alpha_n^2}{2 \beta_n^2} \left(\frac{C_p}{x} - 1 \right)^2 \right\} \quad \text{pro } x > 0$$

a $f_R(x) = 0$ jinak.

Zcela analogická situace nastává v případě b), kdy se využije asymptotického chování výběrové směrodatné odchylky a to

$$P \left\{ \frac{s_n - a_{n-1} \sigma}{\sigma b_{n-1}} < x \right\} \xrightarrow{n \rightarrow \infty} N(0, 1).$$

Pak průměrná směrodatná odchylka z k podskupin dává odhad pravděpodobnosti

$$P \left\{ \frac{\hat{C}_p}{C_p} < x \right\} \doteq 1 - \Phi \left(\frac{a_{n-1} \sqrt{k}}{b_{n-1}} \left(\frac{1}{x} - 1 \right) \right),$$

pak tedy

$$f_s(x) \doteq \frac{a_{n-1} \sqrt{k}}{b_{n-1} x^2} C_p \exp \left\{ -\frac{k a_{n-1}^2}{2 b_{n-1}^2} \left(\frac{C_p}{x} - 1 \right)^2 \right\}$$

pro $x > 0$ a $f_s(x) = 0$ jinak. Zde

$$a_{n-1} = \sqrt{\frac{2}{n-1} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}} \doteq \sqrt{1 - \frac{1}{2(x-1)}},$$

$$b_{n-1} = \sqrt{1 - a_{n-1}^2}.$$

U třetího případu c) je situace odlišná, zde lze odvodit snadno hustotu rozdělení odhadu \hat{C}_p pomocí χ^2 -rozdělení, které má veličina

$$\frac{k(n-1)}{\sigma^2} (\hat{\sigma}_I)^2.$$

Na základě tohoto pak odpovídající hustota pro odhad \hat{C}_p s pomocí $\hat{\sigma}_I$ má tvar

$$f_I(x) = \frac{C_p^{k(n-1)} [k(n-1)]^{\frac{k(n-1)}{2}}}{2^{\frac{k(n-1)}{2}-1} \Gamma(\frac{k(n-1)}{2})} \exp \left\{ -\frac{k(n-1) C_p^2}{2x^2} \right\} \frac{1}{x^{k(n-1)+1}}$$

pro $x > 0$ a $f_I(x) = 0$ jinak.

V případě odhadu ukazatele P_p je situace zcela analogická, pouze se změní počet stupňů volnosti v rozdělení χ^2 z $k(n-1)$ na $kn-1$, jinak se tvar odpovídající hustoty nemění. Pokud bude proces statisticky zvládnut a stabilní, pak odhady \hat{C}_p a \hat{P}_p by se neměly příliš lišit. Pokud se ale objeví zvláštní příčina vyvolávající variabilitu mezi podskupinami, pak s velkou pravděpodobností bude

$$\hat{P}_p < \hat{C}_p.$$

2. MLE ukazatele C_p pro případy a) a b)

Předpokládejme tedy, že máme k dispozici N odhadů ukazatele C_p — $\hat{C}_p^{(1)}, \hat{C}_p^{(2)}, \dots, \hat{C}_p^{(N)}$. Pak jejich sdružená hustota pravděpodobnosti má tvar

$$f_R(x_1, x_2, \dots, x_N) = C_p^N \rho_{n,k}^N \exp \left\{ -\frac{1}{2} \rho_{n,k}^2 \sum_{i=1}^N \left(\frac{C_p}{x_i} - 1 \right)^2 \right\} \prod_{i=1}^N \frac{1}{x_i^2}$$

při použití odhadu $\hat{\sigma}_R$ ve všech N případech. Odpovídající logaritmus věrohodné funkce má pak tvar

$$\ln f_R(x_1, x_2, \dots, x_N) = N \ln C_p + N \ln \rho_{n,k} - \sum_{i=1}^N \ln x_i^2,$$

kde

$$\rho_{n,k} = \frac{\alpha_n}{\beta_n} \sqrt{n}.$$

Při použití odhadu $\hat{\sigma}_s$ pro inherentní směrodatnou odchylku je odpovídající logaritmus věrohodností funkce zcela tentýž s tím rozdílem, že pak

$$\rho_{n,k} = \frac{a_{n-1}}{b_{n-1}} \sqrt{k}.$$

MLE pro C_p se pak získá řešením kvadratické rovnice

$$\frac{N}{C_p} = \rho_{n,k}^2 \sum_{i=1}^N \left(\frac{C_p}{x_i} - 1 \right) \cdot \frac{1}{x_i},$$

která vyplyne z věrohodnostní rovnice

$$0 = \frac{\partial \ln f_R(x_1, x_2, \dots, x_N)}{\partial C_p} = \frac{N}{C_p} - \rho_{n,k}^2 \sum_{i=1}^N \left(\frac{C_p}{x_i} - 1 \right) \cdot \frac{1}{x_i}.$$

Kvadratická rovnice dává 2 řešení

$$\hat{C}_p^{\text{MLE}} = \frac{T_1(x_1, x_2, \dots, x_N) \pm \sqrt{T_2^2(x_1, x_2, \dots, x_N) + 4 \frac{T_2(x_1, x_2, \dots, x_N)}{\rho_{n,k}}}}{2T_2(x_1, x_2, \dots, x_N)}.$$

My ale uvažujeme pouze řešení se znaméním +, protože C_p je vždy kladné.

V případě $N = 1$, což je velice častý případ v praxi, dostaneme odhad

$$\hat{C}_p^{(\text{MLE})} = \frac{1 + \sqrt{1 + \frac{4}{\rho_{n,k}^2}}}{2} \hat{C}_p.$$

Statistiky T_1 a T_2 jsou definovány následovně:

$$T_1(x_1, x_2, \dots, x_N) = \frac{1}{N} \sum_{i=1}^N \frac{1}{x_i}, \quad T_2(x_1, x_2, \dots, x_N) = \frac{1}{N} \sum_{i=1}^N \frac{1}{x_i^2}.$$

Poznámka 1. Je nutno zdůraznit, že se nejedná přesně o MLE odhad, protože vše je založeno pouze na aproximaci sdružené hustoty, která vychází z asymptotického chování \bar{R} resp. \bar{s} . Správně by tedy mělo být AMLE (approximative MLE).

Bude-li tedy počet podskupin dostatečně velký ($k \geq 25$), pak se odhady \hat{C}_p a $\hat{C}_p^{(\text{MLE})}$ nebudou příliš lišit pro obvyklé velikosti podskupin $2 \leq n \leq 10$.

Platí následující věta, která je zcela samozřejmá.

Věta 1. Bude-li $k \rightarrow \infty$, pak pro každé $n \in \mathcal{N}$ a každé $N \in \mathcal{N}$

$$\hat{C}_p^{(\text{MLE})} \xrightarrow[n \rightarrow \infty]{} C_p \quad \text{s. j.}$$

Zajímavá je ale otázka, co se bude dít, když n a k budou fixovány a $N \nearrow \infty$, což je v praxi realističtější případ, který odpovídá zvětšujícímu se počtu krátkodobých odhadů ukazatele C_p získaných postupně z regulačních karet za předpokladu, že proces je stále pod statistickou kontrolou.

Pokud $N \nearrow +\infty$ s fixními n a k , pak máme co činit se zákonem velkých čísel, neboť ve vzorcích pro MLE se objevují výrazy

$$\frac{1}{N} \sum_{i=1}^N \frac{1}{\hat{C}_p^{(i)}}, \quad \frac{1}{N} \sum_{i=1}^N \frac{1}{(\hat{C}_p^{(i)})^2}.$$

Aby platil silný zákon velkých čísel, je nutná a postačující existence odpovídajících středních hodnot, neboť za našich předpokladů se jedná o průměry od i.i.d. veličin. Jde tedy o to, zdali existují

$$E \left\{ \frac{1}{\hat{C}_p^{(i)}} \right\}, \quad E \left\{ \frac{1}{(\hat{C}_p^{(i)})^2} \right\},$$

což vede ke konvergenci či divergenci následujících integrálů:

$$\int_0^\infty \frac{1}{x} f_R(x) dx, \quad \int_0^\infty \frac{1}{x^2} f_R(x) dx,$$

resp. s hustotou $f_s(\cdot)$. Lze se přesvědčit, že oba integrály konvergují, což znamená, že silný zákon velkých čísel platí (viz dále).

3. MLE ukazatele C_p pro případ c)

V tomto případě je odhad ukazatele C_p založen na “pooled standard deviation”, a tudíž odpovídající sdružená hustota pro N nezávislých odhadů $\hat{C}_p^{(1)}, \hat{C}_p^{(2)}, \dots, \hat{C}_p^{(N)}$ má tvar

$$f_i(x_1, x_2, \dots, x_N) = \frac{C_p^{N(k(n-1))} [k(n-1)]^{\frac{Nk(n-1)}{2}}}{2^{N[\frac{k(n-1)}{2}-1]} [\Gamma(\frac{n(k-1)}{2})]^N} \exp \left\{ -\frac{k(n-1) C_p^2}{2} \sum_{i=1}^N \frac{1}{x_i^2} \right\} \frac{1}{\prod_{i=1}^N x_i^{k(n-1)+1}}.$$

Snadným výpočtem dojdeme k rovnici pro MLE ukazatele C_p :

$$\frac{\partial \ln f_I(x_1, x_2, \dots, x_N)}{\partial C_p} = \frac{Nk(n-1)}{C_p} - C_p n(k-1) \sum_{i=1}^N \frac{1}{x_i^2} = 0,$$

což dává řešení rovnice ve tvaru:

$$\hat{C}_p^{(\text{MLE})} = \left(\frac{1}{N} \sum_{i=1}^N \frac{1}{(\hat{C}_p^{(i)})^2} \right)^{-1/2}.$$

Pro nejjednodušší případ $N = 1$ ihned máme $\hat{C}_p^{(\text{MLE})} = \hat{C}_p$.

Poznámka 2. V případě odhadů ukazatele C_p odvozených od \bar{R} a \bar{s} jsou použity jako odhady odpovídajících hustot rozdělení pravděpodobnosti funkce $f_R(\cdot)$ a $f_s(\cdot)$. Aby tyto funkce byly skutečně hustotami pravděpodobnosti, bylo by nutné je vynásobit vhodnými koeficienty, aby integrál z nich přes obor $(0, +\infty)$ byl skutečně roven 1. Tyto koeficienty jsou:

1. pro případ s \bar{R} :

$$\frac{1}{1 - \phi\left(-\frac{\alpha_n}{\beta_n} \sqrt{k}\right)}$$

2. pro případ s \bar{s} :

$$\frac{1}{1 - \phi\left(-\frac{a_{n-1}}{b_{n-1}} \sqrt{k}\right)}.$$

Protože ale tyto koeficienty velice rychle konvergují k 1 při $k \rightarrow \infty$, chyba, které se touto nepřesností dopustíme, je z praktického hlediska vlastností odhadů ukazatele C_p zanedbatelná.

Poznámka 3. Důkaz silné konzistence odhadů ukazatele C_p je založen na silném zákonu velkých čísel pro vzájemně nezávislé a stejně rozdělené náhodné veličiny, kde konečnost střední hodnoty je nutnou a postačující podmínkou. V případech maximálně věrohodných odhadů založených na \bar{R} a \bar{s} se tedy jedná o konečnost středních hodnot

$$E \left\{ \frac{1}{\hat{C}_p} \right\} \quad \text{a} \quad E \left\{ \frac{1}{\hat{C}_p^2} \right\}.$$

Abychom dokázali platnost silného zákona velkých čísel při použití odhadů a) a b), je nutno ověřit konečnost příslušných středních hodnot. Na základě poznámky 2 označme korekci

$$\delta_{n,k} = 1 - \Phi \left(-\frac{\alpha_n}{\beta_n} \sqrt{k} \right),$$

resp.

$$\delta_{n,k} = 1 - \Phi \left(-\frac{a_{n-1}}{b_{n-1}} \sqrt{k} \right),$$

pak tedy užijeme vztah

$$P \left\{ \frac{\hat{C}_p}{C_p} < x \right\} = \delta_{n,k}^{-1} \left[1 - \Phi \left(\frac{\alpha_n}{\beta_n} \sqrt{k} \left(\frac{1}{x} - 1 \right) \right) \right].$$

Z tohoto odvodíme tvar pro hustotu bodového odhadu \hat{C}_p a musíme ověřit konvergenci integrálů

$$E \left\{ \frac{1}{\hat{C}_p} \right\} = \int_0^{+\infty} \delta_{n,k}^{-1} \frac{1}{x} f_R(x) dx$$

a

$$E \left\{ \frac{1}{\hat{C}_p^2} \right\} = \int_0^{+\infty} \delta_{n,k}^{-1} \frac{1}{x^2} f_R(x) dx$$

(pro případ b) místo funkce $f_R(\cdot)$ vystupuje funkce $f_s(\cdot)$).

Označme $\varphi(\cdot)$ hustotu rozdělení $N(0, 1)$. Pak lze poměrně snadným výpočtem dospět k výrazu

$$\begin{aligned} E \left\{ \frac{1}{\hat{C}_p} \right\} &= \delta_{n,k}^{-1} \int_{-\rho_{n,k}}^{\infty} \frac{1}{C_p} \left(\frac{z}{\rho_{n,k}} + 1 \right) \varphi(z) dz = \\ &= \delta_{n,k}^{-1} \int_{-\rho_{n,k}}^{\infty} \frac{z}{C_p \rho_{n,k}} \varphi(z) dz + \frac{\delta_{n,k}^{-1}}{C_p} \int_{-\rho_{n,k}}^{\infty} \varphi(z) dz. \end{aligned}$$

Je ihned vidět, že první integrál konverguje a je pro $k \geq 20$ zanedbatelný, druhý integrál konverguje rovněž k hodnotě $\frac{1}{C_p}$, která zaručuje asymptotickou nestrannost.

V případě $E \left\{ \frac{1}{\hat{C}_p^2} \right\}$ postupujeme zcela obdobně a opět poměrně snadným výpočtem dojdeme k integrálu

$$\begin{aligned} E \left\{ \frac{1}{\hat{C}_p^2} \right\} &= \int_{-\rho_{n,k}}^{\infty} \frac{\delta_{n,k}^{-1}}{C_p^2} \left(\frac{z}{\rho_{n,k}} + 1 \right)^2 \varphi(z) dz = \\ &= \frac{\delta_{n,k}^{-1}}{C_p^2} \int_{-\rho_{n,k}}^{\infty} \frac{z^2}{C_p \rho_{n,k}^2} \varphi(z) dz + \frac{2\delta_{n,k}^{-1}}{C_p^k} \int_{-\rho_{n,k}}^{\infty} \frac{z}{\rho_{n,k}} \varphi(z) dz + \frac{1}{C_p^2}. \end{aligned}$$

První dva integrály konvergují a opět pro $k \geq 20$ je lze pro praxi zanedbat, protože jsou řádu $o(k)$ a $o(k^{1/2})$.

Souhrmně lze tedy psát, že

$$E \left\{ \frac{1}{\hat{C}_p} \right\} = \frac{1}{C_p} + o(k^{1/2}), \quad E \left\{ \frac{1}{\hat{C}_p^2} \right\} = \frac{1}{C_p^2} + o(k^{1/2}).$$

Tím je dokázáno, že pro případy odhadů a) a b) platí konvergence ve smyslu s. j., ale odhady při $N \rightarrow \infty$ jsou mírně vychýlené právě v závislosti na velikosti parametru k .

V případě odhadu c) založeného na sdružené směrodatné odchylce je nutno dokázat konvergenci integrálu

$$E \left\{ \frac{1}{\hat{C}_p^2} \right\} = \int_0^\infty \frac{1}{x^2} f_I(x) dx.$$

Po drobných úpravách dospějeme k integrálu

$$E \left\{ \frac{1}{\hat{C}_p^2} \right\} = \frac{(C_p^2 s)^{s/2}}{2^{s/2-1} \Gamma(s/2)} \int_0^\infty y^{s+1} e^{-\frac{C_p^2 s y^2}{2}} dy,$$

kde $s = k(n-1)$ a snadno zjistíme, že v tomto případě

$$E \left\{ \frac{1}{\hat{C}_p^2} \right\} = \frac{1}{C_p^2}.$$

Integrál tedy konverguje, což dokazuje platnost silného zákona velkých čísel a navíc výše uvedená vlastnost nestrannosti vede k tvrzení, že maximálně věrohodný odhad založený na sdružené směrodatné odchylce je silně konzistentní, tedy

$$\lim_{N \rightarrow \infty} \left(\frac{1}{N} \sum_{i=1}^N \frac{1}{(\hat{C}_p^{(i)})^2} \right)^{-1/2} = C_p \text{ s. j.}$$

Získané výsledky lze tedy shrnout do následující věty.

Věta 2. V případě odhadů a) a b) lze dokázat konvergenci ve smyslu s. j. k hodnotě mírně vychýlené od hodnoty ukazatele C_p , v případě odhadu c) platí silná konzistence.

Poděkování: Tato práce vznikla v rámci projektu 1M08047 MŠMT Centrum pro jakost a spolehlivost ve výrobě.

Reference

- [1] Hald A. (1952) *Statistical Theory with Engineering Application*. John Wiley and Sons, N. Y.
- [2] Michálek J. (2008) *Vlastnosti odhadů ukazatelů způsobilosti*. Request'08, Sborník konference CQR, FSI VUT Brno, 153–162.