INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 23, číslo 3, září 2012

WALDŮV INTERVALOVÝ ODHAD PARAMETRU BINOMICKÉHO ROZDĚLENÍ A JEHO ALTERNATIVY

Martina Litschmannová

Adresa: Ing. Martina Litschmannová, VŠB-TU Ostrava, Fakulta elektrotechniky a informatiky, Katedra aplikované matematiky

E-mail: martina.litschmannova@vsb.cz

Abstrakt: V biomedicínských aplikacích, a nejen v nich, se velmi často setkáváme s poměrně nízkou incidencí některých kategorií sledované proměnné, resp. s malým rozsahem výběru (např. incidence významných komplikací spojených s radioterapii karcinomu prostaty). To jsou situace, které z důvodu porušení předpokladů plynoucích z centrální limitní věty kontraindikují možnost použití standardně využívaného Waldova intervalu spolehlivosti pro výpočet intervalového odhadu relativní četnosti. V literatuře (například: Agresti & Coull (1998, 2000), Anděl, Černý, Charamza a Neustadt (2004), Blyth & Still (1983), Clopper & Pearson (1934), Neyman (1935), Pires (2008), Wald (1939), Wilson (1927)) lze najít zhruba 20 různých alternativních metod umožňujících stanovit intervalový odhad parametru binomického rozdělení (angl. „confidence interval for binomial proportion“). Příspěvek je věnován přehledu a srovnání vybraných typů intervalových odhadů na základě statistik používaných k hodnocení jejich vlastností.

1. Úvod

Na Katedře aplikované matematiky, FEI, VŠB-TU Ostrava se, zejména v souvislosti se spoluprácí s Fakultní nemocnicí v Ostravě – Porubě, setkáváme s požadavky na analýzu medicínských dat. Lékaři řeší spoustu zajímavých, často i složitých problémů, které vedou na více či méně sofistikované statistické metody a modely. Analyzovaný soubor pacientů lze většinou považovat za náhodný výběr. Jedním z dílčích úkolů pak obvykle bývá odhad relativního zastoupení (incidence) sledovaných jevů – například stupňů onemocnění v populaci nemocných, jednotlivých stupňů intenzity nežádoucích účinků léčby, apod. Ze statistického hlediska se jedná o odhad ukazatelů extenzity, tj. o odhad parametru binomického rozdělení.

V biomedicínských aplikacích se velmi často setkáváme s poměrně nízkou incidencí některých variant sledovaných jevů, resp. s malým rozsahem výběru. To jsou situace, které z důvodu porušení předpokladů plynoucích z centrální

limitní věty kontraindikují možnost použití standardně využívaného Waldova intervalu spolehlivosti pro výpočet intervalového odhadu ukazatelů extenzity.

V odborných statistických kruzích jde o starý a dobře známý problém. V literatuře (např. Wilson (1927), Clopper & Pearson (1934), Blyth & Still (1983), Agresti & Coull (1998, 2000), Brown, Cai & DasGupta (2001), Pires (2008)) lze najít více než 20 různých alternativních metod umožňujících stanovit intervalový odhad (dále IO) parametru binomického rozdělení (angl. „confidence interval for binomial proportion"). Přestože je známo, že mnohé z těchto alternativních přístupů poskytují lepší výsledky než Waldův interval, použití tohoto intervalu ve výuce i ve statistické praxi přetrvává. V tomto příspěvku je ve stručnosti uveden přehled vybraných typů intervalových odhadů a statistik používaných k hodnocení jejich vlastností a metody jsou porovnány s ohledem na možnost použití v biomedicínských aplikacích vyznačujících se nízkým rozsahem výběru a nízkou incidencí sledovaného jevu.

2. Jak posoudit kvalitu intervalu spolehlivosti?

Předpokládejme, že X je počet úspěchů v n nezávislých opakováních pokusu, v němž úspěch nastane s pravděpodobností π. Náhodná veličina X má tedy binomické rozdělení s parametry n a π ($n = 1, 2, \ldots;\ 0 \le \pi \le 1$), tj.

$$P(X = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}, \qquad x = 0, 1, \ldots, n.$$

Budeme srovnávat oboustranné intervaly spolehlivosti $\langle \pi_D, \pi_H \rangle$ pro parametr π se spolehlivostí $1-\alpha$ při pevném známém n. Nejpoužívanějším nástrojem k vyšetření vlastností intervalu spolehlivosti je pravděpodobnost jeho pokrytí $C(n, \pi)$. **Pravděpodobnost pokrytí** (angl. „coverage probability") parametru π binomického rozdělení $Bi(n; \pi)$ je definována jako

$$C(n, \pi) = P(\pi \in IO) = \sum_{x=0}^{n} I_x(\pi, x)\binom{n}{x}\pi^x(1-\pi)^{n-x}, \tag{2.1}$$

kde $I_x(\pi, x) = \begin{cases} 1, & \pi \in \langle \pi_D(x,n), \pi_H(x,n) \rangle \\ 0, & \pi \notin \langle \pi_D(x,n), \pi_H(x,n) \rangle \end{cases}$ je tzv. indikátor pokrytí (angl. „coverage flag").

Přestože očekáváme, že pravděpodobnost pokrytí parametru π binomického rozdělení $Bi(n; \pi)$ bude pro všechny používané IO blízká specifikované úrovni spolehlivosti $1-\alpha$ (tzv. nominální pravděpodobnosti pokrytí), realita je, jak bude ukázáno, jiná. Z tohoto pohledu rozlišujeme dva typy intervalů. Intervaly, jejichž minimální pravděpodobnost pokrytí ($\min\{C(n, \pi) | n \in \mathbb{N}, \pi \in$

$\langle 0;1\rangle\}$) je menší než nominální pravděpodobnost pokrytí $1-\alpha$, jsou nazývány **liberální** a intervaly, jejichž minimální pravděpodobnost pokrytí je větší než nominální, se nazývají **konzervativní**.

V souvislosti s pravděpodobností pokrytí jsou v literatuře uváděny rovněž další parametry umožňující srovnání různých IO: střední pokrytí (angl. „mean of coverage probability“) a střední kvadratická chyba pokrytí (angl. „root mean square error of coverage probability“). Má-li pravděpodobnost π rovnoměrné rozdělení na intervalu $\langle 0;1\rangle$, pak je střední hodnota pravděpodobnosti pokrytí, tzv. **střední pokrytí**, definována jako

$$MC(n) = \int_0^1 C(n,\pi)\,\mathrm{d}\pi \tag{2.2}$$

a střední kvadratická chyba pokrytí je definována vztahem

$$RMSE(n) = \sqrt{\int_0^1 \left(C(n,\pi) - (1-\alpha)\right)^2\,\mathrm{d}\pi}. \tag{2.3}$$

Dalším důležitým měřítkem chování intervalového odhadu je jeho délka. Pro hodnocení kvality odhadu bude používána délka intervalového odhadu (angl. „expected length“) a střední délka intervalového odhadu (angl. „mean expected length“). Délka intervalového odhadu parametru π binomického rozdělení $Bi(n;\pi)$ je

$$EL_{n,\pi}(\textit{délka IO}) = \sum_{x=0}^{n} \Big(\pi_H(x,n) - \pi_D(x,n)\Big)\binom{n}{x}\pi^x(1-\pi)^{n-x}, \tag{2.4}$$

kde $\pi_D(x,n)$ a $\pi_H(x,n)$ jsou dolní a horní mez příslušného intervalového odhadu. **Střední délku intervalového odhadu** pak vypočteme jako

$$MEL(n) = \int_0^1 EL_{n,\pi}(\textit{délka IO})\,\mathrm{d}\pi. \tag{2.5}$$

3. Přehled vybraných intervalových odhadů parametru binomického rozdělení

3.1. Waldův interval

Zaměřme se nejprve na nejčastěji uváděný a stále ještě obecně používaný intervalový odhad parametru binomického rozdělení, který je založen na aproximaci normálním rozdělením:

$$\left\langle p - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}; p + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right\rangle, \tag{3.1}$$

kde $p = \frac{x}{n}$ a z_α je α-kvantil normovaného normálního rozdělení. (Poznámka: Tento standardně používaný interval publikoval Laplace v roce 1812. Vzhledem k tomu, že je založen na Waldově testu, bývá nazýván Waldův, později také standardní interval.)

Ve vztahu (3.1) je pro zpřesnění často prováděna korekce na spojitost (dále cc, Blyth & Still, 1983). Waldův IO pak lze uvést ve tvaru

$$\left\langle p - \frac{1}{2n} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}; p + \frac{1}{2n} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right\rangle. \tag{3.2}$$

Mezi známé nevýhody Waldova IO patří skutečnosti, že:

- V případě, že pozorujeme malý počet úspěchů nebo neúspěchů, může dolní mez Waldova IO vyjít záporná, popř. horní mez větší než 1. Při prezentaci Waldova IO je poté vhodné tyto meze korigovat (viz 3.3).

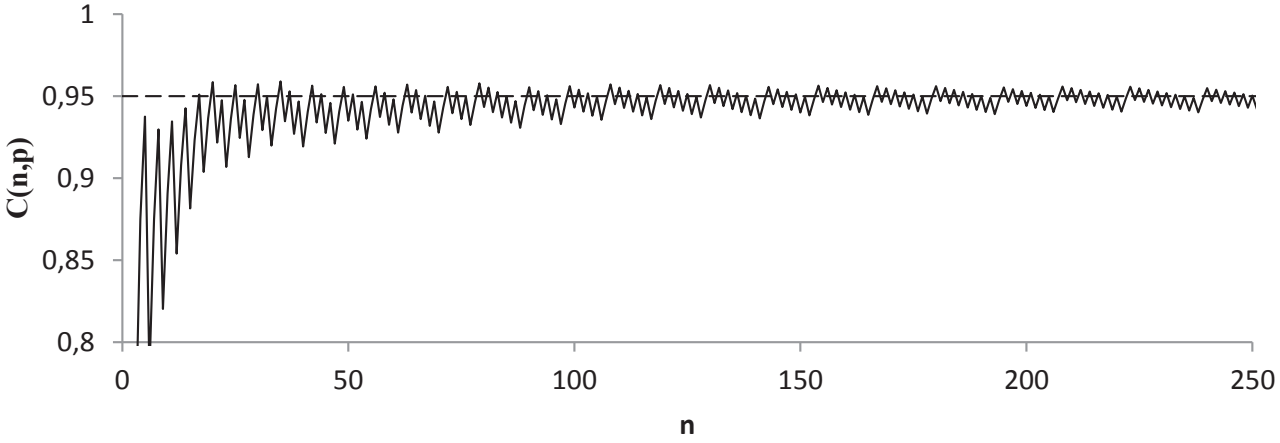
$$\begin{aligned} \pi_D &= \max\left\{p - \frac{1}{2n} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}; 0\right\}, \\ \pi_H &= \min\left\{p + \frac{1}{2n} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}; 1\right\}. \end{aligned} \tag{3.3}$$

- Pozorujeme-li $X = 0$ nebo $X = n$ úspěchů, degeneruje Waldův IO na jeden bod. Někteří statistikové (např. Vollset, 1993) doporučují v těchto hraničních případech nahradit meze Waldova IO mezemi Clopperovými-Pearsonovými (3.10). Takto upravený Waldův interval pak bývá používán jak s korekcí na spojitost, tak bez ní. Meze Waldova-Clopperova-Pearsonova odhadu (dále nazývaného jako „Waldův-Clopperův interval s cc“), v nichž je aplikována korekce na spojitost, jsou dány vztahy

$$\pi_D = \begin{cases} 0, & x = 0 \\ \max\left\{p - z_{1-\frac{\alpha}{2}}\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}} - \frac{1}{2n}; 0\right\}, & 0 < x < n \\ \left(\frac{\alpha}{2}\right)^{\frac{1}{n}}, & x = n, \end{cases} \tag{3.4}$$

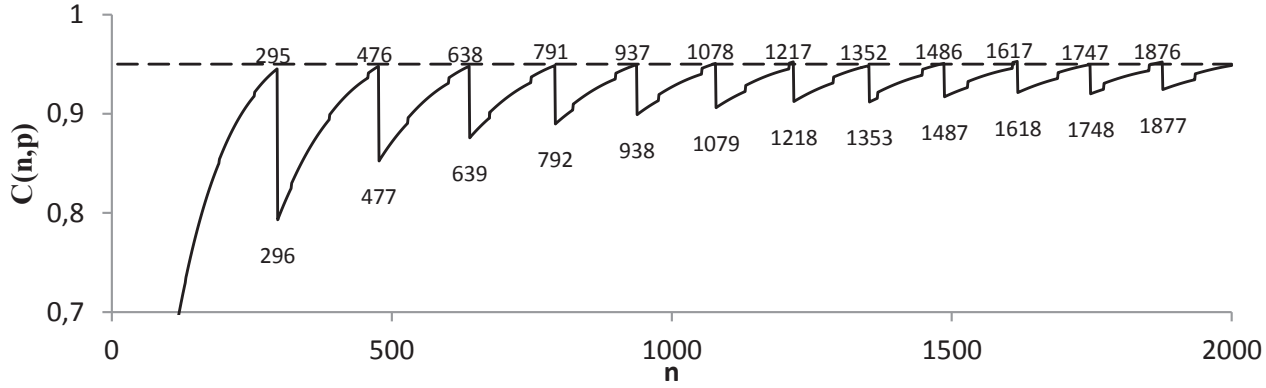
$$\pi_H = \begin{cases} \left(\frac{\alpha}{2}\right)^{\frac{1}{n}}, & x = 0 \\ \max\left\{p + z_{1-\frac{\alpha}{2}}\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}} + \frac{1}{2n}; 0\right\}, & 0 < x < n \\ 1, & x = n. \end{cases}$$

V sérii článků věnovaných metodám odhadu relativní četnosti (Brown, Cai & DasGupta, 2001, Agresti & Coull, 1998) je poukazováno na skutečnost, že pravděpodobnost pokrytí $C(n;\pi)$ Waldova intervalu obsahuje silné oscilace (viz Obr. 1). V literatuře bývají uváděna různá doporučení, kdy by mohl být Waldův odhad používán: $np(1-p) > 5$, $n \cdot \min(p, 1-p) > 5$, $n\pi(1-p) > 5$, $n \cdot \min(\pi, 1-\pi) > 5$. Místo hodnoty 5 pak někteří autoři uvádějí hodnotu 9 nebo 10.



Obrázek 1: Oscilace pokrytí Waldova intervalu ($1-\alpha = 0{,}95$, $\pi = 0{,}5$)

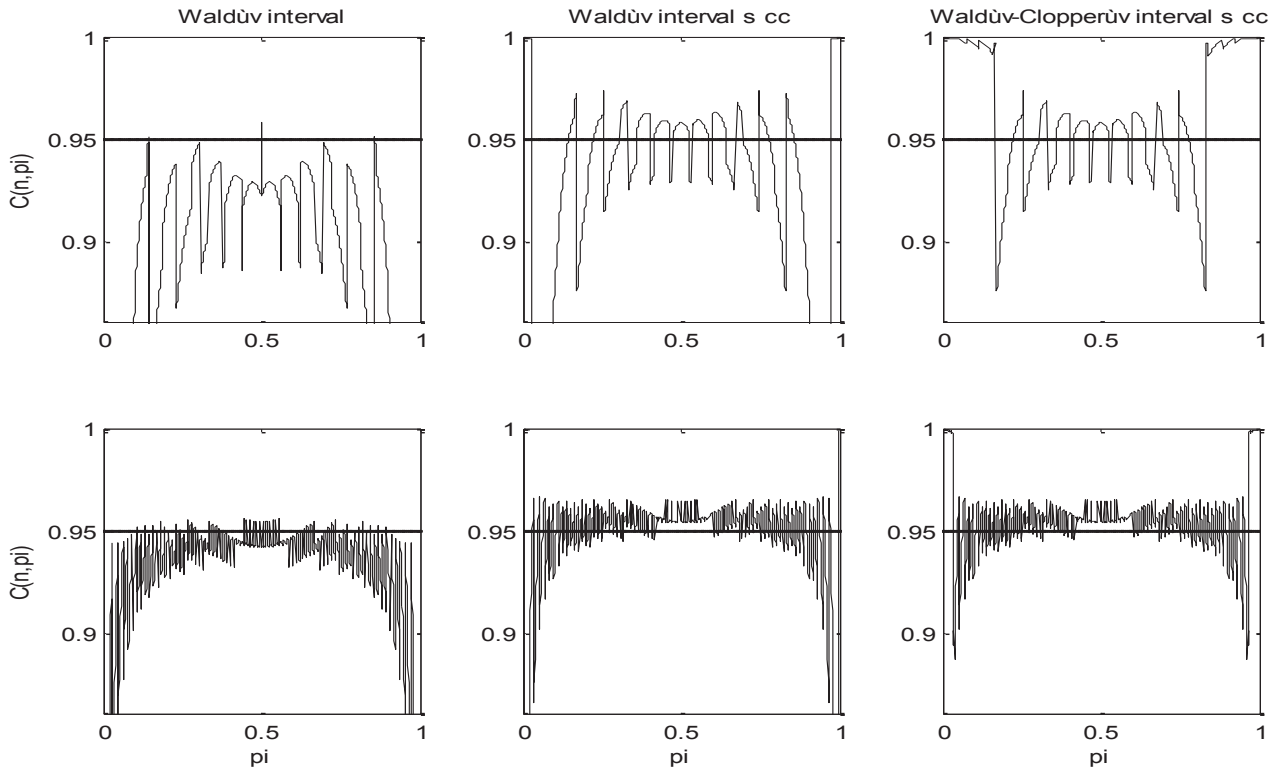
Jak již bylo uvedeno, v biomedicínských aplikacích potřebujeme často odhadovat nízké hodnoty pravděpodobnosti π. Obr. 2 ukazuje oscilace pravděpodobnosti pokrytí Waldova intervalu pro $\pi = 0{,}01$. Pravděpodobnost pokrytí roste až do rozsahu výběru 295, kdy nabývá hodnoty 0,9452. Pro rozsah výběru 296 dojde k prudkému poklesu pravděpodobnosti pokrytí na 0,7930. Následují další oscilace. Srovnáme-li Obr. 1 a Obr. 2, můžeme vidět, že nízká pravděpodobnost π má za následek pozdější vznik výrazných oscilací.



Obrázek 2: Oscilace pokrytí Waldova intervalu ($1-\alpha = 0{,}95$, $\pi = 0{,}01$)

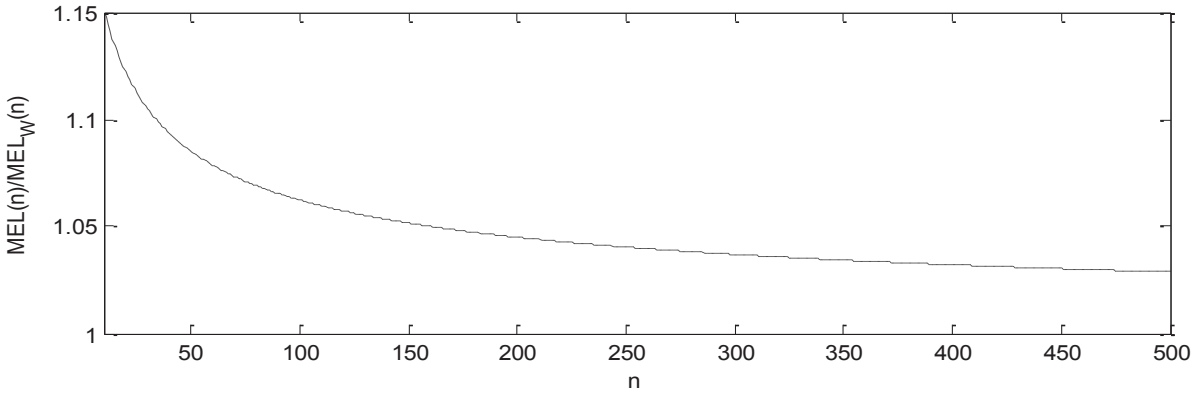
Waldův interval lze označit jako liberální – většina dvojic $(n;\pi)$, tzv. „nešťastné“ dvojice, má pravděpodobnost pokrytí $C(n;\pi)$ nižší než nominální pravděpodobnost pokrytí (obvykle 0,95). Existuje však také několik tzv. „šťastných“ dvojic, jejichž pravděpodobnost pokrytí $C(n;\pi)$ nominální pravděpodobnost pokrytí převyšuje (viz Obr. 1, Obr. 2). Ukazuje se tak, že s rostoucím rozsahem výběru nedochází automaticky k nárůstu pravděpodobnosti pokrytí. Například pro parametr binomického rozdělení $\pi = 0{,}5$ je pro rozsah výběru $n = 20$ pravděpodobnost pokrytí 0,959, zatímco pro dvojnásobný rozsah výběru ($n = 40$) je pravděpodobnost pokrytí pouze 0,919.

Srovnáním pokrytí Waldova intervalu (3.2), Waldova intervalu s cc (3.3) a Waldova-Clopperova intervalu s cc (3.4) (viz Obr. 3) dojdeme k zajímavým závěrům. Zatímco pokrytí Waldova intervalu se jeví jako zcela nevyhovující (pro většinu hodnot π je pokrytí mnohem menší než nominální hodnota), pokrytí Waldova intervalu s cc již osciluje kolem nominální hodnoty, pouze pro hodnoty π blízké 0 nebo 1 je velmi nízké. Nevyhovující pokrytí v blízkosti krajních hodnot π se pak daří výrazně zlepšit zavedením Clopperových-Pearsonových mezí ve Waldově-Clopperově intervalu s cc.



Obrázek 3: Srovnání pokrytí Waldova intervalu, Waldova intervalu s cc a Waldova-Clopperova intervalu s cc v závislosti na π pro spolehlivost odhadu $1 - \alpha = 0{,}95$, $n = 20$ (nahoře) a $n = 100$ (dole).

Srovnáme-li tyto tři intervaly z hlediska střední délky intervalového odhadu (viz Obr. 4), vidíme, že střední délka Waldova-Clopperova intervalu nepřekročí 1,15 násobek střední délky intervalu Waldova.



Obrázek 4: Poměr střední délky Waldova-Clopperova IO a Waldova IO v závislosti na rozsahu výběru n

3.2. Wilsonův (skórový) intervalový odhad

Další zajímavou variantu pro intervalový odhad relativní četnosti vycházející z aproximace normálním rozdělením uvedl v roce 1927 Wilson (Wilson, 1927). Ekvivalentními úpravami

$$\lim_{n\to\infty} P\left(z_{\frac{\alpha}{2}} \leq \frac{p-\pi}{\sqrt{\pi(1-\pi)}}\sqrt{n} \leq z_{1-\frac{\alpha}{2}} \right) = 1-\alpha, \tag{3.5}$$

tj. invertováním skórového testového kritéria $\frac{p-\pi}{\sqrt{\pi(1-\pi)}}\sqrt{n}$, lze dosáhnout tvaru

$$\lim_{n\to\infty} P(\pi_D \leq \pi \leq \pi_H) = 1-\alpha, \tag{3.6}$$

kde

$$\begin{aligned} \pi_D &= \frac{2np + z^2_{1-\frac{\alpha}{2}} - z_{1-\frac{\alpha}{2}}\sqrt{4np(1-p) + z^2_{1-\frac{\alpha}{2}}}}{2\left(n + z^2_{1-\frac{\alpha}{2}}\right)}, \\ \pi_H &= \frac{2np + z^2_{1-\frac{\alpha}{2}} + z_{1-\frac{\alpha}{2}}\sqrt{4np(1-p) + z^2_{1-\frac{\alpha}{2}}}}{2\left(n + z^2_{1-\frac{\alpha}{2}}\right)}. \end{aligned} \tag{3.7}$$

Výhodou Wilsonova intervalu oproti Waldovu intervalu je skutečnost, že jeho meze leží vždy v intervalu $\langle 0;1\rangle$ a zároveň tento interval nikdy nedegeneruje na jeden bod. Především však Wilsonův interval pokrývá skutečnou hodnotu π mnohem lépe než Waldův IO (viz Obr. 5). Blyth a Still (1983) uvádějí alternativní aproximaci s korekcí na spojitost (dále „**Wilsonův interval s cc**“), kde se intervalový odhad bere jako množina π, pro které platí

$$\lim_{n\to\infty} P\left(z_{\frac{\alpha}{2}}\sqrt{\frac{\pi(1-\pi)}{n}} \le (p-\pi) \le z_{1-\frac{\alpha}{2}}\sqrt{\frac{\pi(1-\pi)}{n}}\right) = 1-\alpha. \tag{3.8}$$

Tomu odpovídá intervalový odhad s mezemi

$$\begin{aligned}
\pi_D &= \begin{cases} \dfrac{2np+z^2_{1-\frac{\alpha}{2}} - z_{1-\frac{\alpha}{2}}\sqrt{4np(1-p)+z^2_{1-\frac{\alpha}{2}}}}{2\left(n+z^2_{1-\frac{\alpha}{2}}\right)}, & x>0, \\ 0, & x=0, \end{cases} \\
\pi_H &= \begin{cases} \dfrac{2np+z^2_{1-\frac{\alpha}{2}} + z_{1-\frac{\alpha}{2}}\sqrt{4np(1-p)+z^2_{1-\frac{\alpha}{2}}}}{2\left(n+z^2_{1-\frac{\alpha}{2}}\right)}, & x<n, \\ 1, & x=n. \end{cases}
\end{aligned} \tag{3.9}$$

Použití korekce na spojitost vede ke konzervativnímu intervalovému odhadu (viz Obr. 5).

3.3. Clopperův-Pearsonův interval

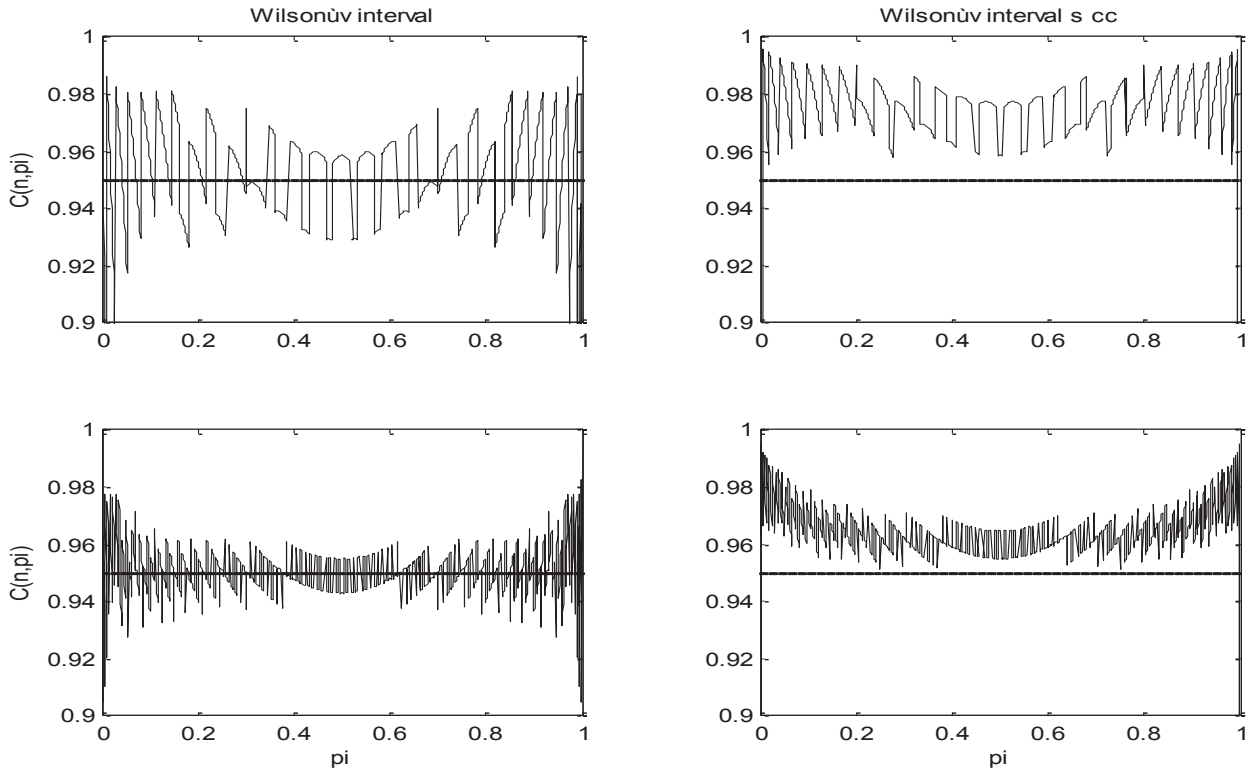
Poměrně známou alternativou k Waldovu intervalu je Clopperův-Pearsonův interval. Clopper a Pearson (1934) uvedli ve svém článku exaktní vztahy pro meze intervalového odhadu pro případy, kdy pozorujeme $X = 0$ nebo $X = n$ úspěchů v n pokusech.

$$\pi_D(0) = 0, \quad \pi_D(n) = \left(\frac{\alpha}{2}\right)^n, \quad \pi_H(0) = 1-\left(\frac{\alpha}{2}\right)^n, \quad \pi_H(n) = 1 \tag{3.10}$$

V ostatních případech jsou π_D, π_H řešením rovnic

$$\sum_{i=X}^{n}\binom{n}{i}\pi_D^i(1-\pi_D)^{n-i} = \frac{\alpha}{2}, \quad \sum_{i=0}^{X}\binom{n}{i}\pi_H^i(1-\pi_H)^{n-i} = \frac{\alpha}{2}. \tag{3.11}$$

Vzhledem k tomu, že je tento poměrně často používaný intervalový odhad založen přímo na binomickém rozdělení, nikoliv na jeho aproximaci, je



Obrázek 5: Srovnání pokrytí Wilsonova intervalu a Wilsonova intervalu s cc v závislosti na π pro spolehlivost odhadu $1-\alpha = 0{,}95$, $n = 20$ (nahoře) a $n = 100$ (dole).

často označován jako „exaktní“. Pro výpočet π_D a π_H dle (3.11) je nutno použít iterační metody a výpočet se tak stává numericky náročným. Mnohem výhodnější je použít možnosti vyjádření binomického rozdělení pomocí Fisherova-Snedecorova rozdělení uvedeného např. v Anděl (1993). Pak lze meze intervalového odhadu $\langle \pi_D, \pi_H \rangle$ pro $X \neq 0$, n určit jako

$$\begin{aligned} \pi_D &= \frac{x}{x + (n-x+1)F^{-1}_{2(n-x+1),2x}\left(1-\frac{\alpha}{2}\right)}, \\ \pi_D &= \frac{(x+1)F^{-1}_{2(x+1),2(n-x)}\left(1-\frac{\alpha}{2}\right)}{n-x+(x+1)F^{-1}_{2(x+1),2(n-x)}\left(1-\frac{\alpha}{2}\right)}, \end{aligned} \tag{3.12}$$

kde $F^{-1}_{m,n}(\alpha)$ je α kvantil Fisherova-Snedecorova rozdělení s m, n stupni volnosti. Rovněž lze ukázat, že meze intervalového odhadu $\langle \pi_D, \pi_H \rangle$ pro $X \neq 0$,

n lze určit pomocí beta rozdělení (Rutledge, Wagner, 1999). Pak

$$\pi_D = \begin{cases} 0, & x = 0 \\ 1 - beta^{-1}_{n-x+1,x}\left(1 - \frac{\alpha}{2}\right), & 0 < x < n \\ \left(\frac{\alpha}{2}\right)^{\frac{1}{n}}, & x = n, \end{cases}$$

$$\pi_H = \begin{cases} 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{n}}, & x = 0 \\ 1 - beta^{-1}_{n-x,x+1}\left(\frac{\alpha}{2}\right), & 0 < x < n \\ 1, & x = n, \end{cases} \tag{3.13}$$

kde $beta^{-1}_{a,b}(\alpha)$ je α kvantil beta rozdělení s parametry a, b. (Poznámka: Další „exaktní“ intervalové odhady, tj. odhady parametru binomického rozdělení vycházející z binomického rozdělení, lze najít například v Sterne (1954), Crow (1956), Clunies-Ross (1958), Blyth & Still (1983) a Reiczigel (2003). Zmíněné IO určujeme pomocí numerických metod. Jejich stanovení je pro naše potřeby výpočetně náročné, a proto nejsou do dalších analýz zařazeny.)

Na Obr. 6 lze vidět, že Clopperův-Pearsonův interval je silně konzervativní, tj. že pokrytí tohoto intervalu pro všechna (n, π) převyšuje nominální hodnotu $1-\alpha$. Cenou za vysoké pokrytí je velká šířka Clopperova-Pearsonova odhadu (viz Obr. 15, Obr. 16).

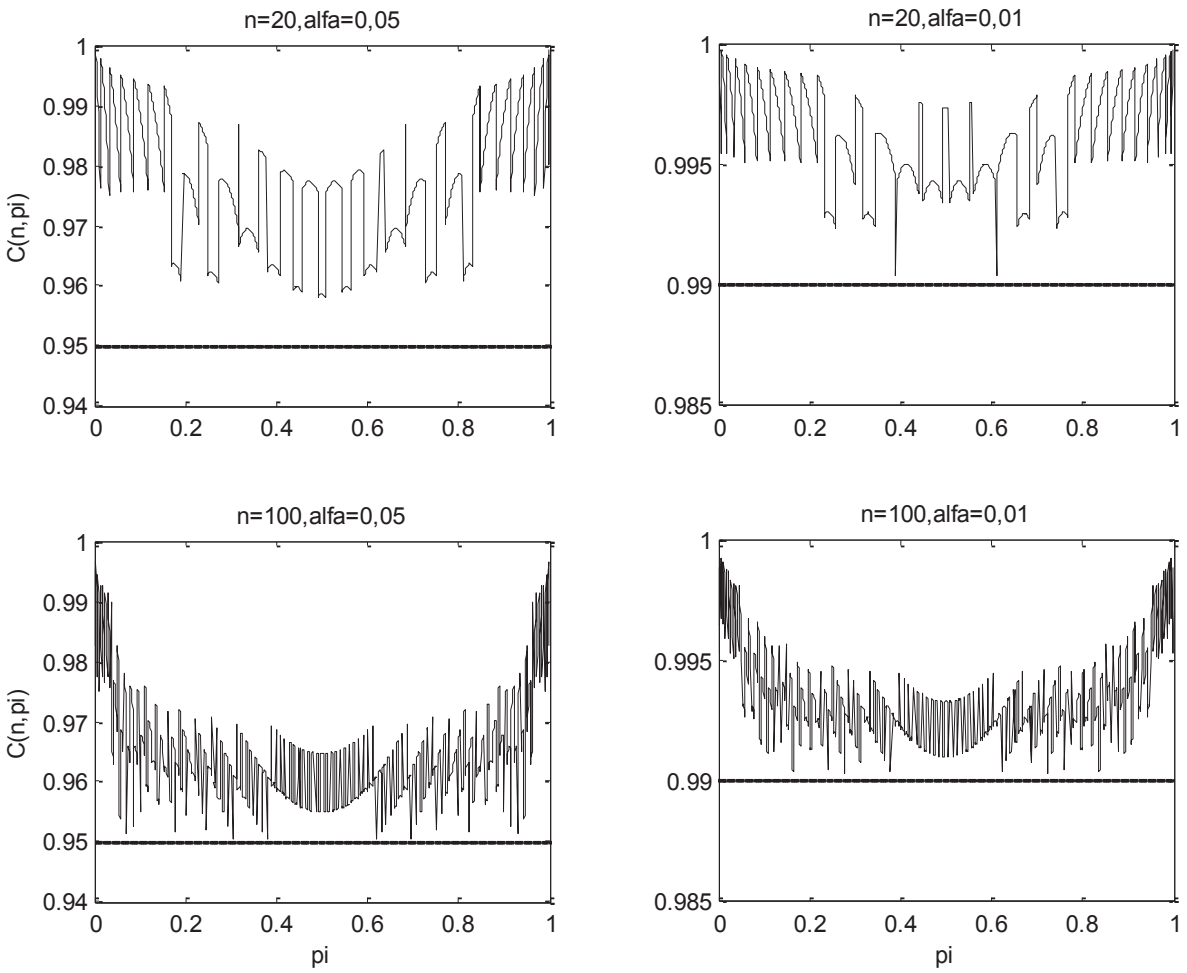
3.4. Arcsinový interval

Další prezentovaná alternativní metoda pro nalezení intervalového odhadu parametru π binomického rozdělení je založena na využití arcsinové transformace $Y = \arcsin\sqrt{X/n}$ pro stabilizaci rozptylu v binomickém rozdělení (viz např. Bickel a Docksum, 1977). Meze pro arcsinový intervalový odhad parametru binomického rozdělení π jsou

$$\pi_D = \begin{cases} \sin^2\left(\arcsin\sqrt{\frac{x}{n}} - \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right), & x > 0 \\ 0, & x = 0, \end{cases}$$

$$\pi_H = \begin{cases} \sin^2\left(\arcsin\sqrt{\frac{x}{n}} + \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right), & x < n \\ 1, & x = n. \end{cases} \tag{3.14}$$

Anscombe (1948) ukázal, že arcsinová transformace ve tvaru $Y = \arcsin\sqrt{(8nX+3)/(8n+6)}$ je oproti transformaci $Y = \arcsin\sqrt{X/n}$ stabil-

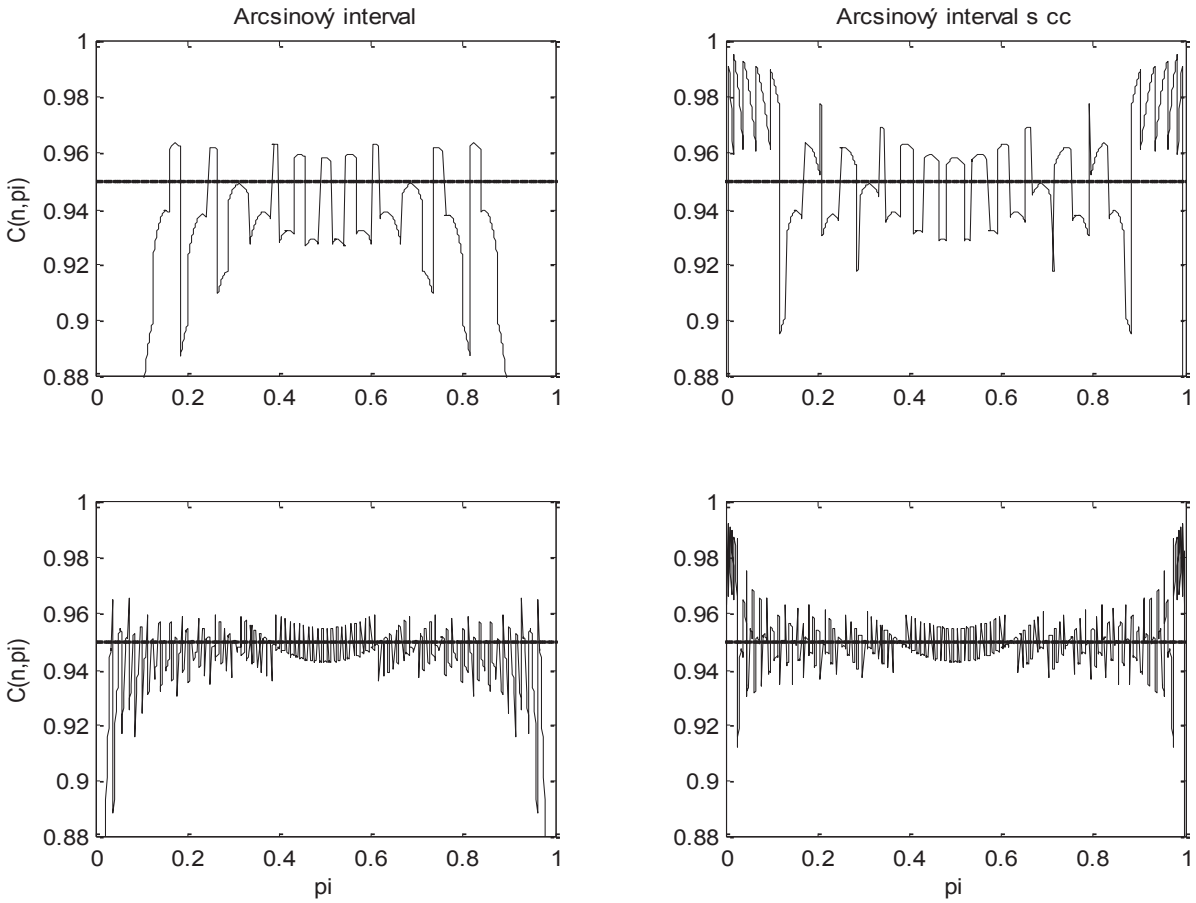


Obrázek 6: Závislost pokrytí $C(n,\pi)$ na π pro $n = 20; 100$, $1-\alpha = 0{,}95; 0{,}99$

nější, tj. má menší rozptyl. S využitím této transformace lze odvodit intervalový odhad s mezemi

$$\pi_D = \begin{cases} \sin^2\left(\arcsin\sqrt{\frac{8x+6}{8n+3}} - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{4n+2}}\right), & x > 0 \\ 0, & x = 0, \end{cases}$$
$$\pi_H = \begin{cases} \sin^2\left(\arcsin\sqrt{\frac{8x+6}{8n+3}} + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{4n+2}}\right), & x < n \\ 1, & x = n. \end{cases} \tag{3.15}$$

Intervalový odhad s mezemi (3.15) budeme dále nazývat „**arcsinový interval II**“. Na Obr. 7 lze sledovat významný vliv korekce na spojitost na pokrytí arcsinového intervalu zejména v oblasti krajních hodnot π.



Obrázek 7: Srovnání pokrytí arcsinového intervalu a arcsinového intervalu II v závislosti na π pro spolehlivost odhadu $1-\alpha = 0{,}95$, $n = 20$ (nahoře) a $n = 100$ (dole).

3.5. Adjustovaný Waldův (Agrestiho-Coullův) intervalový odhad

Agresti a Coull (1998) navrhli, zvláště pro výuku, aproximaci Wilsonova intervalu. Všimněme si, že střed Wilsonova intervalového odhadu (3.7) je

$$\pi_S = \frac{np}{n + z^2_{1-\frac{\alpha}{2}}} + \frac{\frac{z^2_{1-\frac{\alpha}{2}}}{2}}{n + z^2_{1-\frac{\alpha}{2}}} = \frac{x + \frac{z^2_{1-\frac{\alpha}{2}}}{2}}{n + z^2_{1-\frac{\alpha}{2}}},$$

kde x je počet pozorovaných úspěchů v n pokusech. Označme

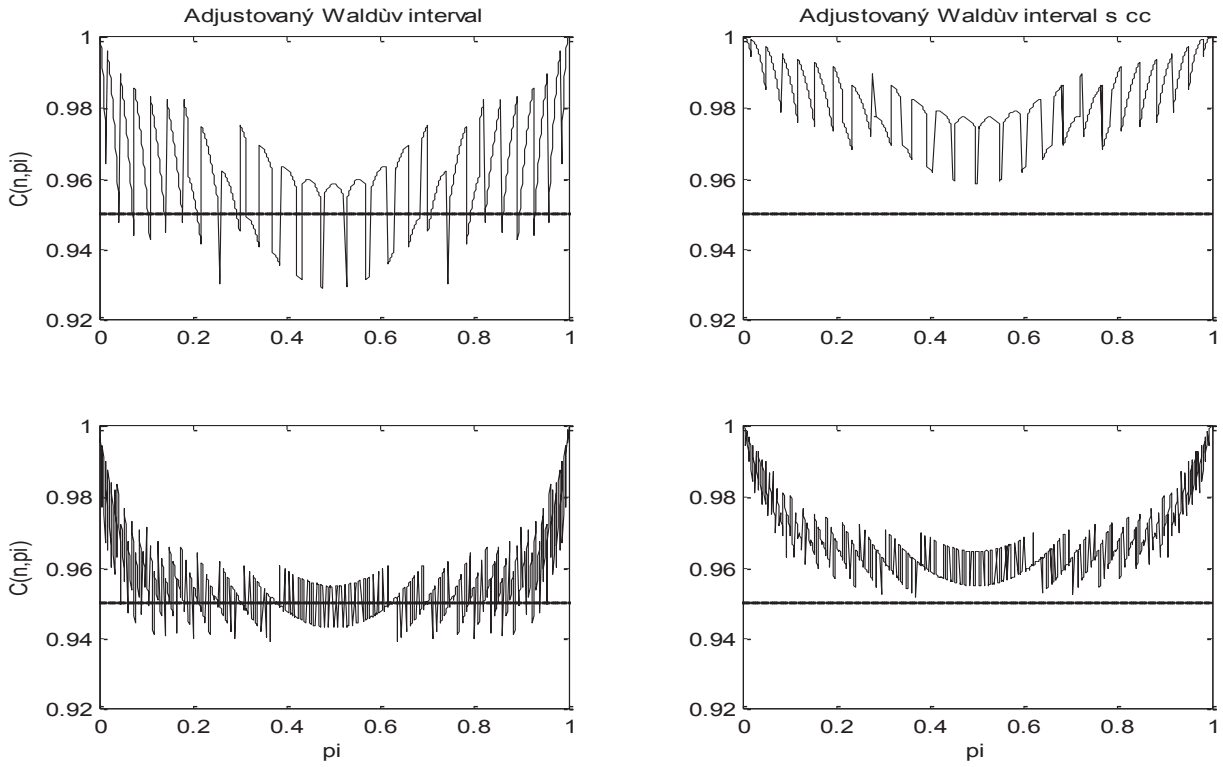
$$x_a = x + \frac{z^2_{1-\frac{\alpha}{2}}}{2}, \qquad n_a = n + z^2_{1-\frac{\alpha}{2}}, \qquad p_a = \frac{x_a}{n_a}.$$

Adjustovaný Waldův interval, podle autorů rovněž nazývaný Agrestiho-Coullův interval, se počítá jako Waldův interval, s tím rozdílem, že se místo

n použije n_a a $p = \frac{X}{n}$ se nahradí p_a. Pak je oboustranný adjustovaný Waldův $100(1-\alpha)\%$ interval

$$\left\langle p_a - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_a(1-p_a)}{n_a}}; p_a + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_a(1-p_a)}{n_a}}\right\rangle. \tag{3.16}$$

Za zmínku stojí, že adjustovaný Waldův interval je mnohdy uváděn pouze pro spolehlivost 95%. V tomto případě je využito aproximace $z_{1-\frac{\alpha}{2}} \doteq 1{,}96 \doteq 2$. Pak $x_a = x{+}2$ a $n_a = n{+}4$, což bývá prezentováno jako přidání dvou „úspěchů“ a dvou „neúspěchů“ k výběrovému souboru. Meze adjustovaného Waldova intervalu mohou, stejně jako u Waldova intervalu, ležet mimo interval $\langle 0; 1\rangle$, interval však nikdy nedegeneruje na jeden bod.



Obrázek 8: Srovnání pokrytí adjustovaného Waldova intervalu a adjustovaného Waldova intervalu s cc v závislosti na π pro spolehlivost odhadu $1-\alpha = 0{,}95$, $n = 20$ (nahoře) a $n = 100$ (dole).

Stejně jako Waldův interval, lze také adjustovaný Waldův interval zpřesnit aplikací korekce na spojitost, korekce na interval $\langle 0; 1\rangle$ a využitím Cloppe-rových-Pearsonových mezí pro případy, kdy je pozorováno 0 nebo n výskytů události. Takto korigované meze intervalového odhadu (dále nazývaného jako

„**adjustovaný Waldův interval s cc**“) lze určit dle vztahů

$$\pi_D = \begin{cases} 0, & x = 0 \\ \max\left\{p_a - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_a(1-p_a)}{n_a}} - \frac{1}{2n_a}; 0\right\}, & 0 < x < n \\ \left(\frac{\alpha}{2}\right)^{\frac{1}{n}}, & x = n, \end{cases}$$

$$\pi_H = \begin{cases} 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{n}}, & x = 0 \\ \max\left\{p_a + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_a(1-p_a)}{n_a}} + \frac{1}{2n_a}; 1\right\}, & 0 < x < n \\ 1, & x = n. \end{cases} \tag{3.17}$$

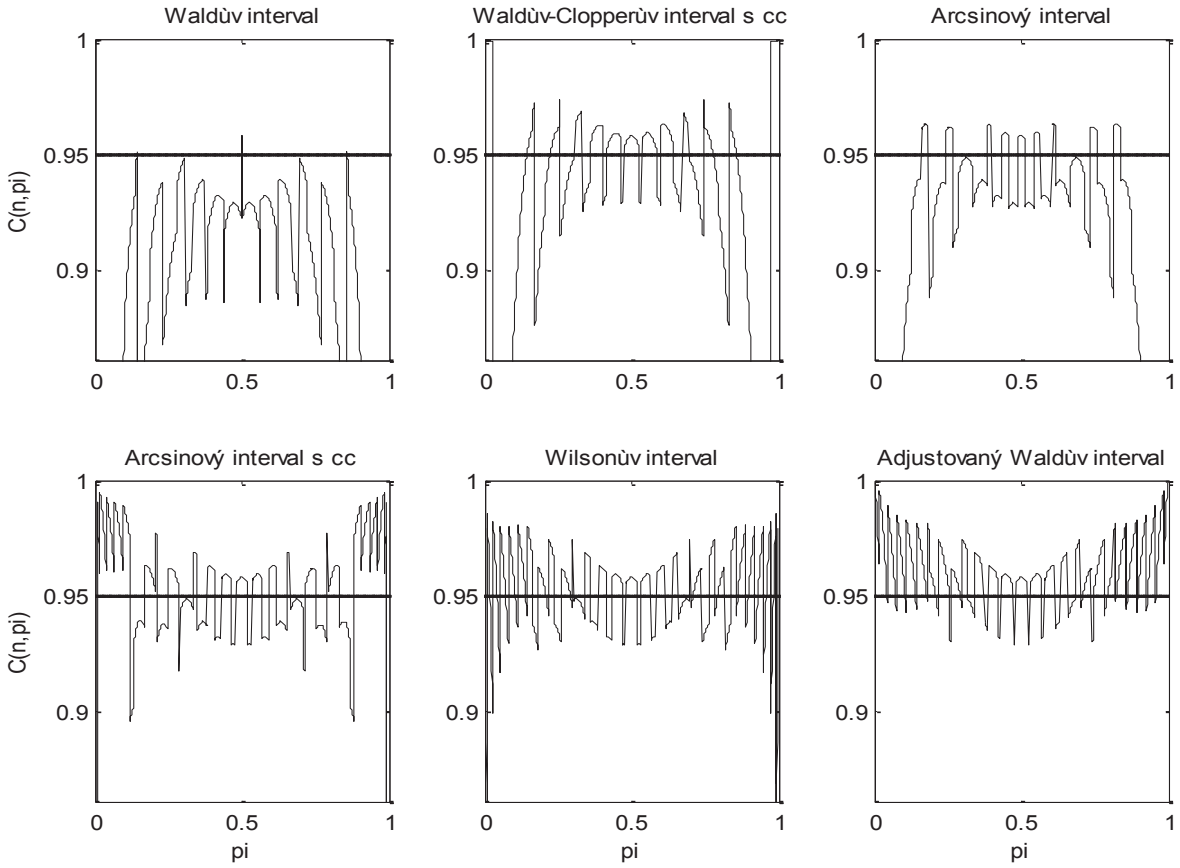
Komplexnímu srovnání analyzovaných liberálních i konzervativních intervalových odhadů parametru π binomického rozdělení pro nízké rozsahy výběru je věnována následující kapitola.

4. Porovnání analyzovaných intervalových odhadů

Srovnáním pravděpodobnosti pokrytí Waldova intervalu a ostatních liberálních intervalových odhadů popsaných v předcházejících kapitolách lze dojít k následujícím závěrům: v případě, že sledujeme pokrytí pro nízký rozsah výběru (viz Obr. 9), vidíme, že o něco lepší výsledky než Waldův interval poskytuje interval arcsinový, popř. Waldův-Clopperův interval s cc. Přiblížení se pravděpodobnosti pokrytí k nominální hodnotě $1 - \alpha$ je u těchto intervalů rychlejší, počet „šťastných“ dvojic $(n; \pi)$ je vyšší. Posuzujeme-li analyzované intervaly pouze z hlediska pokrytí, jeví se jako vhodný Wilsonův interval, jehož pravděpodobnost pokrytí osciluje kolem nominální hodnoty $1 - \alpha$ pro všechna π na intervalu $\langle 0; 1\rangle$, arcsinový interval s cc zajišťující vyšší pokrytí pro hodnoty π blízké 0 nebo 1, resp. adjustovaný Waldův interval vykazující velmi přijatelné minimum pravděpodobnosti pokrytí.

Srovnáme-li průběhy pravděpodobnosti pokrytí analyzovaných liberálních intervalových odhadů pro $\pi = 0{,}01$, $1 - \alpha = 0{,}95$ a $10 \leq n \leq 1000$ (Obr. 10), dojdeme k obdobným závěrům jako v případě analýzy pravděpodobnosti pokrytí v závislosti na skutečné pravděpodobnosti π. Nejlepší pokrytí (a to i pro malé rozsahy výběru) lze očekávat u arcsinového intervalu s cc, Wilsonova intervalu a adjustovaného Waldova intervalu.

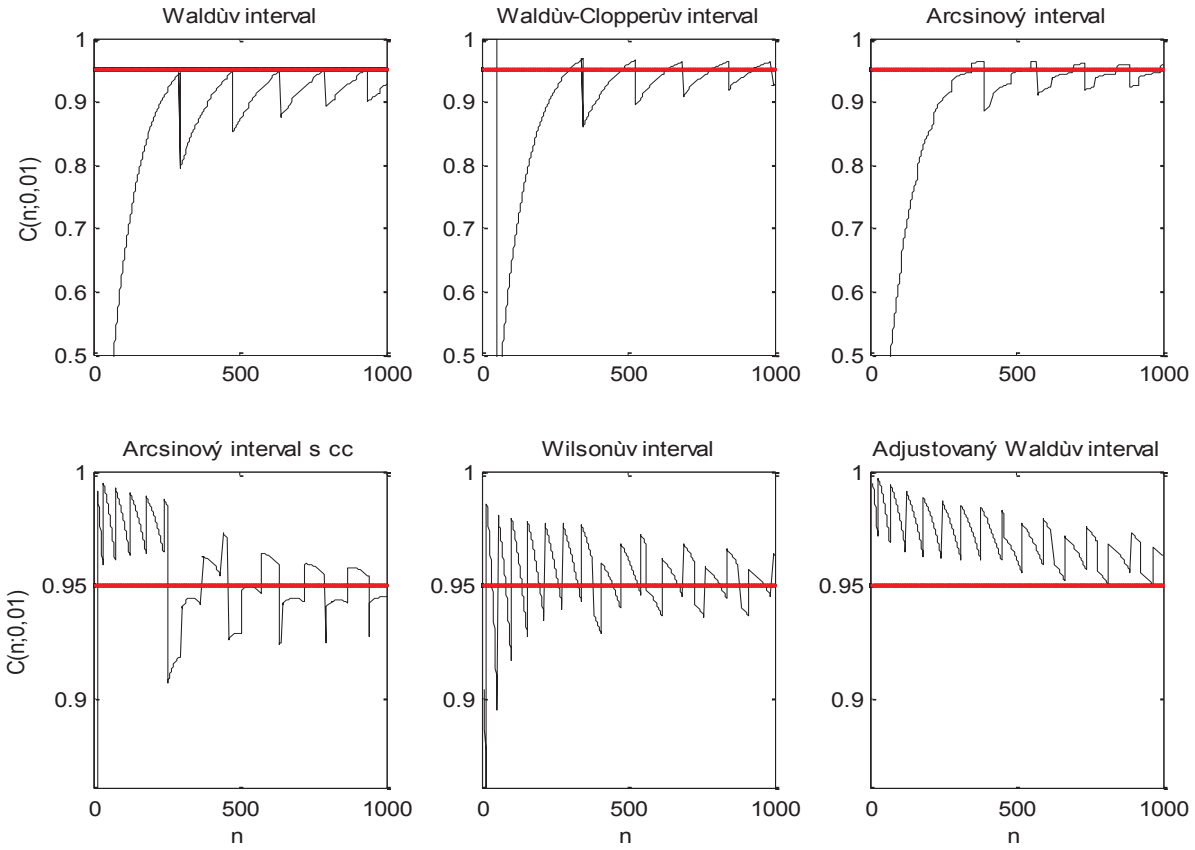
Pro srovnání analyzovaných liberálních odhadů z hlediska jejich očekávané délky bylo použito porovnání poměrů střední délky jednotlivých IO a střední délky Waldova IO v závislosti na n (viz Obr. 11). Nejmenší střední délku ze všech analyzovaných liberálních IO vykazuje pro $n > 12$ arcsinový interval.



Obrázek 9: Srovnání pokrytí analyzovaných liberálních intervalů pro $n = 20$, $1 - \alpha = 0{,}95$

Jeho pokrytí (viz Obr. 12) však není optimální. Preferujeme-li u odhadu jeho minimální délku, lze doporučit arcsinový interval II, který nepřekračuje střední délku Waldova IO o více než 4 %. Pro $n > 25$ klesne rozdíl mezi střední délkou Waldova IO a arcsinového intervalu II na méně než 1 %. Ještě lepších výsledků dosahuje Wilsonův interval, který vykazuje stejnou střední délku jako Waldův interval pro všechna $n = 10, 11, \ldots, 100$, a jeho pokrytí osciluje kolem nominální hodnoty. Adjustovaný Waldův interval, který má nejvyšší minimální pokrytí z analyzovaných intervalů, lze z hlediska střední délky rovněž považovat za vhodný.

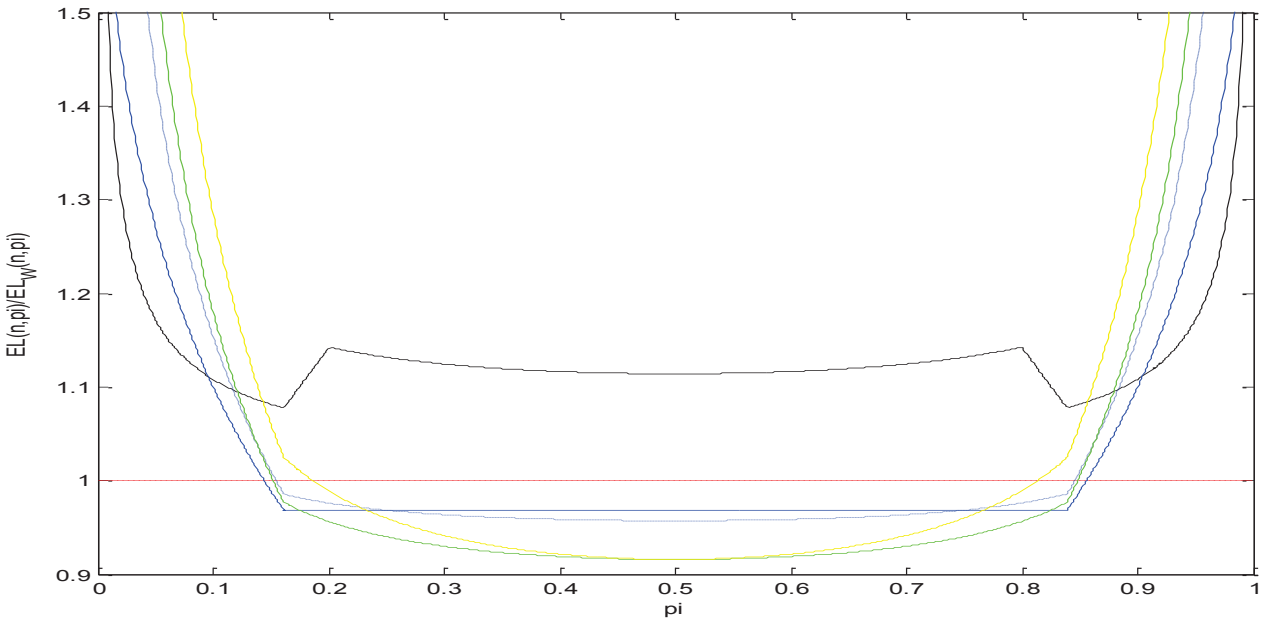
Rozdíl mezi střední délkou Waldova IO a adjustovaného Waldova intervalu se pohybuje mezi 4 % a 1 %. Výrazně největší střední délku z analyzovaných IO má Waldův-Clopperův interval s cc převyšující střední délku IO až o 15 %. Z Obr. 11 lze usuzovat na to, že v oblasti π blízkých 0 nebo 1 je při daném rozsahu výběru n délka Waldova IO výrazně nejnižší. Druhou nejmenší délku pak lze pozorovat u Waldova-Clopperova IO.



Obrázek 10: Srovnání pokrytí analyzovaných liberálních odhadů pro $\pi = 0{,}01$, $1 - \alpha = 0{,}95$ a $10 \leq n \leq 1000$

Potřebujeme-li zajistit konzervativní pokrytí, lze volit mezi Clopperovým-Pearsonovým IO, adjustovaným Waldovým intervalem s cc a Wilsonovým intervalem s cc. Tyto IO garantují minimálně nominální pokrytí parametru binomického rozdělení (tj. jejich pokrytí neklesne pod nominální hodnotu pro žádné π na intervalu $\langle 0; 1 \rangle$). Pokrytí těchto intervalů lze srovnat na základě Obr. 13 a Obr. 14. Jak lze očekávat, cenou za vysoké hodnoty pokrytí je velká střední délka těchto intervalových odhadů.

Zatímco střední délka liberálních odhadů nepřekročila ve většině případů střední délku Waldova intervalu o více než 4 %, střední délka konzervativních odhadů překračuje pro rozsah výběru $n = 10$ střední délku Waldova intervalu o 15 % až 22 % (viz Obr. 16). Při $1 - \alpha = 0{,}95$ má pro $10 \leq n \leq 32$ nejmenší střední délku Wilsonův odhad s cc. Pro $n \geq 33$ má nejmenší střední délku Clopperův-Pearsonův interval, rozdíl ve střední délce Clopperova-Pearsonova intervalu a Wilsonova intervalu s cc lze však pro $n \geq 33$ považovat za minimální. Srovnáme-li průběh očekávané délky konzervativních odhadů v závislosti na π pro $n = 20$ (viz Obr. 15), lze opět pozorovat srovnatelnou očeká-

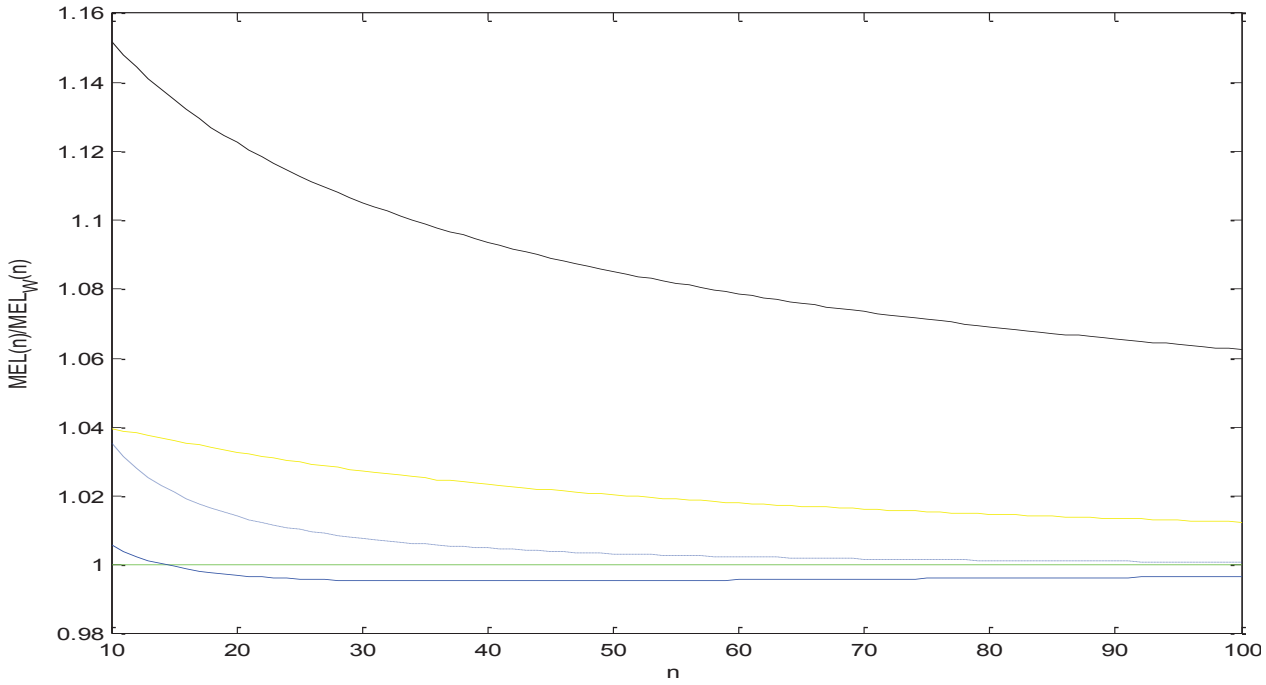


Obrázek 11: Srovnání poměrů délky analyzovaných liberálních odhadů k délce Waldova intervalu pro $n = 20$, $1 - \alpha = 0{,}95$ (Waldův-Clopperův odhad s cc – černá, arcsinový odhad – modrá plná, arcsinový odhad II – modrá čerchovaná, Wilsonův odhad – zelená, adjustovaný Waldův odhad – žlutá).

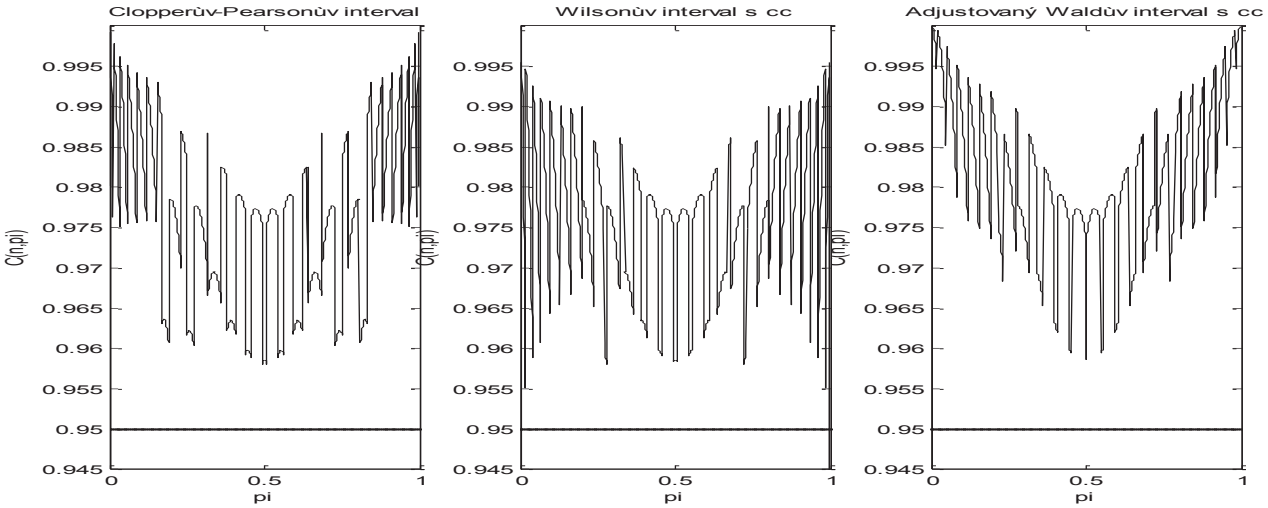
vanou délku Clopperova-Pearsonova intervalu a Wilsonova intervalu s cc pro hodnoty π blízké 0 nebo 1. V okolí $\pi = 0{,}5$ se pak srovnává očekávaná délka všech analyzovaných IO.

5. Software pro výpočet intervalového odhadu parametru binomického rozdělení

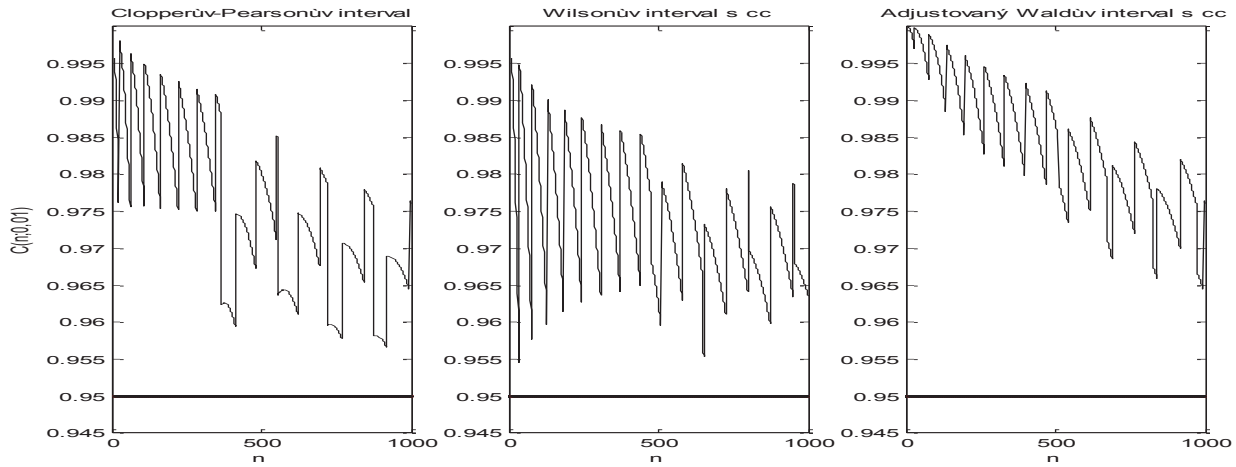
Požadavek na odhad parametru binomického rozdělení se v biomedicínské praxi vyskytuje velmi často. Srovnáme-li ve statistické praxi nejčastěji používané programy (SAS, S-PLUS, SPSS, STATGRAPHICS, R), dojdeme k následujícím závěrům. SAS 9.2 umožňuje výpočet Waldova, adjustovaného Waldova, exaktního Clopperova-Pearsonova, Jeffreysova a Wilsonova skórového intervalu, přičemž u Clopperova-Pearsonova a Jeffreysova intervalu je dolní mez IO nastavena pro $x = 0$ na nulu a pro $x = n$ na jedničku. S-PLUS 8 poskytuje Waldův, exaktní Clopperův-Pearsonův a Wilsonův IO (i s korekcí na spojitost). SPSS až do verze 17 neumožňoval výpočet intervalových odhadů parametru binomického rozdělení. SPSS 18 nabízí Waldův, Jeffreysův a Clopperův-Pearsonův IO. Statgraphics Centurion používá exaktní Clopperův-Pearsonův IO a pro volně šiřitelný software R byla v současné



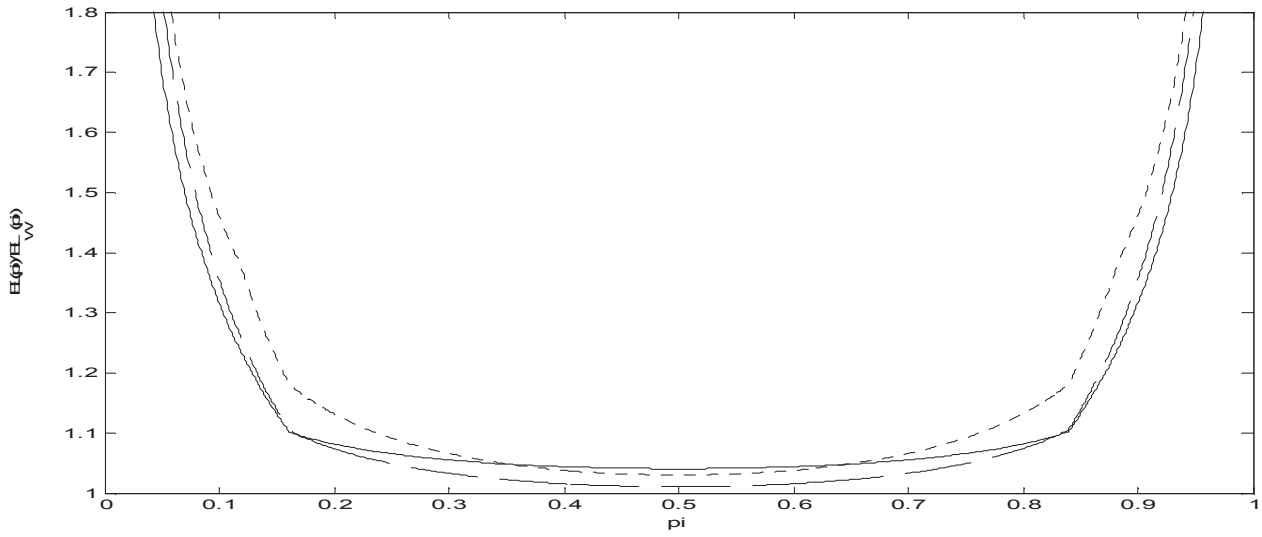
Obrázek 12: Srovnání poměrů střední délky analyzovaných liberálních odhadů ke střední délce Waldova intervalu v závilosti na n pro $1 - \alpha = 0{,}95$ (Waldův-Clopperův odhad s cc – černá, arcsinový odhad – modrá plná, arcsinový odhad II – modrá čerchovaná, Wilsonův odhad – zelená, adjustovaný Waldův odhad – žlutá).



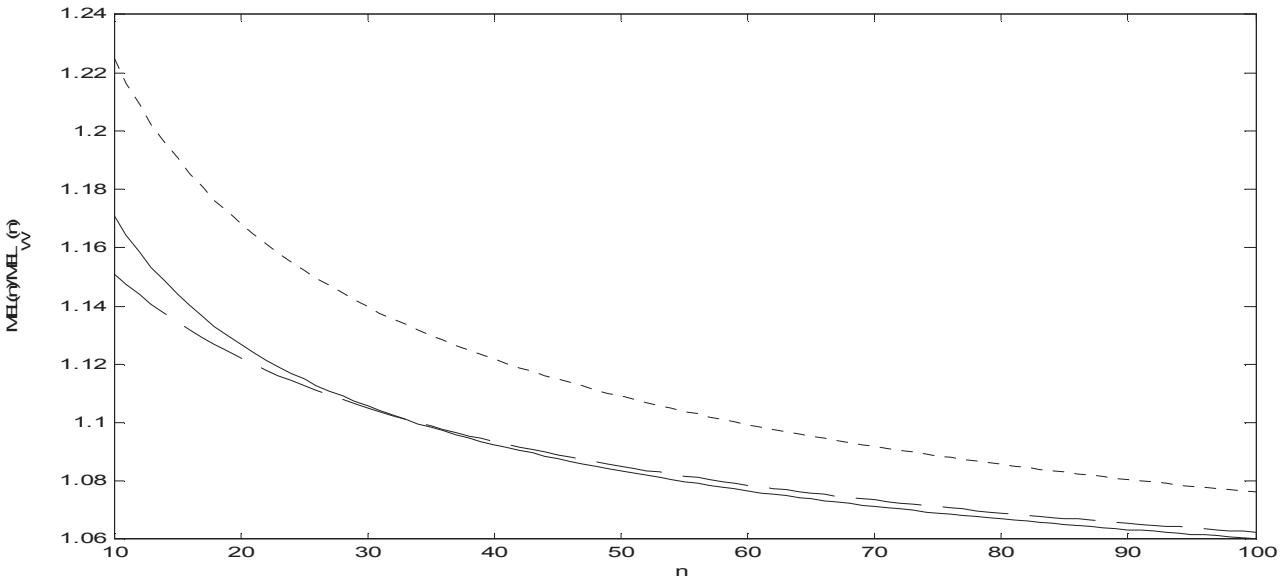
Obrázek 13: Srovnání pokrytí analyzovaných konzervativních intervalů pro $n = 20$, $1 - \alpha = 0{,}95$



Obrázek 14: Srovnání pokrytí analyzovaných konzervativních intervalů pro $\pi = 0{,}01$, $1 - \alpha = 0{,}95$ a $10 \leq n \leq 1000$



Obrázek 15: Srovnání poměrů délky analyzovaných konzervativních odhadů k délce Waldova intervalu pro $n = 20$, $1 - \alpha = 0{,}95$ (Clopperův-Pearsonův interval – plná, Wilsonův interval s cc – čerchovaná, adjustovaný Waldův interval – tečkovaná)



Obrázek 16: Srovnání poměrů střední délky analyzovaných konzervativních odhadů ke střední délce Waldova intervalu v závilosti na n pro $\pi = 0{,}01; 1 - \alpha = 0{,}95$ (Clopperův-Pearsonův interval – plná, Wilsonův interval s cc – čerchovaná, adjustovaný Waldův interval – tečkovaná)

	A	B	C	D	E	F	G	H	I	J
12										
13	rozsah výběru n:		34							
14	počet úspěchů x:		2							
15	spolehlivost odhadu 1-α:		0,95							
16										
17	Výsledky:									
18										
19	Bodový odhad parametru π:		0,0588							
20										
21	Liberální intervalové odhady									
22	Název			Odhad se spolehlivostí 0,95		Délka odhadu				
23	Adjustovaný Waldův interval			<0,0065; 0,2007>		0,1942				
24	Wilsonův interval			<0,0163; 0,1909>		0,1747				
25	Arcsinový interval s cc			<0,0143; 0,192>		0,1777				
26	Waldův-Clopperův interval s cc			<0; 0,1379>		0,1379				
27	Arcsinový interval			<0,0059; 0,1611>		0,1552				
28	Waldův interval			<0; 0,1379>		0,1379	Minimální požadovaný rozsah výběru pro použití Waldova intervalu je 161.			
29										
30										
31	Konzervativní intervalové odhady									
32	Název			Odhad se spolehlivostí 0,95		Délka odhadu				
33	Clopperův-Pearsonův interval			<0,0072; 0,1968>		0,1896				
34	Wilsonův interval s cc			<0,0103; 0,2106>		0,2003				
35	Adjustovaný Waldův interval s cc			<0,0065; 0,2007>		0,1942				
36										

Instrukce a výpočet | Použité vzorce

Obrázek 17: Náhled na výpočetní applet IO_binom.xlsx – výsledky

době vytvořena makra pro výpočet osmi různých IO parametru binomického rozdělení (Dorai-Raj, 2011). Vzhledem k tomu, že lékaři obvykle nemají pro základní analýzu svých dat k dispozici statistický software, byl pro usnadnění aplikace těchto intervalů v praxi navržen výpočetní applet IO_binom.xlsx (Obr. 17), který umožňuje automaticky výpočet všech intervalových odhadů analyzovaných v předcházejících kapitolách. Pro návrh appletu byl použit běžně dostupný software Microsoft Excel, přičemž pro využití appletu lze využít rovněž volně dostupné OpenOffice.org. Intervalové odhady jsou v appletu, stejně jako v analýzách prováděných v předkládané práci, rozděleny na liberální a konzervativní. Liberální odhady jsou pak seřazeny podle velikosti pokrytí pro nízké pozorované relativní četnosti výskytu události. Jako doplňková informace je pro všechny intervaly stanovena délka IO.

Poděkování: Tato práce byla podporována z FEECS VŠB – Technická Univerzita Ostrava (číslo projektu SP 2012/108) a také Ministerstvem školství, mládeže a tělovýchovy České republiky (číslo projektu 1M06047).

Literatura

[1] Agresti, A.; Coull, B. A. (1998), Approximate is better than „exact“ for interval estimation of binomial proportions, *The American Statistician*, Vol. 52, pp. 119–126.

[2] Agresti, A.; Caffo, B. (2000), Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures, *The American Statistician*, Vol. 54, pp. 280–288.

[3] Anděl, J. (1993), *Statistické metody*, Matfyzpress, Praha, Vydavatelství Matematicko-fyzikální fakulty Univerzity Karlovy v Praze.

[4] Anděl, M.; Černý, R.; Charamza, P., Neustadt, J. (2004), Přehled metod odhadu statistické chyby ve výběrových šetřeních, *Informační Bulletin České statistické společnosti*, číslo 2–3, ročník 15, ISSN 1210–8022.

[5] Anscombe, F. J. (1948), The Transformation of Poisson, Binomial and Negative-Binomial Data, *Biometrika*, Vol. 35, pp. 246–254.

[6] Bickel, P. J.; Docksum, K. A. (1977). *Mathematical Statistics*. San Francisco, Holden Day.

[7] Blyth, C. R.; Still, H. A. (1983), Binomial confidence intervals, *Journal of the American Statistical Association*, Vol. 78, pp. 108–116.

[8] Brown, D.; Cai, T.; Dasgupta, A. (2001), Interval Estimation for a Binomial Proportion, *Statistical Science*, Vol. 16, Issue 2, pp. 101–133.

[9] Clopper, C. J.; Pearson, E. S. (1934), The use of confidence or fiducial limits illustrated in the case of the binoval, *Biometrika*, Vol. 26, pp. 404–413.

[10] Clunies-Ross, C. W. (1953), Interval estimation for the parameter of binomial proportions, *Biometrika*, Vol. 45, pp. 275–279.

[11] Crow, E. L. (1956), Confidence interval for a proportion, *Biometrika*, Vol. 43, pp. 423–435.

[12] Dorai-Raj, S. (2011), R-Documentation: Binomial Confidence Interval, [cit. 2011-09-25], dostupný na World Wide Web: `http://rss.acs.unt.edu/Rdoc/library/binom/html/binom.confint.html`.

[13] Neyman, J. (1935), On the problem of confidence limits, *The Annals of Mathematical Statistics*, Vol. 6, pp. 111–116.

[14] Pires, A. M.; Conceicao, A. (2008), Interval estimators for binomial proportion: Comparison of twenty methods, *Statistical Journal*, Vol. 6, pp. 165–197.

[15] Reiczigel, J. (2003), Confidence intervals for the binomial parameter: some new considerations, *Statistics in Medicine*, Vol. 22, pp. 611–621.

[16] Rutledge, J.; Warner, B. (1999), *Using the Beta Distribution on Confidence Intervals for Proportions*, [cit. 2011-07-15], více na World Wide Web: `http://www.data-vision.biz/extra/CI_Proportion.PDF`.

[17] Vollset, S. E. (1993), Confidence intervals for a binomial proportion, *Statistics in Medicine*, Vol. 12, pp. 809–824, doi: 10.1002/sim.4780120902

[18] Wald, A.; Wolfowitz, J. (1939), Confidence limits for continuous distribution functions, *The Annals of Mathematical Statistics*, Vol. 10, pp. 105–118.

[19] Wilson, E. B. (1927), Probable inference, the law of succession, and statistical inference, *Journal of the American Statistical Association*, Vol. 22, pp. 209–212.

PRAVDĚPODOBNOSTNÍ ROZDĚLENÍ V MICROSOFT EXCEL 2010

Luboš Marek

Adresa: Luboš Marek, Katedra statistiky a pravděpodobnosti, Fakulta informatiky a statistiky, Vysoká škola ekonomická v Praze, nám. W. Churchilla 4, 138 00, Praha 3

E-mail: marek@vse.cz

Poděkování: Příspěvek je součástí řešení projektu GAČR P402/12/G097 „DYME – Dynamické modely v ekonomii“.

Abstract: This paper deals with new probability functions in Microsoft Excel 2010. The Czech version of Microsoft Excel 2003 has a different list of probability functions than the 2010 version. Because the author often has to deal with wrong use of these functions, he decided to write this paper. The possibilities of Microsoft Excel in computation of basic probability functions are comparable with statistical packages. Achieved results in Excel are the same as in Stagraphics Centurion, etc. The number of offered probability distributions is obviously much smaller. There was an unification of the function names. So that the function name consists of distribution label, followed by dot and the text DIST (for distribution or probability function or density function) or INV (for the inversed distribution function). The function syntax is now consistent and uniformed.

1. Úvod

Programový balík Microsoft (dále MS) Office, jehož součástí je i tabulkový kalkulátor MS Excel, je dnes instalován téměř na každém osobním počítači. MS Excel je nezastupitelný při práci s daty, a navíc je velmi dobře vybaven i statistickými procedurami a funkcemi. Pomineme-li trochu nešťastný překlad některých termínů z angličtiny (je zjevné, že překladatel nebyl statistik), v mnoha základních funkcích či procedurách si nezadá i se specializovanými statistickými programy. To se týká i práce s pravděpodobnostními rozděleními a jim příslušejícími funkcemi (distribuční funkce, pravděpodobnostní funkce, hustota pravděpodobnosti a kvantily).

Již před několika roky jsem napsal článek o pravděpodobnostních rozděleních v MS Excel (verze 2003), neboť jsem se často setkával se špatnou aplikací pravděpodobnostních funkcí v Excelu. Protože od té doby uběhla řada let a MS Excel se objevil v dalších verzích, reaguji v tomto článku

i na poměrně výraznou změnu v této oblasti. Došlo totiž ke změnám názvu všech pravděpodobnostních funkcí a u mnoha z nich se změnil i jejich obsah. Navíc přibyly některé nové funkce, které nahrazují ty starší, přičemž byly tyto funkce rozšířeny (např. počítají oproti starším verzím nejen hodnoty distribuční funkce, ale i hodnoty hustoty pravděpodobnosti či kvantilů). Z důvodů zpětné kompatibility jsou v Excelu platné i dřívější syntaxe funkcí. To znamená, že tyto starší funkce sice fungují, ale již je v nabídce funkcí nenalezneme. Ukázkou může být např. funkce BINOM.DIST, která nahradila funkci BINOMDIST, přičemž stará i nová verze fungují úplně stejně. Naproti tomu např. ve starší verzi Excelu funkce TDIST počítá něco úplně jiného než funkce T.DIST v nové verzi. Došlo též ke sjednocení názvu funkcí, neboť nyní je název důsledně složen z označení rozdělení, pak následuje tečka a označení DIST pro distribuční funkci (resp. pravděpodobnostní funkci či hustotu) či INV (pro označení inverzní funkce k funkci distribuční, tedy pro kvantilovou funkci). Do označení funkcí tak byl vnesen určitý řád a jednotnost.

V tomto článku uvedeme všechny funkce pravděpodobnostních rozdělení ve verzi MS Excel 2010. V první řadě zhodnotíme nabídku pravděpodobnostních rozdělení v tomto programu. Dále popíšeme rozdíl v syntaxi těchto funkcí oproti verzi 2003 a okomentujeme veškeré změny. Uvedeme též nové funkce či rozšíření stávajících. Všimneme si též kvality překladu do češtiny a posoudíme přesnost výpočtů.

2. Diskrétní rozdělení

V oblasti diskrétních (nespojitých) rozdělení obsahuje MS Excel následující rozdělení, u kterých zároveň uvádíme název příslušné funkce.

Jedná se tedy o naprosto základní typy rozdělení, navíc ne vždy je možné spočítat kvantily. To ale není žádné neštěstí, neboť kvantily jsme schopni poměrně snadno spočítat z hodnot pravděpodobnostní funkce. Podívejme se nyní na jednotlivá rozdělení podrobněji. Je třeba ještě uvést, že distribuční funkce je v Excelu definována jako $F(x) = P(X \leq x), \forall x \in R$.

Tabulka 1: Přehled funkcí pro nespojitá rozdělení v MS Excel 2010

Rozdělení	**Pr. a distribuční funkce**	**Kvantily**
Binomické	BINOM.DIST	BINOM.INV
Negativně binomické	NEGBINOM.DIST	—
Poissonovo	POISSON.DIST	—
Hypergeometrické	HYPGEOM.DIST	—

Tabulka 2: Přehled funkcí pro nespojitá rozdělení v MS Excel 2003

Rozdělení	**Dist. funkce**	**Prav. funkce**	**Kvantily**
Binomické	BINOMDIST	BINOMDIST	CRITBINOM
Negativně binomické	—	NEGBINOMDIST	—
Poissonovo	POISSON	POISSON	—
Hypergeometrické	—	HYPGEOMDIST	—

2.1. Shrnutí syntaxe 1

Rozdělení	*Binomické rozdělení*	*Negativně binomické rozdělení*
$F(x)$	=BINOM.DIST(x;n;π;1)	=NEGBINOM.DIST(x;n;π;1)
$P(x)$	=BINOM.DIST(x;n;π;0)	=NEGBINOM.DIST(x;n;π;0)
x_P	=BINOM.INV(n;π;P)	—

Rozdělení	*Poissonovo rozdělení*	*Hypergeometrické rozdělení*
$F(x)$	=POISSON.DIST(x;λ;1)	=HYPGEOM.DIST(x;n;M;N;1)
$P(x)$	=POISSON.DIST(x;λ;0)	=HYPGEOM.DIST(x;n;M;N;0)
x_P	—	—

Toto shrnutí je pouze hypotetické a v Excelu se takto uvedené funkce nezobrazují. Při vlastním výpočtu na místě parametrů budou jejich konkrétní hodnoty či odkazy na hodnoty v buňkách. Do uvedené tabulky jsme zapsali parametry symbolicky, aby byl na první pohled patrný jejich význam. Drželi jsme se při tom obvyklé symboliky, běžně používané v literatuře. Např. místo HYPGEOM.DIST(x;n;M;N;1) zobrazuje Excel:

HYPGEOMDIST(úspěch; celkem; základ_úspěch; základ_celkem)

2.2. Změny proti verzi MS Excel 2003

- Sjednocení názvů funkcí.
- Přidána distribuční funkce hypergeometrického rozdělení.
- Původní funkce již nejsou v nabídce (jejich přehled je obsažen v tabulce 2), tyto funkce však není třeba ve starších sešitech MS Excel přepisovat na nový název, zpětná kompatibilita funguje.
- Funkce CRITBINOM pro výpočet kvantilů binomického rozdělení byla nahrazena funkcí BINOM.INV.

- Funkce POISSON byla nahrazena funkcí POISSON.DIST.
- Ostatní funkce se liší pouze změnou názvu – většinou má syntaxe tvar: JMENOFUNKCE.DIST.

2.3. Zhodnocení funkcí pro diskrétní rozdělení

- Pravděpodobnostní funkce jsou naprogramovány v obvyklém tvaru, jak je známe z literatury.
- Překlad do češtiny není příliš šťastný – např. pro hypergeometrické rozdělení je parametr M označen jako **Základ_úspěch**, přičemž nápověda je „počet úspěšných pokusů v základním souboru“. To samozřejmě částečně komplikuje výpočet.
- Poměrně malý počet rozdělení.

3. Spojitá rozdělení

V oblasti spojitých rozdělení je nabídka funkcí pro pravděpodobnostní rozdělení daleko bohatší, než je tomu u rozdělení diskrétních. MS Excel obsahuje většinu běžně používaných rozdělení, u kterých opět uvádíme název a syntaxi příslušné funkce jak pro výpočet hodnot distribuční funkce, tak i hodnot hustoty a kvantilů. Další funkce v MS Excel 2010 jsou zmíněny v tabulce 5.

Tabulka 3: Přehled funkcí pro spojitá rozdělení v MS Excel 2010

Rozdělení	**Dist. funkce a hustota**	**Kvantily**
Normální	NORM.DIST	NORM.INV
Normované normální	NORM.S.DIST	NORM.S.INV
Logaritmicko-normální	LOGNORM.DIST	LOGNORM.INV
Exponenciální	EXPON.DIST	—
Weibullovo	WEIBULL.DIST	—
Studentovo (t)	T.DIST	T.INV
Fisher-Snedecorovo (F)	F.DIST	F.INV
Chí-kvadrát (χ^2)	CHISQ.DIST	CHISQ.INV
Beta	BETA.DIST	BETA.INV
Gama	GAMMA.DIST	GAMMA.INV

Tabulka 4: Přehled funkcí pro spojitá rozdělení v MS Excel 2003

Rozdělení	**Dist. funkce**	**Hustota**	**Kvantily**
Normální	NORMDIST	NORMDIST	NORMINV
Normované-normální	NORMSDIST	—	NORMSINV
Logaritmicko normální	LOGNORMDIST	—	LOGINV
Exponenciální	EXPONDIST	EXPONDIST	—
Weibullovo	WEIBULL	WEIBULL	—
Studentovo (t)	TDIST	—	TINV
Fisher-Snedecorovo (F)	FDIST	—	FINV
Chí-kvadrát (χ^2)	CHIDIST	—	CHIINV
Beta	BETADIST	—	BETAINV
Gama	GAMMADIST	GAMMADIST	GAMMAINV

Tabulka 5: Přehled dalších funkcí pro spojitá rozdělení v MS Excel 2010

Rozdělení	$1 - F(x)$	$P(\|X\| > x)$	$x_{1-P/2}$	x_{1-P}
Studentovo	T.DIST.RT	T.DIST.2T	T.INV.2T	—
Fisher-Snedecorovo	F.DIST.RT	—	—	F.INV.RT
Chí-kvadrát	CHISQ.DIST.RT	—	—	CHISQ.INV.RT

3.1. Shrnutí syntaxe 2

Rozdělení	*Normální*	*Normované normální*
$F(x)$	=NORM.DIST(x;μ;σ;1)	=NORM.S.DIST(x;1)
$f(x)$	=NORM.DIST(x;μ;σ;0)	=NORM.S.DIST(x;0)
x_P	=NORM.INV(P;μ;σ)	—

Rozdělení	*Logaritmicko-normální*	*Exponenciální*
$F(x)$	=LOGNORM.DIST(x;μ;σ;1)	=EXPON.DIST(x;λ;1)
$f(x)$	=LOGNORM.DIST(x;μ;σ;0)	=EXPON.DIST(x;λ;0)
x_P	=LOGNORM.INV(P;μ;σ)	—

Rozdělení	*Weibullovo*	*t-rozdělení*
$F(x)$	=WEIBULL.DIST(x;α;β;1)	=T.DIST(x;n;1)

$f(x)$	=WEIBULL.DIST(x;α;β;0)	=T.DIST(x;n;0)
x_P	—	=T.INV(P;n)

Rozdělení	*F-rozdělení*	*Chí-kvadrát*
$F(x)$	=F.DIST(x;n;m;1)	=CHISQ.DIST(x;n;1)
$f(x)$	=F.DIST(x;n;m;0)	=CHISQ.DIST(x;n;0)
x_P	=F.INV(x;n;m)	=CHISQ.INV(P;n)

Rozdělení	*Beta*	*Gama*
$F(x)$	=BETA.DIST(x;α;β;1;a;b)	=GAMMA.DIST(x;α;β;1)
$f(x)$	=BETA.DIST(x;α;β;0;a;b)	=GAMMA.DIST(x;α;β;0)
x_P	=BETA.INV(P;α;β;a;b)	=GAMMA.INV(P;α;β)

3.2. Shrnutí syntaxe 3

Rozdělení	*t-rozdělení (t)*	*F-rozdělení (F)*	*Chí-kvadrát (χ^2)*
$1 - F(x)$	=T.DIST.RT(x;n)	=F.DIST.RT(x;m;n)	=CHISQ.DIST.RT(x;n)
x_{1-P}	—	=F.INV.RT(P;m;n)	=CHISQ.INV.RT(P;n)
$P(\|X\| > x)$	=T.DIST.2T(x;n)	—	—
$x_{1-P/2}$	=T.INV.2T(P;n)	—	—

Pro syntaxi a hlavně pro význam jednotlivých parametrů platí stejné závěry jako pro diskrétní rozdělení. Opět jsou v tabulce zapsány pouze symbolicky, aby byl na první pohled jasný jejich význam.

3.3. Změny proti verzi MS Excel 2003

- Změny jsou v zásadě stejné jako u diskrétních rozdělení.
- Sjednocení názvů funkcí.
- Počet pravděpodobnostních rozdělení se nezměnil.
- Byly přidány nové funkce – viz přehled Syntaxe 3. Zároveň je však třeba upozornit, že některé funkce byly ve starších verzích MS Excel, ale měly jiný název – komentář viz dále v tabulce 6.
- U všech rozdělení je možnost spočítat hodnotu hustoty pravděpodobnosti (ve verzi 2003 byla tato možnost pouze u 4 rozdělení).
- Rozdíly v názvu funkcí jsou patrné z tabulky 3 a tabulky 4.

Tabulka 6: Přehled „problémových“ změn

Rozdělení	Excel 2003	Význam	Excel 2010	Význam
t-rozdělení (t)	TDIST	$1 - F(x)$	T.DIST	$F(x)$
	TINV	$x_{1-P/2}$	T.INV	x_P
	—	—	T.INV.2T	$x_{1-P/2}$
	—	—	T.DIST.RT	$1 - F(x)$
F-rozdělení (F)	FDIST	$1 - F(x)$	F.DIST	$F(x)$
	FINV	x_{1-P}	F.INV	x_P
	—	—	F.DIST.RT	$1 - F(x)$
	—	—	F.INV.RT	x_{1-P}
Chí-kvadrát (χ^2)	CHIDIST	$1 - F(x)$	CHISQ.DIST	$F(x)$
	CHIINV	x_{1-P}	CHISQ.INV	x_P
	—	—	CHISQ.DIST.RT	$1 - F(x)$
	—	—	CHISQ.INV.RT	x_{1-P}

- Původní funkce již nejsou v nabídce, tyto funkce však není třeba ve starších sešitech MS Excel přepisovat na nový název, zpětná kompatibilita funguje.
- Ostatní funkce se liší pouze změnou názvu – většinou má syntaxe tvar: JMENOFUNKCE.DIST, případně JMENOFUNKCE.INV.
- Funkce s téměř stejným názvem (až na tečku) mají v různých verzích různý význam. Naproti tomu funkce, které počítají stejné hodnoty, mají v obou verzích naprosto jiné názvy – týká se to zejména t-rozdělení, F-rozdělení a chí-kvadrát rozdělení. Přehled těchto „potencionálně problémových“ funkcí je obsažen v tabulce 6.

3.4. Zhodnocení funkcí pro spojitá rozdělení

- Na tabulkový kalkulátor velmi slušný počet základních spojitých rozdělení.
- Vzorce distribuční funkce (resp. hustoty) jsou naprogramovány v obvyklém tvaru, jak je známe z literatury – výjimku tvoří exponenciální rozdělení (převrácená hodnota parametru oproti standardu).
- Překlad do češtiny opět není příliš šťastný.

- Funkce počítající hodnoty $1 - F(x)$ či x_{1-P}, se jeví jako nadbytečné – statistik si bez nich vystačí.
- Při použití funkcí pro spojitá rozdělení je třeba daleko větší obezřetnosti – viz „problémové změny“ v tabulce 6.

4. Závěr

Na závěr je třeba uvést, že všechna uvedená pravděpodobnostní rozdělení (resp. příslušné funkce) byla v Excelu podrobně prozkoumána a přepočítána. Dosažené výsledky (hodnoty pravděpodobnostní funkce, distribuční funkce, hustoty pravděpodobnosti, kvantilů) byly porovnávány s programem Statgraphics Centurion (zkušební verze). Srovnání dopadlo z hlediska přesnosti pro Excel velmi uspokojivě, neboť jsme nezaznamenali žádné rozdíly ve výsledcích v obou programech. Oba programy se vesměs shodovaly i z hlediska vzorců pravděpodobnostních funkcí či hustot.

Lze tedy konstatovat, že MS Excel (verze 2010) je co se týče výčtu popsaných pravděpodobnostních rozdělení zcela srovnatelný s tímto statistickým programem (ten má pochopitelně mnohem širší nabídku pravděpodobnostních rozdělení). Pokud se vyskytly nějaké rozdíly, byly většinou způsobeny tvarem parametrů rozdělení (Excel např. počítá s převrácenou hodnotou parametru oproti programu Statgraphics apod.). Pokud jsme však respektovali tyto odlišnosti, vycházely výsledky výpočtů stejně.

Z hlediska nabídky pravděpodobnostních rozdělení je na tom MS Excel poměrně dobře – obsahuje funkce pro 9 nejpoužívanějších spojitých rozdělení a pro 4 nejčastěji používaná diskrétní rozdělení.

Změny ve verzi 10 programu MS Excel výrazně prospěly – zejména co se týče sjednocení názvů funkcí. Do práce s funkcemi tak byl vnesen řád a jednotnost. Částečně se zlepšilo i označení parametrů funkcí a nápověda, byť stále ještě neodpovídá běžně používané pravděpodobnostní terminologii.

Co se týče využití, má MS Excel dvě nesporné výhody oproti všem stat. programům a těmi jsou dostupnost a úspora nákladů. Je totiž nainstalován téměř na každém osobním počítači a k výpočtu základních funkcí v oblasti pravděpodobnosti tak není potřeba specializovaný statistický program.

Literatura

[1] Nápověda k programu Microsoft Excel.

[2] Nápověda k programu Statgraphics Centurion.

MULTIVARIATE STATISTICAL PROCESS CONTROL
VÍCEROZMĚRNÉ STATISTICKÉ ŘÍZENÍ PROCESŮ

Martin Kovářík

Adresa: Ing. Bc. Martin Kovářík, Ph.D., Univerzita Tomáše Bati ve Zlíně, Fakulta managementu a ekonomiky, nám. T. G. Masaryka 5555, 760 01 Zlín

E-mail: m1kovarik@fame.utb.cz

Abstract: Statistical Process Control (SPC) is a preventive quality control tool, because the early detection of significant deviations of the process from set level provides to exercise of interventions to the process with the aim of desire to keep it at acceptable levels, eventually process improvement. Because the SPC in practice almost observe several variables simultaneously, it is offering for usage of multivariate analysis methods with advantage also in this area. In my article, I will focus on three types of multivariate diagrams. These include Hotelling's T-square statistics, Multivariate Exponentially Weighted Moving Average (MEWMA) and Multivariate Cumulative Sum (MCUSUM).

Keywords: MSPC (Multivariate Statistical Process Control), Hotelling's Control Chart, MEWMA (Multivariate Exponentially Weighted Moving Average), MCUSUM (Multivariate Cumulative Sum), PCA – Control Chart.

Abstrakt: Statistické řízení procesů (Statistical Process Control – SPC) představuje preventivní nástroj řízení kvality, neboť na základě včasného odhalování významných odchylek procesu od předem stanovené úrovně umožňuje vykonávat zásahy do procesu s cílem dlouhodobě jej udržovat na přípustné úrovni, popř. proces zlepšovat. Protože v praxi SPC téměř vždy sledujeme několik proměnných současně, nabízí se použití metod vícerozměrné analýzy s výhodou i v této oblasti. Ve svém příspěvku se zaměřím na tři druhy vícerozměrných diagramů. Mezi ně patří Hotellingova statistika T-kvadrát, vícerozměrné exponenciálně vážené průměry (MEWMA) a vícerozměrné kumulované součty (MCUSUM).

Klíčová slova: MSPC (vícerozměrné statistické řízení procesů), Hotellingův regulační diagram, MEWMA (vícerozměrné exponenciálně vážené průměry), MCUSUM (vícerozměrné kumulované součty), PCA – regulační diagram.

Úvod

Statistické řízení procesu (SPC) představuje zpětnovazební systémové ovládání procesu na základě informace o výkonu procesu ve formě údajů zjištěných při vlastní regulaci. Proces ovlivňovaný pouze systémem náhodných

příčin (Chance Causes) má charakter statisticky zvládnutého procesu a takový proces je predikovatelný. Naproti tomu přítomnost zvláštních příčin (nazývaných také vymezitelné příčiny – Assignable Causes) vyvolává v procesu neočekávané změny. Tyto typy příčin je nutné identifikovat. Teorie statistické regulace procesu vychází z existence variability jako imanentní vlastnosti každého procesu a příčiny jeho neopakovatelnosti. I za relativně stálých podmínek objektivně působí na proces, a tím i na jeho výstupy, mnoho vlivů, které způsobují, že nelze vytvořit dva zcela totožné produkty. Tyto rušivé vlivy však lze studovat a vytvářet podmínky k tomu, aby variabilita procesu byla stabilní a pohybovala se ve svých přirozených mezích, při jejichž znalosti by bylo možné předvídat chování procesu v budoucnosti. Menší variabilita procesu znamená:

- stejnoměrnější výrobu;
- menší pravděpodobnost výskytu neshodných produktů;
- menší rozsah kontroly a nižší náklady na kontrolu a zkoušení;
- nižší náklady vyvolané poruchami procesu, produkováním odpadu a jednotek vyžadujících přepracování;
- větší počet spokojených zákazníků.

Princip SPC vychází z členění variability na variabilitu vyvolanou náhodnými (přirozenými) příčinami a variabilitu vyvolanou příčinami vymezitelnými (identifikovatelnými).

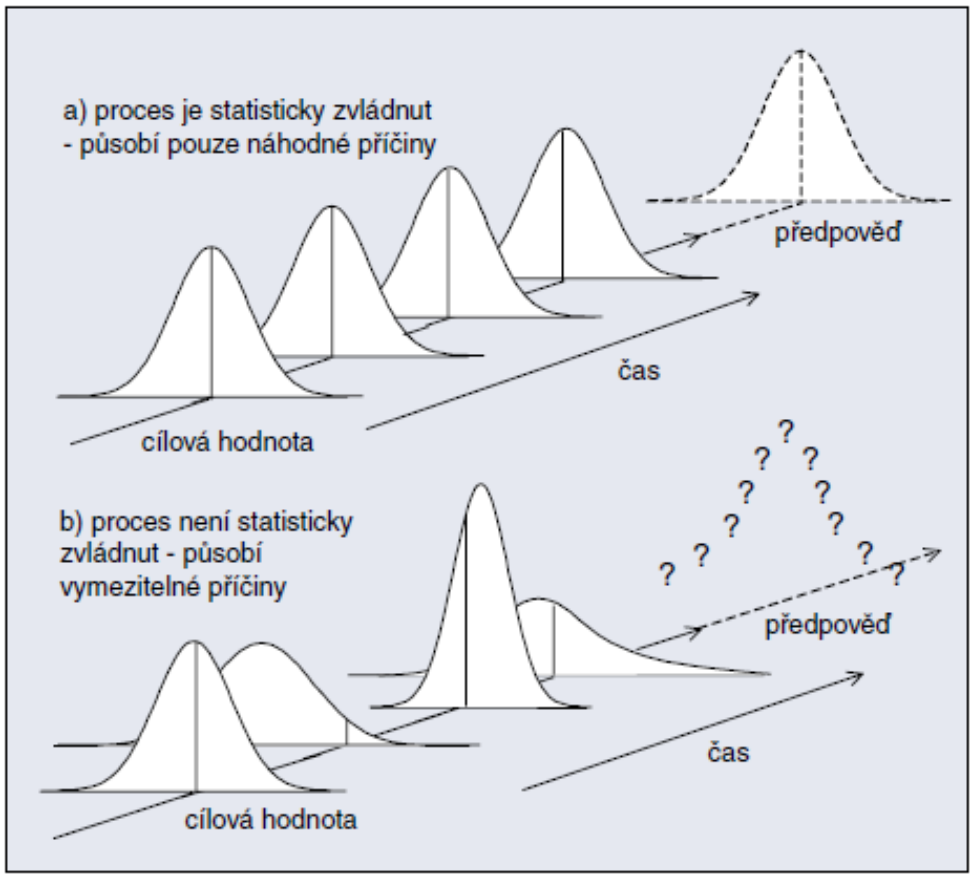
K odstranění vlivu těchto příčin obvykle stačí pouze lokální zásah osoby přímo zodpovědné za vykonávání činnosti v rámci daného procesu. Základním nástrojem SPC je regulační diagram (Control Chart), který objasním níže.

1. Základní charakteristika regulačního diagramu

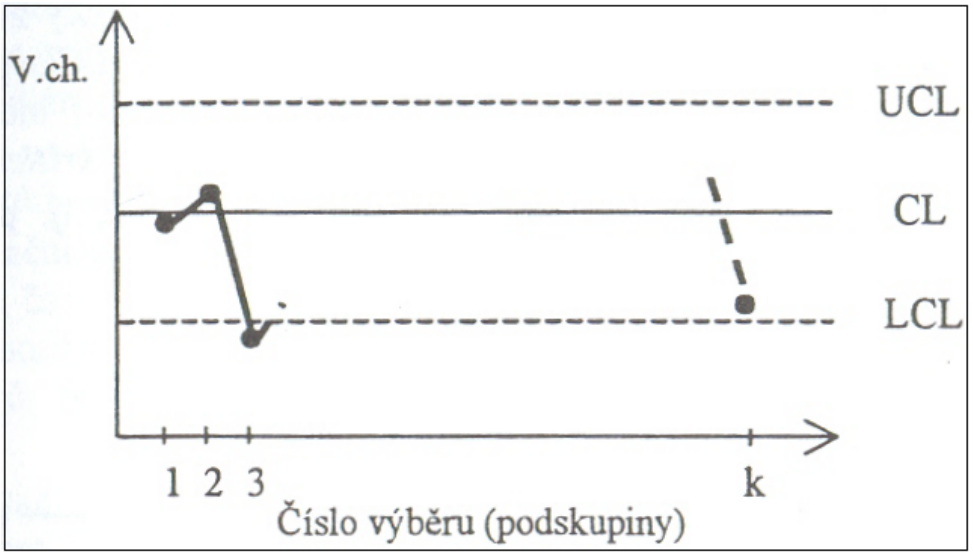
Základním nástrojem SPC je regulační diagram (obr. 2). Je to grafický prostředek zobrazení vývoje variability procesu v čase využívající principů testování statistických hypotéz. Rozhodnutí o statistické zvládnutosti procesu umožňují 3 základní čáry.

CL – střední přímka; odpovídá tzv. referenční (požadované) hodnotě použité znázorňované charakteristiky. Z hlediska účinnosti regulačního diagramu a základního rozhodnutí o statistické zvládnutosti procesu je rozhodující stanovení horní a dolní regulační meze:

- UCL je horní regulační mez (Upper Control Limit),
- LCL je dolní regulační mez (Lower Control Limit).



Obrázek 1: Náhodné a vymezitelné příčiny variability



V. ch. = výběrová charakteristika použitá jako testové kritérium v daném regulačním diagramu (např. $\overline{x},\ R,\ s, \ldots$).

Obrázek 2: Základní struktura regulačního diagramu

Těmto regulačním mezím se také říká akční meze. Vymezují pásmo působení pouze náhodných příčin variability a jsou základním rozhodovacím kritériem, zda učinit regulační zásah do procesu či nikoliv. V některých aplikacích se zakreslují do regulačního diagramu další meze nazývané výstražné meze: UWL (Upper Warning Limit – horní výstražná mez) a LWL (Lower Warning Limit – dolní výstražná mez). Pásmo, které vymezují tyto meze, je vždy užší než pásmo mezi akčními mezemi, nejčastěji $\pm 2\sigma$ od CL.

1.1. Interpretace regulačního diagramu

Pro interpretaci regulačního diagramu platí obecně základní pravidlo:

a) Leží-li všechny body uvnitř UCL a LCL, je proces pokládán za statisticky zvládnutý a není vyžadován žádný zásah do procesu.

b) Leží-li některý bod mimo regulační mez UCL nebo LCL, je proces pokládán za statisticky nezvládnutý, je vyžadována identifikace vymezitelné příčiny této odchylky a přijetí opatření s cílem úplné či alespoň částečné eliminace vymezitelného vlivu.

Použijí-li se i meze výstražné, mohou nastat kromě uvedené základní situace ještě další dvě situace:

1) Některý bod leží uvnitř výstražných mezí – lze předpokládat, že proces je ve statisticky zvládnutém stavu a není třeba žádného zásahu.

2) Některý bod leží mezi UWL a UCL, resp. mezi LWL a LCL. V této situaci se doporučuje postupovat následovně: Ihned bez ohledu na kontrolní interval se provede další výběr. Jestliže nový bod, odpovídající tomuto bezprostřednímu výběru, leží mezi výstražnými mezemi, není třeba do procesu zasahovat. Jestliže však i tento nový bod leží mimo výstražné meze, je to signál, že na proces s velkou pravděpodobností působí vymezitelná příčina a je nutné provést regulační zásah.

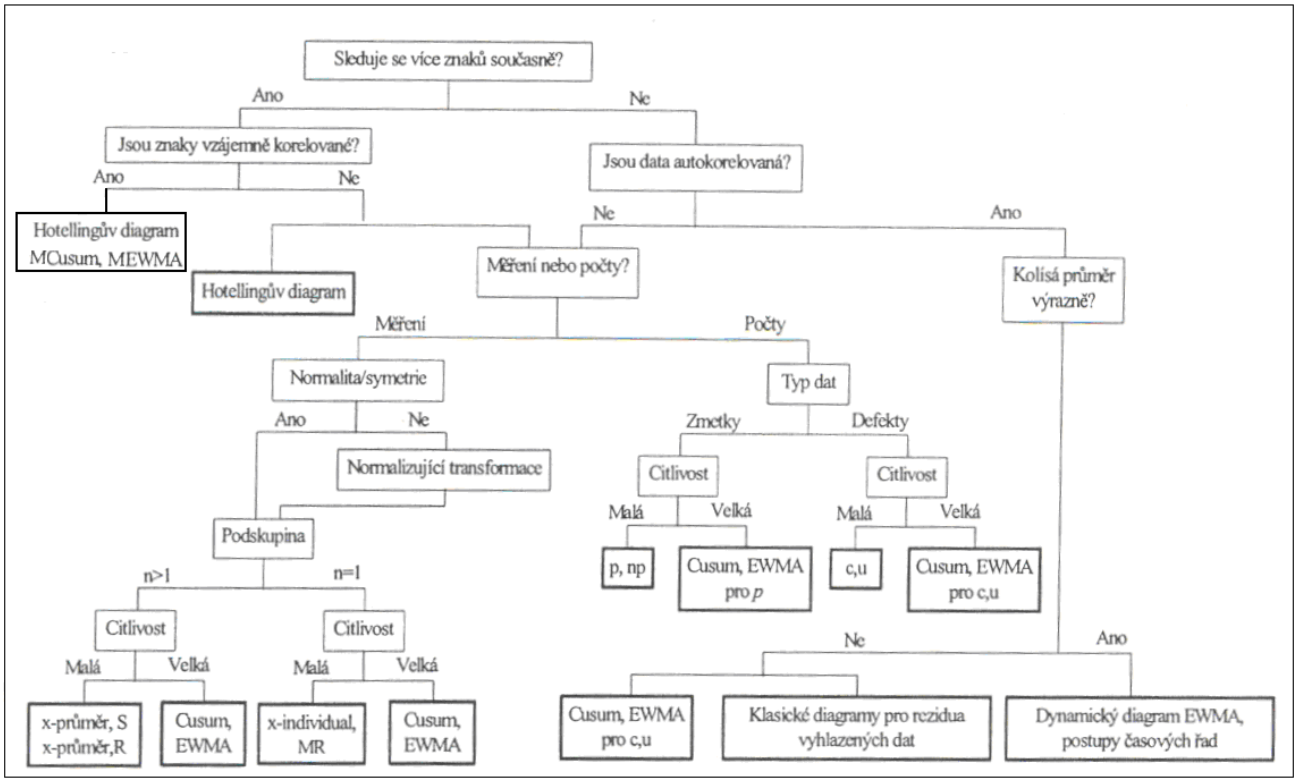
1.2. Obecný postup sestrojení a analýzy regulačního diagramu

Dosavadní poznatky o regulačních diagramech můžeme stručně shrnout do devíti základních kroků, které je nutné provádět bez ohledu na použitou metodu SPC. Jsou to tyto kroky:

1. Volba regulované veličiny.
2. Sběr a záznam dat.
3. Ověření předpokladů o datech.
4. Volba rozsahu výběru.

5. Volba vhodného regulačního diagramu.
6. Výpočet hodnot zvoleného testového kritéria (výběrové charakteristiky) pro jednotlivé výběry.
7. Ověření a zajištění statistické zvládnutosti procesu.
8. Ověření a zabezpečení způsobilosti procesu.
9. Vlastní regulace procesu.

Na následujícím obrázku je zobrazen postup při výběru regulačního diagramu, ve kterém jsou zakomponovány i vícerozměrné regulační diagramy.



Obrázek 3: Postup při výběru regulačního diagramu

2. Vícerozměrné regulační diagramy

Při použití jednorozměrných, například Shewhartových, regulačních diagramů se posuzuje stabilita střední hodnoty μ, případně rozptylu σ^2. Sleduje-li se tímto způsobem m proměnných, máme informace o $m + m = 2m$ statistických parametrech. Avšak m-tice náhodných proměnných chápaná jako m-rozměrná náhodná veličina je charakterizována nejméně vektorem středních hodnot μ a kovarianční maticí σ, tedy celkem $m+m^2$ parametry. Přehlédli bychom tedy řadu parametrů, které mohou přispět k posouzení stability.

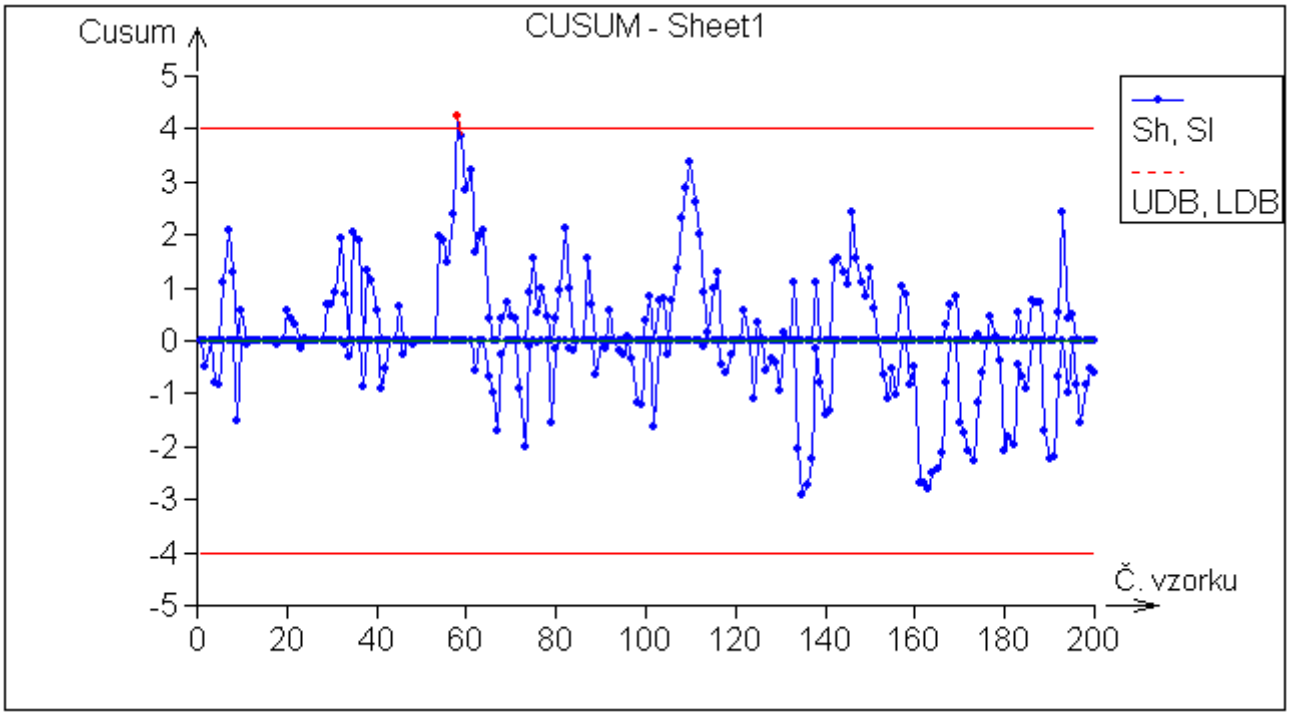
Tento nedostatek izolovaných jednorozměrných regulačních diagramů řeší vícerozměrné regulační diagramy a další nástroje sledování vícerozměrné stability. Nyní se zaměřím na tři druhy vícerozměrných diagramů. Mezi ně patří Hotellingova statistika T-kvadrát, vícerozměrné exponenciálně vážené průměry (MEWMA) a vícerozměrné kumulované součty (MCUSUM).

2.1. Metoda kumulovaných součtů (CUSUM)

Regulační diagramy CUSUM se pro svou větší citlivost na změny procesu (na rozdíl od Shewhartových diagramů) využívají v případě, že je třeba zajistit rychlé a ekonomicky nenáročné odhalení náhlých, určitou dobu působících odchylek, které nejsou většího rozsahu, ale způsobují odchylky od cílové hodnoty. Na ose x se vynáší pořadí výběru k, na ose y pak kumulativní součet odchylek zvolené výběrové charakteristiky, pro který platí,

$$C_k = \sum_{j=1}^{k} (\overline{x}_j - \mu_0) = C_{k-1} + (\overline{x}_k - \mu_0) \tag{1}$$

$C_0 = 0$, kde k je pořadí výběru a $\overline{x}_j$ je výběrový průměr z hodnot regulované veličiny v j-tém výběru. Vyhodnocení diagramu CUSUM se provádí buď pomocí rozhodovacího intervalu nebo tzv. V-masky.



Obrázek 4: Diagram CUSUM

Jedná se o diagram s pamětí a všechny odchylky mají stejnou váhu. Paměť je tady neomezená a rovnoměrná.

MCUSUM diagramy jsou umístěny do dvou hlavních kategorií. V první kategorii je směr posunu (nebo změny) považován za známý (směr specifických režimů), zatímco druhý směr posunu je považován za neznámý (směrově invariantní systémy). MCUSUM regulační diagramy jsou široce používány v průmyslu, protože jsou silné a snadné na použití.

2.1.1. Porovnání účinnosti klasického Shewhartova diagramu pro $(\overline{x})$ s diagramem CUSUM pro $(\overline{x})$ při $\alpha = 0{,}0027$ a při $\alpha = 0{,}01$

Mějme proces, kde hodnoty regulované veličiny pocházejí z normálního rozdělení s parametry $\mu_0 = 100\,\text{mm}$ a $\sigma_0 = 20\,\text{mm}$. Při těchto hodnotách je proces statisticky zvládnutý. Mezi 15. a 16. výběrem však dochází ke kritickému posunu střední hodnoty na $\mu = 110$ a proces se stává statisticky nezvládnutým. Úkolem je ověřit účinnost diagramu CUSUM pro výběrové průměry ve srovnání s účinností ekvivalentního klasického Shewhartova regulačního diagramu pro výběrové průměry.

A. Porovnání účinnosti klasického Shewhartova diagramu pro $(\overline{x})$ s diagramem CUSUM pro $(\overline{x})$ při $\alpha = 0{,}0027$

Nejdříve se sestrojí pro analyzovaný proces Shewhartův regulační diagram $(\overline{x})$ s rizikem $\alpha = 0{,}0027$. Střední přímka a regulační meze pro tento diagram byly stanoveny následovně:

$CL = \mu_0 = 100\,\text{mm}$

$UCL = CL + 3 \cdot \sigma_0/\sqrt{n} = 100 + 20/\sqrt{5} = 126{,}83\,\text{mm}$,

$LCL = CL - 3 \cdot \sigma_0/\sqrt{n} = 100 - 20/\sqrt{5} = 73{,}17\,\text{mm}$.

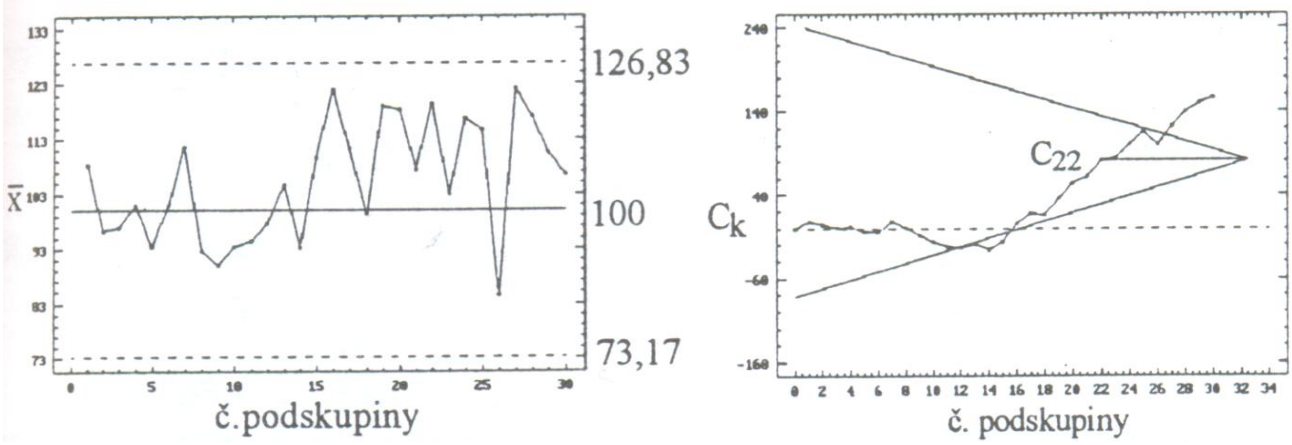
Tento klasický Shewhartův diagram je na obr. 5a.

Neuvažujeme-li nenáhodná seskupení, pak diagram na obr. 5a říká, že proces lze pokládat za statisticky zvládnutý, i když v něm došlo ke kritickému posunu střední hodnoty $\gamma = 10\,\text{mm}$.

Jestliže $\alpha = 0{,}0027$, pak $ARL(0) = 1/0{,}0027 = 370{,}57$ výběrů pro klasický Shewhartův diagram. Pravděpodobnost odhalení změny střední hodnoty γ sestrojeným klasickým Shewhartovým diagramem pro výběrové průměry lze určit následovně:

$$\begin{aligned}(1-\beta) &= P(X < LCL) + P(X > UCL) = F(73{,}17) + (1 - F(126{,}83)) = \\ &= 1 + \Phi\left(\frac{73{,}17-110}{20/\sqrt{5}}\right) - \Phi\left(\frac{126{,}83-110}{20/\sqrt{5}}\right) = 0{,}030145\end{aligned}$$

Riziko chybějícího signálu je pak $\beta = 1 - (1 - \beta) = 0{,}969855$.



a) Klasický Shewhartův diagram b) Diagram CUSUM

Obrázek 5: Regulační diagramy ($\overline{x}$) pro
statisticky nezvládnutý proces (α = 0,0027)

Průměrný počet výběrů mezi okamžikem vzniku kritické odchylky a okamžikem jejího odhalení v diagramu $ARL\,(10) = 1/\,(1-\beta) = 1/0{,}030145 = 33{,}2$ výběry. V průměru tedy odhalí sestrojený Shewhartův diagram ($\overline{x}$) odchylku střední hodnoty o velikosti 10 mm po 33 výběrech od vzniku odchylky. Diagram CUSUM pro výběrové průměry odhalí tuto odchylku již po 6 výběrech od vzniku odchylky (obr. 5b).

B. Porovnání účinnosti klasického Shewhartova diagramu pro ($\overline{x}$) **s diagramem CUSUM pro** ($\overline{x}$) **při** $\alpha = 0{,}01$

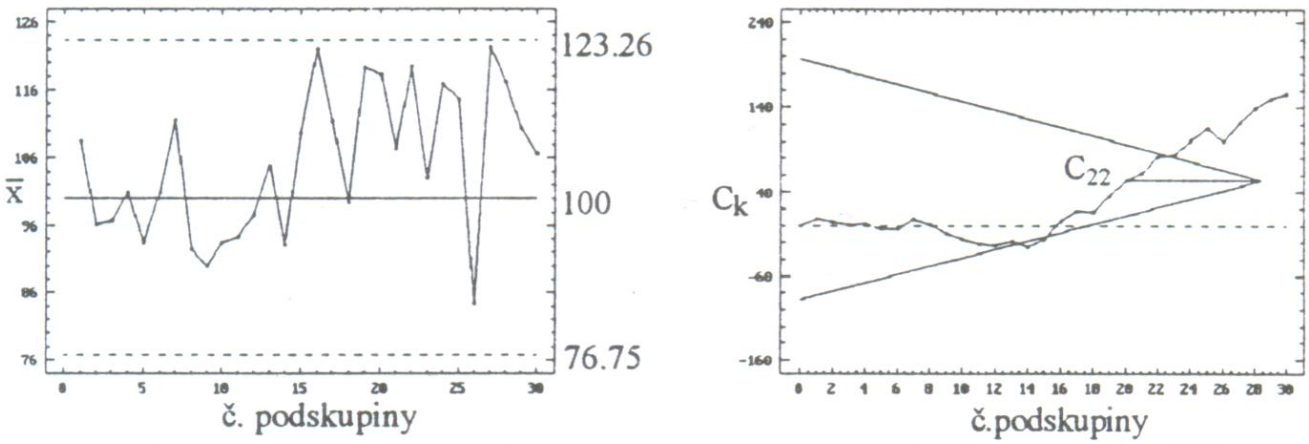
Jestliže budeme řešit stejnou úlohu pro větší riziko zbytečného signálu α = 0,01, dostaneme tyto výsledky:

Klasický Shewhartův regulační diagram pro výběrové průměry (viz obrázek 6a): CL = 100 mm, UCL = 123,26 mm, LCL = 76,75 mm, $ARL(0)$ = 100 výběrů, $(1-\beta) = 0{,}0695$, $ARL\,(10) = 14{,}4$ výběrů.

Obr. 6a ukazuje, že ani diagram s menší vzdáleností mezi regulačními mezemi nesignalizuje kritický posun střední hodnoty rozdělení regulované veličiny.

Diagram CUSUM pro výběrové průměry s V-maskou (viz obr. 6b): $d = 8{,}5$, $\theta = 14^{\circ}$. Diagram na obr. 6b signalizuje kritickou změnu střední hodnoty již po 4 výběrech od okamžiku vzniku změny střední hodnoty. To znamená, že i pro riziko zbytečného signálu $\alpha = 0{,}01$ je regulační diagram účinnější v odhalení dané kritické odchylky střední hodnoty než klasický Shewhartův diagram.

Pro malé kritické změny jsou účinnější diagramy CUSUM a tato relativní účinnost roste s poklesem hodnoty rizika zbytečného signálu α.



a) Klasický Shewhartův diagram b) Diagram CUSUM

Obrázek 6: Regulační diagramy ($\overline{x}$) pro statisticky nezvládnutý proces (α = 0,01)

2.1.2. Volba mezi klasickým Shewhartovým regulačním diagramem a CUSUM diagramem

V této části si na závěr shrneme situace, kdy je vhodnější použít CUSUM diagram a kdy naopak klasický Shewhartův diagram.

CUSUM diagramy se volí v těchto situacích:

- malá nebo středně veliká odchylka; není-li co nejdříve odhalena, vede k relativně vysokým ztrátám spojeným s produkcí neshodných produktů,
- náklady na kontrolu jsou relativně vysoké,
- jednoduchost regulačních postupů není rozhodující.

Klasickým Shewhartovým diagramům by se měla dát přednost v situacích, kdy např.:

- je třeba zajistit jednoduchost regulačních postupů,
- náklady na kontrolu a zkoušení nejsou vysoké,
- ztráty spojené s produkcí neshodných produktů nejsou vysoké.

2.2. Diagramy EWMA

Obdobně jako diagram CUSUM uvedený v předchozím odstavci, také diagram EWMA se hodí pro situace, kdy v procesu dochází k náhlým malým, ale přetrvávajícím změnám procesu a hodnoty sledovaného znaku jakosti nejsou závislé. Na rozdíl od klasických diagramů závisí regulační meze na okamžiku

výběru. Střední přímka CL se stanoví ze vztahu

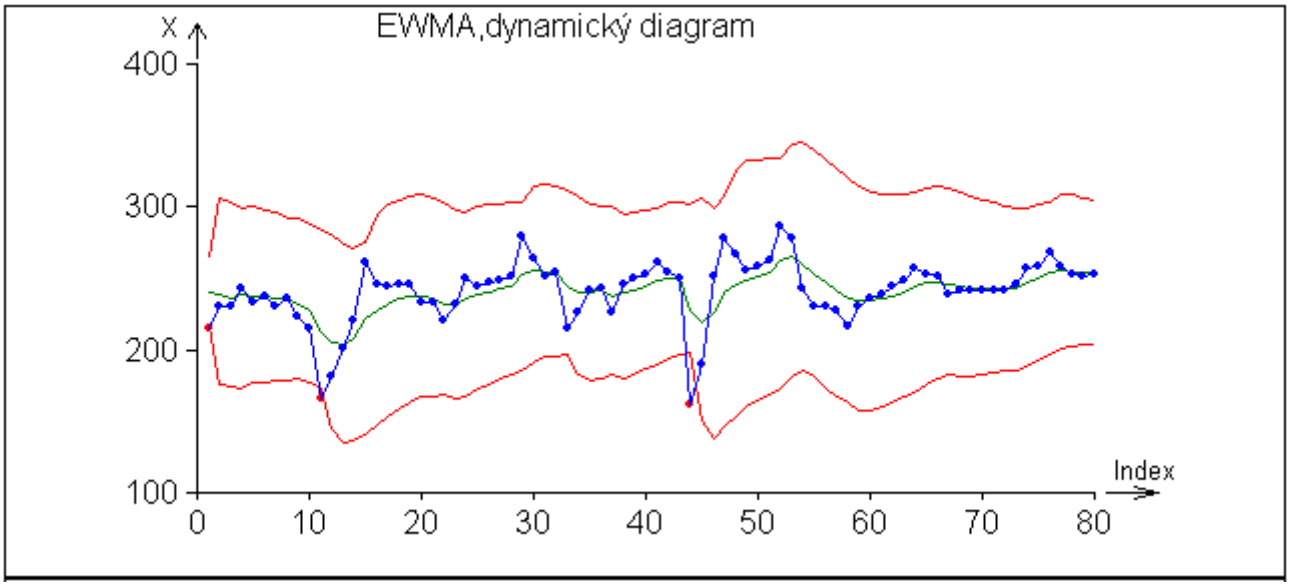
$$CL = \mu_0. \tag{2}$$

Regulační meze se pak určí ze vztahů

$$UCL = CL + K \cdot \sigma_{EWMA}, \tag{3}$$

$$LCL = CL - K \cdot \sigma_{EWMA}, \tag{4}$$

kde K je konstanta pro stanovení regulačních mezí při zvoleném riziku α a σ_{EWMA} se vypočte ze speciálního vztahu využívajícího parametr zapomínání λ. Ukázka takového diagramu je na obr. 7.



Obrázek 7: Dynamický diagram EWMA

Diagramy EWMA patří mezi diagramy s neomezenou nerovnoměrnou pamětí. Ve vícerozměrném případě, lze rozšířit tento vzorec

$$Z_i = \Lambda X_i + (1 - \Lambda) Z_{i-1} \tag{5}$$

kde Z_i je i-tá EWMA statistika, X_i je i-tý vektor pozorování, pro $i = 1, 2, ..., n$. Z_0 je vektor hodnot z historických dat. Λ je diag $(\lambda_1, \lambda_2, \ldots, \lambda_p)$, která je diagonální matice s $\lambda_1, \lambda_2, \ldots, \lambda_p$ na hlavní diagonále, a p je počet proměnných, které představují počet prvků v každém vektoru.

Bylo prokázáno (Lowry a kol., 1992), že (k, l)-tého prvku kovarianční matice i-tého EWMA, Σ_{zi}, je

$$\Sigma_{zi}(k, l) = \lambda_k \lambda_l \frac{\left[1 - (1 - \lambda_k)^i (1 - \lambda_l)^i\right]}{[\lambda_k + \lambda_l - \lambda_k \lambda_l]} \sigma_{k,l} \tag{6}$$

kde $\sigma_{k,l}$ je (k,l)-tý element Σ, kovarianční matice $\mathbf{X}$.

Pokud $\lambda_1 = \lambda_2 = \ldots = \lambda_p = \lambda$, potom se výše uvedený výraz zjednodušuje na

$$\Sigma_{zi}(k,l) = \frac{\lambda}{2-\lambda}\left[1-\left(1-\lambda^{2i}\right)\right]\Sigma, \tag{7}$$

kde Σ je kovarianční matice vstupních dat.

2.2.1. Případová studie

Jako vedoucí výroby hraček chceme sledovat hmotnost (v gramech) a délku (v cm) jednoho z dílů našich hraček. Nasbírali jsme 4 vzorky každý den po dobu 20 dnů, viz obr. 8. Vzhledem k tomu, že hmotnost a délka spolu korelují, a chceme zjistit malé posuny v těchto proměnných, můžeme vytvořit vícerozměrný EWMA diagram, viz obr. 9.

Interpretace výsledků

Všechny body leží pod horní regulační mezí, což naznačuje, že rozdíly v hmotnosti a délce jsou v průběhu času způsobeny společnými příčinami. Pro vytvoření diagramu MEWMA jsem použil software Minitab 14.

2.3. Hotellingův diagram

Ne vždy se na jednom produktu sleduje pouze jeden znak jakosti. Pro případ, že chceme sledovat více znaků jakosti na jednom produktu, je velmi vhodné použít jako nástroj statistické regulace procesu Hotellingův diagram, a to zejména v případě, kdy znaky jakosti jsou vzájemně korelované a použití samostatných klasických Shewhartových regulačních diagramů pro jednotlivé znaky jakosti by vedlo k nesprávným závěrům. Testovým kritériem je v tomto případě jednorozměrná Hotellingova statistika T, jejíž maticový zápis u regulačních diagramů pro výběrové průměry lze vyjádřit jako:

$$T_j^2 = n \cdot (\overline{x}_j - \overline{\overline{x}})^T C^{-1} (\overline{x}_j - \overline{\overline{x}}), \tag{8}$$

pro $j = 1, 2, \ldots, k$, kde n = rozsah podskupiny; $\overline{x}_j$ = vektor výběrových průměrů všech znaků jakosti v j-tém výběru; $\overline{\overline{x}}$ = vektor, pomocí něhož odhadujeme hodnoty μ pro simultánně sledované znaky jakosti; C kovarianční matice.

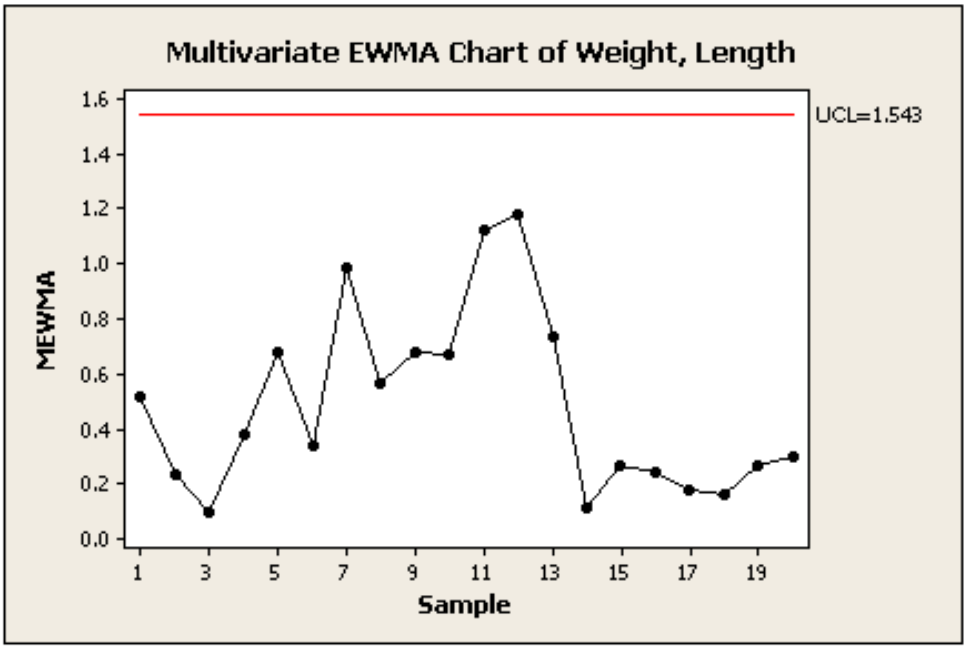
Každá hodnota T_j^2 je porovnávána s regulační mezí UCL, která se vypočte ze vztahu:

$$UCL = \left(\frac{k \cdot n \cdot m - k \cdot m - n \cdot m + m}{k \cdot n - k - m + 1}\right) \cdot F_{(m,k\cdot n-k-m+1)}(\alpha), \tag{9}$$

Toys.MTW ***

↓	C1	C2	C3	C4	C5	C6	C7
	Day	Weight	Length	Defects	Sample	Rejects	Inspected
1	1	10,10	2,54	9	110	8	200
2	1	10,15	2,56	11	101	13	200
3	1	10,11	2,55	2	98	7	200
4	1	10,12	2,55	5	105	8	200
5	2	10,12	2,54	15	110	5	200
6	2	10,14	2,57	13	100	13	200
7	2	10,08	2,50	8	98	7	200
8	2	10,10	2,53	7	99	12	200
9	3	10,09	2,50	5	100	27	200
10	3	10,15	2,56	2	100	10	200
11	3	10,14	2,55	4	102	12	200
12	3	10,11	2,53	4	98	6	200
13	4	10,07	2,49	2	99	10	200
14	4	10,13	2,53	5	105	9	200
15	4	10,12	2,52	5	104	13	200
16	4	10,11	2,52	2	100	7	200

Obrázek 8: Data pro případovou studii



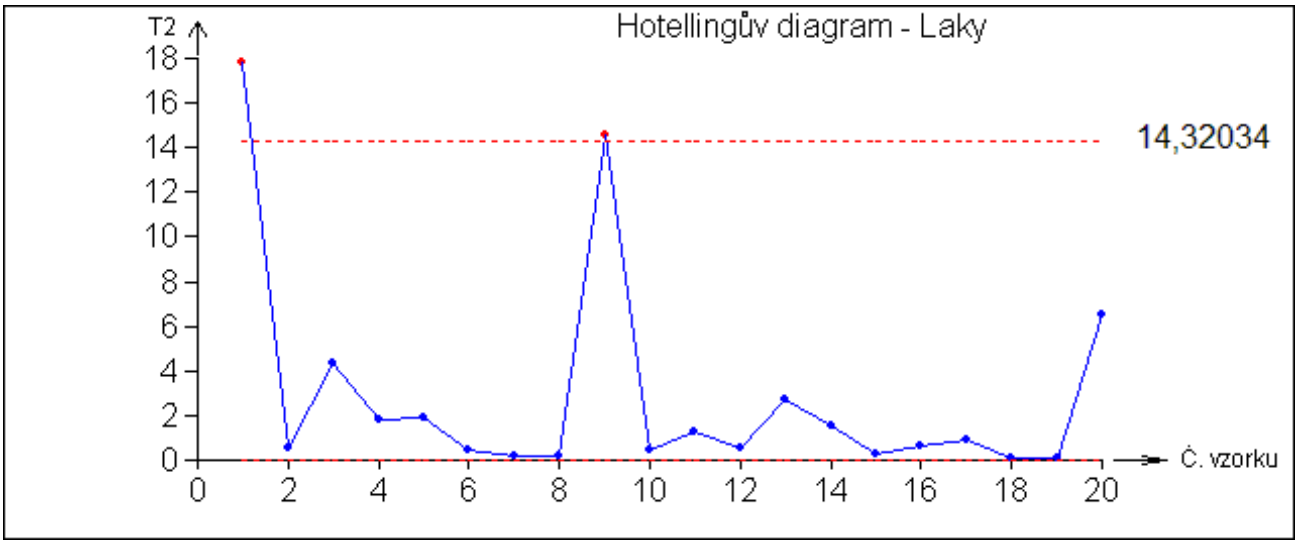
Obrázek 9: Diagram MEWMA

kde $F_{(m,k\cdot n-k-m+1)}(\alpha)$ je kritická hodnota Fischerova-Snedecorova rozdělení. Hotellingův diagram má pouze horní regulační mez.

Při užití Hotellingova diagramu předpokládáme vícerozměrné normální rozdělení u sledované veličiny.

2.3.1. Případová studie

Tato případová studie vychází z mojí diplomové práce. Na následujících obrázcích jsou klasické Shewhartovy diagramy pro sledované znaky kvality u tavných lepidel (viskozita a teplostálost spoje). Tyto diagramy nesignalizují žádnou podstatnou odchylku. Na obr. 10 je Hotellingův diagram pro všechny tři veličiny, který odhaluje výrazné překročení regulační meze na začátku, uprostřed a na konci měřícího intervalu.

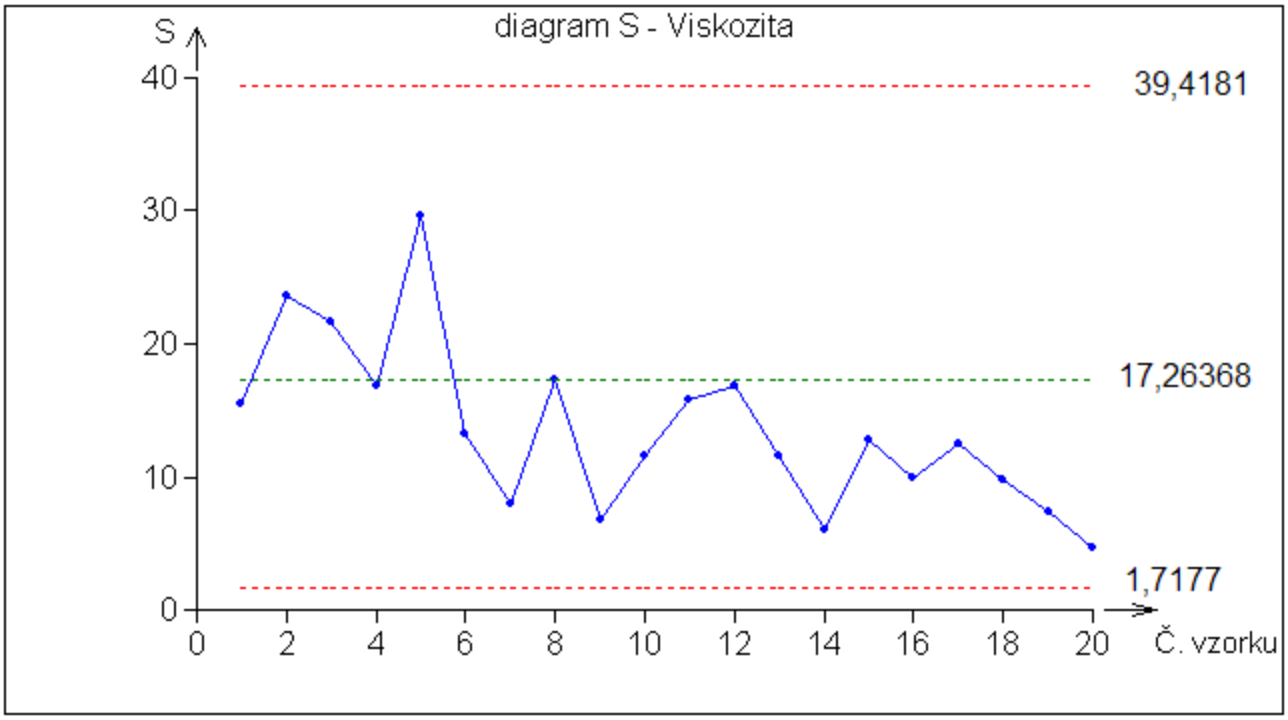


Obrázek 10: Hotellingův diagram

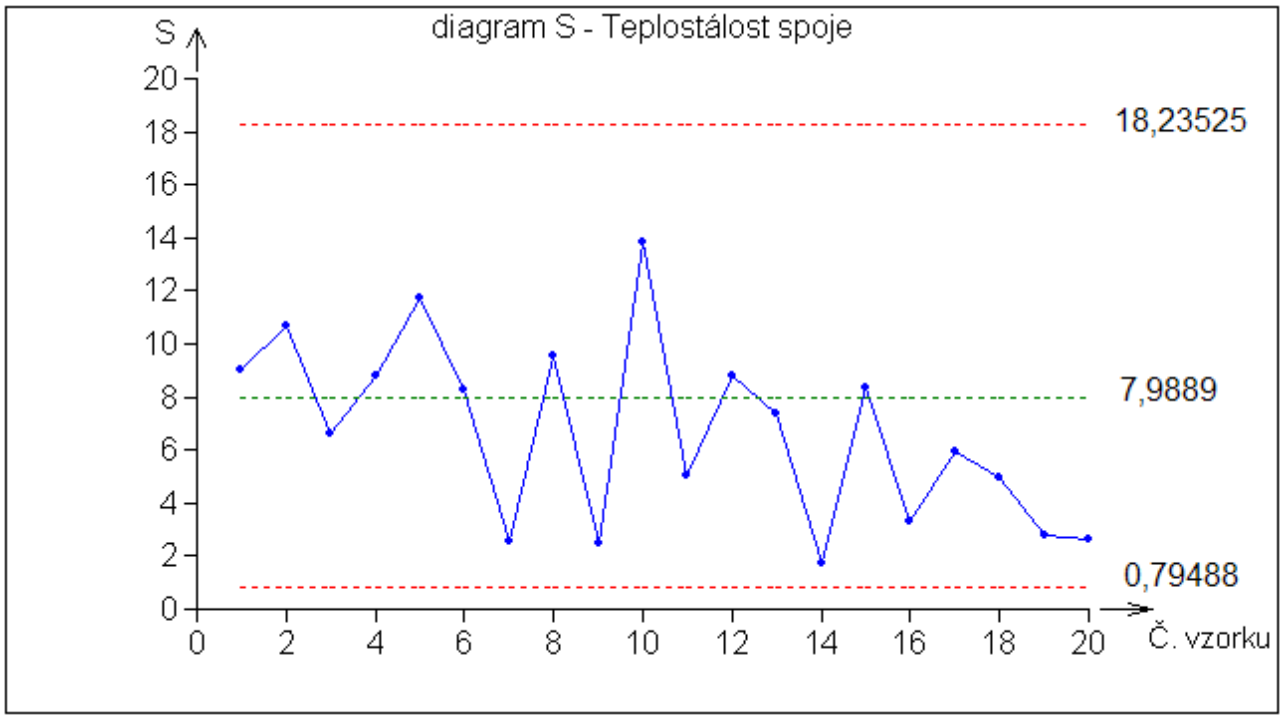
Byly sestrojeny rozptylové regulační diagramy (S) pro každý sledovaný znak jakosti (viskozita a teplostálost spoje) s rizikem zbytečného signálu α = 0,0027 (riziko je zvoleno tak, aby bylo kompatibilní s rizikem α zvoleným pro Hotellingův diagram).

U diagramu S pro viskozitu (obr. 11) nebylo porušeno žádné z pravidel a proces vykazuje značnou stabilitu, jelikož nebyly překročeny regulační meze. Proto zkusím další diagramy pro odhalení nestability při překročení regulačních mezí.

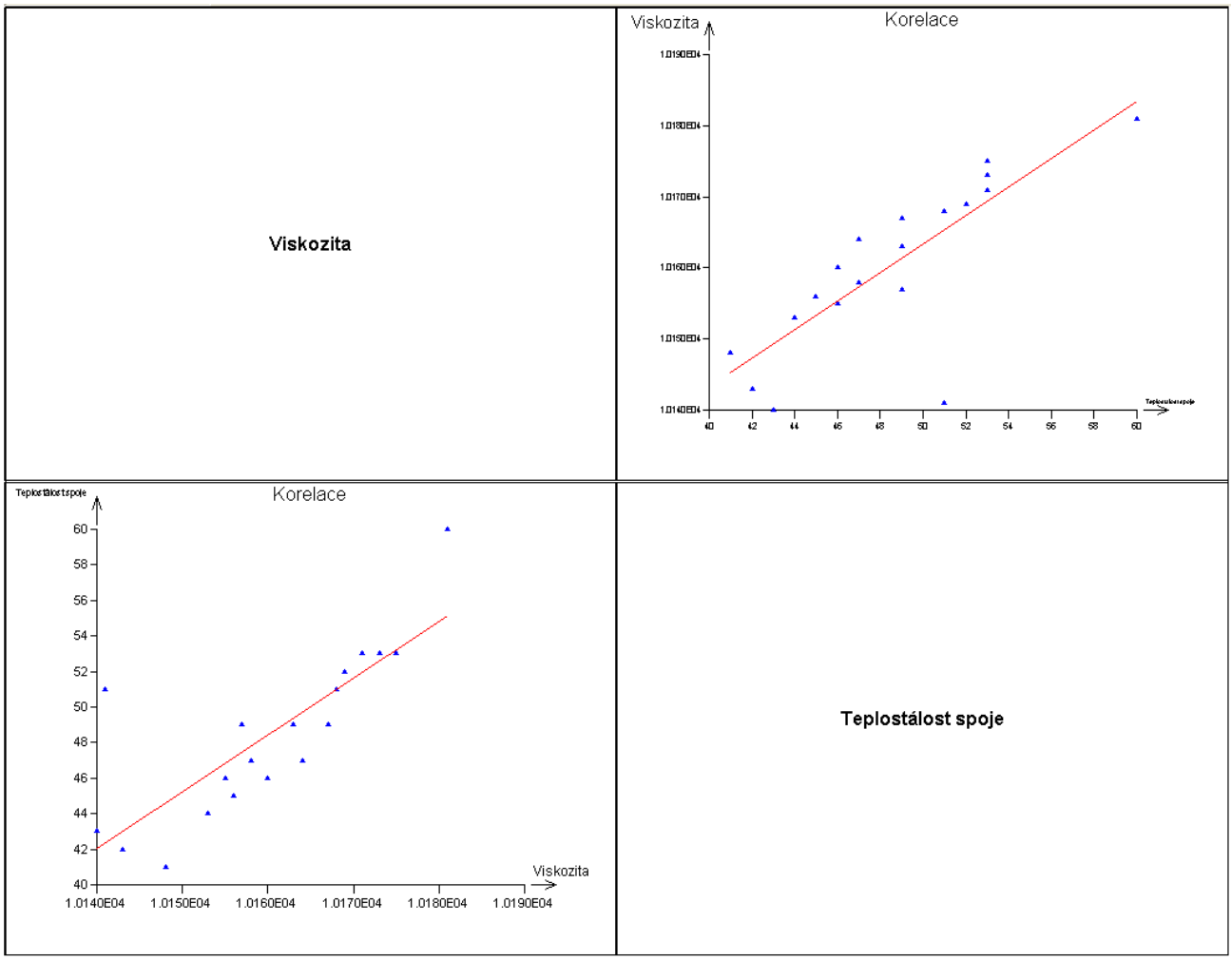
U diagramu S pro teplostálost spoje (obr. 12) nedošlo k porušení žádného z pravidel, což ukazuje na značnou stabilitu procesu. Avšak Hotellingův diagram teprve odkryje nedostatky předchozích diagramů a objasní interpretaci výsledků.



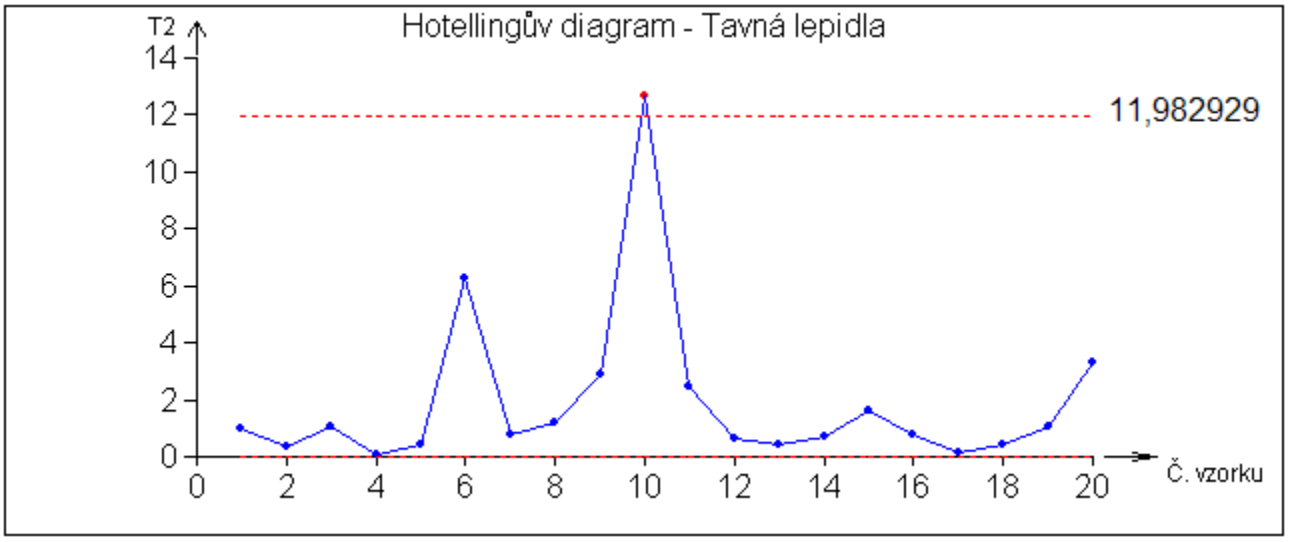
Obrázek 11: Regulační diagram S pro viskozitu



Obrázek 12: Regulační diagram S pro teplostálost spoje



Obrázek 13: Graf korelace pro oba dva sledované znaky jakosti (QC Expert)

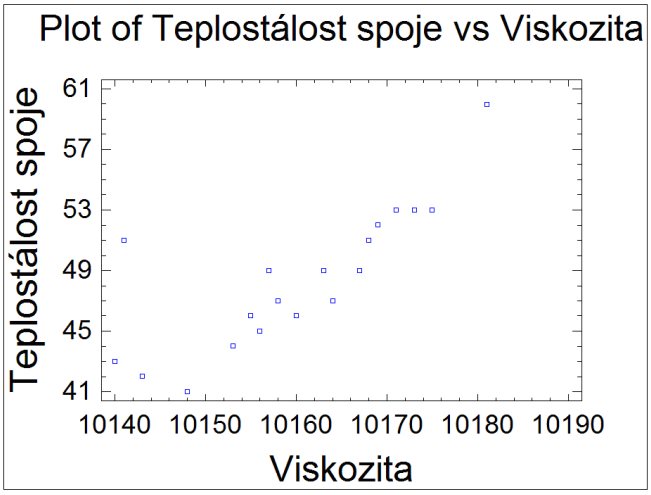


Obrázek 14: Hotellingův diagram

Mezi znaky jakosti viskozitou a teplostálostí spoje existuje korelační vztah, který je znázorněn v následujícím obrázku 15 i s body, jenž se vymykají převažujícímu trendu.

Ověření korelace obou znaků jakosti bylo provedeno pomocí grafu korelace. Následující výstupy byly realizovány jak ve statistickém softwaru QC Expert, tak i Statgraphics Plus. Korelační grafy jsem sestrojil pomocí výběrových průměrů jednotlivých podskupin obou sledovaných znaků jakosti. Korelační analýza: Párová korelace (0,80123), parciální korelace (0,80123), Spearmanova korelace (0,87218). Trojnásobná korelační analýza jasně ukazuje na korelační vztah mezi viskozitou a teplostálostí spoje. Z grafu je patrná vysoká pozitivní korelace obou sledovaných znaků jakosti. Proto je vhodné použít pro statistickou regulaci daného procesu Hotellingův regulační diagram a samostatné regulační diagramy ($\overline{x}$) pro jednotlivé znaky jakosti.

V posledním kroku jsem zjišťoval, zda konstrukce Hotellingova diagramu pomůže odhalit odchylky, které předchozí regulační diagramy nezaznamenaly. Podezřelost z těchto odchylek je patrná z korelačních grafů pro jednotlivé znaky jakosti u tavných lepidel.



Obrázek 15: Graf korelace pro oba dva sledované znaky jakosti (Statgraphics)

Hotellingův diagram pro všechny parametry u tavných lepidel

Následovat bude Hotellingův diagram, pro jehož sestrojení jsem musel stanovit hodnoty testového kritéria T_j^2. Graf tohoto diagramu, viz obr. 14, jsem sestrojil pomocí výběrových průměrů jednotlivých podskupin měřených znaků jakosti, vektoru středních hodnot a kovarianční matice pro stanovení testového kritéria T_j^2.

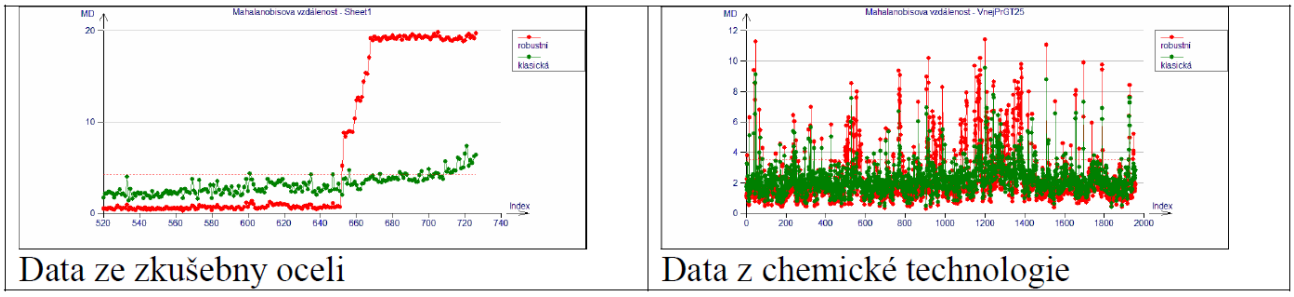
Analyzuji-li Hotellingův diagram, zjistím, že hodnota T^2_{10} překračuje regulační mez. U výběru č. 10 jde o výraznou odchylku. Pokud bychom zkonstruovali stejné regulační diagramy jako u teplostálosti spoje, tak bychom zjistili, že ani jeden z nich u 10. výběru nesignalizuje statisticky nezvládnutelný proces. Rozpor mezi výpovědí Hotellingova diagramu a samostatných regulačních diagramů pro jednotlivé znaky jakosti podporuje tvrzení, že pro korelovaná data je nutné použít Hotellingův diagram a ne pouze samostatné regulační diagramy pro jednotlivé znaky jakosti.

Naopak, pokud bychom zkonstruovali regulační diagram ($\overline{x}$) pro teplostálost spoje, signalizoval by statisticky nezvládnutý stav u 6. výběru, ale v Hotellingově diagramu tato odchylka signalizována není. To podporuje doporučení, aby současně s Hotellingovým diagramem byly vedeny a analyzovány i samostatné regulační diagramy pro jednotlivé znaky jakosti. Jestliže však znaky jakosti sledované simultánně na jednom produktu nejsou korelovány, dávají oba postupy (tj. Hotellingův diagram a samostatné regulační diagramy pro jednotlivé znaky jakosti) přibližně stejné výsledky.

2.4. Robustní Hotellingovy diagramy

Robustifikace regulačního diagramu spočívá v robustních odhadech polohy (vektoru průměru) a kovarianční matice. Robustní odhady nejsou ovlivněny vybočujícími a netypickými daty tolik jako klasické odhady, jako jsou například průměry atd.

Jako robustních technik odhadů lze použít M-odhady založené na iterativních výpočtech váženého průměru pomocí vlivových funkcí. Použití robustních regulačních diagramů přinese značné zostření a vyšší citlivost diagnostiky výchylek v procesu, jak je ilustrováno na následujících grafech, kde zelené body představují klasický a červené robustní Hotellingův diagram.



Data ze zkušebny oceli

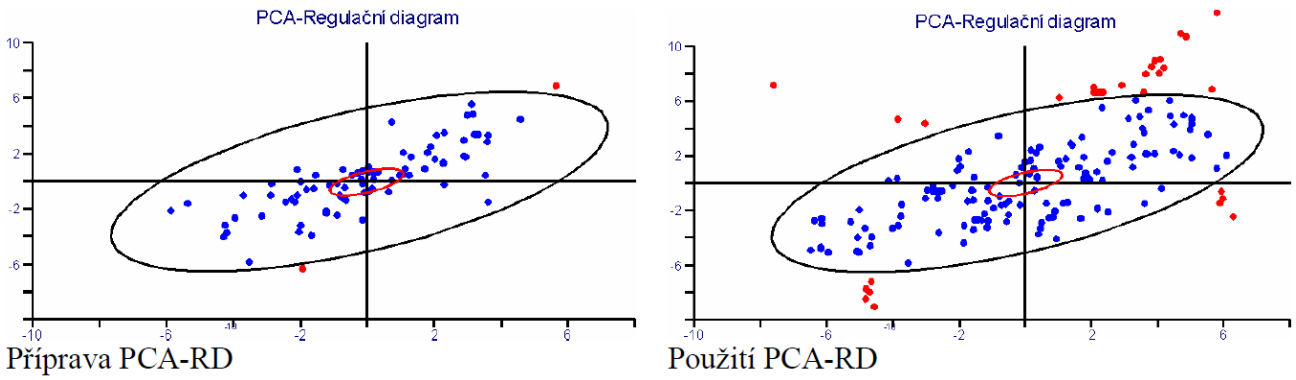
Data z chemické technologie

Obrázek 16: Srovnání klasického a robustního Hotellingova diagramu

2.5. PCA – regulační diagram

Další z možností, které nabízí vícerozměrná analýza, je metoda PCA (Principal Component Analysis – Metoda hlavních komponent). Zde lze využít výrazné snížení dimenze, případně použití pouze prvních dvou hlavních komponent k popisu procesu. Často se podaří do prvních dvou komponent promítnout podstatnou část informace i o mnohorozměrném procesu. Pak stačí využít projekce do těchto dvou komponent s limitním elipsoidem odpovídajícím 99,73 % kvantilu normálního rozdělení. Tento model lze pak využít jako diagnostický nástroj k identifikaci neobvyklých měření.

Výhodou tohoto postupu proti Hotellingovu diagramu je vyšší stabilita, nevýhodou použití neúplné informace, absence časové osy a obtížná interpretace příčiny bodů ležících vně elipsoidu.



Obrázek 17: PCA – regulační diagram

3. Závěr

Problémy monitorování procesu, ve kterém se sleduje několik korelovaných proměnných současně, jsou společně označovány jako vícerozměrné statistické řízení procesů (MSPC). Předkládaný příspěvek má dvě části. V první části jsem se zaměřil na charakterizování statistického řízení kvality a popis regulačního diagramu. Ve druhé části svého příspěvku jsem diskutoval tři nejčastěji používané druhy vícerozměrných diagramů. Mezi ně patří Hotellingova statistika T-kvadrát, vícerozměrné exponenciálně vážené průměry (MEWMA) a vícerozměrné kumulované součty (MCUSUM). Každou diskutovanou část vícerozměrných diagramů jsem pro lepší pochopení čtenáře zakončil případovou studií.

Použitá a doporučená literatura

[1] Cézová E. (2008), Ekonomicko-statistický návrh regulačního diagramu, sborník konference *Request'08,* CQR, VUT Brno.

[2] ČSN ISO 7870 *Regulační diagramy – Obecné pokyny a úvod.* Praha: Český normalizační institut, 1995.

[3] ČSN ISO 8258 *Shewhartovy regulační diagramy.* Praha: ČNI, 1993.

[4] Hebák, P. a kol. *Vícerozměrné statistické metody.* 1. vyd. Praha: Informatorium, spol. s r. o., 2004. 236 s. ISBN 80-7333-025-3.

[5] Kupka, K. *Statistické řízení jakosti.* 1. vyd. Pardubice: TriloByte, 2001. 191 s. ISBN 80-238-1818-X.

[6] Königová, M. a kol. *Matematické a statistické metody v informatice.* 1. vyd. Praha: Státní pedagogické nakladatelství, n. p., 1988. 192 s. IČ 14-556-88.

[7] Meloun, M.; Militký, J. *Kompendium statistického zpracování dat.* 2. vyd. Praha: Academia, nakladatelství Akademie věd České republiky, 2006. 982 s. ISBN 80-200-1396-2.

[8] Meloun, M.; Militký, J.; Hill, M. *Počítačová analýza vícerozměrných dat v příkladech.* 1. vyd. Praha: Academia, nakladatelství věd České republiky, 2005. 450 s. ISBN 80-200-1335-0.

[9] Meloun, M.; Militký, J. *Statistická analýza experimentálních dat.* 2. vyd. Praha: Academia, nakladatelství Akademie věd České republiky, 2004. 953 s. ISBN 80-200-1254-0.

[10] Minitab 14 – Help.

[11] Noskievičová, D. Automatizovaná výroba a SPC. In *Automatizace,* 2001, číslo 7–8, strana 5–9.

[12] Tošenovský, J.; Noskievičová, D. *Statistické metody pro zlepšování jakosti.* 1. vyd. Ostrava: Montanex, a. s., 2000. 362 s. ISBN 80-7225-040-X.

[13] Zimmerman, S. M., et al. *Statistical Quality Control Using Excel.* ASQ 2003. 249 s. ISBN 0873895665.

[14] Healy, J. D. *A note on multivariate CUSUM procedures.* Technometrics, 1987, Vol. 29, pp. 409–412.

[15] Kalgonda, A. A., Kulkarni, S. R. *Multivariate quality control chart for autocorrelated processes.* Journal of Applied Statistics. 2004, Vol. 31, pp. 317–327.

[16] Lowry, C. A., Woodall, W. H., Champ, C. W., Rigdon, S. E. *Multivariate exponentially weighted moving average control chart.* Technometrics, 1992, Vol. 34, pp. 46–53.

[17] Runger, G. C. *Multivariate statistical process control for autocorrelated processes.* Intl. Journal of production Research, 1996, Vol. 34, pp. 1715–1724.

[18] Bass, I. *Six Sigma Statistics with Excel and Minitab.* 1. vydání. Nakladatelství The McGraw-Hill Companies, Inc., United States of America, 2007. 374 s. ISBN 978-0-07-148969-0.

[19] Bass, I.; Lawton, B. *Lean Six Sigma Using SigmaXL and Minitab.* 1. vydání. Nakladatelství The McGraw-Hill Companies, Inc., United States of America, 2009. 362 s. ISBN 978-0-07-162621-7.

[20] BISSELL, D. *Statistical Methods for SPC and TQM.* 1. vydání. Nakladatelství Chapman and Hall, London 1994. 373 s. ISBN 0-412-39440-5.

[21] Kovářík, M. *Projekt zavedení statistické regulace jakosti v podniku Tegü Vuko, s. r. o.* Diplomová práce. Zlín: UTB, FaME, 2007. Bez ISBN.

[22] Fuchs, C.; Kenett, Ron S. *Multivariate Quality Control.* 1. vydání. Nakladatelství Marcel Dekker, Inc, 1998, New York. 212 s. ISBN 0-8247-9939-9.

[23] English, J. R.; Taylor, G. D. *Process capability analysis – a robustness study.* International Journal of Product Research, 31, 1621–1635, 1993.

[24] Kovářík, M. *Vícerozměrné statistické řízení procesů.* XII. ročník mezinárodní konference MEKON 2010. Technická univerzita Ostrava, Ekonomická fakulta. 3. – 4. února 2010, Ostrava. ISBN 978-80-248-2165-8.

[25] Škop, M. *Od regulačních diagramů k Six Sigma.* Řízení jakosti – Automa, 2001. Číslo 7–8.

[26] Fabian F.; Horálek V.; Křepela J.; Michálek J.; Chmelík V.; Chodounský J.; Král J.: *Statistické metody řízení jakosti.* Praha, ČSJ, 2007. ISBN 978-80-02-01897-1.

[27] Hůlová M.; Jarošová E. *Statistické metody v managementu kvality, environmentu a bezpečnosti.* 2. vyd. Praha: Ediční oddělení VŠE, 2001. 119 s. ISBN 80-245-0251-8.

[28] Chambers David S.; Wheeler D. J. *Understanding Statistical Process Control.* 2nd edition. USA: SPC Press, Inc., 1992. 300 s. ISBN 0-945320-13-2.

[29] Chandra, M. Jeya. *Statistical Quality Control.* 1. vydání. Nakladatelství CRC Press, LLC., United States of America, 2001. 280 s. ISBN 0-8493-2347-9.

[30] Mason, Robert L.; Young, John C. *Multivariate Statistical Process Control with Industrial Applications.* 1. vydání. Vydalo The American Statistical Association and the Society for Industrial and Applied Mathematics, Philadelphia, 2002. 263 s. ISBN 0-89871-496-6.

[31] Montgomery, Douglas C. *Introduction to Statistical Quality Control.* 6. vydání. Nakladatelství John Wiley & Sons, Inc, 2009. 734 s. ISBN 978-0-470-16992-6.

[32] Oakland, John S. *Statistical Process Control.* 6th edition. USA: SPC Press, Inc., 2008. 472 s. ISBN 0-7506-5766-9.

[33] Ryan, Thomas P. *Statistical Methods for Quality Improvement.* 2. vydání. Nakladatelství John Wiley & Sons, Inc. United States of America, 2000. 555 s. ISBN 0-471-19775-0.

[34] Stapenhurst, T. *Mastering Statistical Process Control.* 1. vydání. Nakladatelství Charon Tec Pvt. Ltd, Chennai, India. United Kongdom 2005. 460 s. ISBN 0-7506-6529-7.

[35] Wheeler, D. J. *Advanced Topics in Statistical Process Control: The Power of Shewhart's Charts.* 2nd edition. USA: SPC Press, Inc., 2004. 470 s. ISBN 978-0945320630.

[36] Zmatlík, J. Trendy pro manažery, Ekonomika a management: Shewhartovy regulační diagramy a jejich účinnost. *Automatizace.* 2006, roč. 49, č. 2, s. 74.

KDE STUDENTI HLEDAJÍ INFORMACE

Marta Žambochová

Adresa: FSE UJEP, KMS, Moskevská 54, CZ-400 96, Ústí nad Labem

E-mail: marta.zambochova@ujep.cz

Poděkování: Tato práce byla podporována grantem IGA 45 206 15 0001 01.

Abstract: Modern era creates increased pressure on individual education of people and puts increasing emphasis on the most effective acquisition of knowledge. It is important to know the popularity of different methods of knowledge transfer for providers of this information. We conducted a survey among people over 15 years and we found out how and where they get their information and knowledge. We made some classification of respondents from different perspectives based on this survey. We performed the classification of respondents namely in terms of age, size of the village, the educational attainment, type of the education, and the field of the education. We conducted the classification of respondents also in terms of popularity of different information sources. We used two types of classification – Cluster Analysis and Classification Trees.
Keywords: Education, Information Sources, Cluster Analysis, Classification Trees.

Abstrakt: Moderní doba vytváří stále větší tlak na individuální vzdělávání člověka a tím klade stále větší důraz na co nejefektivnější získávání vědomostí. Pro poskytovatele těchto informací je důležité znát oblibu jednotlivých způsobů předávání znalostí. Uskutečnili jsme průzkum mezi lidmi staršími 15 let a zjišťovali, jak a kde získávají informace a vědomosti. Na základě tohoto průzkumu jsme provedli klasifikaci respondentů z různých hledisek, jmenovitě z hlediska věku, velikosti obce bydliště, dosaženého vzdělání, jeho oboru a typu, a zároveň z pohledu oblíbenosti různých informačních zdrojů. Použili jsme dva základní typy klasifikace – shlukovou analýzu a klasifikační stromy.
Klíčová slova: Vzdělávání, informační zdroje, shluková analýza, klasifikační stromy.

1. Úvod

Hlavní motivací našeho výzkumu byla analýza alternativních možností financování terciálního školství. Primárně jsme se zaměřili na školné placené studenty. Jedním z našich cílů byl průzkum zájmu studentů ochotných platit určitou formu školného, a to především studentů zahraničních a studentů

celoživotního vzdělávání. Jako jednu ze základních oblastí, kde je možno studenty oslovit ohledně zvýšení jejich zájmu a ochoty platit, jsme uvažovali oblast podpory studentů ve studiu. A právě tímto tématem se zabývá článek. Mezi hlavní otázky průzkumu patřily:

- identifikační údaje
  - věk
  - pohlaví
  - velikost bydliště
- informace o vzdělání
  - výše dokončeného vzdělání
  - prospěch v rámci střední školy
  - počet neúspěšných vysokoškolských studií
  - (případná) současná vysoká škola
  - převažující obor vzdělání
- využívání zdrojů k získávání informací
  - internetové vyhledávače
  - Wikipedie
  - intranetové zdroje vlastní školy
  - učebnice a skripta
  - jiné knihy
  - odborné časopisy a články
  - přímá výuka (škola)
  - přímá výuka (doučování a kroužky)
  - jiné zdroje

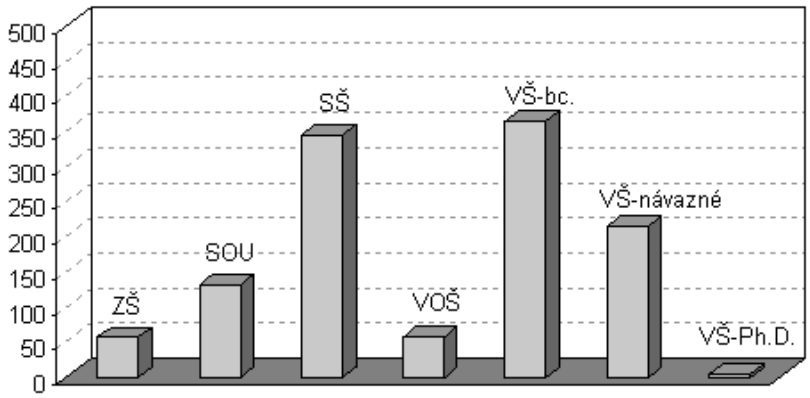
2. Data

Na dotazník odpovědělo 1073 respondentů starších 15 let. Jejich vzdělanostní struktura je zřejmá z obrázku 1. Je zřejmé, že v průzkumu převažují středoškoláci a bakaláři. Nejméně respondentů bylo s doktorským vzděláním.

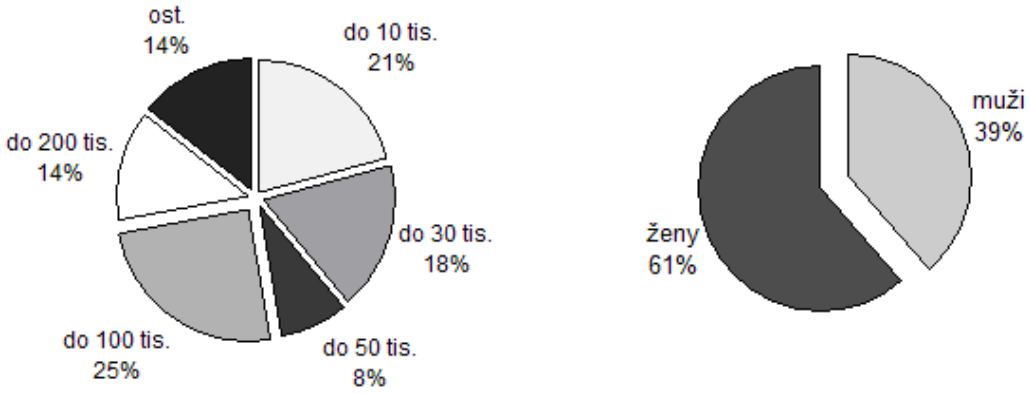
Z grafů na obrázku 2. je vidět struktura respondentů z pohledu velikosti obce bydliště a dle pohlaví. Nejvíce respondentů pochází z obcí s počtem obyvatel mezi 50 a 100 tisíci obyvatel. Nejméně respondentů je z obcí s počtem obyvatel v rozmezí 30 až 50 tisíc. Mezi respondenty bylo 61 % žen a 39 % mužů.

Z obrázku 3. je zřejmá struktura respondentů z hlediska zaměření jejich studia. Většina respondentů je humanitního zaměření a asi třetina respondentů je technického zaměření. Ostatní směry zaměření jsou jen minoritní.

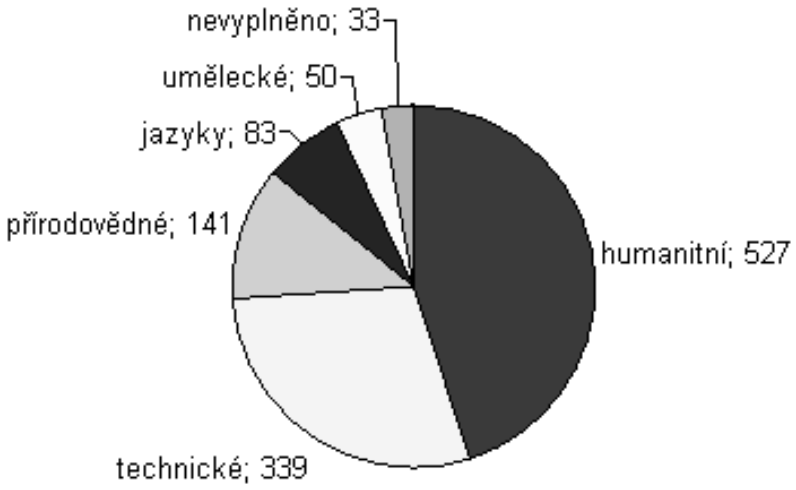
V tabulce 1. jsou shrnuty souhrnné údaje o prospěchu respondentů na střední škole. Tabulka obsahuje četnosti respondentů daných vlastností. Kaž-



Obrázek 1: Vzdělanostní struktura respondentů



Obrázek 2: Struktura respondentů dle velikosti obce bydliště a dle pohlaví



Obrázek 3: Struktura respondentů z hlediska studijního zaměření

Tabulka 1: Prospěch respondentů na střední škole

	Hum.	**Přír.**	**Tech.**	**Cizí j.**	**Min**	**Max**
Nebyly	99	125	159	54	276	276
Podprůměr	89	123	214	200	6	390
Průměr	555	649	514	582	263	476
Nadprůměr	430	276	286	337	628	31

dý respondent měl uvést svůj prospěch ve stupnici podprůměr – průměr – nadprůměr, a to jednak v oblasti humanitních (bez jazyků), přírodovědných, technických předmětů a cizích jazyků, případně měl respondent uvést, že daný obor předmětů na střední škole nestudoval. Z tabulky je dle očekávání zřejmé, že ve všech typech předmětů převládá průměrné hodnocení, nejlepší výsledky mají studenti v oblasti humanitních předmětů a naopak nejhorší v oblasti technických předmětů. Dále jsou v tabulce uvedeny souhrnné údaje o nejlepším a nejhorším hodnocení daného respondenta. Z těchto údajů je zřejmé, že 6 respondentů uvedlo ve všech případech podprůměrné hodnocení a naopak 31 respondentů uvedlo ve všech případech hodnocení nadprůměrné.

Nejvíce respondentů-vysokoškoláků pocházelo z Univerzity J. E. Purkyně v Ústí nad Labem, dále z Univerzity Karlovy v Praze a Českého vysokého učení technického v Praze.

3. Zpracování dat

Oblíbenost zdrojů informací

Ve zpracování dat jsme se nejprve zabývali sledováním oblíbenosti jednotlivých zdrojů informací. Respondentům bylo nabídnuto osm různých typů zdrojů a pro každý z nich měli dotázaní uvést míru oblíbenosti ve stupnici 0 až 10, kde 0 znamenala, že respondent daný typ nevyužívá nikdy, a 10 znamenala nejvyšší míru obliby. Pracovali jsme s ordinálními veličinami, proto byl použit Friedmanův test, který je založen na pořadí hodnot. Viz [3] či [2]. Nulovou hypotézou byla nezávislost míry obliby na typu zdroje, čili srovnatelná úroveň obliby v rámci všech nabízených zdrojů. Výsledná p-hodnota $1{,}2 \cdot 10^{-14}$ poukazuje na zamítnutí nulové hypotézy, tedy míra oblíbenosti se u jednotlivých nabízených zdrojů informací významně liší. V tabulce 2. jsou uvedena průměrná pořadí jednotlivých zdrojů.

Z výše uvedeného je vidět, že největší oblibu má internet, následován učebnicemi a přímou výukou. Naopak jako nejméně užitečná byla uváděna mimoškolní výuka a interní internetové učební materiály daných škol. Dále

Tabulka 2: Průměrná pořadí sledovaných zdrojů informací dle Friedmanova testu

Informační zdroj	**Průměrné pořadí**
Internet	6,21
Učebnice a skripta	5,30
Přímá výuka	5,23
Ostatní knihy	4,94
Odborné časopisy	4,47
Wikipedie	4,03
Intranetové zdroje vlastní školy	3,25
Mimoškolní výuka	2,56

měli respondenti možnost uvést jiné využívané zdroje informací. Mezi těmito ostatními zdroji byly nejčastěji uváděny dokumentární pořady v médiích, kolegové, spolužáci a odborné semináře.

Klasifikace

Při provádění klasifikace respondentů jsme nejprve použili shlukovou analýzu, která se řadí ke klasifikačním metodám „učení bez učitele“. Shluková analýza (Cluster analysis) [1], [4] se zabývá podobností datových objektů. Řeší dělení množiny objektů do několika předem nespecifikovaných skupin (shluků, clusters) tak, aby si objekty uvnitř jednotlivých shluků byly co nejvíce podobny a objekty z různých shluků si byly podobny co nejméně. Shlukovou analýzu lze provádět mnoha různými metodami. Jednotlivé metody se od sebe liší jednak různými způsoby určování podobnosti objektů (měr podobnosti) a jednak způsoby shlukování (např. hierarchické a nehierarchické). Při výběru metody shlukové analýzy záleží jednak na tom, zda máme k dispozici přímo zdrojová data či agregované údaje (např. tabulku četností, či matici podobností). Pokud máme k dispozici zdrojová data, je výběr metody závislý na typu proměnných (nominální, ordinální, kvantitativní proměnné). V našem případě jsme pracovali s ordinálními proměnnými vyjadřujícími míru obliby jednotlivých informačních zdrojů. Tyto proměnné nabývaly hodnot 0 až 10. Pro zpracování našich dat nebyl z důvodu relativně velkého počtu objektů, a tím i malé přehlednosti výsledků, vhodný výběr hierarchické metody. Vhodnější se jevil algoritmus k-průměrů. Nejvhodnější metodou pro zpracování dat byla dvoukroková (two-step) metoda.

Princip algoritmu k-průměrů je uveden například v [1] či [4]. Principy dvoukrokové shlukové analýzy jsou uvedeny například v [4]. Tato metoda využívá algoritmu BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), který je blíže popsán v [7] či [8]. V statistickém systému SPSS je od verze 11.5 imlementována metoda two-steps.

Rozhodovací stromy se řadí do skupiny metod učení s učitelem, kde se rozhodovací pravidla pro zařazení objektů do tříd vytváří na základě učící (trénovací) množiny. Různé typy rozhodovacích stromů jsou velmi rozšířenou skupinou stromů, kterých se využívá v datových modelech. Rozhodovací stromy jsou struktury, které rekurzivně rozdělují zkoumaná data dle určitých rozhodovacích kritérií. Kořen stromu reprezentuje celý populační soubor. Vnitřní uzly stromu reprezentují podmnožiny populačního souboru. V listech stromu můžeme vyčíst hodnoty vysvětlované proměnné. Využívají se dva typy rozhodovacích stromů, a to klasifikační stromy (v každém listu je přiřazení třídy) a regresní stromy (v každém listu je přiřazení konstanty – odhad hodnoty vysvětlované proměnné).

Pro vytváření rozhodovacích stromů bylo vyvinuto velké množství algoritmů. Nejvíce používané jsou CART, ID3, C4.5, AID, CHAID a QUEST, viz např. [16] či [5]. Pro práci jsme využili tři typy, jejichž algoritmy jsou implementovány ve statistickém systému SPSS, a to CART, CHAID a QUEST.

Nejdříve jsme provedli shlukovou analýzu, a to jednak dvoukrokovou metodu a jednak metodu k-průměrů, obě jsme zpracovávali v systému SPSS.

Dvoukroková metoda vytvořila dva následující shluky:

- 1. shluk
  - 506 respondentů, kteří
  - vůbec nevyužívají intranet ani Wikipedii,
  - nevyužívají intranet a učebnice jen mírně.
- 2. shluk
  - 659 respondentů, kteří
  - využívají intranet,
  - nevyužívají intranet, ale velmi využívají učebnice či znají Wikipedii.

Dále jsme vytvořili novou proměnnou týkající se příslušnosti ke shluku. Tuto proměnnou jsme použili jako vysvětlovanou proměnnou při tvorbě klasifikačního stromu. Za vysvětlující proměnné jsme zvolili následující faktory:

- věk,
- pohlaví,
- velikost místa bydliště,

- dokončené vzdělání,
- prospěch v rámci SŠ,
- počet neúspěšných VŠ studií,
- převažující oborové vzdělání.

Vytvořili jsme klasifikační strom pomocí metod QUEST, CHAID a CRT, všechny v systému SPSS. Nejlepší kvalitu měl strom vytvořený pomocí metody CRT. Jeho hodnota risk estimate byla 0,23.

Na základě takto vytvořeného klasifikačního stromu jsme zjistili reprezentativní vlastnosti respondentů přiřazených k jednotlivým shlukům, a to:

- 1. shluk
  - starší lidé,
  - lidé středního věku mající nižší vzdělání technického či uměleckého zaměření.
- 2. shluk
  - lidé mladší 30 let mající alespoň vyšší odbornou školu,
  - mladí lidé s nižším vzděláním přírodovědného, humanitního či jazykovědného zaměření.

Dále jsme provedli shlukovou analýzu pomocí metody k-průměrů. Nejlépe vyšla kvalita při vytvoření dvou shluků, které vypadaly následovně:

- 1. shluk
  - 380 respondentů, kteří
  - neupřednostňují přímou výuku.
- 2. shluk
  - 785 respondentů, kteří
  - upřednostňují přímou výuku a učebnice,
  - nevyužívají intranet, ale velmi využívají učebnice, či znají Wikipedii.

Opět jsme dále vytvořili klasifikační stromy pomocí metod QUEST, CRT a CHAID, jejichž vysvětlovanou proměnnou byla příslušnost ke shluku a vysvětlující proměnné byly vybrány stejně jako v předchozím případu. V tomto případě vyšel nejlépe strom vytvořený pomocí algoritmu QUEST, jehož hodnota risk estimate byla 0,285. Reprezentativní vlastnosti respondentů přiřazených k jednotlivým shlukům byly následující:

- 1. shluk
  - starší lidé,
  - lidé mladší a středního věku mající nižší vzdělání a byli podprůměrní v humanitních předmětech.

- 2. shluk
  - lidé mladší a středního věku mající vyšší vzdělání,
  - lidé mladší a středního věku mající nižší vzdělání, ale byli alespoň průměrní v humanitních předmětech.

Metoda k-průměrů dala ještě dobrý výsledek v případě vytváření tří shluků, které lze popsat následovně:

- 1. shluk
  - 358 respondentů, kteří
  - neupřednostňují přímou výuku.
- 2. shluk
  - 577 respondentů, kteří
  - upřednostňují přímou výuku, hodně využívají internet a znají Wikipedii.
- 3. shluk
  - 232 respondentů, kteří
  - upřednostňují přímou výuku, ale internet využívají jen průměrně.

I v tomto případě jsme následně vytvořili klasifikační strom pomocí všech výše zmíněných metod. V tomto případě měl nejlepší kvalitu strom vytvořený pomocí metody QUEST, jehož hodnota risk estimate byla 0,315. Reprezentativní vlastnosti respondentů přiřazených k jednotlivým shlukům byly:

- 1. shluk
  - starší lidé,
  - muži mladší a středního věku mající nižší vzdělání technického či uměleckého směru.
- 2. shluk
  - lidé středního věku mající vyšší vzdělání,
  - mladší lidé mající nižší vzdělání.
- 3. shluk
  - mladší ženy uměleckého a humanitního zaměření.

4. Závěr

V průzkumu jsme oslovili větší množství respondentů napříč věkovými kategoriemi i vzděláním. U respondentů jsme sledovali oblibu jednotlivých informačních zdrojů a faktory, které potencionálně tuto oblibu ovlivňují. Data jsme dále zpracovali jednak pomocí vybraných testů hypotéz, ale také pomocí různých typů klasifikace, a to shlukové analýzy a klasifikačních stromů. Výsledky našeho průzkumu můžeme shrnout do následujících závěrů.

Nejoblíbenějším zdrojem informací se jeví internet, následován je učebnicemi a přímou výukou. Mladší muži a muži středního věku mající nižší vzdělání technického či uměleckého zaměření se vyhýbají přímé výuce, na rozdíl od mladých žen humanitního a uměleckého zaměření, které přímou výuku upřednostňují. Vzdělanější mladší lidé dají přednost internetu před intranetem. Učebnice jsou preferovány napříč celým spektrem respondentů.

Je tedy zřejmé, že studenti stále preferují učení se z učebnic a skript. Ne zcela důvěřují vlastním internetovým výukovým stránkám školy. V případě internetu upřednostňují veřejné webovské stránky. Tento fakt by zasloužil hlubší analýzu prozkoumávající příčinu tohoto jevu. Není vyloučeno, že touto příčinou je nedostatečná kvalita internetových výukových stránek dané školy.

Reference

[1] Hebák, P.; Hustopecký, J.; Pecáková, I.; Plašil, M.; Průša, M.; Řezanková, H.; Vlach, P.; Svobodová, A. (2007) *Vícerozměrné statistické metody [3].* 2. vyd. Informatorium, Praha, 272 s.

[2] Novák, I.; Pecáková, I. (2001) *Měření souvislostí kategoriálních proměnných.* Statistika, 2001, roč. 38, č. 1, 35 – 48.

[3] Řezanková, H. (2010) *Analýza dat z dotazníkových šetření.* 2. uprav. vyd., Professional Publishing, Praha, 217 s.

[4] Řezanková, H.; Húsek, D.; Snášel, V.: (2009) *Shluková analýza dat.* Professional Publishing, Praha, 220 s.

[5] Timofeev R. (2004) *Classification and Regression Trees (CART) Theory and Applications.* Master thesis, CASE-Center of Applied Statistics and Economics, Humboldt University, Berlin.

[6] Wilkinson, L. (1992) *Tree Structured Data Analysis: AID, CHAID and CART.* Sun Valley, ID, Sawtooth/SYSTAT Joint Software Conference.

[7] Zhang, T.; Ramakrishnan, R.; Livny, M. (1996) *BIRCH: An Efficient Data Clustering Method for Very Large Databases.* ACM SIGMOD Record, Vol. 25. No. 2, 103 – 114.

[8] Zhang, T.; Ramakrishnan, R.; Livny, M. (1997) *OBIRCH: A New Data Clustering Algorithms and Its Applications.* Journal of Data Mining and Knowledge Discovery, Vol. 1, No. 2, 141 – 182.

COOPERATION ON PUBLICATIONS AND SOCIAL NETWORK ANALYSIS

SPOLUPRÁCE NA TVORBĚ PUBLIKACÍ A ANALÝZA SOCIÁLNÍCH SÍTÍ

Nikola Kaspříková

Address: Katedra matematiky, Vysoká škola ekonomická v Praze, Nám. W. Churchilla 4, 130 67 Praha 3, Czech Republic

E-mail: `data@tulipany.cz`

Abstract: This paper reports on results of data analysis of bibliographic database of scholarly publications of authors affiliated with particular institution. Patterns of cooperation on authorship of publications are investigated using social network analysis (SNA) tools. Definitions of selected concepts used within SNA framework are recalled and an application of a couple of basic SNA tools is shown.

Keywords: Social Network Analysis, Closeness Centrality, Collaboration on Authorship, Publications, `sna`.

Abstrakt: Článek je věnován analýze bibliografické databáze odborných publikací autorů z vybraného pracoviště. Spolupráce na tvorbě publikací je zkoumána pomocí nástrojů pro analýzu sociálních sítí. Jsou připomenuty některé pojmy, se kterými se v oblasti sociálních sítí pracuje, a je ukázáno použití několika základních prostředků pro popis sítí.

Klíčová slova: Analýza sociálních sítí, balíček `sna`.

1. Introduction

Social network analysis (SNA) has been one of the major research tools used in sociology and social psychology for a long time. Introduction of a sociogram as early as in 1930's marked the beginning of sociometry (Wasserman and Faust, 1994). Concept of opinion leaders and other concepts used within social network analysis framework are still of major importance even nowadays, when the number of participants in various web-based online networking communities is increasing and SNA is also used to support business decisions in finance or telecommunication companies, among others.

SNA framework provides tools for efficient description of structure of relations in a group investigated and may be also used for identification of most popular or influential actors. SNA methods are often combined with other data analysis tools, such as text mining (Bohn et al., 2011). One of a most

common applications of SNA framework is analysis of coauthorship networks, see e. g. (Said et al., 2008) or (Newman, 2001).

Professionally maintained bibliographic databases may be supposed to represent rather reliable source of information on collaboration of authors on publications authorship, even though there may occur some data quality issues too. Analysis of cooperation on publications produced at University of Economics in Prague within years 2007 – 2009 is addressed in this paper.

2. Material and methods

2.1. Source data and problem representation

We analyse collaboration on authorship of publications published within years 2007, 2008 and 2009 by authors affiliated with University of Economics in Prague. The dataset includes 2384, 2589 and 2904 publications respectively, which gives 7877 publications in total. The three years time window should be sufficiently long for collecting enough data cases and at the same time should be short enough so that the group of authors is more or less the same.

Strictly speaking, it is not possible to distinguish authors precisely, because strings in the database, which are supposed to represent names of authors, may not be unique for all authors and at the same time it is also possible that some authors even change their name within the three years time window considered. Nevertheless data quality issues hopefully do not make any major trouble in this analysis.

Regarding data preparation step, no authors have been dropped out from the analysis, even if some of the coauthors may not be affiliated with the institution considered. If a publication has just one author, then it is certainly authored by an insider, as otherwise it would not have appeared in the database.

We do not use any special node attributes, even though information on number of publications written by particular author as a single author could have been used as a node attribute in the analysis.

We take authors (and coauthors) to be represented by nodes in a graph and there is an (undirected) edge in the graph between node B and node A if and only if A and B have published together. In our approach, we consider coauthorship relation as symmetric and we do not use any weights for edge evaluation to express strength of cooperation on authorship of publications.

2.2. Methods

We use standard and mostly descriptive tools and concepts of social network analysis and we refer the reader to (Wasserman and Faust, 1994) and (Butts, 2008) for further details.

Basic graph level characteristics of social networks include density, connectedness, reciprocity and transitivity. Density of a graph refers to number of edges in the graph expressed as a proportion of the maximum possible number of edges. It is natural that larger social networks have lower density, because the number of possible edges increases rapidly with the number of nodes in a graph and at the same time the number of connections which each person can maintain is usually limited. Connectedness refers to density within reachability graph. Another basic social network characteristic is its reciprocity. Edgewise reciprocity is the proportion of edges which are reciprocated. We consider coauthorship relation as symmetric in this analysis, so we will not investigate reciprocity. Interesting measure of social network is transitivity, which is the fraction of connected triplets of vertices which also form triangles, so it refers to probability that there is a tie between A and C for A, B and C such that there is a tie between A and B and a tie between B and C.

Centrality, which may be used as a measure of prestige of a node in the network, is a node-level characteristic, as opposed to centralization, which is a graph-level property. The centralization of a graph G with set of nodes V for centrality measure c is given by (Butts, 2008):

$$C(G) = \sum_{i=1}^{|V|} \Big[\big(\max_{v \in V} c(v, G) \big) - c(v_i, G) \Big],$$

which is in effect equal to the difference between the maximum and mean centrality scores multiplied by the number of vertices in the network. Sometimes it is more convenient to work with normalized centralization, which is obtained by dividing C(G) by its maximum across all graphs of the same order as G.

We will use degree, betweenness and closeness as measures of centrality (Butts, 2008). Degree may be interpreted as a measure of activity of a node in the network. Actors with high betweenness scores may be considered to have good control over information flow in the network, as betweenness refers to number of shortest paths between nodes in the network, which go through the particular node. Actors with high closeness scores (i. e. actors with low average of geodesic (the shortest path) distances to all other nodes) have good access to other members of the network.

We use `sna` package for social network analysis (Butts, 2008 and 2010) in R environment for statistical computing (R Development Core Team, 2011) for calculation of network characteristics.

3. Results and discussion

3.1. Basic results and graph characteristics

Authors have published 5.2 publications on average and a publication has 1.45 authors on average. Graph resulting from data under investigation has 1521 nodes (representing (co)authors) and 4183 undirected edges.

Centralization with regard to degree is 0.04 and centralization with respect to betweenness is 0.2. Connectedness of the network is 0.58 and graph transitivity is 0.61, which is quite high value when compared with coauthorship networks described in (Newman, 2001).

The network has 106 components. Size of the largest component is 1154, which represents 76 % of actors. The second largest component has only 22 nodes. When considering finite geodesic distances, mean is just 6.6 and this result supports the idea that even large networks are some sort of a small world. The highest finite geodesic distance is 18. These properties of our network are in agreement with results of other analyses of authorship collaboration networks (Newman, 2001), which report a little higher percentage of the largest component. Higher transitivity and a little smaller coverage of the group by the largest component in our network in comparison with other collaboration networks is interesting, possible explanations for this may include shorter time window considered or the fact that we have taken authors affiliated with the same institution and their publications not just in peer-reviewed journal so that it may seem natural that cooperation on publication authorship may be driven to more extent by acquaintance relationships then by true professional needs.

3.2. Node-level characteristics

Distribution of prestige index based on degree in the network is skewed, with majority of actors having degree 3 or lower. Mean degree is 5.5, i. e. an average author has collaborated with some 6 authors when considering all publications of the author. One author in the database has degree 65, which is by far the highest degree in the network. Taking degree value as a measure for assessment of authorship leadership, the author with degree value of 65 can be reckoned as an authorship leader within the group investigated.

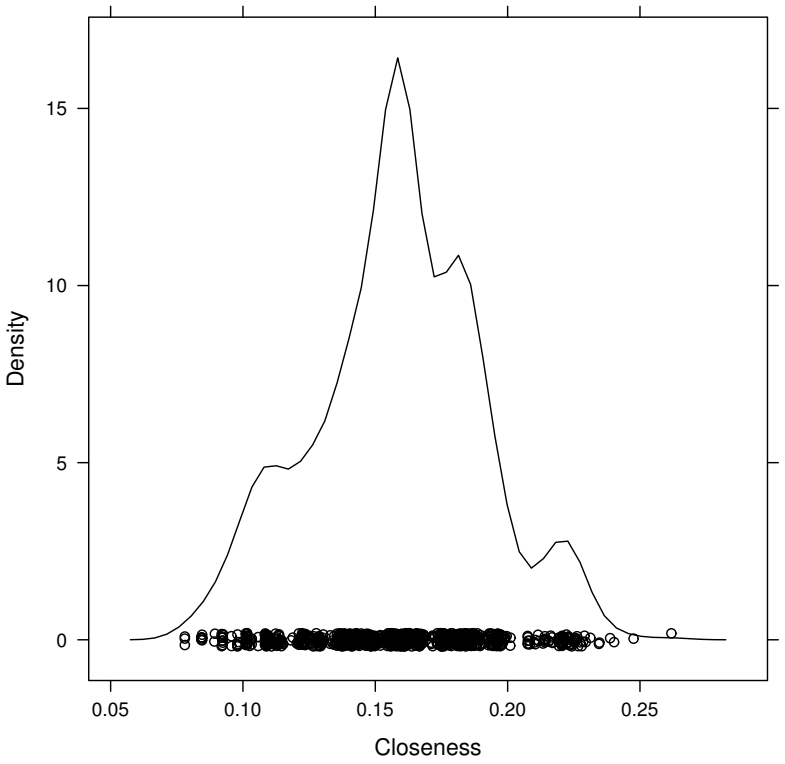


Fig. 1. Density estimate of closeness in the largest component

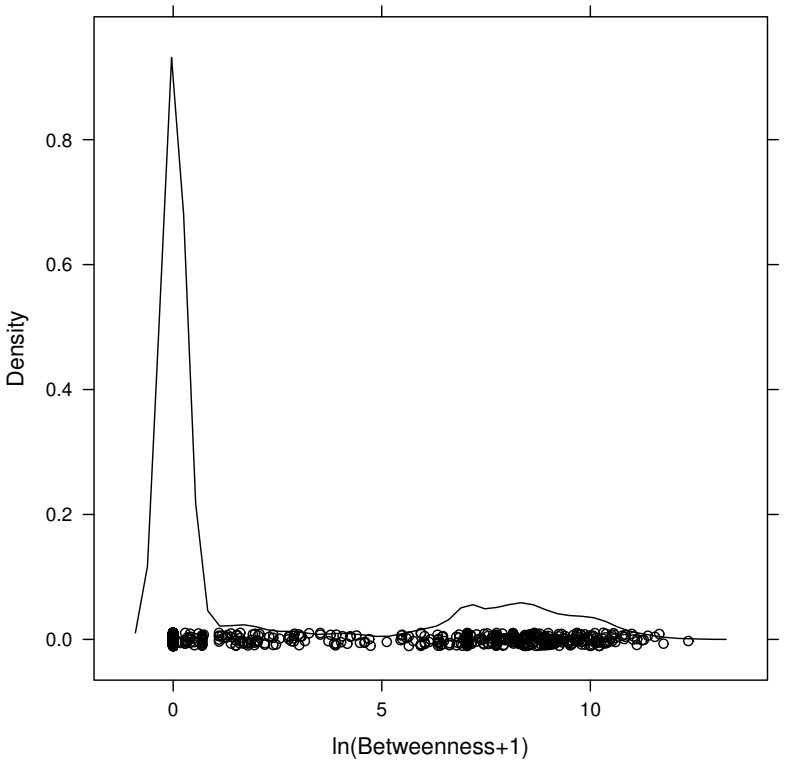


Fig. 2. Betweenness in the network

For calculation of closeness scores (see Fig. 1) we considered just the largest component. Betweenness scores (see Fig. 2) were calculated for all actors and members of the largest component showed higher betweenness scores. A cluster of authors in the largest component who have comparatively high closeness centrality (over 0.2) has been identified. This group represents approximately 5 % of all authors.

4. Conclusion

Analysis of the social network of authors linked by coauthorship relations was performed, showing basic possibilities of application of social network analysis techniques for network description. Mean finite geodesic distance in the network is 6.6, the largest component in the network includes more than 3/4 of actors and network has rather high transitivity. A group of authors in the largest component who have comparatively high closeness centrality has been identified.

Bibliography

[1] Bohn, A.; Feinerer, I.; Hornik, K.; Mair, P. (2011). *Content-based SNA of mailing lists.* The R Journal, 3(1):11–18, June 2011.

[2] Butts, C. T. (2008). *Social Network Analysis with sna.* Journal of Statistical Software, 24(6).

[3] Butts, C. T. (2010). *sna: Tools for Social Network Analysis.* R package version 2.2-0. `http://cran.R-project.org/package=sna`

[4] Newman, M. E. J. (2001). *Who is the best connected scientist? A study of scientific coauthorsip networks.* Phys. Rev. E64.

[5] R Development Core Team (2011). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL: `http://www.R-project.org/`

[6] Said, Y. H.; Wegman, E. J.; Sharabati, W. K.; Rigsby, J. T. (2008). *Social networks of author–coauthor relationships.* Computational Statistics & Data Analysis 52.

[7] Wasserman, S.; Faust, K. (1994). *Social Network Analysis: Methods and Applications.* Cambridge: Cambridge University Press.

STATISTIKA PRO NESTATISTIKY
STATISTICS FOR NON-STATISTICIANS

Hana Skalská

Adresa: Fakulta informatiky a managementu, Univerzita Hradec Králové, Rokitanského 62, 500 03 Hradec Králové 3

E-mail: hana.skalska@uhk.cz

Abstrakt: Článek shrnuje pohledy na aplikace statistiky a na její výuku pro nestatistiky. Znalost získaná procesem učení a rozvoj koncepčního uvažování jsou pro praxi přínosnější než učení se specifickým metodám. Cíle výuky musí být předem formulovány a měly by být provázeny vhodnými pedagogickými principy, které pomohou eliminovat obavy ze statistiky. Aktivizace studentů, technologické nástroje (software) a reálná data umožní lépe pochopit statistiku a rozvíjet statistické uvažování a zobecňování

Klíčová slova: Výuka statistiky, motivace, obava ze statistiky, zpětná vazba.

Abstract: Some difficulties with teaching statistics for non-statisticians are summarized and explained. Constructing knowledge through the process of learning is more important in these courses than teaching specific techniques. Learning outcomes formulated carefully and completed by appropriate pedagogical principles can reduce anxiety to statistics. Use of technological tools and real data can support conceptual understanding. Active involvement of the students and feedback on their performance help to the students to develop their statistical thinking and reasoning.

Keywords: Teaching Statistics, Motivation, Anxiety to Statistics, Feedback.

Úvod

Statistika a pravděpodobnost jsou součástí metodologie aplikovaných věd, mají místo ve výzkumu a praxi různých oborů. Poskytují specifické nástroje umožňující popis chování jevů reálného světa a souvislostí mezi nimi, srovnání rozhodovacích možností, odhad rizika, stanovení variability a neurčitosti dějů.

Principy statistického a pravděpodobnostního myšlení vedou k utváření úsudku a zobecňování, pomáhají rozvíjet kritické a objektivní uvažování, umožňují minimalizovat riziko zkreslení skutečnosti nesprávným nebo zjednodušeným vyhodnocením kvantitativní informace. Tyto způsobilosti a styl uvažování jsou důležité při rozhodování, očekávají se jako samozřejmé od absolventa vysoké školy.

Článek je zaměřený na vysokoškolskou výuku statistiky pro studenty nestatistiky, jejichž hlavním oborem studia není statistika nebo matematika. Shrnuje faktory, které působí na výuku statistiky u nestatistiků a které mohou objasnit jevy interpretované jako nezájem nestatistiků o statistiku. Důvodem je snaha nalézt vysvětlení některých problémů výuky, shrnout možnosti motivace a aktivizace studentů a pokusit se o nadhled nad vlastním přístupem k výuce.

1. Přístupy k výuce statistiky u nestatistiků

Výuka statistiky pro nestatistiky v České republice nemá v jednotlivých oborech stejnou váhu. Obsahově se kurzy částečně liší mezi obory, rozsah výuky je různý nejen mezi obory, ale pro podobné obory je odlišný i mezi univerzitami. Úvodní kurzy nepředpokládají hlubší znalost statistiky a pravděpodobnosti z nižších stupňů vzdělání.

Výukové cíle, které závisí na úrovni kurzu, vedou od získávání základních znalostí statistiky přes pochopení souvislostí mezi pojmy (rozvoj statistického uvažování) až po obecnější znalosti a rozvoj statistického myšlení. *Očekávaným efektem studia pokročilejších kurzů je porozumění možnostem statistiky a statistickému zobecňování, jejich aplikace, schopnost navrhnout vhodnou metody analýzy určitého problému, ověřit předpoklady a vysvětlit výsledky analýzy.*

Představu o cílech výuky statistiky se nedaří naplnit u všech studentů. V rámci České statistické společnosti je výuce věnována zvýšená pozornost od roku 1999, jak svědčí tématické sborníky [2], [3], [4] a další navazující práce, z nichž některé budou zmíněny.

Obecnější shrnutí a názory na výuku přináší práce Hebáka [7], který se zabývá postavením statistiky ve výuce a shrnuje názory na výuku nestatistiků. *Doporučuje méně matematického výkladu, důraz na nadhled a potlačení detailů, kvalitní výklad a kvalitní studijní literaturu. Jako nutný předpoklad vytknul hluboké znalosti přednášejících. Není nakloněn výuce u počítače ani vysvětlování vzorců nebo detailů výpočetních postupů.*

Anděl [1] popisuje výsledky hodnocení výuky statistiky u studentů matematiky, ale nestatistiků. Podle citované studie *studenti hodnotili pravděpodobnost a statistiku jako neobtížnější předmět v daném semestru. Stejní studenti nejvíce cenili výuku v předmětech, které měly jasnou organizaci, předmět naučil odbornému vyjadřování a kde byli studenti vybízeni k diskusi.* Autor na základě toho zdůrazňuje **význam přístupu vyučujícího k výuce**.

Práce Hindlse a Hronové [8] se věnuje *kritickým místům* výuky pro nestatistiky. Jedním z kritických míst jsou pojmy, které statistik považuje za

základní, ale studenti je nechápou (například míry variability). Dalším je *zbytečný formalismus a kvazi–reálné příklady* (popis operací s náhodnými jevy, zbytečné zavádění některých teoretických rozdělení), *velké množství metod a technik* (různých typů testů). Autoři zdůrazňují *význam charizmatického vyučujícího* a zmiňují *problém strachu* studentů ze statistiky, kterou si někteří studenti ztotožňují s matematikou.

2. Příčiny neporozumění statistice

Kvaszová [11] *vysvětluje neporozumění statistice Piagetovým modelem kognitivního vývoje.* Úrovně znalostí chápe jako stádia procesu poznání, která na sebe navazují. Stádium intrafigurální (INTRA) je obdobím vyjasnění základních pojmů a kvalitativního porozumění problematice. Ve stádiu interfigurálním (INTER) se vytváří a zdokonaluje proces formalizace, který umožní nahradit kvalitativní porozumění kvantitativním popisem. V procesu transfigurálním (TRANS) probíhá zobecnění a upřesnění logické stavby. Stádia INTRA a INTER jsou založena na názornosti, stádium TRANS zahrnuje jednotící pohled a vyžaduje *schopnost zobecnění.*

Vynechání nebo urychlení stádia INTRA může předurčit neporozumění kvantitativnímu popisu a zobecnění v dalším výkladu. Výzkum u vysokoškoláků ukázal, že studenti často nerozumí kvalitativnímu vyjádření jevu (fáze INTRA). Pokud se zcela vynechá tato etapa objasnění základních poznatků, jeví se pojmy vyšších úrovní poznání studentům nepřirozené. Naučí se pracovat formálně, dosazují do vzorců a nesnaží se vzorcům rozumět.

Podle Kvaszové, *pokud nejsou pojmy dostatečně pochopeny (když se urychlí stádium INTRA) a vyučující zavede pojmy z vyšších úrovní, pak student neporozumí logice myšlenkového vývoje. Nepochopí styl uvažování a vytváří si pocit nejistoty až méněcennosti k dané vědě. Důležitá je zpětná vazba, pomocí které lze kontrolovat míru porozumění u studentů.*

Obavy ze statistiky. Řada zahraničních autorů se zabývala faktorem úzkosti a obav ze statistiky, který se považuje za jeden z možných zdrojů nežádoucích efektů při výuce statistiky. Prokazují, že obavy ze statistiky ve výuce negativně ovlivňují vztah k předmětu, snižují vůli používat statistiku v praxi, vedou k neochotě analyzovat a interpretovat data a **vytvářejí bariéru budoucímu využívání statistiky**.

Vnějším projevem obav je odkládání statistiky do vyšších semestrů nebo neúspěšnost při jejím studiu. Výzkumy uvádějí tyto příznaky u 50 % – 70 % studentů nestatistiků. Měřením obav ze statistiky u nestatistiků a jejich důsledky se zabýval Onwuegbudzie [14], jeho závěry o vlivu úzkosti a obav na přístup ke statistice potvrdily i studie dalších autorů. Současné práce stu-

dují příčiny tohoto jevu a navrhují opatření. Hsu a kol. [9] empiricky ukazují, že výuka práce se softwarem, který nemá velké nároky na uživatele (SPSS), pozitivně ovlivňuje postoje k užitečnosti statistiky a nezvyšuje obavy z předmětu. Naopak, obavy ze statistiky mají negativní vliv na mínění o užitečnosti statistiky a na ochotu učit se používat software při studiu statistiky.

Lacasse a Chiocchio [12] vycházejí z různých dimenzí obav a úzkostí ze statistiky a odvozují možnost zvýšit efektivitu výuky diferencovaným přístupem a přizpůsobením procesu výuky. Perepiczka a kol. [15] potvrzují negativní korelaci mezi úzkostí a efektivitou výuky. Pro snížení obavy a pocitu úzkosti doporučují *aktivní strategie vyučujícího*, které budou obavy potlačovat. Roli vyučujícího a přístup ke studentům označují jako velmi podstatné a doporučují *entuziasmus vyučujícího, pozitivní atmosféru, diskuse se studenty o tématu, testy, kterými lze ověřit správnost uvažování, zpětnou vazbu, ocenění dílčích pokroků ve znalostech.*

3. Vliv technologií na statistiku

Změny statistiky.

Statistika se změnila s technologickým vývojem [6]. Důsledkem je částečný odklon od matematiky při její výuce. Ve většině aplikovaných oborů je statistika pomocnou vědou, její použití není téměř vázáno na specialistu [5]. Uvádí se, že dnes většinu statistických výpočtů provádějí nestatistici.

Podle Browna a Kasse [5] se současná výuka statistiky příliš soustředila na množství technik a setrvávají v ní anachronismy, které studenty nemusí zaujmout. Měla by více směřovat k principiálnímu uvažování. Proto [5] doporučují klást **větší důraz na statistické myšlení a důslednou aplikaci tohoto principu**, samotné zvládnutí technik by nemělo být cílem kurzu. Výuka by se měla soustředit na **přehledy metod a jejich srovnání, na koncepční uvažování a na analýzu postupů využitelných pro sběr dat,** predikci a vědecké zobecňování.

Úlohy převažující v praxi vyžadují znalost metod sběru dat, popisu dat a jejich vizualizace, predikce, zobecňování. Význam mají metodiky předcházející statistické analýze. Jejich vynechání předpokladem, že jsou studentům známé, může vést k názoru odtrženosti statistiky od praktických problémů a domněnce, že užití statistiky je úzce omezeno například jen na výzkumné oblasti, kde je příprava dat řízena experimentem.

Motivace nestatistika. Zvyšuje se tempo, jakým narůstá množství dat, a očekává se, že také poroste zájem o statistiku. Při analýze dat se ale stále více využívají nestatistické metody nebo metody založené na výpočetní statistice. Praktické problémy se často vyznačují velkým množstvím dat a někdy

datovými typy, pro které nejsou čistě statistické postupy vhodné. Statistické metody a pravděpodobnost se uplatňují jen u části z nich. Klasické pojetí kurzů statistiky, které uvažuje data připravená ve strukturované podobě, zjednodušuje pohled na proces analýzy vynecháním důležitých kroků. Nereflektuje očekávání studentů, že porozumí řešení reálných problémů. Je nutné hledat pro výuku *reálné úlohy, které lze obsahově co nejvíce integrovat s oborem studia. Takové úlohy mají větší naději, že nebudou považované pouze za početní trénink, ale povedou k zamyšlení nad způsobem řešení problému, jeho pochopení a integraci znalostí z různých oblastí.*

Na univerzity nastupují studenti generace Z, pro kterou jsou samozřejmé sociální sítě, ipod, internet, virtuální výpočetní prostředí, apod. Nepoznali svět bez těchto možností, proto považují informace za kdykoliv dostupné a vybírají si takové, které považují za potřebné. Z ostatních preferují informace, u kterých je *srozumitelným a přesvědčivým způsobem prezentována jejich užitečnost* a dovedou si představit jejich využitelnost.

Rizika pro výuku statistiky. Technologie změnily zvyklosti lidí a způsob uvažování. *Důsledkem spoléhání se na technologie může být snížená vůle nebo schopnost chápat význam a možnosti popisu světa pomocí kvantitativní informace v klasické formě.* Praxe zrychluje rozvoj statistiky zaměřením na nové oblasti bádání. Některé metody nevyžadují pouze odbornou znalost, ale také zručnost v používání technologií (při velkém objemu dat).

Příkladem jsou analýzy zvyklostí uživatelů webu, podobnost uživatelů sociálních sítí, kategorizace textových dokumentů, rozpoznání pokusu o podvodné jednání, rozpoznání dat v síti s nebezpečným obsahem (hacker), detekce změn v datech, hodnocení kvality dat, rozpoznání spamu, analýza účinnosti webu apod. Podobné úlohy jsou zajímavé pro studenty. Statistický kurz může připravit základ pro uvažování o podobných problémech, v kurzu to vyžaduje přizpůsobivost obsahu a multidisciplinární pojetí výuky statistiky.

Předpoklady o studentech. Možnost studovat na univerzitě má dnes větší podíl populace než dříve. Jedná se o celosvětový trend. Moore [13] uvádí, že přibližně dvě třetiny absolventů středních škol v roce 1997 v USA pokračovaly v bakalářském studiu na vysoké škole a mezi studenty tedy nejsou jenom nejlepší. Varuje před možným poklesem úrovně univerzitního vzdělání. Důsledkem jsou vyšší počty studentů v kurzech statistiky, nemožnost jejich diferenciace podle zájmu. Situace u nás je obdobná, podle [10] se zapsalo do bakalářských programů v roce 2008 zhruba 60 % studentů odpovídající věkové kohorty (nárůst 10 % oproti roku 2005). Větší počty studentů ve výuce a jejich postoje ke studiu vyžadují velkou komunikační dovednost vyučujících.

4. Vlastní zkušenost

Uvedená hlediska odpovídají zkušenosti s výukou Aplikované statistiky (APSTA) a Stochastického modelování (STOMO) pro studenty magisterského studia v oborech aplikovaná informatika a informační management.

Kurz APSTA (testování hypotéz, analýza rozptylu, vícerozměrná lineární regrese a časové řady) předpokládá znalosti z úvodního kurzu Pravděpodobnost a statistika (PSTA) bakalářského studia.

Na kurzy PSTA a APSTA navazuje STOMO (modelování, simulace, generování pseudonáhodných čísel, statistické testy generátorů, modely dynamických diskrétních dějů a demografického procesu). Všechny tři kurzy jsou šestikreditové, jednosemestrové, ukončeny zkouškou, mají týdenní čtyřhodinovou dotaci na přímou výuku (přednáška a cvičení) a hodinu týdně na semestrální práci. Ve výuce se používá statistický software a Microsoft Excel, simulace na webu, ve STOMO někteří studenti programují vlastní aplikace.

Kurz STOMO rozšiřuje aplikace statistických testů. Seznamuje s vybranými typy modelů a metod (Markovovy řetězce, model obnovy, úmrtnostní tabulka). Seminární práci studenti předkládají jako individuální nebo skupinový projekt. V semestru píší dva průběžné testy. Od určitého výsledku semestrálních testů a projektu získají bonifikaci k výsledku zkouškového testu.

Pro seminární práce podporujeme i řešení vlastních témat studentů, potom konzultují předem zaměření práce a způsob řešení. Nápadité práce jsou zpřístupněny pro motivaci v dalších letech výuky tohoto předmětu prostřednictvím LMS Blackboard.

Typickou odezvou studentů po zkoušce STOMO je nabytí dojmu, že konečně pochopili statistiku. Více zaujmou postupy, pro které dovedou najít vlastní využití. Studenti Aplikované informatiky dosahují o něco lepších výsledků u zkoušky.

Pro udržení zájmu je nutné v průběhu času provádět aktualizace úloh a doporučených témat pro projekty. Pozorované rozdíly mezi přístupem studentů ke kurzu APSTA (který obsahuje více teorie a nových konceptů a studenti mají častější obavy z neporozumění) a STOMO odpovídají názorům uvedeným výše a popisovaným v literatuře. Zdá se, že studenti nemají problém s kurzem, který přináší určitý nadhled nad metodami, které poznali v předchozí výuce. Je však nutné (v intencích uvedených doporučení) kurz dynamicky přizpůsobovat a vyhledávat nové aplikace, nové řešené problémy a typové úlohy.

5. Doporučení a závěr

Příspěvek je motivován hledáním ověřených principů výuky, podnětů pro aktualizaci obsahu a pro předcházení stereotypům ve výuce. Mezi hlavní zásady patří aktivizace a vedení k porozumění možnostem statistiky. Snahou je předcházet obavám ze statistiky, bez snižování požadavků na výukový cíl. Shrnout lze tato doporučení:

- Nepodcenit různou úroveň zkušeností studentů s popisem světa. Vysvětlovat možnosti a nástroje kvantitativního popisu dějů.
- Uvědomit si možnost obavy ze statistiky, která může negativně ovlivnit vztah ke statistice a její aplikaci v budoucnu. Obavy lze zmírnit přístupem vyučujícího.
- Význam mají komunikační schopnosti vyučujícího, jeho entuziasmus, povzbuzení a podněty k uvažování, atmosféra ve výuce, promyšlená příprava výuky, přizpůsobení stylu výuky.
- Aktivizovat formou diskusí a slovních testů, ověřovat porozumění pojmů a metod.
- Stanovit předem pravidla a požadavky výuky. Neklást nereálné požadavky.
- Do obsahu kurzu nezařazovat dlouhý seznam témat a pojmů.
- Vzorce lze uvádět a vysvětlovat, ale nezkoušet je.
- Používat software ve výuce. Usnadňuje pohled na data, umožňuje rozumět možnostem statistiky.
- Úlohy, které se tématicky vztahují k oboru studia, motivují a pomáhají povzbudit zájem o statistiku.
- Podpořit pochopení některých konceptů statistiky pomocí simulačních úloh (interaktivních na webu, nebo vytvářených studenty).
- Podporovat a umožnit studentům zpětnou vazbu s vyučujícím, dbát na interakci se studenty během výuky, dávat jim odezvu o správnosti uvažování při řešení zadaných úkolů a problémů.

Poděkování: Tato práce vznikla s částečnou podporou grantu REFIMAT CZ.1.072/2.2.2.00/15.0016.

Literatura

[1] Anděl J. (2010) *Statistika a počítače, studenti a učitelé.* Informační Bulletin České statistické společnosti **22**, 8 – 16.

[2] Antoch J., Dohnal G., Malý M., Eds. (2000) *STAKAN I – II.* Sborník prací semináře STAKAN (15. – 17. 9. 1999). Česká statistická společnost, Praha, 100 s.

[3] Antoch J., Štěpán J., Eds. (2002) *Výuka statistiky v České republice I.* Sborník prací semináře v Praze (22. 11. 2002). Matfyzpress, Praha, 114 s.

[4] Antoch J., Dohnal G., Štěpán J., Eds. (2004) *Výuka statistiky v České republice II.* Sborník prací semináře STAKAN (23. – 25. 5. 2003). Česká statistická společnost, Praha, 132 s.

[5] Brown E. N., Kass R. E. (2009) *What is statistics?* The American Statistician **63**, 105 – 110.

[6] Efron, B. (2007). *The future of statistics.* Zdroj dostupný na WWW: `http://www-stat.stanford.edu/~brad/talks/future.pdf`, cit. 29. 10. 2011.

[7] Hebák P. (2007) *Učíme statistiku.* Informační Bulletin České statistické společnosti **18**, 6 – 24.

[8] Hindls R., Hronová S. (2005) *Jak výuka odrazuje nestatistiky od statistiky.* Statistika **42**, 168 – 172.

[9] Hsu, M. K., et al. (2009) *Computer attitude, statistics anxiety, and self-efficacy on statistical software adoption behavior: An empiricical study of online MBA learners.* Computers in Human Behavior **25**, 412 – 420.

[10] Koucký J. (2009) *Kolik máme vysokoškoláků?* Aula **17**, 5 – 18.

[11] Kvaszová M. (2009) *Proč nám nerozumějí.* Informační Bulletin České statistické společnosti **20**, 10 – 18.

[12] Lacasse C., Chiocchio F. (2005) *Anxiety towards statistics: Further developments and issues.* 66th ACPA, Montreal. Zdroj dostupný na: `http://www.mapageweb.umontreal.ca/chiocchf/pub/999046_lacasse_chiocchio_handout.pdf/`, cit. 1. 11. 2011.

[13] Moore, D.S. (2001) *Undergraduate Programs and the Future of Academic Statistics.* The American Statistician **55**, 1 – 6.

[14] Onwuegbudzie, A. J., et al. (2000) *Factors associated with achievement in educational research courses.* Research in Schools **7**, 53 – 65.

[15] Perepiczka M., Chandler N., Becerra M. (2011) *Relationship between graduate students' statistics self-efficacy, statistics anxiety, attitude towards statistics, and social support.* Research and Practice **1**, 99 – 108.

[16] Wilkinson L. (2008) *The future of statistical computing.* Technometrics **50**, 418 – 435.

MEDICI, LÉKAŘI A STATISTIKA

PHYSICIANS, STUDENTS OF MEDICINE, AND STATISTICS

Josef Tvrdík

Adresa: Ostravská univerzita, Přírodovědecká fakulta,
Katedra informatiky, 30. dubna 22, 701 03 Ostrava

E-mail: josef.tvrdik@osu.cz

Poděkování: Tento příspěvek byl podporován Interní grantovou agenturou OU z projektu SGS/13/PřF/2012.

Abstrakt: Článek se zabývá spoluprací statistika s lékařem na aplikacích statistiky v medicínském výzkumu, a to jak z pohledu praktického, tak i etického. Stručně je popsán také obsah a metody výkladu kurzu o základech statistiky pro studenty prvního ročníku Lékařské fakulty.

Klíčová slova: Aplikace statistiky, klinická data, spolupráce s lékaři, výuka statistiky pro mediky.

Abstract: Cooperation of statistician and physician is addressed both from the pragmatic and the ethic view. The content of the statistical course for medical students and the way of teaching are also briefly described.

Keywords: Applied Statistics, Clinical Data, Cooperation with Physicians, Statistical Course for Students of Medicine.

1. Úvod

Tento příspěvek je stručným záznamem a dodatečným komentářem ke sdělení přednesenému na konferenci Stakan 2011 na přelomu září a října v Železné Rudě. Tam bylo anotováno jako „krátké sdělení o dlouhých zkušenostech ze spolupráce s lékaři ve statistickém zpracování dat a krátkých zkušenostech z vyučování statistiky v prvním ročníku Lékařské fakulty Ostravské univerzity“.

Sdělení vyvolalo překvapivý ohlas u publika, s diskusí pokračující i po ukončení sekce. Pár kolegů (nebylo jich mnoho) se na mne obrátilo s přáním, abych obsah sdělení sepsal, což jsem původně ani neměl v úmyslu. Ale hlas lidu, hlas boží, navíc Bulletin je časopis, kde za článek jsou autorovi a tedy i jeho pracovišti přidělovány nyní tak žádané body ovlivňující bytí a nebytí katedry, fakulty i univerzity, tak se i z důvodů pragmatických pokouším o písemné sdělení.

K aplikacím statistiky a spolupráci s lékaři jsem se dostal víceméně shodou náhod. Od začátku osmdesátých let minulého století jsem byl zaměstnán ve výpočetní laboratoři Krajské hygienické stanice v Ostravě a zabýval jsem se implementací laboratorních informačních systémů na mini- a mikropočítače. Protože nikoho specializovaného na statistické zpracování dat tam neměli, spadly aplikace statistiky převážně na mne a statistika pak zabírala zhruba třetinu mé pracovní náplně. K tomu se postupně přidávala spolupráce s doktory i z jiných zařízení a od té doby pokračuje navzdory času a mému přechodu do školství. Spolupracující lékaři se během těch více jak třiceti let obměňovali a obměňují, někteří bohužel nevratně, ale spolupráce na biostatistických aplikacích prozatím přetrvává.

2. Role lékaře a role statistika

Lékař z kliniky či obyčejné nemocnice, který se kromě své lékařské práce pokouší i o výzkum a potřebuje statisticky zpracovat svá data, se dostává do pro něj naprosto nezvyklé role. Je vlastně v pozici pacienta, který potřebuje pomoci řešit svůj sice ne zdravotní, ale výzkumný problém. Naléhavost jeho problému je často srovnatelná s naléhavostí zdravotního problému pacienta. Na výsledek spěchá (konference se blíží, téma je žhavé a ve světě na něm pracují jiné týmy, takže datum odeslání článku je důležité atd.). Navíc statistické znalosti lékařů jsou zhruba srovnatelné s našimi pacientskými znalostmi medicíny: víme, že nás něco bolí, a způsob léčení si představujeme jednoduše – rychle a bez našeho většího úsilí. Lékař je dokonce v ještě obtížnější situaci než nemocný pacient v tom, že statistika pro spolupráci hledá obtížněji. Nepošle ho k nám žádný „obvoďák“ a ani „ordinační hodiny“ statistika nejsou nikde na internetu.

Naopak, statistik je v pozici „poskytovatele péče“, má tedy roli lékaře a může se začít rozpomínat na to, jak tuto roli hráli lékaři, k nimž on jako pacient přišel. Prosím, zapomeňme na případnou špatnou zkušenost vlastní nebo známou z doslechu, kdy doktor nemá čas něco srozumitelně vysvětlovat nebo se chová s nepříjemnou povýšeností pána nad naším organismem a způsobem jeho léčení. Chovejme se tak, jak bychom si přáli, aby se k nám choval lékař, když k němu přicházíme jako pacienti. Zamysleme se raději nad tím, co lékař přicházející se svými daty očekává a potřebuje:

- Především potřebuje dobře porozumět datům, se kterými přichází. To ostatně nutně potřebujeme i my, pokud je máme statisticky zpracovat. A není to vůbec samozřejmá věc ani pro jednu zúčastněnou stranu. Obtíž může činit např. rozpoznat to, co jsou párová měření a co jsou nezávislé skupiny, nebo někdy lze z veličin v datech získat odvozenou

veličinu, která je pro daného lékaře velmi užitečná, a on neví, že hodnoty této veličiny je možné vhodnou transformací dat získat.

- Pochopit výsledky statistické analýzy. Počítačový výstup obvykle není tím podkladem, který je dostatečný pro pochopení výsledků lékařem a jejich publikaci.
- Poradit s přehlednou prezentací statistických výsledků. Lékařské časopisy mívají zažité způsoby takové prezentace, např. některé preferují intervaly spolehlivosti, jiné úroveň významnosti p dosaženou v testech, někde chtějí směrodatnou odchylku, jinde střední chybu průměru atd.

O co bychom se měli snažit při statistickém zpracování dat:

- neuškodit,
- být trpěliví a důkladní při zjišťování „anamnézy“,
- užívat srozumitelný jazyk,
- nenechat klienta bezradného,
- poslat ke specialistovi, když je potřeba.

Nejčetnějšími statistickými úlohami, které řeší „venkovský statistický obvoďák“, jsou problémy s dvěma výběry a párová porovnání, vše za různých okolností, tj. spojité veličiny, proporce atd. Občas se analyzují kontingenční tabulky, většinou jen dvourozměrné, někdy se využije analýza rozptylu (výjimečně i repeated measures), někdy korelace a regrese (poměrně často i logistická) a sem tam i analýza přežití. Z uvedeného výčtu je zřejmé, že tyto metody pokrývá každý běžný statistický programový systém, např. NCSS [2] užívaný u nás. Rozhodně to nejsou žádné výzvy pro objevování nových statistických obzorů a cesta k ocenění ve vědecké komunitě. Ale to není nic překvapivého, vždyť celá statistika byla vymyšlena pro to, aby byla aplikována, tj. aby sloužila pro jiné obory. Už dlouho jsem přesvědčen, že aplikace statistiky není žádná věda, ale spíše řemeslo nebo snad inženýrství. A základní pracovní inženýrskou metodou je kompromis, hlavní zásadou „lepší je nepřítelem dobrého“ a vyřešení inženýrského problému nebo zhotovení řemeslného výrobku je požadováno v zadaném termínu. To všechno se hodí při statistickém zpracování dat.

Naskýtá se přirozená otázka, proč by se statistik měl věnovat něčemu tak neatraktivnímu jako je spolupráce s lékaři. Odpověď je jednoduchá: je to užitečné a bez statistika to prozatím kvalitně nejde. Profesor Komenda říkával, že při aplikacích by statistik měl „vzít rozum do hrsti a mít oči na štopkách“, a to je právě to, co je na aplikacích statistiky zajímavé a často i zábavné. Nelze očekávat, že by se statistik na občasných aplikacích proslavil

nebo zbohatl. Pokud je výsledek publikován, tak se statistikovo jméno ztrácí v dlouhém seznamu ostatních autorů a časopis se týká oboru, který téměř žádný statistik nemá důvod číst. Finanční odměna za občasnou statistickou aplikaci je krajně nespolehlivý zdroj příjmů. Pokud vůbec je výzkum součástí nějakého grantu, navrhovatel na potřebu statistické analýzy často zapomene a neplánuje tedy žádné peníze na statistika. Někdy se stane, že jako odměna ke statistikovi doputuje láhev původně patrně vložená do zdravotnictví vděčným pacientem. Spolupracuje-li statistik s nějakým lékařem delší dobu, může se stát jeho pacientem požívajícím výhod většího výběru termínu návštěvy u lékaře a při návštěvě ordinace se nehovoří jen o jeho zdravotních potížích, ale i o statistickém zpracování dat, což překryje zbytečné prožívání zdravotních potíží a působí příznivě na pacientovu psychiku. Největší odměnou za aplikace statistiky je dvojí pocit užitečnosti, a to užitečnosti oboru aplikace a užitečnosti statistice, která byla pro aplikace stvořena, máme ji rádi (jinak bychom ji tak dlouho a vytrvale nestudovali), a která nás také jakž takž živí.

3. Výuka statistiky pro mediky

K výuce statistiky pro studenty prvního ročníku medicíny jsem přišel také shodou náhod. Po dlouhém úsilí vedení Ostravské univerzity a významných regionálních veličin byla v roce 2010 akreditována výuka na Lékařské fakultě. V přípravě akreditace jsem se trochu podílel na návrhu obsahu předmětu Lékařská biofyzika a informatika I. Ten mají studenti v zimním semestru prvního ročníku a jeho součástí je i šest přednášek a osm cvičení věnovaných statistice. Při přípravě předmětu jsem se domníval, že učit ho bude někdo jiný. Podmínkou akreditační komise však bylo, že přednášejícími ve všech předmětech mohou být pouze docenti a vyšší hodnosti, a tak prst náhody ukázal na mne.

Hned v roce 2010 byla do prvního ročníku přijata stovka studentů a já stál před dobrodružstvím, jak připravit přednášky o základech statistiky, které by mohly být pro studenty srozumitelné a od statistiky je neodradily. Aby jim pomohly nejen k získání zápočtu z tohoto předmětu, ale aby studentům zůstala i nějaká představa o možnostech a obsahu statistiky na delší dobu, navzdory tomu, že během dlouhých šesti let studia medicíny bude statistika mocně překryta spoustou jiných předmětů, které budou muset absolvovat.

Obsah každého základního kurzu statistiky je víceméně jasný, ve stručnosti ho lze vymezit pojmy: statistická data, popisná statistika, pravděpodobnost, náhodná veličina, rozdělení, náhodný výběr, odhady (zejména pochopení intervalů spolehlivosti), testování hypotéz, korelace a regrese. To, o čem jsem přemýšlel, byla forma, jak učit. Malý rozsah hodin přednášek a rozměry

posluchárny (dlouhá a poměrně úzká místnost s tabulí překrývanou promítacím plátnem) vylučovaly můj oblíbený způsob přednášení s křídou a tabulí. Nezbylo než připravit promítané prezentace [3], které by studenty základními pojmy provedly a přitom udržely jejich zájem. Pomáhal jsem si dost neobvyklými prostředky. Např. výklad o vzniku a struktuře dat začíná příběhem o zjišťováním data příchodu Járy Cimrmana do Liptákova [1, str. 11–12], kde lze ukázat nepřesnosti měření a jakýsi náznak potřeby intervalů spolehlivosti. Metody popisné statistiky jsou vysvětlovány na simulovaných datech o příjmu pacientů do nemocnice. Pak už bohužel legrace trochu ubývá, ale výklad je stále zaměřen na porozumění základním myšlenkám a pojmům než na dril přesných definic a přísných postupů. Závěr poslední přednášky připomíná, že je nemožné za těch pár hodin se naučit statistiku. Důležité je pochopit její základní myšlenky a možnosti, umět formulovat svůj problém a obrátit se na statistika, pokud úloha vyžaduje netriviální statistické dovednosti.

Po dvouletých zkušenostech si netroufám tvrdit, že zvolený přístup je úspěšný. Kromě četnosti zápočtů udělených na cvičení je zatím jediným kritériem účast na přednáškách odhadnutá pohledem do posluchárny. Účast studentů má sice v průběhu semestru sestupný trend, ale nejde k nule, i na poslední přednášku přišla vždy zhruba třetina všech studentů. Jaké budou statistické znalosti dnešních studentů za pár let, až statistiku budou potřebovat, se teprve uvidí.

4. Závěr

Text se zabývá spoluprací statistika a lékaře při aplikacích statistiky v lékařském výzkumu a výukou statistiky pro studenty Lékařské fakulty. Nepřináší žádné převratné poznatky, svou formou se pohybuje někde mezi stručnou zprávou a úvahami založenými na subjektivních zkušenostech, nikoliv na tvrdých datech. Nezbývá než doufat, že přispěje k zamyšlení a výhledově snad i ke zvýšení intenzity a kvality spolupráce statistiků na aplikacích statistiky v lékařském výzkumu.

Reference

[1] Cimrman, J.; Smoljak, L.; Svěrák, Z. (1992) *Posel z Liptákova,* Paseka.

[2] Hintze J. (2001) *NCSS and PASS, Number Cruncher Statistical System*, Kaysville, Utah, `http://www.ncss.com/`

[3] Tvrdík J. (2011) *Výuka – studijní materiály,* `http://www1.osu.cz/~tvrdik/down/vyuka.html`

REDUCTION OF TOTAL COST OF A COMPANY USING OPTIMALISATION METHOD

Alena Kolčavová

Address: Mgr. Alena Kolčavová, Ph.D., Tomas Bata University in Zlin, Faculty of Management and Economics, Mostní 5139, 76001 Zlín

E-mail: kolcavova@fame.utb.cz

Abstract: The article deals with utility of optimalisation methods in Czech companies. It analysis the present situation and uncovers reasons that prevent introduction of these methods in practice. On the basis of a financial analysis advantageousness of these methods is proved in a small company dealing with goods distribution.

Keywords: Optimalisation Methods, Route 66, WinQSB, Operations Research, Network Analysis, Transportation Problems.

1. Reason for the Topic Selection and the Present State of the Problems Solved

I cannot omit to mention at least in brief how operational reserach was used in the past, because this can show the best drawbacks and causes of today's state. I am going to concentrate on the domestic environment in particular.

History of the operational research dates back to 1930s and 1940s and it is connected with the names of G. B. Dantzig and L. Kantorowitz, the Nobel Prize winners for economics. At that time socialism was the ruling system in the Eastern Block, i.e. the buying and selling prices of all raw materials and services were identical and long-term planning of production offered itself. That time was ideal for utilization of linear methods of programming.

A rapid development of this discipline during the 2nd World War was caused by needs of the military industry. Especially methods of project management and analysis were developed. Special teams analyzing complex strategical and tactical military operations were made up in the USA and Great Britain.

In the post-war period, when there was significant development of computer science, the application of quantitative methods in decision-making increased due to easier input data processing.

The year 1989 brought, among others, an aversion to everything that was planned and organized. No wonder, because the whole national economy had been managed centrally, by means of 5-year plans, time schedules and plans of development were practically on each noticeboard. At that time, when

everything was given in a form of directives, when prices were uniform, there was practically no competition, no variety, and no possibility of choice.

Our generation became a witness of a historically unique period, when all the existing orders and hierarchy were broken. The things considered unthinkable before became a reality, all of us were at the same starting point, had the same opportunities, possibilities and conditions. Political membership or connections were no more decisive, only the spirit of enterprise and readiness to risk were important. It was a period of euphoria, when those who had had practically nothing before could move to the head with their enterprise and vice versa.

At those dramatic times full of changes there was no time, thought of, and either a need to deal with long-term planning. Everything developed in motion and intuitively, the imperfect legal system only stood by.

The year 2000 brought about not only the beginning of a new millenium but also a change of situation in our economy. The wild waters calmed down, the confused situation settled. Competition started to wake up and it was not so easy any more to win recognition on the market as it used to be in the past.

Now there is coming a time when each company is forced to leave the intuitive approach to its management. And just this is the sphere of action of the quantitative methods in decision making, which could help companies to optimize their production, manage resources, time and costs. There is a wide range of possibilities how to use these methods in different spheres of enterprise.

After the Czech Republic joins the European Union the competitive pressures on domestic companies will even increase, and unfamiliarity with possibilities of the company optimum management can be fatal for many companies.

Most managers are aware of the changed situation, but in spite of this, project planning and management are not used even in the cases where they offer themselves. The question is, what is the reason? Possibly unwillingness to leave the running style of management, and then it is a question of time when the company will not be able to keep pace with competition any more. Another reason can consist in unfamiliarity with possibilities provided by quantitative methods in decision-making. In this case my work could be an inspiration to companies.

Operations research deals with coordination of many professions – from a foreman at the plant, who is able to describe individual steps of the manufacture, including the necessary quantitative data. Through a designer, who can phase individual steps of the situation under change, a programmer, who

designs the corresponding software, up to the manager, who can decide the range of the planned changes on the basis of mathematical model outputs. Each link of the chain is of the same importance, its absence results in distortion of the real situation and results of the model do not have the correct informative value.

An essential part of the project management is also time analysis of any project. It is surprising that it is not utilised even in those spheres in which its applicability just offers itself – namely in civil engineering.

If we follow every order as a project, such a project can be followed from several points of view – time, cost and sources. Chaos that exists in most civil engineering companies is caused by non-existence of coordination of work of individual profession groups of workers. And it is just the time analysis of the project that offers introduction of clarity, keeping link of individual operations and time schedule of the whole project. The whole project can be checked from the point of view of cost and a uniform spread-over of sources even in more projects at the same time.

And what knowledge should have the person elaborating a proposal of a change? This is of course a key problem – absence of specialists who would be able to introduce optimalisation methods in practice. Such a person should be creative, should be able to apply theoretical knowledge in practice, should self-educate and follow novelties in his branch.

Without invention at work, required quality is not guaranteed. For this reason theoretical knowledge is not decisive, but it is the real interest in being the best in his branch. This is valid in each profession.

There are many spheres that could prove the hypothesis that using a suitable optimalisation method the total cost of the company can be cut.

2. Case Study – Reduction of Total Cost of a Company Using Optimalisation Method

For an illustration I have chosen an XY company which produces and distributes baked goods in its region. I concentrated on a possibility to reduce transport costs by means of optimalisation of the existing distribution routes. The company realises 10 routes, but not all of them are suitable for optimalisation. A detailed attention has been paid to route No. 9 and its optimalisation. To be able to specify the total cost saving of the transport route after its optimalisation, optimalisation of other routes, on which it is effective and possible, has to be done. Routes 1, 2 and 4 with the total designation Town are excluded from the group of routes. At these routes the cost saving cannot

Table 1: Comparison of the existing and optimum length of the route No. 9

	Distribution 1	Distribution 2	Total
Present Situation	—	—	100.00 km
Additional information: According to documents of the transport section			
Little Method	55.60 km	36.00 km	**91.60 km**
Additional information: According to solution using programme WinQSB			

be calculated because distances between individual customers of the baked goods are not available.

Optimalisation can be done by means of Little method of branches and borders, which is a demanding method from the point of view of calculation. Thanks to today's development of information technologies a specialised software WinQSB can be used, module Network Modeling, version 1.00 for Windows.

Distances between individual customers can be specified accurately by means of the programme Route 66 (see the maps on the next page).

2.1. Financial Evaluation of Optimalisation of the Route No. 9

Detailed calculation of optimalisation carried out on the Route No. 9 showed a saving of 8.4 km per day, which means a drop by 2,134 km per year.

On the basis of the data on the total number of kilometres on the existing and the designed route and the calculated cost per 1 km of the drive, financial evaluation for the given time period (day, week, month, year) can be made. The calculation will be based on the data from the planning calendar, where one week = 5 working days, one month = 21 working days, and one year = 254 working days.

Costs on the Route No. 9 per 1 day dropped by 129.61 CZK from 1,543 CZK to 1,413.39 CZK, which represents a cost reduction by 8.4 %. The cost saving following from optimalisation of the Route No. 9 makes 32,921.45 CZK per year.

2.2. Comparison of Results of Optimalisation and the Present Distribution

Table 3 shows distances on the present routes and on their optimised forms (km). Most of the routes consists of two distributions, only the route No. 7

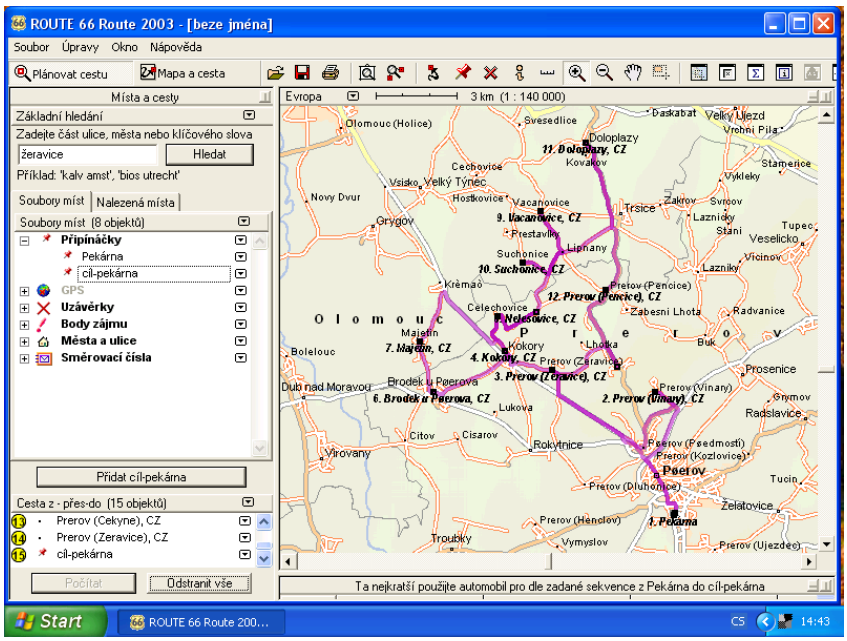


Figure 1: Present Route. Distance: 74 km, time: 2 h 35 min, fuel: 26.3 litres. Přerov – Bakery, Vinary, Žeravice, Kokory, Nelešovice, Brodek at Přerov, Majetín, Čelechovice, Vacanovice, Suchonice, Doloplazy, Penčice, Čekyně, Žeravice, Přerov – Bakery.

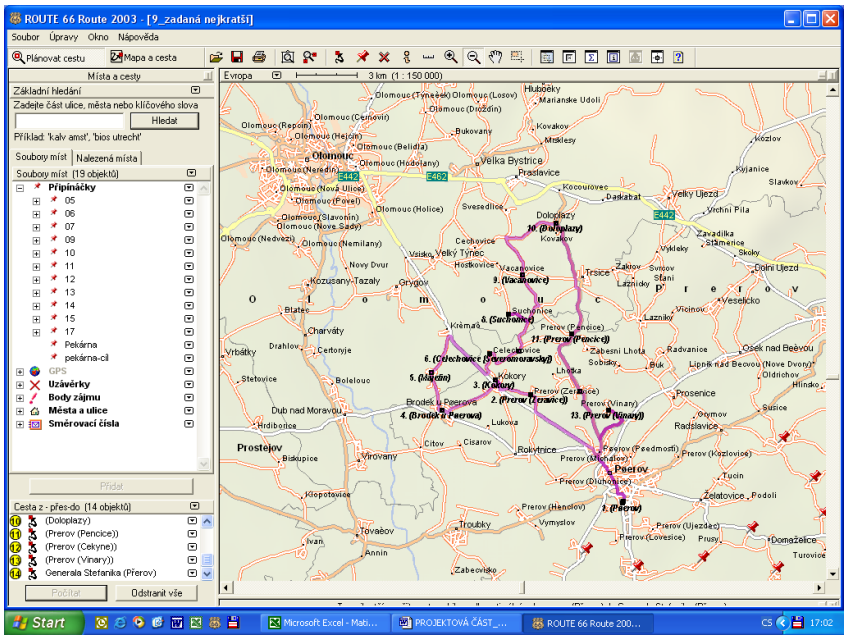


Figure 2: Designed Optimum Route, distance: 55.6 km, time: 1 h 58 min, fuel: 20.04 litres. Přerov – Bakery, Žeravice, Kokory, Brodek, Majetín, Čelechovice, Nelešovice, Suchonice, Vacanovice, Doloplazy, Penčice, Čekyně, Vinary, Přerov – Bakery.

Table 2: Specification of cost on Route No. 9 today and after the designed adaptation and calculation of differences in costs.

	Situation		Cost Saving
	Present	Designed	
Cost per 1 km of drive	15.43 CZK		
Number of driven kms per 1 day	100 km	91.6 km	−8.4 %
Cost per 1 day	1 543.00 CZK	1 413.39 CZK	129.61 CZK
Cost per 1 week (5 working days)	7 715.00 CZK	7 066.94 CZK	648.06 CZK
Cost per 1 month (Avg. 21 working days)	32 403.00 CZK	29 681.15 CZK	2 721.85 CZK
Cost per 1 years (254 working days)	391 922.00 CZK	359 000.55 CZK	32 921.45 CZK

has one distribution, on the other hand the route No. 10 has 3 distributions. Drive in the locality is not excluded from the length of routes, as it was made in the previous part on the route No. 9. The saving in driven kilometres in the last column is a difference between the present and the designed situation.

The total daily saving on all routes makes 282.1 km. This table of lengths of all routes will be used as a basis for financial evaluation of the present situation, designed situation and calculation of savings in the CZK.

As it is evident from Table 4, the total saving of 282.1 km on all routes brought a financial saving of 4,352.8 CZK per day, which represents a drop by 18.3 % in the transport costs. This daily saving is not negligible, especially if it is recalculated to one week, one month, and one year.

3. Conclusion

In the beginning of the case study a hypothesis was made that using an optimalisation method cost of transport of products can be reduced together with the total cost of the company. A detailed analysis defined the present situation of distribution routes as well as cost of the transport route. The present situation was compared with the designed situation (we have arrived at it using optimalisation method). The conclusions are the following:

- Detailed calculation of optimalisation made on the Route No. 9 showed a saving of 8.4 km per day, which means a drop by 2,134 km yearly.

Table 3: Comparison of lengths of individual routes at the present distribution and after optimalisation (km) per 1 working day.

** After optimalisation the route is longer in the number of kilometres compared to the present route, because connecting roads of lower class, which in fact the vehicle driver uses on the Route No. 15, cannot be included in the calculation.

Route		Present	Distribution			Lisation	Saving
No.	Name	route (km)	1st	2nd	3rd	(km)	(km)
5	Kojetín	125.0	53.9	38.3	–	92.2	32.8
6	Lipník	100.0	40.5	31.6	–	72.1	27.9
7	Val. Meziříčí	205.0	160.4	–	–	160.4	44.6
9	Brodek u Př.	110.0	55.6	36.0	–	91.6	18.4
10	Moštěnice	141.0	34.5	51.0	8.9	94.4	46.6
11	Tovačov	91.0	38.3	34.4	–	72.7	18.3
12	Troubky	142.0	74.3	62.9	–	137.2	4.8
13	Buk	154.0	94.4	47.6	–	142.0	12.0
14	Pavlovice	137.0	48.6	49.2	–	97.8	39.2
15	Bystřice p/H.	156.0	109.4	55.5	–	164.9	**
17	Hranice	180.0	88.5	54.0	–	142.5	37.5
Total		1541.0	798.4	460.5	8.9	1267.8	282.1

Table 4: Specification of cost of all routes today and after the designed adaptation and calculation of differences in cost for different periods.

	Situation (in CZK)		Cost
	Present	Designed	Saving
Cost per 1 km of drive	15.43		
Total number of driven kms per 1 day	1541 km	1258.9 km	−18.3 %
Cost per 1 day	23 777.63	19 424.83	4 352.80
Cost per 1 week (5 working days)	118 888.15	97 124.14	21 764.02
Cost per 1 month (Avg. 21 working days)	499 330.23	407 921.37	91 408.86
Cost per 1 years (254 working days)	6 039 518.02	4 933 906.06	1 105 611.96

- The cost on Route No. 9 per 1 day dropped by 129.61 CZK from 1,543 CZK to 1,413.39 CZK – which means a cost reduction by 8.4 %. The cost saving following from optimalisation of the Route No. 9 makes 32,921.45 CZK yearly.
- The total daily saving on all routes after optimalisation is 282.1 km.
- The saving of 282.1 km together on all routes brought a financial saving of 4,352.8 CZK per day, which means a drop by 18.3 % in costs.
- Optimalisation of all routes brings a financial saving in the amount of 1,105,611.96 CZK per year, which means cost reduction by 18.3 %.

These results have confirmed the hypothesis on cost saving.

Printed Sources

[1] Jablonský, J. *Operations Research.* 2nd edition, University of Economics, Prague 1998. ISBN 80-7079–597-2.

[2] Jablonský, J. *Operations Research – Quantitative Models for Economic Decision-making.* 1st edition. Prague: Professional Publishing, 2002. 323 pp. ISBN 80-86419-23-1.

[3] Kolčavová, A. *Vybrané optimalizační metody a jejich využitelnost v praxi (in Czech).* Zlín. CEED. 2011. 116 p. ISBN 978-80-87301-04-3.

[4] Lawrence, J., Pasternack, B. *Applied Management Science.* New York: John Wiley, 1998. 665 pp. ISBN 0-471-13776-6.

[5] Malovaná, E. *Projekt optimalizace distribučních cest firmy XY, s. r. o. (in Czech).* Diplomová práce, UTB FaME, Zlín, 2004. 82 p.

Internet Sources

[6] Berka, M. *Eulerovy a Hamiltonovy cykly (in Czech).* [online] [cit. 2012-02-16] Available on: `http://www.berkovi.cz/milan/berka/o/zaklady.htm`

[7] Berka, M. *Metoda větví a hranic – Algoritmus Littla (in Czech).* [online] [cit. 2012-02-16] Available on: `http://www.berkovi.cz/milan/berka/o/grafy.htm`

Software Products

[8] *Program Route 66.* For Microsoft Windows, Windows 98, Windows NT, Windows ME, Windows 2000, Windows XP. Made by the company and Data Solutions B.V. Version 3.3.0 Copyright 2002, the Route 66 logo and Route 66 are registered trademarks, Engine version 3.3.0; Copyright 1993–2002 Route 66 [cit. 2012-05-11].

[9] *WinQSB – Network Modeling.* Version 1.00. Copyright Yih-Long Chang. [cit. 2012-04-11].

ANALÝZA VZTAHŮ ORDINÁLNÍCH PROMĚNNÝCH APLIKOVANÁ NA ÚROVNĚ KOMPETENCÍ ABSOLVENTŮ VYSOKÝCH ŠKOL

ANALYSIS OF ORDINAL VARIABLE RELATIONSHIPS APPLIED TO COMPETENCE LEVELS OF GRADUATES

Hana Řezanková, Renáta Kunstová

Adresa: Vysoká škola ekonomická v Praze,
nám. W. Churchilla 4, 130 67 Praha 3

E-mail: hana.rezankova@vse.cz, renata.kunstova@vse.cz

Poděkování: Práce na tomto článku byla podpořena granty Grantové agentury České republiky P202/10/0262, P403/10/0092 a výzkumným záměrem MSM6138439910.

Abstract: In the paper we investigate influence of a number of ordinal variables categories on values of dependence, agreement, similarity and sameness coefficients. These different types of relationships are represented by well-known coefficients tau-b, kappa and cosine measure and by a newly proposed competence coefficient. This measure assigns a greater weight to better evaluations. Suitability of coefficients is illustrated by the analysis of competence levels evaluated by graduates in the REFLEX 2006 and REFLEX 2010 surveys. Graduates evaluated a level which was achieved by them and a level required by an employer. Calculations were performed both for original scales (seven levels in the survey from 2006 and ten levels in the survey from 2010) and for the recoded three-level scale. From comparison of results obtained by individual coefficients it followed that a newly proposed competence coefficient is robust in relation to a change of category numbers when it was used for competence ordering.
Keywords: Graduates, Competence Levels, Ordinal Variables, Analysis of Relationships, Dependence Measures, Agreement Measures, Similarity Measures, Sameness Measures.

Abstrakt: V příspěvku zkoumáme vliv počtu kategorií ordinálních proměnných na hodnoty koeficientů závislosti, souhlasu, podobnosti a shody. Tyto různé typy vztahů jsou reprezentovány jak dobře známými koeficienty tau-b, kappa a kosinovou mírou, tak nově navrženým kompetenčním koeficientem, který přiřazuje větší váhu lepším hodnocením. Vhodnost koeficientů je ilustrována na analýze úrovní kompetencí hodnocených absolventy v šetřeních

REFLEX 2006 a REFLEX 2010. Absolventi hodnotili úroveň, kterou u jednotlivých kompetencí dosáhli, a úroveň požadovanou zaměstnavatelem. Výpočty byly provedeny jednak pro původní škály úrovní (sedmiúrovňovou v šetření z roku 2006 a desetiúrovňovou v šetření z roku 2010), a jednak pro překódovanou tříúrovňovou škálu. Z porovnání výsledků získaných podle jednotlivých koeficientů vyplynulo, že nově navržený kompetenční koeficient je při využití pro uspořádání kompetencí podle jeho hodnot robustní vůči změně počtu kategorií.
Klíčová slova: Absolventi vysokých škol, úrovně kompetencí, ordinální proměnné, analýza vztahů, míry závislosti, míry souhlasu, míry podobnosti, míry shody.

1. Úvod

S ordinálním typem proměnných se často setkáváme při vyhodnocení dat z dotazníkových šetření, která jsou realizována za účelem sociologických, marketingových či některých dalších podobně koncipovaných výzkumů. Respondenti jsou dotazováni na názory a hodnocení (např. výrobků či služeb), přičemž nabízené odpovědi jsou na ordinální (pořadové) škále uspořádány buď od negativního po pozitivní hodnocení (resp. stanovisko), či naopak. Jinými příklady ordinálních proměnných jsou odpovědi, kdy respondent zařazuje kvantitativní hodnotu do některého ze stanovených intervalů, nejčastěji pokud ji odhaduje (např. průměrné měsíční výdaje domácnosti na potraviny, nebo částku, kterou by byl ochoten zaplatit za určitý výrobek či službu).

V tomto článku se zaměříme na proměnné vyjadřující hodnocení či uspořádaná stanoviska (stupně souhlasu či nesouhlasu s určitými výroky). Tato hodnocení mohou být vyjádřena pomocí různého počtu kategorií, v praxi jsou využívány škály od třístupňových po desetistupňové (obvykle z lichých třístupňové, pětistupňové a sedmistupňové, ze sudých čtyřstupňové a desetistupňové). Častým typem analýzy je zkoumání závislosti, případně jiných vztahů dvou proměnných. Protože u ordinálních proměnných má smysl pořadí hodnot, můžeme zkoumat, jak se se zvyšujícími se hodnotami jedné proměnné mění hodnoty druhé proměnné (zda se také zvyšují, nebo snižují, případně zůstávají přibližně stejné), můžeme tedy zkoumat korelaci mezi pořadími.

Motivací pro zkoumání prezentované v dalším textu bylo použití různých počtů úrovní sledovaných kompetencí při opakovaném výzkumu absolventů vysokých škol. Za předpokladu, že by ve dvou obdobích byly sledovány stejné kompetence, pak se nabízí otázka, zda má různý počet úrovní vliv na hodnoty koeficientů vyjadřujících vztahy mezi dvěma proměnnými.

Problematiku budeme ilustrovat na datových souborech pořízených v rámci projektů REFLEX 2006 (mezinárodní šetření, do kterého byla zapojena i Česká republika) a REFLEX 2010 (šetření realizované v ČR), viz [5] a [6]. Tyto soubory obsahují odpovědi absolventů vysokých škol z let 2001 a 2002 (první šetření) a 2005 a 2006 (druhé šetření). Do analýz pro účely tohoto článku byly zahrnuty odpovědi absolventů magisterského studia vybraných fakult ekonomického zaměření (každá vysoká škola, resp. fakulta, která se do šetření zapojila, má k dispozici údaje získané od svých absolventů; způsoby výběrů absolventů se na jednotlivých fakultách mohly lišit, návratnost kromě ochoty zapojit se do šetření závisela na tom, zda se podařilo dotazník absolventovi doručit na platnou adresu).

V obou výzkumech absolventi pro každou sledovanou kompetenci (dovednost, schopnost či znalost) uváděli jednak úroveň, kterou dosáhli, a jednak úroveň požadovanou zaměstnavatelem. Kromě toho, že můžeme sledovat vztahy mezi různými kompetencemi, můžeme tedy pro jednotlivé kompetence zkoumat vztah úrovně dosažené a úrovně požadované zaměstnavatelem (párové hodnoty z pohledu absolventa). Dotazníky v obou šetřeních byly podobně sestavené, avšak například v případě hodnocení úrovní kompetencí se lišil počet úrovní. V šetření v roce 2006 to bylo sedm úrovní, zatímco v roce 2010 jich bylo deset (pro porovnatelnost však bylo mnohem více na závadu, že v obou obdobích byly sledovány jiné kompetence – i když byly některé podobné obsahem, měly jiný název, což mohlo ovlivnit hodnocení jejich úrovně z hlediska různého subjektivního chápání obsahu dané problematiky jednotlivými absolventy).

2. Způsoby hodnocení vztahů mezi ordinálními proměnnými a proměnnými se stejným počtem kategorií

Jak již bylo zmíněno v úvodu, jednou z možností, jak hodnotit vztahy mezi dvěma ordinálními proměnnými, je *korelační analýza*. Pro proměnné vyjadřující pořadí lze využít buď koeficient Spearmanův[1] nebo Kendallův (Kendallovo tau-b)[2]. Kendallův korelační koeficient je založen na *počtu konkordantních, diskordantních a vázaných párů* – kolik existuje párů respondentů, v nichž jeden respondent hodnotí oba ukazatele vyšší nebo nižší úrovní než druhý, kolik existuje párů respondentů hodnotících ukazatele rozdílně, tj. jeden vyšší a druhý nižší úrovní, a u kolika párů respondentů je některý

[1] Spearman, C. The proof and measurement of association between two things. *Amer. J. Psychol.* **15,** 1904: 72–101.

[2] Kendall, M. A new measure of rank correlation. *Biometrika* **30** (1–2), 1938: 81–89. doi:10.1093/biomet/30.1-2.81.

z ukazatelů hodnocen stejnou úrovní. Uvedené počty párů jsou též základem pro výpočty některých jiných koeficientů, jako jsou Kendallovo tau-c[3], gama[4,5,6,7] či Somersovo d[8] – vzájemnou závislost hodnotí symetrická varianta Somersova d počítaná jako harmonický průměr z asymetrických variant; její hodnoty jsou jen málo odlišné od hodnot Kendallova tau-b, které je geometrickým průměrem asymetrických variant.

Pro hodnocení vztahu mezi ordinálními proměnnými je možné použít také míry pro nominální proměnné, ovšem pro jiné typy vztahů než je závislost. Protože proměnné nabývají v rámci jednoho šetření stejných kategorií, můžeme zkoumat *stupeň souhlasu*, vztahující se k výskytu stejných kategorií. K tomuto účelu slouží Cohenovo kappa[9] – porovnává četnosti na diagonále kontingenční tabulky s teoretickými četnostmi v případě nezávislosti, které jsou základem pro chí-kvadrát test.

Při aplikacích některých vícerozměrných metod (shlukové analýzy či vícerozměrného škálování) je potřeba vyjít z matice vzdáleností či podobností. Vzhledem k omezeným možnostem některých programových systémů z hlediska speciálních měr pro ordinální proměnné je v praxi používána *míra podobnosti* pro kvantitativní proměnné, kterou je například kosinová míra (bývá implementován též Pearsonův korelační koeficient, který je však mírou závislosti již zastoupenou koeficientem tau-b). Níže navíc odvodíme speciální míru pro ordinální proměnné.

V dalším textu se budeme zabývat především vzájemným vztahem dvou proměnných. Zkoumání omezíme pouze na některé koeficienty, konkrétně Kendallovo tau-b jako míru vzájemné závislosti pro ordinální proměnné (doporučovanou pro proměnné se stejným počtem kategorií), Cohenovo kappa jako míru souhlasu pro proměnné se stejnými kategoriemi a kosinovou míru

[3]Kendall, M. *Rank Correlation Methods.* Charles Griffin & Company Limited, 1948.

[4]Goodman, L. A., Kruskal, W. H. Measures of association for cross classifications. *Journal of the American Statistical Association* **49** (268), 1954: 732–764.

[5]Goodman, L. A., Kruskal, W. H. Measures of association for cross classifications, II: Further discussion and references. *Journal of the American Statistical Association* **54** (285), 1959: 123–163.

[6]Goodman, L. A., Kruskal, W. H. Measures of association for cross classifications, III: Approximate sampling theory. *Journal of the American Statistical Association* **58** (302), 1963: 310–364.

[7]Goodman, L. A., Kruskal, W. H. Measures of association for cross classifications, IV: Simplification of asymptotic variances. *Journal of the American Statistical Association* **67** (338), 1972: 415–421.

[8]Somers, R. H. A new asymmetric measure of association for ordinal variables. *American Sociological Review* **27**, 1962: 799–811.

[9]Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20** (1), 1960: 37–46. doi:10.1177/001316446002000104.

podobnosti. Podrobněji o hodnocení vztahů mezi proměnnými viz např. [1], [3] a [4]. Pokud vyjdeme z četností v kontingenční tabulce, pak je známo, že když se ve čtvercových tabulkách nenulové sdružené četnosti kombinací kategorií nacházejí pouze na diagonále (všichni respondenti by udávali hodnotu dosažené úrovně stejnou jako požadované úrovně), všechny tři uvedené koeficienty nabývají hodnoty jedna. Do ilustračního srovnání bude navíc zahrnuto Kendallovo tau-c, které hodnoty 1 nabývá pouze v případě shodných četností na diagonále (tento koeficient je doporučován v případě, kdy tabulky nejsou čtvercové, tedy pokud proměnné nabývají různého počtu kategorií).

V tabulce 1 jsou pro ilustraci uvedeny hodnoty zmíněných koeficientů pro některá vybraná dvourozměrná rozdělení četností pro proměnné obsahující odpovědi na třístupňové škále. Je v ní uvedeno devět takových případů, zaznamenaných ve sloupcích tabulky, přičemž celkový rozsah souboru je vždy stejný (90 jednotek – respondentů). V prvních šesti případech je rozdělení četností pro jednotlivé proměnné rovnoměrné, tj. každá proměnná pro každou ze svých kategorií nabývá četnosti 30.

V prvním z uvedených případů jsou sdružené četnosti rovnoměrně rozloženy. Bez ohledu na kategorie jedné z proměnných nabývá druhá proměnná stejných rozdělení četností. To je situace odpovídající nezávislosti proměnných jak v případě nominálních, tak v případě ordinálních proměnných. Hodnoty koeficientů *tau-b*, *tau-c* i *kappa* jsou proto nuly.

Z hlediska zhodnocení souladu dosažených a požadovaných úrovní kompetencí nás spíše zajímá, jak jsou si úrovně podobné, případně zda dosažené úrovně jsou alespoň takové, jako požadované. V prvním případě můžeme použít výše zmíněnou *kosinovou míru*. Předpokládejme, že kategorie jsou označeny pořadovými čísly od jedničky. V tom případě je index řádku (sloupce) shodný s pořadovou hodnotou. Nechť R označuje počet kategorií řádkové proměnné v kontingenční tabulce, S označuje počet kategorií sloupcové proměnné (pro čtvercovou tabulku $R = S$), n_{ij} označuje sdruženou četnost, n_{i+} značí řádkovou marginální četnost a n_{+j} sloupcovou marginální četnost. Kosinová míra podobnosti je vyjádřená vztahem

$$\cos = \frac{\sum_{i=1}^{R}\sum_{j=1}^{S} n_{ij} ij}{\sqrt{\sum_{i=1}^{R} n_{i+} i^2 \sum_{j=1}^{S} n_{+j} j^2}}.$$

Při rovnoměrném rozložení sdružených četností se pro jednu třetinu statistických jednotek hodnoty obou proměnných shodují a pro ostatní statistické jednotky jsou si vždy určitým způsobem podobné. Při použití kosinové míry

pro první variantu rozdělení četností v tabulce 1 je celková podobnost dvou proměnných vyjádřena hodnotou 0,857.

Pokud by místo pořadí kategorií bylo uvažováno pořadí hodnot odvozené od rozsahu souboru, pak by míra podobnosti byla vyjádřena jako

$$\cos\,\mathrm{rank} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{S} n_{ij} r_i r_j}{\sqrt{\sum_{i=1}^{R} n_{i+} r_i^2 \sum_{j=1}^{S} n_{+j} r_j^2}},$$

kde r_i (r_j) je průměrné pořadí pro i-tou (j-tou) kategorii. Například pokud se kategorie 1, 2 a 3 vyskytují všechny 30krát, tak kategorie 1 má průměrné pořadí 15,5, kategorie 2 průměrné pořadí 45,5 a kategorie 3 průměrné pořadí

Tabulka 1: Hodnoty vybraných koeficientů závislosti, souhlasu, podobnosti a shody pro vybraná dvourozměrná rozdělení četností

(i,j)	$n_{ij}(1)$	$n_{ij}(2)$	$n_{ij}(3)$	$n_{ij}(4)$	$n_{ij}(5)$	$n_{ij}(6)$	$n_{ij}(7)$	$n_{ij}(8)$	$n_{ij}(9)$
(1,1)	10	30	0	0	0	0	28	50	10
(1,2)	10	0	0	15	30	0	28	0	0
(1,3)	10	0	30	15	0	30	28	0	0
(2,1)	10	0	0	15	30	30	1	0	0
(2,2)	10	30	30	0	0	0	1	30	30
(2,3)	10	0	0	15	0	0	1	0	0
(3,1)	10	0	30	15	0	0	1	0	0
(3,2)	10	0	0	15	0	30	1	0	0
(3,3)	10	30	0	0	30	0	1	10	50

Míra	$n_{ij}(1)$	$n_{ij}(2)$	$n_{ij}(3)$	$n_{ij}(4)$	$n_{ij}(5)$	$n_{ij}(6)$	$n_{ij}(7)$	$n_{ij}(8)$	$n_{ij}(9)$
tau-b	0	1	−1	−0,417	0,333	−0,333	0	1	1
tau-c	0	1	−1	−0,417	0,333	−0,333	0	0,852	0,852
kappa	0	1	0	−0,500	0	−0,500	0	1	1
cos	0,857	1	0,714	0,786	0,929	0,786	0,857	1	1
cos rank	0,775	1	0,551	0,663	0,888	0,663	0,855	1	1
K_{sym}	0,222	0,667	0,222	0	0,333	0	0,122	0,519	0,815
$K_{Y\|X}$	0,519	0,667	0,556	0,444	0,556	0,333	0,652	0,519	0,815

75,5. Pro první variantu rozdělení četností v tabulce 1 je celková podobnost dvou proměnných vyjádřena hodnotou 0,775.

Zaměříme-li se pouze na *shodu* hodnot, pak můžeme zjišťovat podíl četností na diagonále na celkovém počtu statistických jednotek. V případě, že vyšší hodnota znamená vyšší úroveň a naším cílem je získat vyšší ohodnocení shody u ukazatelů s větším podílem vyšších hodnot, můžeme tento podíl násobit aritmetickým průměrem úrovní, váženým sdruženými četnostmi na diagonále, a dělit maximální možnou hodnotou tohoto součinu. Tato nově navržená míra shody (dále *kompetenční koeficient*) je vyjádřena vzorcem

$$K_{\mathrm{sym}} = \frac{\frac{\sum_{i=1}^{R} n_{ii}}{n} \cdot \frac{\sum_{i=1}^{R} i \cdot n_{ii}}{\sum_{i=1}^{R} n_{ii}}}{R} = \frac{\sum_{i=1}^{R} i \cdot n_{ii}}{nR}.$$

Při rovnoměrném rozdělení sdružených četností v kontingenční tabulce pro výše uvedený příklad je hodnota tohoto koeficientu 0,222.

Pokud by nás zajímaly nejen četnosti shodných úrovní, ale také počty absolventů, kteří hodnotí svoji dosaženou úroveň výše, než je úroveň požadovaná, můžeme vzít jako váhy odpovídající četnosti. Bude-li dosažená úroveň sloupcovou proměnnou, pak lze její vztah k řádkové proměnné, vyjadřující požadovanou úroveň, ohodnotit pomocí asymetrické varianty kompetenčního koeficientu. Ten vypočteme podle vzorce

$$K_{Y|X} = \frac{\frac{\sum_{i=1}^{R}\sum_{j\geq i}^{S} n_{ij}}{n} \cdot \frac{\sum_{i=1}^{R}\sum_{j\geq i}^{S} j \cdot n_{ij}}{\sum_{i=1}^{R}\sum_{j\geq i}^{S} n_{ij}}}{R} = \frac{\sum_{i=1}^{R}\sum_{j\geq i}^{S} j \cdot n_{ij}}{nR}.$$

Pro první uvedenou variantu sdružených četností v tabulce 1 je hodnota této varianty kompetenčního koeficientu rovna 0,519.

Případ s výskytem pouze shodných kategorií u obou proměnných, a tudíž výskytem nenulových četností pouze na hlavní diagonále, je v tabulce 1 označen jako varianta (2), případ s výskytem nenulových četností pouze na vedlejší diagonále je označen jako varianta (3). Ve variantě (4) jsou nenulové četnosti ve všech políčkách mimo diagonálu. Následují dvě varianty se všemi marginálními četnostmi 30 a pouze třemi nenulovými políčky. Poslední tři varianty jsou již s odlišnými marginálními četnostmi, jedna je příkladem nezávislosti a v posledních dvou jsou nenulové četnosti pouze na diagonále, ale četnosti se liší. Je to případ, kdy koeficient tau-c nenabývá hodnoty 1.

Pro každou variantu jsou uvedeny hodnoty vybraných koeficientů. Pro koeficienty tau-b a tau-c jsou uvedeny příklady na přímou lineární závislost,

zahrnující i maximální hodnotu koeficientů (hodnotu 1), nepřímou lineární závislost, zahrnující i minimální hodnotu koeficientů (hodnotu –1), a na lineární nezávislost, indikovanou hodnotou 0. Koeficient kappa nabývá své maximální hodnoty 1 ve stejném případě jako koeficient tau-b, tj. při výskytu nenulových sdružených četností pouze na hlavní diagonále. Pro případy, kdy jsou naopak na diagonále pouze nuly, je výsledkem záporná hodnota. Hodnota 0 je výsledkem tehdy, pokud se součet zjištěných sdružených četností na diagonále rovná součtu četností očekávaných v případě nezávislosti. V daných případech je to jednak lineární nezávislost, jednak úplná nepřímá lineární závislost, a dále slabá lineární závislost.

Kosinová míra nabývá své maximální hodnoty 1 ve stejných situacích jako je tomu u koeficientů tau-b a kappa. Při porovnání aplikací této míry na pořadí kategorií a pořadí hodnot je nejnižší hodnota v obou případech dosažena pro stejné rozložení sdružených četností. Nově navržený kompetenční koeficient v základní (symetrické) variantě nabývá hodnoty 0, pokud se na diagonále vyskytují pouze nuly. V ostatních případech nabývá kladných hodnot, nejvyšší hodnoty v případě nejvyšší četnosti pro shodu v nejvyšší úrovni.

3. Porovnání hodnocení kompetencí s využitím vybraných koeficientů a různých počtů úrovní

Předpokládejme, že cílem analýzy odpovědí respondentů je získat pořadí kompetencí z hlediska ohodnocení vztahu dosažených a požadovaných úrovní. V této části se zaměříme na to, zda na získaná pořadí má vliv různý počet úrovní. Vyjdeme přitom z analýzy datových souborů pořízených v rámci zmíněných projektů REFLEX 2006 a REFLEX 2010. V prvním případě jsme měly k dispozici data od 363 absolventů magisterského studia, kteří sdělili své názory týkající se úrovní všech nebo některých ze sledovaných kompetencí, ve druhém to bylo 592 absolventů. Protože někteří respondenti se k některým kompetencím nevyjádřili, do jednotlivých výpočtů je z prvního šetření zahrnuto 360–363 dvojic odpovědí hodnotících dosažené a požadované úrovně 19 kompetencí a z druhého šetření to bylo 591–592 dvojic pro dosažené a požadované úrovně 24 kompetencí.

Zaměřily jsme se pouze na koeficienty hodnotící vzájemný vztah, které mohou být použity i pro vytvoření matice vztahů (pro dosažené nebo pro požadované úrovně) potřebné pro některé metody vícerozměrné analýzy (vícerozměrné škálování a shlukovou analýzu). Vzájemnou závislost dvou proměnných jsme hodnotily pomocí Kendallova tau-b (doporučovaného pro čtvercové tabulky), souhlas pomocí Cohenova kappa, podobnost pomocí kosinové

míry pro pořadí kategorií a shodu pomocí symetrického kompetenčního koeficientu (dále jen kompetenční koeficient).

Pro obě období jsme zahrnuly jak původní škálu (7 úrovní pro šetření z roku 2006 a 10 úrovní pro šetření z roku 2010), tak překódovanou tříúrovňovou škálu. V obou skupinách jsme u každého koeficientu a u každé varianty počtu úrovní vybraly tři kompetence s nejvyšším ohodnocením vztahu. V tabulce 2 jsou pro lepší přehlednost uvedeny pouze kompetence, které se podle některého z aplikovaných koeficientů umístily od prvního do třetího místa. Pro rok 2006 to je pět kompetencí a pro rok 2010 je to šest kompetencí.

Zatímco kompetence umístěné na prvních třech pozicích z hlediska závislosti (vyjádřené pomocí hodnot koeficientu tau-b) dosažených a požadovaných úrovní se pro různý počet úrovní v obou obdobích liší, z hlediska hodnot kosinové míry a kompetenčního koeficientu se na prvních třech místech umístily v obou obdobích tři stejné kompetence. Tyto tři kompetence jsou stejné u obou měr, až na to, že u kosinové míry je v případě sedmi úrovní v šetření z roku 2006 hodnota na třetím místě shodná u dvou kompetencí.

Další „nevýhodou“ měr závislosti je to, že pomocí výsledné hodnoty nelze rozpoznat, zda jsou více zastoupeny nižší nebo vyšší úrovně. Na základě dat z šetření z roku 2010 byly získány nejvyšší hodnoty koeficientu tau-b (při obou variantách počtu úrovní) u *schopnosti použít základní výzkumné postupy svého oboru.* Tato kompetence je charakterizována největším podílem nejnižších úrovní, a to jak pokud jde o dosažené, tak i o požadované úrovně (aritmetické průměry úrovní obou typů jsou nejnižší ze všech kompetencí).

Na rozdíl od této vlastnosti měr závislosti pro ordinální proměnné jsou na tom míry pro hodnocení jiných typů vztahů lépe. Podle koeficientu kappa je sice zmíněná kompetence mezi prvními třemi, ale až na třetím místě. Hodnota kosinové míry je u této kompetence druhá nejnižší a hodnota kompetenčního koeficientu zcela nejnižší (v obou případech bez ohledu na počet úrovní).

Některé poznatky, ke kterým jsme dospěly, sice nelze zobecnit, nicméně naznačují výhody a nevýhody jednotlivých koeficientů, které zastupují různé způsoby hodnocení vztahů dvou ordinálních proměnných se stejným počtem kategorií. Hodnoty kosinové míry v 86 sledovaných případech (43 kompetencí v obou šetřeních pro dvě varianty počtu úrovní) byly z intervalu od 0,935 do 0,996 (obě mezní hodnoty byly získány pro tři úrovně z roku 2006). Nejnižší hodnota byla zjištěna u kompetence s nejnižší průměrnou dosaženou úrovní. Vzhledem k tomuto úzkému intervalu byly pro několik kompetencí získány stejné hodnoty.

Kompetenční koeficient je přímo navržen, aby „zvýhodňoval“ kompetence s vyšším podílem vyšších shodných hodnot. U něj se tedy očekává, že bude nabývat vyšších hodnot u kompetencí s vyšším hodnocením. Vzhledem k tomu,

Tabulka 2: Hodnocení vztahu dosažených a požadovaných úrovní vybraných kompetencí z šetření v roce 2006 (horní tabulka) a 2010 (dolní tabulka) – postupně koeficienty závislosti, souhlasu, podobnosti a shody

Koeficienty, počet úrovní	**tau-b**		**kappa**		**kosinová míra**		**kompetenční**	
Název kompetence	3 ú.	7 ú.	3 ú.	7 ú.	3 ú.	7 ú.	3 ú.	7 ú.
Schopnost rychle si osvojit nové znalosti	0,338	**0,474**	0,291	0,291	**0,986**	**0,984**	**0,848**	**0,450**
Schopnost koordinovat činnosti	0,452	**0,539**	**0,372**	**0,319**	0,982	**0,982**	0,811	0,421
Schopnost mobilizovat pracovní kapacity druhých	**0,465**	0,460	0,347	0,164	0,950	0,956	0,545	0,222
Schopnost používat PC a internet	**0,573**	0,427	**0,507**	**0,359**	**0,996**	**0,989**	**0,934**	**0,579**
Schopnost připravovat písemné podklady, zprávy	**0,485**	**0,540**	**0,473**	**0,369**	**0,984**	**0,982**	**0,848**	**0,472**

Koeficienty, počet úrovní	**tau-b**		**kappa**		**kosinová míra**		**kompetenční**	
Název kompetence	3 ú.	10 ú.	3 ú.	10 ú.	3 ú.	10 ú.	3 ú.	10 ú.
Znalost podmínek pro využití odborných metod a teorií v praxi	**0,639**	0,615	**0,599**	0,439	0,970	0,977	0,564	0,320
Schopnost použít základní výzkumné postupy svého oboru	**0,664**	**0,663**	**0,556**	**0,455**	0,955	0,962	0,436	0,255
Dovednost pracovat s informacemi	0,507	0,615	0,456	**0,472**	**0,989**	**0,989**	**0,807**	**0,453**
Dovednost identifikovat a řešit problémy	0,501	**0,621**	0,456	0,437	**0,989**	**0,989**	**0,803**	**0,435**
Schopnost přizpůsobit se změněným okolnostem, podmínkám	**0,612**	0,610	**0,595**	**0,455**	**0,988**	**0,988**	**0,805**	**0,429**
Schopnost pracovat v interkulturním / mezinárodním prostředí	0,595	**0,620**	0,435	0,386	0,964	0,962	0,616	0,361

že větší podíl shodných úrovní se vyskytuje v případě menšího počtu úrovní, obor hodnot koeficientu se liší pro různé počty úrovní. Pro deset úrovní byly vypočteny hodnoty z intervalu od 0,231 do 0,453, pro sedm úrovní z intervalu od 0,182 do 0,579 a pro tři úrovně z intervalu od 0,436 do 0,934.

Z dalších poznatků zjištěných na základě dostupných dat uvádíme, že pro nižší počet úrovní byly ve většině případů zjištěny také vyšší hodnoty koeficientu kappa (17 hodnot z 19 a 23 z 24) a nižší hodnoty koeficientu tau-b (13 hodnot z 19 a 14 z 24).

4. Závěr

Při analýze úrovní kompetencí hodnocených absolventy různých fakult podobného zaměření jsme komentovaly vlastnosti vybraných koeficientů určených k hodnocení vztahů dvou proměnných. Posuzovaly jsme známé koeficienty pro hodnocení vzájemné závislosti, souhlasu a podobnosti a nově navržený koeficient pro hodnocení shody. Zaměřily jsme se na vztahy úrovní kompetencí dosažených absolventy a úrovní požadovaných zaměstnavateli, přičemž oba typy úrovní hodnotili sami absolventi. Protože v šetřeních ze dvou období byly použity různé počty úrovní, výpočty byly provedeny jednak pro původní sedmi či desetiúrovňovou škálu, jednak pro překódovanou tříúrovňovou škálu. Pro obě šetření dohromady tak bylo získáno 86 hodnot pro každý sledovaný koeficient.

Při zkoumání vztahu proměnných zmíněného charakteru je důležité zaměřit se na podobnost proměnných. Speciální míry podobnosti byly navrženy a jsou v praxi používány pro nominální a kvantitativní proměnné. V případě ordinálních proměnných jsou obvykle jejich kategorie označeny pořadovými čísly a pak se postupuje jako v případě kvantitativních dat. Pro ohodnocení podobnosti jsme použily kosinovou míru.

Jednou ze sledovaných vlastností koeficientů byl vliv změny počtu úrovní na pořadí kompetencí z hlediska hodnot těchto koeficientů. V posuzovaných případech u kosinové míry a nově navrženého kompetenčního koeficientu došlo ke změně pořadí jen minimálně (obvykle šlo o sousední pořadí a velmi blízké hodnoty koeficientů). Také pořadí získané pomocí kosinové míry a kompetenčního koeficientu bylo velmi podobné (viz předchozí komentář). Zjištěná robustnost těchto koeficientů vůči počtu úrovní je výhodou vůči koeficientům, které buď přímo měří intenzitu závislosti, nebo posuzují sdružené četnosti vzhledem k četnostem očekávaným v případě nezávislosti.

Výhodou kompetenčního koeficientu je dále to, že přiřazuje větší váhu vyšším úrovním. Na rozdíl od kosinové míry jsou jeho hodnoty různorodější, což umožňuje jednotlivé kompetence více odlišit. Tento koeficient by mohl

být využíván pro přípravu matice podobností – vstupní matice pro některé metody vícerozměrné analýzy, jako je shluková analýza a vícerozměrné škálování.

Literatura

[1] Pecáková, I. *Statistika v terénních průzkumech.* 2. vyd. Praha: Professional Publishing, 2011.

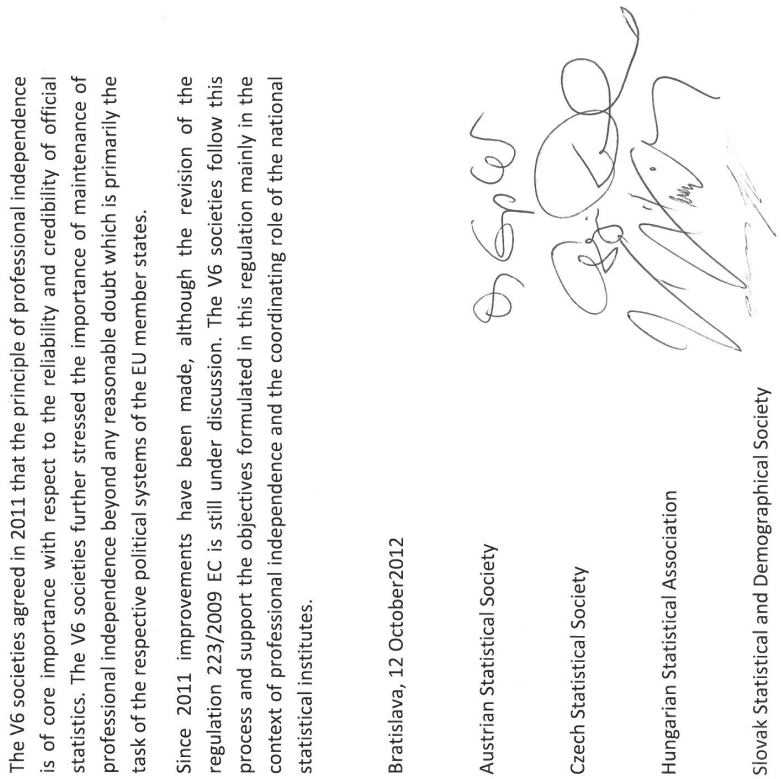
[2] Řehák, J.; Řeháková, B. *Analýza kategorizovaných dat v sociologii.* Praha: Academia, 1986.

[3] Řezanková, H. *Analýza dat z dotazníkových šetření.* 3. dopl. vyd. Praha: Professional Publishing, 2011.

[4] Řezanková, H.; Húsek, D.; Snášel, V. *Shluková analýza dat.* 2. rozšíř. vyd. Praha: Professional Publishing, 2009.

[5] Středisko vzdělávací politiky UK. *Mezinárodní projekt Reflex.* [online]. Praha: SVP, Univerzita Karlova. [cit. 2011-11-19].
`www.strediskovzdelavacipolitiky.info/default.asp?page=reflex`

[6] Středisko vzdělávací politiky UK. *Reflex 2010.* [online]. Praha: SVP, Univerzita Karlova. [cit. 2011-11-19]. Dostupné na webové stránce:
`www.strediskovzdelavacipolitiky.info/default.asp?page=svp&KID=85`



Joint statement of the V6 societies 2012

The V6 societies agreed in 2011 that the principle of professional independence is of core importance with respect to the reliability and credibility of official statistics. The V6 societies further stressed the importance of maintenance of professional independence beyond any reasonable doubt which is primarily the task of the respective political systems of the EU member states.

Since 2011 improvements have been made, although the revision of the regulation 223/2009 EC is still under discussion. The V6 societies follow this process and support the objectives formulated in this regulation mainly in the context of professional independence and the coordinating role of the national statistical institutes.

Bratislava, 12 October2012

Austrian Statistical Society

Czech Statistical Society

Hungarian Statistical Association

Slovak Statistical and Demographical Society

Na setkání představitelů národních statistických společností bylo zformulováno a společně podepsáno zmíněné prohlášení (více viz následující článek).

STRETNUTIE ŠTATISTICKÝCH SPOLOČNOSTÍ V BRATISLAVE

Peter Mach

12. 10. 2012 sa v Bratislave uskutočnilo stretnutie predstaviteľov štatistických spoločností stredoeurópskeho regiónu. Štatistické spoločnosti z regiónu sa pravidelne raz ročne stretávajú, aby sa navzájom informovali o svojej činnosti a diskutovali o otázkach spoločného záujmu. Bolo to už ôsme stretnutie spoločností, ktoré sa druhý raz koná v Bratislave. Na stretnutí sa zúčastnili zástupcovia štatistických spoločností z Rakúska, Česka, Maďarska a Slovenska. Zástupca Slovinskej štatistickej spoločnosti sa pripojil k časti rokovania prostredníctvom telekonferencie, predstavitelia Rumunskej spoločnosti sa z vážnych pracovných dôvodov ospravedlnili.

Stretnutie otvoril predseda Slovenskej štatistickej a demografickej spoločnosti Jozef Chajdiak, ktorý privítal účastníkov a poďakoval sa predstaviteľom ŠÚ SR za vytvorenie podmienok pre uskutočnenie stretnutia.

Na úvod stretnutia predstavitelia Štatistického úradu SR prezentovali slovenský štatistický systém a projekt Elektronické služby Štatistického úradu SR, ktorý je súčasťou Operačného programu Informatizácia spoločnosti. Predsedníčka ŠÚ SR Ľudmila Benkovičová vo svojom príhovore poukázala na aktuálne problémy oficiálnej štatistiky: *„Tým kľúčovým, je na jednej strane neustály rast požiadaviek na kvalitu a rozsah štatistických výstupov, na strane druhej neustály pokles disponibilných finančných zdrojov zo štátneho rozpočtu“* – konštatovala predsedníčka ŠÚ SR a dodala, že riešenie treba hľadať v racionalizácii a zvyšovaní efektívnosti štatistického systému, ako aj v nových spôsoboch zberu štatistických údajov, ktoré sú finančne najnáročnejšou fázou procesov štatistického zisťovania.

Zástupcovia jednotlivých spoločností informovali o činnosti svojich spoločností. V rámci tohto bodu informovali zástupcovia Maďarskej štatistickej spoločnosti, že na polovicu novembra pripravujú slávnostnú konferenciu pri príležitosti 90. výročia založenia ich spoločnosti. Zástupca Slovinskej štatistickej spoločnosti informoval o činnosti svojej spoločnosti prostredníctvom telekonferencie. Súčasne ponúkol usporiadanie budúcoročného stretnutia. Táto ponuka bola s potešením prijatá.

V rámci diskusie o otázkach spoločného záujmu sa hovorilo najmä o aktuálnych otázkach európskeho štatistického systému. Už na predchádzajúcom stretnutí bola veľká pozornosť venovaná otázke profesionálnej nezávislosti oficiálnej štatistiky. Zúčastnené spoločnosti potvrdili, že táto otázka je naďalej

dôležitá najmä v súvislosti s pripravovanou revíziou európskej štatistickej legislatívy a podporujú princíp profesionálnej nezávislosti a koordinačnú úlohu národných štatistických úradov v národných štatistických systémoch.

Účastníci stretnutia dostali tiež krátky leták o činnosti Federácie európskych národných štatistických spoločností (FENStatS). Viac informácií o jej činnosti je na `http://www.fenstats.eu/`.

Popoludní prijala účastníkov stretnutia v reprezentačných priestoroch Primaciálneho paláca prvá námestníčka primátora hlavného mesta SR Ing. Viera Kimerlingová.

Fotografie zo stretnutia si môžete pozrieť na:
`http://www.facebook.com/media/set/?set=a.488901401134892.118294.100000451094271&type=1&l=86c234499b` a na:
`http://www.bratislava.sk/vismo/galerie2.asp?id_org=700000&id_galerie=5010159`

Peter Mach
podpredseda SŠDS pre medzinárodné styky

Zľava: Branislav Bleha (SK), Lörinc Soós (HU), Ján Luha (SK), Gejza Dohnal (CZ), Peter Mach (SK), Éva Laczka (HU), Karol Pastor (SK), Hana Řezanková (CZ), Jozef Chajdiak (SK), Joachim Lamel (AT), Margit Epler (AT), Ľudmila Ivančíková (SK).

Obsah

Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo.

Časopis je zařazen do seznamu Rady pro výzkum, vývoj a inovace, více viz server `http://www.vyzkum.cz/`.

Předseda společnosti: prof. RNDr. Gejza Dohnal, CSc.
ÚTM FS ČVUT v Praze, Karlovo náměstí 13, 121 35 Praha 2
E-mail: `gejza.dohnal@fs.cvut.cz`

Redakční rada: prof. Ing. Václav Čermák, DrSc. (předseda), prof. RNDr. Jaromír Antoch, CSc., doc. Ing. Josef Tvrdík, CSc., RNDr. Marek Malý, CSc., doc. RNDr. Jiří Michálek, CSc., doc. RNDr. Zdeněk Karpíšek, CSc., prof. Ing. Jiří Militký, CSc., prof. RNDr. Gejza Dohnal, CSc.

Technický redaktor: Ing. Pavel Stříž, Ph.D., `pavel@striz.cz`
Informace pro autory jsou na stránkách `http://www.statspol.cz/`

DOI: 10.5300/IB, http://dx.doi.org/10.5300/IB
ISSN 1210–8022 (Print), ISSN 1804–8617 (Online)