SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

2/2014
ročník/volume 24

Recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov.

Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 1

Typ článku/Type of article: vedecký článok/scientific article

Strany/Pages: 3 – 17

Dátum vydania/Publication date: 15. apríl 2014/April 15, 2014

Eva KOTLEBOVÁ
Katedra štatistiky Fakulty hospodárskej informatiky, Ekonomická univerzita v Bratislave

Ivan LÁSKA
TREXIMA Bratislava, s. r. o.

VYUŽITIE BAYESOVSKÉHO PRÍSTUPU PRI ODHADE PODIELU A MOŽNOSTI JEHO APLIKÁCIE V EKONOMICKEJ PRAXI

UTILIZATION OF THE BAYESIAN APPROACH IN POPULATION PROPORTIONS´ ESTIMATION AND POSSIBILITIES OF ITS APPLICATION IN BUSINESS PRACTICE

ABSTRAKT
Príspevok sa zaoberá bodovým odhadom parametra π binomického rozdelenia. V porovnaní so štandardným prístupom, v ktorom sa na jeho odhad používa výberový podiel, bayesovský prístup zohľadňuje aj iné dostupné informácie o skúmanom probléme. Na bodový odhad podielu sa v bayesovskej štatistike využíva konjugovaný systém binomické/beta.
V príspevku sa ako kritérium kvality bodového odhadu používa stredná kvadratická chyba. Ukázali sme, že sa prostredníctvom vhodnej voľby parametrov apriórneho rozdelenia vždy dá nájsť taký interval, na ktorom je stredná kvadratická chyba bayesovského bodového odhadu menšia ako stredná kvadratická chyba klasického bodového odhadu. Navrhli sme algoritmus na stanovenie bayesovského odhadu na vopred stanovenom intervale.

ABSTRACT
The article deals with the point estimation of binomial proportion. In comparison with the classical approach, bayesian approach takes into account another source of information about the problem. In bayesian statistics, the conjugated family binomial/beta is used for the proportion´s estimation.
In the article, the mean square error was used as the estimation´s criterion of quality. We proved, that if prior parameters´ values are appropriately determined, there exists interval, within which the bayesian point estimation has smaller mean square error than the classical one. We designed algorithm for evaluating the bayesian point estimation.

KĽÚČOVÉ SLOVÁ
bodový odhad, bayesovský prístup, apriórne rozdelenie, konjugovaný systém, stredná kvadratická chyba

KEY WORDS
point estimation, bayesian approach, prior distribution, conjugate family, mean square error

1. ÚVOD
V ekonomickej praxi je mimoriadne dôležité vedieť čo najpresnejšie predvídať kvantitatívny rozmer ekonomických javov, aby sme vedeli stanoviť optimálnu

stratégiu pre akúkoľvek činnosť. Odhad podielu štatistických jednotiek s určitou vlastnosťou nie je len abstraktnou matematickou kategóriou, ale má aj mnohé konkrétne ekonomické podoby. Napríklad podiel nekvalitných výrobkov z produkcie vyrobenej na novej výrobnej linke môže byť rozhodujúcim kritériom na posúdenie vhodnosti nákupu takejto výrobnej linky, ale aj na stanovenie dodatočných záruk pre spotrebiteľa nad štandardný rámec vyplývajúci zo zákona. Alebo podiel nespokojných zákazníkov z tých, ktorí si kúpili konkrétny produkt, môže byť podkladom na zmenu parametrov produktu, prípadne servisu spojeného s kúpou produktu. Uvedené príklady majú isté spoločné črty: sledovaný podiel sa nedá presne vypočítať (okrem iného preto, že by sme museli započítať aj budúcich klientov alebo výrobky, ktoré sa ešte len budú vyrábať), môžeme ho len odhadnúť na základe údajov, ktoré získame z reprezentatívnej vzorky – výberového súboru. Pomocou metód štatistickej indukcie sa dá sledovaný podiel pomerne jednoducho odhadnúť jedným číslom – bodom alebo intervalom s vopred zvolenou spoľahlivosťou. Je zrejmé, že aj keď ide o (relatívne malé) číslo z intervalu (0;1), aj nepatrná nepresnosť v jeho odhade môže mať výrazný vplyv na výsledný ekonomický efekt. Počet výrobkov sa môže počítať na státisíce, počet zákazníkov v desiatkach tisíc, takže nepresnosť pri odhade podielu čo len o jednu stotinu môže absolútne počty posunúť o tisícky alebo stovky. A to ešte nebol započítaný finančný efekt jednotlivých javov. Presný odhad podielu je teda kľúčovou otázkou pri riešení rôznych ekonomických úloh a je potrebné venovať mu adekvátnu pozornosť. Náš príspevok sa zaoberá najmä teoretickým aspektom riešenia nastoleného problému, pričom z našich zistení rezultuje návrh konkrétneho algoritmu na stanovenie kvalitného bodového odhadu uvedeného parametra.

2. VLASTNOSTI BODOVÉHO ODHADU A ICH VZÁJOMNÁ SÚVISLOSŤ

Bodovým odhadom neznámeho parametra Θ rozumieme takú výberovú charakteristiku U_n, ktorá má isté požadované vlastnosti. Okrem neskreslenosti, konzistentnosti a výdatnosti, ktoré sa považujú za „štandardné“ a sú uvedené v každej učebnici štatistiky, sa pri podrobnejších analýzach zvyčajne požadujú ešte aj postačujúcosť a robustnosť [10]. My sme sa zaoberali prvou a treťou z týchto vlastností, pretože sú najdôležitejšie a istým spôsobom prepojené, pričom táto súvislosť v praxi nie vždy zaručuje pozitívny efekt. Uvedieme najskôr definície týchto vlastností.

Výberovú charakteristiku U_n považujeme za neskreslený odhad parametra Θ, ak platí: $E(U_n) = \Theta$ (stredná hodnota výberovej štatistiky sa rovná odhadovanému parametru).

Výberovú charakteristiku U_n považujeme za výdatný odhad parametra Θ, ak má spomedzi všetkých neskreslených odhadov najmenší rozptyl.

Neskreslenosť je teda nevyhnutnou podmienkou pre výdatnosť. To znamená, že ak je bodový odhad čo len nepatrne skreslený, automaticky nemôže byť ani výdatným odhadom. Takéto prepojenie dvoch uvedených vlastností však môže viesť k uprednostneniu charakteristiky s neúmerne veľkou variabilitou pred charakteristikou, ktorá je síce nepatrne skreslená, ale vzhľadom na malú variabilitu môžu všetky jej hodnoty byť bližšie k skutočnej hodnote odhadovaného parametra ako u preferovanej charakteristiky.

Na elimináciu možnosti nastania opísanej situácie je rozumné vziať do úvahy obidve uvažované vlastnosti a za optimálny bodový odhad považovať taký, ktorý nie je priveľmi skreslený a súčasne nemá priveľkú variabilitu. Takto voľne formulovanú požiadavku presnejšie vyjadruje stredná kvadratická chyba odhadu (mean square error), ktorá je súčtom rozptylu a druhej mocniny skreslenia uvažovaného odhadu:

$$MSE(U_n) = E\left[(\Theta - U_n)^2\right] = D(U_n) + \Delta_n^2 . \quad (1)$$

V uvedenom vzorci $D(U_n)$ označuje rozptyl charakteristiky, ktorá je bodovým odhadom, a $\Delta_n = E(U_n) - \Theta$ označuje skreslenie odhadu. Za lepší odhad považujeme ten, ktorý má nižšiu kvadratickú chybu.

Naším cieľom je teda nájsť taký bayesovský bodový odhad parametra π, ktorého stredná kvadratická chyba je menšia ako stredná kvadratická chyba klasického bodového odhadu.

3. PRINCÍP BAYESOVSKEJ ŠTATISTIKY

Bayesovská štatistika predstavuje alternatívny prístup k riešeniu problémov štatistickej indukcie. V porovnaní s klasickým prístupom berie do úvahy popri údajoch z výberového súboru aj ďalší druh informácie, ktorá pochádza z iných zdrojov (z predchádzajúcich zisťovaní, podobných prieskumov, dlhodobých skúseností erudovaných odborníkov atď.). Názov *apriórna informácia* súvisí s tým, že ju máme k dispozícii spravidla ešte pred samotným výberovým zisťovaním.

Z matematického hľadiska je bayesovská štatistika o niečo komplikovanejšia ako klasická induktívna štatistika. Odhadovaný parameter sa totiž považuje za náhodnú premennú, čo si vyžaduje o niečo vyššiu mieru abstraktného myslenia ako pri klasickom prístupe, v ktorom sa odhadovaný parameter považuje za jedno číslo. Rozdelenie uvažovanej náhodnej premennej sa aktualizuje vplyvom údajov z výberového súboru.

Apriórnu predstavu o odhadovanom parametri modelujeme pomocou apriórneho rozdelenia. Zapracovanie údajov z výberového zisťovania vedie k tzv. aposteriórnemu rozdeleniu, ktoré je podkladom na induktívne závery.

Bayesovská induktívna štatistika vychádza z Bayesovej vety o podmienenej pravdepodobnosti. Popri diskrétnej verzii sa často uvádza spojitá verzia:

$$f_\Theta(\theta \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \theta) \cdot f_\Theta(\theta)}{\int_\Theta f(\mathbf{x} \mid \theta) \cdot f_\Theta(\theta)\, \mathbf{d}\theta}, \quad (2)$$

v ktorej

$f_\Theta(\theta)$ označuje apriórnu hustotu odhadovaného parametra Θ,

$f_\Theta(\theta \mid \mathbf{x})$ označuje aposteriórnu hustotu parametra Θ,

$f(\mathbf{x} \mid \theta)$ je funkcia vierohodnosti.

Na odvodenie rôznych pravidiel a vzťahov v bayesovskej štatistike sa viac využíva zjednodušený tvar, v ktorom je vzťah rovnosti nahradený vzťahom proporcionality ($\propto$):

$$f_{\Theta}(\theta \mid \mathbf{x}) \propto f(\mathbf{x} \mid \theta) \cdot f_{\Theta}(\theta), \tag{3}$$

ktorý jednoduchšie a názornejšie ukazuje súvislosti medzi uvedenými rozdeleniami – v aposteriórnej hustote sú obsiahnuté obidve informácie, na základe ktorých bola vytvorená: apriórna informácia, reprezentovaná apriórnou hustotou, a údaje pochádzajúce z náhodného výberu, reprezentované funkciou vierohodnosti.

Ak je apriórne rozdelenie a aposteriórne rozdelenie rovnakého typu, hovoríme, že tvoria konjugované rozdelenie k rozdeleniu, z ktorého pochádza náhodný výber. Využitie konjugovaných systémov hustôt ([3], [10], [11]) umožňuje pomerne jednoduché stanovenie hodnôt parametrov aposteriórneho rozdelenia, pretože existujú odvodené vzorce, do ktorých stačí dosadiť parametre apriórneho rozdelenia a hodnoty niektorých výberových charakteristík.

Medzi najznámejšie a najčastejšie využívané konjugované systémy patria tieto (v názve je ako prvé uvedené rozdelenie, z ktorého pochádza náhodný výber, druhé je apriórne, resp. aposteriórne rozdelenie odhadovaného parametra): binomické/beta, Poissonovo/gama a normálne/normálne.

Bodovým odhadom parametra Θ je niektorá charakteristika aposteriórneho rozdelenia – najčastejšie je to jeho stredná hodnota, niekedy sa však na odhad používa medián, prípadne modus aposteriórneho rozdelenia (konkrétny výber charakteristiky závisí od tzv. stratovej funkcie [11], [14]).

V príspevku sme sa venovali prvému z uvedených konjugovaných systémov, pretože ten sa používa na odhad parametra π binomického (alternatívneho) rozdelenia.

4. BAYESOVSKÝ ODHAD PODIELU S VYUŽITÍM STREDNEJ KVADRATICKEJ CHYBY ODHADU

V ďalšom texte budeme využívať symboliku zodpovedajúcu konjugovanému systému binomické/beta – namiesto všeobecného označenia parametra Θ budeme používať symbol π.

Najskôr uvedieme základné poznatky o danom konjugovanom systéme (vzorce, ktoré platia pre uvedený konjugovaný systém, sú odvodené napríklad v [3], [7], [11]):

Ak výber pochádza z binomického rozdelenia ($X \approx Bi(n;\pi)$) s odhadovaným parametrom π, pričom apriórne rozdelenie tohto parametra je rozdelenie beta s parametrami a, b ($\pi \approx Be(a;b)$), tak aposteriórne rozdelenie parametra π je tiež beta rozdelenie s parametrami a', b' ($\pi / \mathbf{x} \approx Be(a';b')$), pre ktoré platia vzťahy

$$a' = a + x, \tag{4}$$

$$b' = b + n - x\,, \tag{5}$$

kde n označuje rozsah výberového súboru a x je počet pokusov, v ktorých nastala uvažovaná udalosť. Základné charakteristiky (apriórneho) rozdelenia beta sú dané vzťahmi

$$E(\pi) = \frac{a}{a+b}\,, \tag{6}$$

$$D(\pi) = \frac{ab}{(a+b)^2(a+b+1)}\,. \tag{7}$$

Potom pre strednú hodnotu aposteriórneho rozdelenia platí

$$E(\pi / \mathbf{x}) = \frac{a+x}{a+x+b+n-x} = \frac{a+x}{a+b+n}\,. \tag{8}$$

Tento vzťah je bayesovským bodovým odhadom podielu π v základnom súbore (označujeme $\hat{\pi}_B$).

Keďže v príspevku porovnávame kvalitu klasického a bayesovského odhadu, pripomenieme niektoré vzťahy, ktoré sa týkajú klasického bodového odhadu podielu π v základnom súbore, ktorým je výberový podiel *p*. Pre túto výberovú charakteristiku platí

$$E(p) = \pi\,, \tag{9}$$

$$D(p) = \frac{\pi \cdot (1-\pi)}{n}\,. \tag{10}$$

Ako sme už uviedli, ako kritérium porovnania odhadov sme stanovili strednú kvadratickú chybu. Pre klasický odhad – výberový podiel *p* – sa rovná rozptylu tejto charakteristiky, pretože tento odhad je neskreslený, takže druhý sčítanec vo vzťahu (1) má hodnotu 0:

$$MSE(p) = D(p) = \frac{\pi \cdot (1-\pi)}{n}\,. \tag{11}$$

Stredná hodnota bayesovského bodového odhadu $\hat{\pi}_B = \frac{a+x}{a+b+n}$ má tvar

$$E(\hat{\pi}_B) = E\left(\frac{a+x}{a+b+n}\right) = \frac{a+n\pi}{a+b+n}\,, \tag{12}$$

z ktorého je zrejmé, že ide o skreslený odhad, pretože výraz sa nerovná odhadovanému parametru π.

Rozptyl premennej $\hat{\pi}_B$ možno vyjadriť takto:

$$D(\hat{\pi}_B) = D\left(\frac{a+x}{a+b+n}\right) = \frac{1}{(a+b+n)^2} \cdot \left[D(a) + D(x)\right] =$$

$$= \frac{1}{(a+b+n)^2} \cdot (0 + n\pi(1-\pi)) = \frac{n\pi(1-\pi)}{(a+b+n)^2} . \qquad (13)$$

Keďže pre kladné čísla a,b platí

$$\frac{n\pi(1-\pi)}{(a+b+n)^2} < \frac{\pi(1-\pi)}{n} , \qquad (14)$$

bayesovský bodový odhad má menší rozptyl ako klasický bodový odhad. Pomocou vzťahov (12) a (13) možno vyjadriť strednú kvadratickú chybu bayesovského odhadu $\hat{\pi}_B$ takto:

$$MSE(\hat{\pi}_B) = \frac{n\pi(1-\pi)}{(a+b+n)^2} + \left(\frac{a+n\pi}{a+b+n} - \pi\right)^2 = \frac{n\pi(1-\pi)}{(a+b+n)^2} + \left(\frac{a-(a+b)\pi}{a+b+n}\right)^2 . \qquad (15)$$

Zaoberali sme sa hľadaním takých okolností, za ktorých je táto veličina menšia ako stredná kvadratická chyba výberového podielu (klasického bodového odhadu), teda za ktorých je bayesovský bodový odhad podielu kvalitnejší ako klasický. Stanovenú požiadavku možno formálne vyjadriť v tvare

$$MSE(\hat{\pi}_B) < MSE(p) , \qquad (16)$$

resp.

$$\frac{n\pi(1-\pi)}{(a+b+n)^2} + \left(\frac{a-(a+b)\pi}{a+b+n}\right)^2 < \frac{\pi \cdot (1-\pi)}{n} . \qquad (17)$$

Ako vidíme, vo vzťahu je viacero premenných, ktorých hodnoty môžu zaručiť jeho platnosť. My sme sa sústredili na nájdenie podmienok pre parametre a,b apriórneho rozdelenia, ktoré môžeme vo väčšine prípadov ako jediné relevantne ovplyvniť.

Modelovanie apriórnej predstavy o odhadovanom parametri π pomocou beta rozdelenia nie je také jednoduché, ako to vyzerá na prvý pohľad (toto rozdelenie je mimoriadne flexibilné, čo je na jednej strane výhodné, ale na druhej strane aj nepatrná zmena hodnoty niektorého parametra môže dosť výrazne zmeniť jeho podobu). Ak by sme presne poznali strednú hodnotu a rozptyl tohto rozdelenia, riešením sústavy dvoch rovníc o dvoch neznámych (vzťahy (6) a (7)) by sme bez problémov mohli vypočítať hodnoty jeho parametrov. Spravidla vieme pomerne presne navrhnúť strednú hodnotu, ale predstava o variabilite už nie je taká jasná. Pritom existuje nekonečný počet dvojíc a,b, ktoré vedú k rovnakej strednej hodnote

apriórneho rozdelenia beta, ale variabilita každého z týchto rozdelení je iná. Pokúsili sme sa nájsť optimálne z nich.

Najskôr sme zisťovali, či vôbec existuje nejaké apriórne rozdelenie, pri ktorom je nerovnosť (17) splnená. Ukázalo sa, že ak by sme si zvolili parametre a, b tak, aby sa stredná hodnota apriórneho rozdelenia rovnala hodnote odhadovaného parametra π ($\frac{a}{a+b} = \pi \Rightarrow b = \frac{a}{\pi} - a$), nerovnosť by platila:

$$MSE(\hat{\pi}_B) = \frac{n\pi(1-\pi)}{(a+\frac{a}{\pi}-a+n)^2} + \left(\frac{a-\frac{a}{\pi}\cdot\pi}{a+\frac{a}{\pi}-a+n}\right)^2 = \frac{n\pi(1-\pi)}{(\frac{a}{\pi}+n)^2} = \frac{n\pi^3(1-\pi)}{(a+n\pi)^2} = \frac{n\pi^3(1-\pi)}{n^2(\frac{a}{n}+\pi)^2} =$$

$$= \frac{\pi(1-\pi)}{n}\cdot\frac{\pi^2}{(\frac{a}{n}+\pi)^2} < \frac{\pi(1-\pi)}{n}$$ pre kladné hodnoty a, n.

Samozrejme, podmienka, podľa ktorej apriórna stredná hodnota sa priamo rovná odhadovanému parametru, je čisto teoretická, stačila nám však na potvrdenie faktu, že apriórne rozdelenie s požadovanou vlastnosťou vždy existuje.

Naše ďalšie úvahy súviseli s hľadaním takých apriórnych rozdelení, ktoré (popri splnení požadovanej nerovnosti) mali strednú hodnotu odlišnú od π.

Kvôli jednoduchším odvodeniam sme namiesto dvojice parametrov a, b pracovali s premennými

$s = \frac{a}{a+b}$ (stredná hodnota apriórneho rozdelenia),

$q = \frac{a}{s}$ (tento parameter je nepriamo úmerný rozptylu apriórneho rozdelenia).

Pri tomto označení platí:

$$a = qs \tag{18}$$

$$b = q - qs \tag{19}$$

$$a + b = q\,. \tag{20}$$

Premenné s, q dosadíme do nerovnosti (17):

$$\frac{n\pi(1-\pi)}{(q+n)^2} + \left(\frac{qs-q\pi}{q+n}\right)^2 < \frac{\pi(1-\pi)}{n}\,. \tag{21}$$

Po niekoľkých jednoduchých úpravách nerovnosti dostaneme vyjadrenie

$$(s-\pi)^2 < \pi(1-\pi)\cdot\frac{q+2n}{nq}, \tag{22}$$

ktoré môžeme zapísať aj v tvare

$$|s-\pi| < \sqrt{\pi(1-\pi)\cdot\frac{q+2n}{nq}}. \tag{23}$$

Posledná nerovnosť názorne ukazuje, že požadovaná podmienka (17) je splnená pre také apriórne rozdelenia, ktorých stredná hodnota s leží v intervale

$$\left(\pi-\sqrt{\pi(1-\pi)\cdot\frac{q+2n}{nq}};\pi+\sqrt{\pi(1-\pi)\cdot\frac{q+2n}{nq}}\right). \tag{24}$$

Ak by sme namiesto nerovnosti (21) upravovali rovnosť, výsledkom by boli krajné body uvedeného intervalu, v ktorých sa stredné kvadratické chyby oboch odhadov (klasického a bayesovského) rovnajú.

Získali sme tak určitú predstavu o jednom (transformovanom) parametri s apriórneho rozdelenia.

Z uvažovanej rovnice možno vyjadriť aj premennú q:

$$q=\frac{2n\pi(1-\pi)}{n(s-\pi)^2-\pi(1-\pi)}. \tag{25}$$

Táto rovnosť platí na hraniciach intervalu (24), teda v prípade, že sa stredné kvadratické chyby oboch odhadov rovnajú.

Poznanie súvislostí medzi všetkými premennými nám umožnilo nájsť pomerne jednoduchý algoritmus na určenie bayesovského bodového odhadu na základe jednoduchej apriórnej predstavy a výberových údajov.

5. POSTUP VÝPOČTU BAYESOVSKÉHO BODOVÉHO ODHADU

Ako sme už uviedli, apriórne predstavy o odhadovanom parametri π bývajú často problematické z hľadiska variability. Oveľa jednoduchšie je predstaviť si interval, v ktorom hľadaný parameter „určite“ leží. Naším cieľom je dosiahnuť, aby celý tento interval bol pokrytý intervalom, na ktorom je bayesovský bodový odhad lepší ako klasický.

Ak teda (apriórne) predpokladáme, že hľadaný parameter π leží v nejakom intervale $(\pi_{\min},\pi_{\max})$, budeme stred tohto intervalu považovať za strednú hodnotu s apriórneho rozdelenia.

Ďalej vyčíslime hodnotu q zo vzťahu (25), do ktorého za n dosadíme aktuálny rozsah výberového súboru a za π jednu z hraníc intervalu $(\pi_{\min}, \pi_{\max})$ – tú, ktorá je vzdialenejšia od stredu intervalu (0;1) (dôvod takejto voľby je obsahom ďalšej časti). Tak bude apriórne rozdelenie jednoznačne definované a na určenie bayesovského bodového odhadu možno použiť vzťah (8).

Načrtnutý postup možno realizovať v týchto krokoch:

1. Na základe apriórnej predstavy stanovíme interval $(\pi_{\min}, \pi_{\max})$.

2. Vypočítame strednú hodnotu apriórneho rozdelenia: $s = \frac{\pi_{\min} + \pi_{\max}}{2}$.

3. Vypočítame hodnotu parametra q zo vzťahu $q = \frac{2n\pi(1-\pi)}{n(\pi - s)^2 - \pi(1-\pi)}$, pričom za π dosadíme tú hranicu intervalu $(\pi_{\min}; \pi_{\max})$, ktorá je vzdialenejšia od čísla 0,5.

4. Vyčíslime hodnoty parametrov apriórneho rozdelenia: $a = qs; b = q - qs$.

5. Určíme bayesovský bodový odhad parametra π : $\hat{\pi}_B = \frac{a + x}{a + b + n}$.

Uvedený algoritmus možno skrátiť, ak v jednotlivých vzorcoch postupne podosadzujeme jednotlivé premenné (postupujeme odzadu). Dostaneme jeden vzorec

$$\hat{\pi}_B = \frac{\pi_{\max} + \pi_{\min} + \frac{x}{n}\left[\frac{n(\pi_{\max} - \pi_{\min})^2}{4\pi(1-\pi)} - 1\right]}{\frac{n(\pi_{\max} - \pi_{\min})^2}{4\pi(1-\pi)} + 1}, \qquad (26)$$

do ktorého treba dosadiť 5 premenných: x, n zodpovedajú výsledku výberového zisťovania, $\pi_{\min}, \pi_{\max}$ vyplývajú z apriórnej predstavy a za π je potrebné dosadiť vzdialenejšiu hranicu intervalu $(\pi_{\min}, \pi_{\max})$ od 0,5. Odhad (26) má zaručene nižšiu hodnotu strednej kvadratickej chyby na intervale $(\pi_{\min}, \pi_{\max})$ ako výberový podiel.

Na odvodený vzorec sa možno pozrieť aj z iného uhla pohľadu: bayesovský prístup sa pri riešení štatistických problémov nevyužíva tak často, ako by bolo žiaduce. Jedným z dôvodov je komplikovaný matematický aparát s vysokou mierou abstrakcie, ktorý bayesovská štatistika využíva. Prezentovaný vzorec možno využiť aj bez toho, že by bolo treba poznať jeho „históriu“ alebo okolnosti, za ktorých vznikal. Stačí ho akceptovať.

5.1 Voľba vhodnej hranice intervalu pri určovaní parametrov apriórneho rozdelenia

V tejto časti zdôvodníme, prečo je potrebné v 3. kroku prezentovaného algoritmu dosadiť za π tú hranicu intervalu ($\pi_{\min}, \pi_{\max}$), ktorá je vzdialenejšia od čísla 0,5.

Označme túto hranicu ako π_1, druhá hranica bude potom π_2. Keďže stredná hodnota apriórneho rozdelenia s je v 2. kroku algoritmu stanovená ako stred intervalu ($\pi_{\min}, \pi_{\max}$), je zrejmé, že je rovnako vzdialená od jeho hraníc ($|s - \pi_{\min}| = |s - \pi_{\max}|$), čo platí aj pre inak označené hranice:

$$|s - \pi_1| = |s - \pi_2| . \tag{26}$$

Rozhodnutie o výbere vhodnej hranice, ktorú treba dosadiť za π v 3. kroku algoritmu, súvisí s priebehom funkcie, ktorej funkčný predpis je uvedený pod odmocninou na pravej strane nerovnosti (23). Aj keď sa v danom výraze vyskytujú rôzne premenné, my ju považujeme za funkciu premennej π, pričom n a q sú parametre. Môžeme ju teda zapísať v tvare

$$f(\pi) = \pi(1-\pi) \cdot \frac{q+2n}{nq} . \tag{27}$$

Ako vidíme, pre pevne zvolené hodnoty parametrov n a q je to konkávna kvadratická funkcia s vrcholom v bode 0,5, z čoho je zrejmé, že pre argument vzdialenejší od bodu 0,5 (na smere vzdialenosti nezáleží) má menšiu funkčnú hodnotu ako v bode, ktorý je bližšie k bodu 0,5. Preto v súlade s naším označením platí

$$\pi_1(1-\pi_1) \cdot \frac{q+2n}{nq} < \pi_2(1-\pi_2) \cdot \frac{q+2n}{nq} , \tag{28}$$

odkiaľ

$$\sqrt{\pi_1(1-\pi_1) \cdot \frac{q+2n}{nq}} < \sqrt{\pi_2(1-\pi_2) \cdot \frac{q+2n}{nq}} . \tag{29}$$

Ako sme už uviedli, po dosadení konkrétneho π do vzťahu (25) dostaneme q, ktoré vedie k bayesovskému bodovému odhadu s rovnakou strednou kvadratickou chybou, ako je stredná kvadratická chyba klasického bodového odhadu (ľavá a pravá strana nerovnice (23) sa rovnajú).

Takže ak za π dosadíme hranicu π_1, platí rovnosť

$$|s - \pi_1| = \sqrt{\pi_1(1-\pi_1) \cdot \frac{q+2n}{nq}} . \tag{30}$$

Z (26), (29) a (30) dostaneme vzťah:

$$|s-\pi_2| = |s-\pi_1| = \sqrt{\pi_1(1-\pi_1)\cdot\frac{q+2n}{nq}} < \sqrt{\pi_2(1-\pi_2)\cdot\frac{q+2n}{nq}}\,, \tag{31}$$

z ktorého vidíme, že podmienka (23) je splnená aj pre druhú hranicu intervalu (π_2). Samozrejme, platí aj pre všetky vnútorné body uvažovaného intervalu, pretože ich vzdialenosť od s je menšia ako polovica šírky intervalu

$$|s-\pi| < |s-\pi_1|\,. \tag{32}$$

Ak by sme za π v 3. kroku algoritmu dosadili hodnotu π_2, rovnosť medzi stranami nerovnice (23) by bola splnená pre π_2:

$$|s-\pi_2| = \sqrt{\pi_2(1-\pi_2)\cdot\frac{q+2n}{nq}}\,. \tag{33}$$

V tom prípade by podmienky (26), (29) a (33) viedli k vzťahu

$$|s-\pi_1| = |s-\pi_2| = \sqrt{\pi_2(1-\pi_2)\cdot\frac{q+2n}{nq}} > \sqrt{\pi_1(1-\pi_1)\cdot\frac{q+2n}{nq}}\,, \tag{34}$$

z ktorého je zrejmé, že v druhom krajnom bode intervalu platí opačná nerovnosť. To znamená, že v takomto prípade interval ($\pi_{\min};\pi_{\max}$) nie je podmnožinou intervalu, na ktorom je bayesovský bodový odhad lepší.

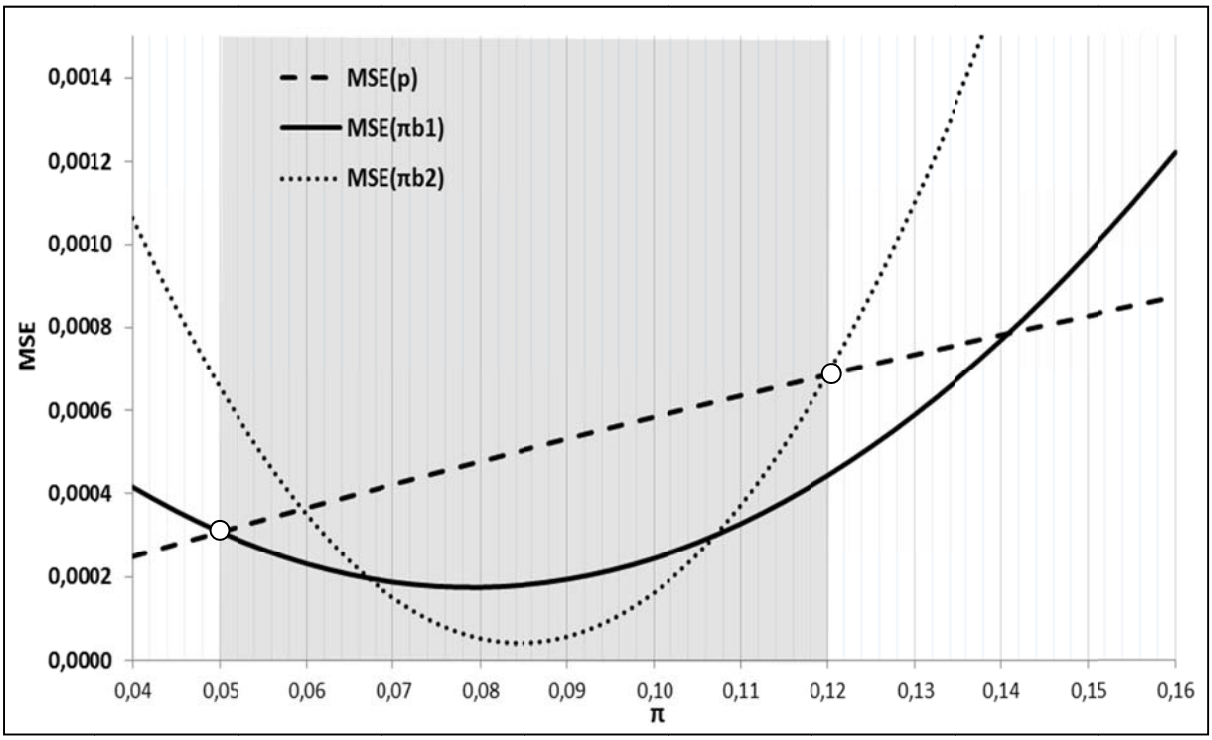
Z uvedeného je zrejmé, že ak chceme na intervale ($\pi_{\min};\pi_{\max}$) dosiahnuť nižšiu hodnotu strednej kvadratickej chyby bayesovského bodového odhadu, ako je stredná kvadratická chyba klasického bodového odhadu, je potrebné v 3. kroku algoritmu dosadiť za π do vzťahu (25) tú hranicu intervalu ($\pi_{\min};\pi_{\max}$), ktorá je vzdialenejšia od čísla 0,5.

Opísané súvislosti sme znázornili graficky pre konkrétne (hypotetické) údaje: $\pi_{\min} = 0{,}05$, $\pi_{\max} = 0{,}12$, $n = 154$, $x = 12$. Na obrázku 1 sú grafy troch rôznych stredných kvadratických chýb: MSE(p) je stredná kvadratická chyba klasického bodového odhadu, MSE(πb1), resp. MSE(πb2) je stredná kvadratická chyba bayesovského bodového odhadu pri dosadení čísla 0,05, resp. 0,12 za π.

Podmienka (23) je splnená pri tých intervaloch, kde grafy stredných kvadratických chýb bayesovského odhadu ležia pod grafom strednej kvadratickej chyby klasického bodového odhadu.

Ako vidíme, voľba vzdialenejšej hranice od 0,5 vedie k splneniu podmienky (23) pri širšom intervale, ako je potrebné. Naproti tomu voľbou druhej hranice intervalu dostaneme interval, ktorý je vlastnou podmnožinou intervalu (0,05; 0,12).

Obrázok č. 1: Porovnanie stredných kvadratických chýb klasického a bayesovského bodového odhadu



Zdroj: vlastné výpočty

6. BAYESOVSKÝ BODOVÝ ODHAD PODIELU ZÁKAZNÍKOV REAGUJÚCICH NA REKLAMNÚ KAMPAŇ

Tento príklad je len ilustratívny, pretože sme nepracovali s reálnymi údajmi. Uvádzané hodnoty však približne zodpovedajú skúsenostiam ľudí pracujúcich v danej sfére.

Predstavme si, že istá spoločnosť zaoberajúca sa predajom výživových doplnkov má v úmysle uviesť na trh nový produkt. Jeho cena nie je zanedbateľná, preto je pre firmu mimoriadne dôležité odhadnúť počet predaných kusov balení. Na propagáciu produktu je navrhnutá reklamná kampaň, ktorá je založená na inom princípe ako doterajšie kampane (preto je aj podstatne drahšia). Kľúčovou otázkou pre firmu je odhadnúť podiel zákazníkov, ktorí zareagujú pozitívne na kampaň (kúpia si daný produkt).

Doterajšie reklamné kampane boli viac alebo menej úspešné – podiel klientov, ktorí si kúpili produkt, sa pohyboval od 15 % do 33 %. Vzhľadom na nové prvky v reklamnej kampani sa očakáva pozitívny ohlas u väčšej časti klientov – marketingoví odborníci spoločnosti odhadli percento úspešnosti vyššie ako 20 %, pričom tento podiel by nemal presiahnuť 35 %.

Uvedené podiely môžu poslúžiť ako minimálna a maximálna hranica intervalu, ktorým modelujeme apriórnu predstavu, teda označíme: $\pi_{\min} = 0{,}2, \quad \pi_{\max} = 0{,}35$.

Ďalej predpokladajme, že prebehla „pokusná“ kampaň – prezentácia, na ktorej sa zúčastnilo 118 náhodne vybraných potenciálnych kupujúcich, pričom 29 z nich si

kúpilo ponúkaný produkt. Tieto údaje môžeme vyjadriť pomocou používanej symboliky: $n = 118, \quad x = 29$.

Štvorica zaznamenaných čísel stačí na výpočet hodnoty bayesovského odhadu pomocou vzorca (26), pričom za π dosadíme dolnú hranicu intervalu (číslo 0,2):

$$\hat{\pi}_B = \frac{0{,}35 + 0{,}2 + \frac{29}{118}\left[\frac{118 \cdot (0{,}35 - 0{,}2)^2}{4 \cdot 0{,}2 \cdot (1 - 0{,}2)} - 1\right]}{\frac{118 \cdot (0{,}35 - 0{,}2)^2}{4 \cdot 0{,}2 \cdot (1 - 0{,}2)} + 1} = 0{,}266355376 .$$

Podľa vypočítanej hodnoty na reklamnú kampaň pozitívne zareaguje 26,64 % oslovených ľudí.

Na porovnanie určíme aj klasický odhad – výberový podiel: $p = \frac{29}{118} = 0{,}24576271$, ktorý zodpovedá 24,58 % úspešnosti.

Ako vidíme, bayesovský odhad je o niečo optimistickejší – je v ňom zohľadnená nielen informácia z kampane, ale aj apriórna predstava, reprezentovaná apriórnou strednou hodnotou – stredom intervalu (0,2; 0,35), teda číslom 0,275.

Pri porovnávaní obidvoch odhadov si treba uvedomiť, že ak sa skutočná hodnota úspešnosti nachádza v intervale (0,2; 0,35), tak bayesovský bodový odhad má nižšiu strednú kvadratickú chybu ako výberový podiel, s čím je spojená menšia pravdepodobnosť výraznejšej chyby pri odhade.

Výsledok, ktorý sme dostali využitím vzorca (26), potvrdzuje známy fakt, ktorý platí v bayesovskej štatistike všeobecne: bayesovský odhad je kompromisom medzi hodnotami, ktoré pochádzajú z výberového zisťovania, a hodnotami reprezentujúcimi apriórnu informáciu. Pritom platí, že čím je rozsah výberového súboru väčší, tým je výsledok bližší hodnote pochádzajúcej z výberového zisťovania, teda ku klasickému bodovému odhadu.

7. ZÁVER

Uvedený príklad ilustroval jednoduchosť algoritmu na stanovenie bayesovského bodového odhadu, ktorý sme navrhli a zdôvodnili v predchádzajúcich častiach. Ako sme už uviedli, vzorec je použiteľný vždy, keď je (okrem výberových údajov) k dispozícii veľmi jednoduchá apriórna predstava o odhadovanom podiele, ktorú nie je potrebné matematicky modelovať nejakým rozdelením. Tým sa do určitej miery eliminuje nevýhoda bayesovského prístupu, ktorou je náročný matematický aparát. Na druhej strane akceptovanie apriórnej informácie v porovnaní s klasickým prístupom vedie k presnejšiemu (teda kvalitnejšiemu) odhadu.

LITERATÚRA

[1] BAKYTOVÁ, H. – HÁTLE, J. – NOVÁK, I. – UGRON, M. Statistická indukce pro ekonomy. Praha: SNTL, ALFA, 1986, ISBN 99-00-00135-X.
[2] BERNARDO, JOSÉ M. – SMITH, ADRIAN F. M. 1995. Bayesian theory. Chichester: John Wiley & Sons Ltd., 640 s., 1995.
[3] BOLSTAD, W. M. 2004. Introduction to Bayesian statistics. New Yersey, USA: John Wiley&sons, Inc. 2004.
[4] FREUND, J. E. 1992. Mathematical statistics. New Jersey: Prentice-Hall, International, 1992.
[5] GARTHWAITE, P. H. – JOLLIFFE, I. – JONES, B. 1995. Statistical Inference. Prentice-Hall International, Inc., 1995.
[6] HORÁKOVÁ, G. – HUŤKA, V.: Teória pravdepodobnosti 1. Bratislava: Vyd. EKONÓM, 2002.
[7] KOTLEBOVÁ, E. 2009. Bayesovská štatistická indukcia v ekonomických aplikáciách. Bratislava: Vyd. EKONÓM, 2009.
[8] LÁSKA, I. 2013. Porovnanie klasického a bayesovského prístupu pri riešení problémov štatistickej indukcie. Bratislava: Diplomová práca (FHI EU).
[9] LEE, PETER, M. 1989. Bayesian statistics. New York: Oxford University Press, 1989.
[10] PACÁKOVÁ, V. a kol. 2012. Štatistická indukcia pre ekonómov. Bratislava: Vyd. EKONÓM, 2012.
[11] PACÁKOVÁ, V. 2004. Aplikovaná poistná štatistika. Bratislava: IURA EDITION, 2004.
[12] PFAFFENBERGER, R. C. – PATTERSON, J. H. 1987. Statistical Methods (For business and Economics). Illinois: IRWIN, 1987.
[13] ŠOLTÉS, E. 2009. Modely kredibility na výpočet poistného. Bratislava: Vyd. EKONÓM, 2009.
[14] TEREK, M. 2003. Úvod do analýzy rozhodovania a bayesovskej indukcie. Bratislava: Vyd. EKONÓM, 2003.

RESUMÉ

V príspevku je prezentovaný bayesovský prístup k bodovému odhadu podielu v základnom súbore. Okrem výberových údajov, ktoré sú jediným zdrojom informácie pri klasickom prístupe, sa berie do úvahy aj informácia pochádzajúca z iných zdrojov ako z výberového zisťovania (tzv. apriórna informácia).
V bayesovskej štatistike sa na odhad podielu bežne využíva konjugovaný systém binomické/beta, ktorý je v príspevku modifikovaný v tom zmysle, že ako kritérium kvality bodového odhadu sa používa stredná kvadratická chyba.
Bolo dokázané, že vždy existuje interval, na ktorom je stredná kvadratická chyba bayesovského odhadu nižšia ako stredná kvadratická chyba klasického odhadu. Tento teoretický poznatok viedol k algoritmu, pomocou ktorého sa dá vypočítať bayesovský bodový odhad na základe jednoduchej apriórnej predstavy, podľa ktorej odhadovaný parameter leží v nejakom konkrétnom intervale. Algoritmus bol prezentovaný na hypotetických údajoch marketingových aktivít firmy zaoberajúcej sa predajom výživových doplnkov.

RESUME

In the article, the bayesian approach to the population proportions´ point estimation is presented. Besides the sample data, which is the only source of information in the

classical statistical inference, bayesian statistics takes into account another information, so-called prior information.
In bayesian statistics, the conjugate family binomial/beta is used for the population proportions´ estimation. This principle is in the article modified in terms of using the mean square error as the main criterion of the quality.
It was proved, that there exists interval, within which the mean square error of the bayesian point estimation is smaller than the mean square error of the classical point estimation (the sample proportion). This knowledge led to the algorithm for evaluating the bayesian point estimation on the base of the simple prior vision (that the population proportion is between particular borders). The algorithm was applied on the hypothetical data of marketing activities of the firm transacted with food additives.

PROFESIJNÉ ŽIVOTOPISY

RNDr. Eva Kotlebová, PhD., *je absolventkou Matematicko-fyzikálnej fakulty Univerzity Komenského v Bratislave (vedecký smer matematika – teória systémov). Po ukončení vysokoškolského štúdia bola tri roky na študijnom pobyte na Katedre štatistiky Fakulty riadenia Vysokej školy ekonomickej v Bratislave. Potom pôsobila niekoľko rokov ako stredoškolská učiteľka matematiky na gymnáziu v Bratislave. Od roku 2003 pracuje na Katedre štatistiky Fakulty hospodárskej informatiky v Bratislave. V roku 2008 ukončila doktorandské štúdium. Venuje sa štatistickej indukcii, bayesovskej štatistike a aplikácii štatistických metód v poisťovníctve.*

Ing. Ivan Láska *ukončil prvý stupeň vysokoškolského štúdia v študijnom odbore aplikovaná matematika. Inžinierske štúdium absolvoval na Fakulte hospodárskej informatiky Ekonomickej univerzity v Bratislave, počas ktorého získal viaceré ocenenia (Cena rektora Ekonomickej univerzity v Bratislave, Cena dekana Fakulty hospodárskej informatiky, 1. miesto vo fakultnom kole ŠVOČ). Od roku 2013 je študentom doktorandského štúdia na tejto fakulte v odbore kvantitatívne metódy v ekonómii. Súčasne pracuje ako analytik v spoločnosti Trexima Bratislava. Venuje sa bayesovskej štatistike, analýzam zamestnanosti a prognózovaniu potrieb trhu práce.*

KONTAKT

eva.kotlebova@gmail.com
ivo.laska@gmail.com