Bayesian Estimation of Probability of Incidences of the Most Serious Oncological Diseases in the Czech Republic

Lucie Kopecká, Viera Pacáková [1]

Abstract

The aim of this article is to estimate the probability of incidences of the most serious two oncological diseases in the three selected five-year age groups in the Czech Republic by using Bayesian theory of credibility, specifically binomial/beta model. These diseases are occurred in early childhood. During the lifetime, these incidences grow rapidly. The next aim of this article is to determine the parameters of a priori beta distribution, specifically in the case that the priory information on estimated parameter of binomial distribution is known. These parameters of beta distribution are determined by using basic characteristics of beta distribution. Possibilities of permanent update of estimates can be used in commercial and health insurance companies. Data on cancer incidences are obtained from database of ÚZIS.

Key words

oncological diseases, Bayesian theory of credibility, binomial/beta model, prior distribution

JEL Classification: C11, C13, I10

1 Úvod

Rakovina tlustého střeva a konečníku a rakovina průdušnice, průdušek a plic patří mezi nejčastěji se vyskytující závažná onkologická onemocnění v České republice. Výskyt těchto onkologických onemocnění je zaznamenán u dětí již v jejich raném věku a s přibývajícím věkem narůstá.

Komerční a zdravotní pojišťovny nabízejí množství produktů týkajících se pojištění závažných onemocnění, které chrání klienta před finančními důsledky způsobenými výskytem těchto onemocnění. Každá pojišťovna musí umět správně odhadnout rizika, aby byla v budoucnu schopna plnit své závazky, a to aplikováním vhodných metod odhadu pravděpodobností. Bayesovská teorie kredibility se jeví jako vhodná metoda stanovování a permanentní aktualizace těchto odhadů, protože se pojišťovny zaměřují na data z vlastních i cizích pojišťoven, a to za minulá období. (Jindrová a Kopecká, 2017; Pacáková, Šoltés a Šoltésová, 2009)

Cílem tohoto článku je odhad pravděpodobností výskytu dvou výše uvedených nejčastějších onkologických onemocnění, a to vždy ve třech vybraných pětiletých věkových skupinách 20-24, 40-44 a 60-64 let v rámci České republiky. Odhady pravděpodobností

[1] Ing. Lucie Kopecká, University of Pardubice, Faculty of Economics and Administration, Department of Mathematics and Quantitative Methods, e-mail: Lucie.Kopecka1@student.upce.cz
prof. RNDr. Viera Pacáková, Ph.D, University of Pardubice, Faculty of Economics and Administration, Department of Mathematics and Quantitative Methods, e-mail: Viera.Pacakova@upce.cz

výskytu těchto onemocnění jsou konstruovány pomocí bayesovského modelu binomické/beta. Dále je věnovaná pozornost stanovení parametrů apriorního rozdělení beta za pomoci základních charakteristik (střední hodnoty a rozptylu) beta rozdělení, tj. v případě, že apriorní informace o odhadovaném parametru binomického rozdělení je známá, podle (Kotlebová, 2009).

2 Data a metodologie

Jak již bylo výše zmíněno, jedním z cílů tohoto článku je odhadnout pravděpodobnosti výskytu rakoviny tlustého střeva a konečníku a rakoviny průdušnice, průdušek a plic ve vybraných pětiletých věkových kategoriích v rámci České republiky. Data o incidencích těchto onkologických onemocnění pocházejí ze stránek ÚZIS ČR, konkrétně jsou čerpána ze stránek Epidemiologie zhoubných nádorů v České republice. Tato databáze poskytuje údaje o incidencích onkologických onemocnění od roku 1977 do roku 2014 za jednotlivé věkové kategorie. Dále data o počtu obyvatel byla čerpána vždy k 1. červenci daného roku ze stránek UNITED NATIONS. Tato data jsou k dispozici od roku 1950 do roku 2015 a také za jednotlivé věkové kategorie. OECD databáze poskytuje data o incidencích rakoviny tlustého střeva a plic pro jednotlivé členské státy, avšak bez ohledu na členění dle věkových kategorií. Jako apriorní informace je považována informace o pravděpodobnostech těchto incidencí v jednotlivých evropských státech, která jsou přepočtena na jednoho obyvatele.

Pro výpočet odhadů pravděpodobností výskytu dvou výše zmíněných onkologických onemocnění byla zvolena bayesovská teorie kredibility, konkrétně model binomické/beta.

2.1 Bayesovská teorie kredibility

Základem bayesovké teorie kredibility je bayesovská teorie odhadu. Bayesovská teorie odhadu nezahrnuje pouze údaje z vlastního výběru, jako je tomu u klasických metod odhadu, ale zahrnuje také jiné dostupné porovnatelné informace, které bývají často k dispozici, jak zmiňuje (Pacáková, 2004; Kotlebová, 2009).

Dalším rozdílem mezi klasickou a bayesovskou statistikou je neznámý parametr θ, který je v případě bayesovské statistiky považován za náhodnou proměnnou, a to na rozdíl od klasické statistiky, kde je tento parametr považován za neznámou konstantu. V případě bayesovské statistiky má tedy parametr θ (náhodná proměnná) rozdělení pravděpodobnosti $f(\theta)$. Toto rozdělení pravděpodobnosti se nazývá apriorní, protože poskytuje první informaci o odhadované parametru θ a tato informace zatím nepochází z vlastního výběrového souboru podle literatury (Pacáková, 2004).

I v případě, že apriorní informace o odhadovaném parametru θ není nejlepší, je možné ji zlepšit pomocí informace aposteriorní, tedy z vlastního výběrového zjišťování. Aposteriorní rozdělení odhadovaného parametru θ využívá informací jak jeho apriorního rozdělení, tak výsledek vlastního výběrového zjišťování a značí se $f(\theta/\boldsymbol{x})$. Jestliže $\boldsymbol{x}$ je náhodný výběr z rozdělení pravděpodobnosti $f(\boldsymbol{x}/\theta)$ a neznámý parametr θ má apriorní rozdělení $f(\theta)$, pak pro aposteriorní rozdělení podle spojité verze bayesovské věty platí vztah:

$$f(\theta/\boldsymbol{x}) = \frac{f(\boldsymbol{x}/\theta)\, f(\theta)}{\int_{\theta} f(\boldsymbol{x}/\theta)\, f(\theta)\, \mathrm{d}\theta} \tag{1}$$

Výraz ve jmenovateli ve vztahu (1) bude označen jako marginální hustota $f(\boldsymbol{x})$, která nezávisí na parametru θ a představuje konstantu, jenž bude pro zjednodušení vynechána. Aposteriorní rozdělení je možné nyní zapsat vztahem:

$$f(\theta / \boldsymbol{x}) \propto f(\boldsymbol{x} / \theta)\, f(\theta) \tag{2}$$

Dá se říci, že aposteriorní rozdělení odhadovaného parametru θ je proporcionální se součinem funkce věrohodnosti a apriorním rozdělením, jak uvádí (Pacáková, 2004).

V případě, že apriorní a aposteriorní rozdělení je stejného typu, ale různých parametrů, jsou tato rozdělení označena jako konjugovaná, a to k rozdělení, ze kterého pochází náhodný výběr. Model binomické/beta patří ke konjugovaným rozdělení využívaných zejména v pojišťovnictví, jak uvádí (Kotlebová, 2009; Pacáková a Kotlebová, 2014).

2.1.1 Model binomické/beta

V modelu binomické/beta je rozdělení beta konjugovaným apriorním rozdělením pro rozdělení binomické s neznámým parametrem θ. Náhodná veličina X má binomické rozdělení s parametrem θ. Podmíněná funkce pravděpodobnosti má tedy tvar:

$$f(x / \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \qquad x = 1,\ldots,n \tag{3}$$

Apriorní rozdělení parametru θ je výše zmíněné beta, a to s parametry α a β. Pro hustotu tohoto rozdělení na intervalu (0,1) platí pro $\theta \in (0,1)$ vztah:

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1} \propto \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1} \tag{4}$$

Po dosazení do vztahu (2) dostaneme hustotu aposteriorního rozdělení vyjádřenou vztahem (5), která vyjadřuje hustotu rozdělení beta s parametry $\alpha^* = \alpha + x$ a $\beta^* = \beta + n - x$: (Kotlebová, 2009; Pacáková, 2004; Pacáková a Kotlebová, 2014)

$$f(\theta / x) \propto \theta^x (1-\theta)^{n-x} \cdot \theta^{\alpha-1} (1-\theta)^{\beta-1} = \theta^{\alpha+x-1} (1-\theta)^{\beta+n-x-1} \tag{5}$$

Bayesovským odhadem θ_B parametru θ bude v tomto případě vzhledem na kvadratickou ztrátu střední hodnota aposteriorního beta rozdělení, která je vyjádřená vztahem:

$$\theta_B = \frac{\alpha + x}{\alpha + \beta + n} \tag{6}$$

Vztah (6) je možné přepsat do tvaru:

$$\theta_B = \frac{n}{\alpha+\beta+n} \cdot \frac{x}{n} + \frac{\alpha+\beta}{\alpha+\beta+n} \cdot \frac{\alpha}{\alpha+\beta} = Z \cdot \frac{x}{n} + (1-Z) \cdot \mu, \tag{7}$$

kde Z vyjadřuje faktor kredibility, $\frac{x}{n}$ maximálně věrohodný odhad parametru θ binomického rozdělení a μ znázorňuje střední hodnotu beta rozdělení. (Gogola, 2013; Pacáková, 2004).

3 Modelování odhadů pravděpodobností výskytu onkologických onemocnění ve vybraných věkových kategoriích

V rámci této kapitoly je již pozornost věnovaná praktickému využití bayesovských odhadů, konkrétně modelu binomické/beta, který se jeví pro odhady pravděpodobností výskytu onkologických onemocnění a tím i pro potřeby pojišťoven jako vhodný. Data získaná ze stránek ÚZIS ČR a UNITED NATIONS jsou čerpaná za období 15 let, a to od roku 2000 do roku 2014.

Důležitou součástí bayesovkých odhadů je stanovení parametrů apriorního rozdělení v tomto případě beta, což vyžaduje znalost apriorní informaci o odhadovaném parametru θ binomického rozdělení. V případě, že tato informace není známá, což v praxi nebývá časté,

využívá se rovnoměrného beta rozdělení, pro něhož jsou parametry α a β rovny 1 pro první rok podle (Kotlebová, 2009). Jde o případ, kdy bayesovský odhad výskytu onkologického onemocnění pro první rok je roven 50 %. Představa, že pravděpodobnost výskytu onkologického onemocnění byla v roce 2000 takto vysoká, však není reálná.

Z pravděpodobností výskytu určitého onkologického onemocnění (rakoviny tlustého střeva nebo plic), které byly odhadnuty na základě dat OECD pro rok 2012 popsaných výše, je možné získat představu o apriorním rozdělení, a to výpočtem střední hodnoty μ^* a rozptylu $(\sigma^*)^2$ těchto pravděpodobností. Parametry apriorního rozdělení beta pro rok 2000 jsou stanoveny pomocí dvou základních charakteristik beta rozdělení, a to střední hodnoty a rozptylu, které jsou vyjádřeny vztahy podle (Kotlebová, 2009):

$$\mu^* = \frac{\alpha}{\alpha + \beta} \tag{8}$$

$$\left(\sigma^*\right)^2 = \frac{\alpha \cdot \beta}{\left(\alpha + \beta\right)^2 \cdot \left(\alpha + \beta + 1\right)} \tag{9}$$

V následujících dvou podkapitolách jsou již konstruovány odhady pravděpodobností výskytu dvou nejčastějších onkologických onemocnění v ČR ve třech vybraných pětiletých věkových kategoriích pomocí modelu binomické/beta, a to pro rok 2015.

3.1 Odhady pravděpodobností výskytu rakoviny tlustého střeva a konečníku

Nyní již jsou odhadovány pravděpodobnosti výskytu rakoviny tlustého střeva a konečníku pro rok 2015 pomocí modelu binomické/beta. Postup výpočtu bayesovských odhadů při využití modelu binomické/beta je popsán v podkapitole 2.1.1 a stanovování parametrů apriorního rozdělení beta pomocí základních charakteristik beta rozdělení je popsáno na začátku kapitoly 3. Za apriorní informaci je brána představa o výskytu rakoviny tlustého střeva v jednotlivých evropských státech v posledním známém roce 2012. V následující tabulce 1 je uveden postup konstrukce těchto odhadů, a to pro věkovou kategorii 20-24 let.

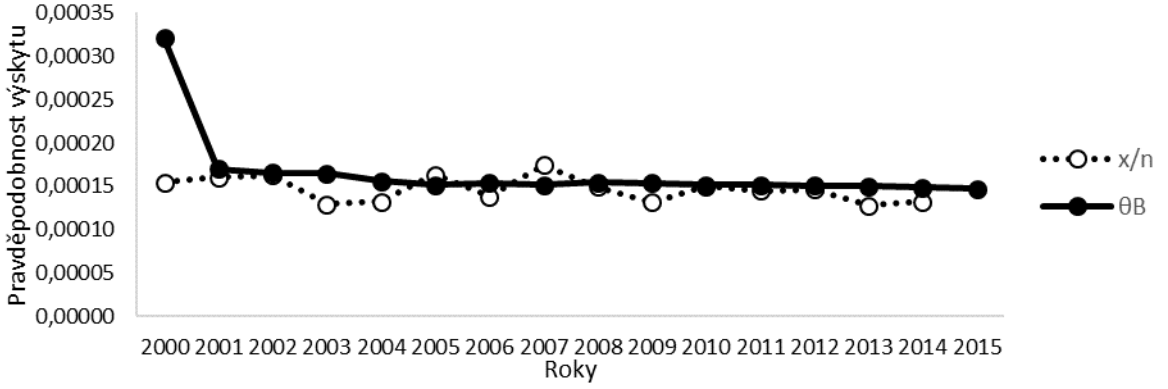
Tabulka 1: Odhady pravděpodobností výskytu rakoviny tlustého střeva a konečníku věkové kategorie 20-24 let

Roky	***n***	***x***	***x/n***	***α***	***β***	θ_B
2000	895 278	7	0,0000078	23	71 541	0,0003205
2001	869 335	1	0,0000012	30	966 812	0,0000310
2002	826 445	3	0,0000036	31	1 836 146	0,0000168
2003	775 603	3	0,0000039	34	2 662 588	0,0000127
2004	730 159	5	0,0000068	37	3 438 188	0,0000107
2005	698 010	7	0,0000100	42	4 168 342	0,0000101
2006	682 500	5	0,0000073	49	4 866 345	0,0000101
2007	680 367	1	0,0000015	54	5 548 840	0,0000097
2008	686 301	0	0,0000000	55	6 229 206	0,0000088
2009	691 891	3	0,0000043	55	6 915 507	0,0000079
2010	691 582	4	0,0000058	58	7 607 395	0,0000076
2011	683 265	13	0,0000190	62	8 298 973	0,0000075
2012	672 760	9	0,0000134	75	8 982 225	0,0000083
2013	658 672	4	0,0000061	84	9 654 976	0,0000087
2014	639 439	7	0,0000109	88	10 313 644	0,0000085
2015				95	10 953 076	**0,0000087**

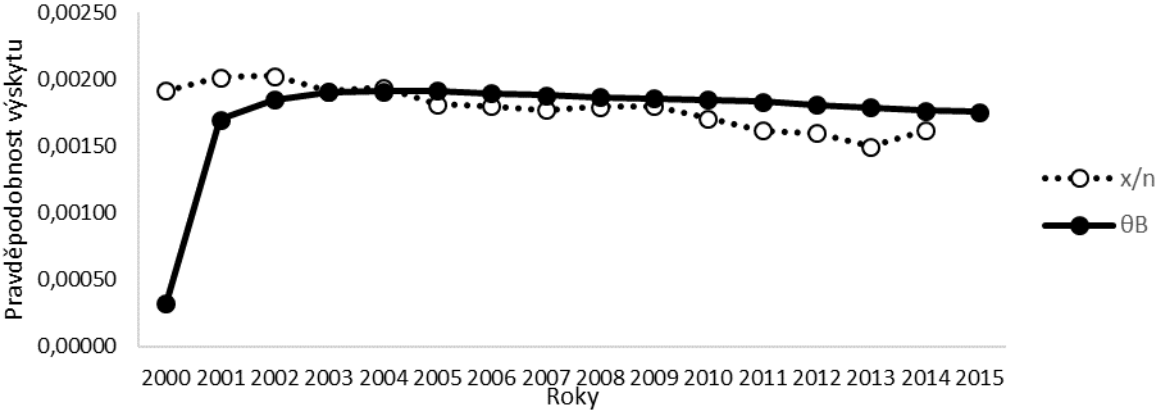
V tabulce 1 označuje sloupec *n* počet obyvatel k 1. červenci daného roku ve věkové skupině 20-24 let a sloupec *x* znázorňuje počet výskytů rakoviny tlustého střeva a konečníku v této věkové skupině. Čtvrtý sloupec x/n vyjadřuje maximálně věrohodný odhad parametru θ binomického rozdělení, který je možné konstruovat pouze pro roky, pro která jsou známá data. Sloupce α a β znázorňují parametry beta rozdělení. Poslední sloupec θ_B vyjadřuje hledaný bayesovský odhad pravděpodobností výskytu rakoviny tlustého střeva a konečníku. Tento odhad má dvě hlavní výhody oproti maximálně věrohodnému odhadu. První z nich je výše zmiňované zachycení apriorní informace, což v případě maximálně věrohodného odhadu není možné. Druhou podstatnou výhodou je ovšem možnost konstrukce odhadu pravděpodobnosti pro rok následující po posledním známém roce, tzn. v tomto případě pro rok 2015. Parametry beta rozdělení jsou pro rok 2000 a věkovou kategorii 20-24 let stanoveny řešením soustavy rovnic (8) a (9). V tomto případě parametr $\alpha = 23$ a $\beta = 71541$. Při těchto parametrech je $\theta_B = 0{,}0003205$ pro rok 2000. Bayesovský odhad pravděpodobnosti výskytu rakoviny tlustého střeva a konečníku ve věkové kategorii 20-24 let klesá do roku 2011, avšak od tohoto roku začíná narůstat a pro rok 2015 tento odhad činí 0,0000087.

Výpočty bayesovských odhadů pomocí modelu binomické/beta pro ostatní uvedené věkové kategorie jsou prováděny analogicky. Na následujících dvou obrázcích 1 a 2 jsou graficky znázorněny odhady pravděpodobností výskytu rakoviny tlustého střeva a konečníku, a to nejprve ve věkové kategorii 40-44 let a následně ve věkové kategorii 60-64 let.

Obrázek 1: Odhady pravděpodobností výskytu rakoviny tlustého střeva a konečníku věkové kategorie 40-44 let



Obrázek 2: Odhady pravděpodobností výskytu rakoviny tlustého střeva a konečníku věkové kategorie 60-64 let



Na obrázku 1 a obrázku 2 je možné porovnat bayesovský odhad θ_B s maximálně věrohodným odhadem x/n. Bayesovský odhad má tendenci přibližovat se k maximálně věrohodnému odhadu v průběhu let, protože apriorní informace (z cizích porovnatelných

rizik) je postupem času vylepšována informacemi aposteriorními (z vlastního výběrového zjišťování). Bayesovský odhad má také hladší průběh na rozdíl od maximálně věrohodného odhadu, což může být považováno za jeho další výhodu. Na základě bayesovského odhadu má pravděpodobnost výskytu rakoviny tlustého střeva a konečníku ve věkových kategoriích 40-44 let a 60-64 let mírně klesající tendenci v průběhu posledních let.

Pro obě tyto věkové kategorie nabývají parametry beta rozdělení pro rok 2000 stejných hodnot jako v případě věkové kategorie 20-24 let. V případě věkové kategorie 40-44 je bayesovský odhad pravděpodobnosti pro rok 2015 roven 0,0001473. Dále pak pro věkovou kategorii 60-64 let bayesovský odhad pravděpodobnosti pro rok 2015 činí 0,0017555.

3.2 Odhady pravděpodobností výskytu rakoviny průdušnice, průdušek a plic

Dalším závažným onkologickým onemocněním, které se vyskytuje již v brzkém věku, je rakovina průdušnice, průdušek a plic. Veškeré výpočty týkající se bayesovských odhadů pravděpodobností výskytu pro jednotlivé věkové kategorie pomocí modelu binomické/beta jsou prováděny analogicky jako v předchozí podkapitole. V tabulce 2 jsou konstruovány odhady pravděpodobnosti výskytu rakoviny průdušnice, průdušek a plic ve věkové kategorii 20-24 let.

Tabulka 2: Odhady pravděpodobností výskytu rakoviny průdušnice, průdušek a plic věkové kategorie 20-24 let

Roky	***n***	***x***	***x/n***	α	β	θ_B
2000	895 278	1	0,0000011	20	65 215	0,0003038
2001	869 335	2	0,0000023	21	960 492	0,0000217
2002	826 445	1	0,0000012	23	1 829 825	0,0000125
2003	775 603	1	0,0000013	24	2 656 269	0,0000090
2004	730 159	1	0,0000014	25	3 431 871	0,0000072
2005	698 010	3	0,0000043	26	4 162 029	0,0000062
2006	682 500	2	0,0000029	29	4 860 036	0,0000059
2007	680 367	1	0,0000015	31	5 542 534	0,0000056
2008	686 301	0	0,0000000	32	6 222 900	0,0000051
2009	691 891	4	0,0000058	32	6 909 201	0,0000046
2010	691 582	1	0,0000014	36	7 601 088	0,0000047
2011	683 265	0	0,0000000	37	8 292 669	0,0000044
2012	672 760	4	0,0000059	37	8 975 934	0,0000041
2013	658 672	0	0,0000000	41	9 648 690	0,0000042
2014	639 439	3	0,0000047	41	10 307 362	0,0000040
2015				44	10 946 798	**0,0000040**

V tabulce 2 je možné vidět klesající trend bayesovských odhadů pravděpodobností výskytu rakoviny průdušnice, průdušek a plic pro věkovou kategorii 20-24 let. Pro tuto věkovou kategorii nabývají parametry beta rozdělení pro rok 2000 řešením soustavy rovnic (8) a (9) hodnot $\alpha = 20$ a $\beta = 65215$. Při těchto parametrech je pro tento rok $\theta_B = 0{,}0003038$, což platí i pro následující dvě věkové kategorie. Bayesovský odhad pro rok 2015 pro tento případ činí 0,0000040. Na obrázku 3 a 4 jsou graficky znázorněny opět maximálně věrohodné i bayesovské odhady pravděpodobností výskytu rakoviny průdušnice, průdušek a plic pro 2 věkové kategorie: 40-44 let a 60-64 let.

Obrázek 3: Odhady pravděpodobností výskytu rakoviny průdušnice, průdušek a plic věkové kategorie 40-44 let

Obrázek 4: Odhady pravděpodobností výskytu rakoviny průdušnice, průdušek a plic věkové kategorie 60-64 let

Na obrázku 3 lze pozorovat klesající trend bayesovských odhadů pravděpodobností ve věkové kategorii 40-44 let. Bayesovský odhad pro rok 2015 činí 0,0000736. V poslední věkové kategorii 60-64 let, pro které jsou odhady pravděpodobnosti znázorněny na obrázku 4, je možné vidět také klesající trend bayesovských odhadů, ale až od roku 2008. Bayesovský odhad pravděpodobnosti pro rok 2015 činí 0,0018003.

4 Závěr

Cílem tohoto článku bylo stanovit odhady pravděpodobností výskytu rakoviny tlustého střeva a konečníku a rakoviny průdušnice, průdušek a plic, a to vždy ve třech vybraných pětiletých věkových skupinách 20-24, 40-44 a 60-64 let v rámci ČR. Pro odhady pravděpodobnosti výskytu těchto onemocnění byl použit bayesovský model binomické/beta. Dále byla věnovaná pozornost stanovení parametrů apriorního rozdělení beta za pomoci střední hodnoty a rozptylu tohoto rozdělení, a to v případě, že apriorní informace o odhadovaném parametru binomického rozdělení je známá. Data byla čerpána ze stránek ÚZIS ČR, UNITED NATIONS a OECD.

Bayesovské odhady mají 3 hlavní výhody oproti maximálně věrohodným odhadům, jak již bylo uvedeno výše. Za první z nich je považována schopnost zahrnout kromě aposteriorní informace (z vlastního výběrového souboru) i informaci apriorní (z cizích porovnatelných rizik). Stanovení vhodné apriorní informace je tedy důležitým krokem pro přesnější konstrukci těchto odhadů. Mezi další výhody patří to, že bayesovské odhady jsou konstruovány pro rok následující po posledním známém roce v našem případě pro rok 2015. Poslední zde zmíněnou výhodou je hladký průběh těchto odhadů oproti maximálně

věrohodným odhadům, což je zjevné z obrázků 1-4. Ve světle těchto výhod se jeví bayesovské odhady vhodnými především pro využití v oblasti pojišťovnictví, konkrétně při pojištění závažných onemocnění.

V případě rakoviny tlustého střeva a konečníku byl zjištěn v posledních letech rostoucí trend bayesovských odhadů ve věkové kategorii 20-24 let. Naopak ve věkových kategoriích 40-44 let a 60-64 let byl tento trend v posledních letech mírně klesající. U rakoviny průdušnice, průdušek a plic jsou ve všech věkových kategoriích zaznamenány klesající trendy bayesovských odhadů v posledních letech. U obou vybraných onkologických onemocnění bylo zjištěno, že dochází k rapidnímu nárůstu pravděpodobnosti výskytu těchto onemocnění mezi jednotlivými věkovými kategoriemi, a to pro odhadovaný rok 2015.

Poděkování

Tento článek byl zpracován s podporou projektu SGS Univerzity Pardubice, Fakulty ekonomicko-správní: SGS_2017_022, „Ekonomický a sociální rozvoj v soukromém a veřejném sektoru“.

Reference

[1] Epidemiologie zhoubných nádorů v ČR (2017). *Epidemiologické analýzy* [online]. Dostupné z: http://www.svod.cz/

[2] Gogola, J. (2013). Spôsob permanentnej úpravy výšky poistného v neživotnom poistení. *E+M Ekonomie a Management,* (4), s. 134-142.

[3] Jindrová, P. a Kopecká, L. (2017). Kvantifikace rizik pro úrazové pojištění. *Scientific Papers of the University of Pardubice - Series D, Faculty of Economics and Administration,* 24(39), s. 75-86.

[4] Kotlebová, E. (2009). *Bayesovská štatistická indukcia v ekonomických aplikaciách*. Bratislava: Ekonóm.

[5] OECD.stat (2017). *Health status: Cancer* [online]. Dostupné z: http://stats.oecd.org/

[6] Pacáková, V. (2004). *Aplikovaná poistná štatistika*. 3. vyd. Bratislava: IURA Edition.

[7] Pacáková, V., Šoltés, E., Šoltésová, T. (2009). Kredibilný odhad škodovej frekvencie. *E+M Ekonomie a Management,* (2), s. 122-126.

[8] Pacáková, V. a Kotlebová, E. (2014). Bayesien Estimation of Event Probability in Accident Insurance. *European Financial Systems, Proceedings of the 11th International Scientific Conference, Brno: Masaryk University,* s. 462-468.

[9] UNITED NATIONS (2017). *Population indicators* [online]. Dostupné z: https://esa.un.org/unpd/wpp/Download/Standard/Population/