Pravděpodobnostní model rozdělení příjmů v České republice[#]

Jitka Bartošová*

Úvod

Zkoumání rozdělení příjmů a jeho porovnávání z různých sociálně-ekonomických a časově-prostorových hledisek je východiskem pro posouzení životní úrovně obyvatelstva, úrovně sociálního zabezpečení a sociální spravedlnosti v rozdělování materiálních hodnot, vytvořených společnostmi. Analýza rozdělení příjmů, vedoucí ke kvantitativnímu vystižení životní úrovně, je zaměřena na charakterizaci příjmové nerovnosti. V minulosti byla tato nerovnost vyjadřována obvykle pomocí Lorenzovy křivky a z ní vypočtených základních charakteristik příjmové diferenciaci, odvozených Paretem, Brescianim, Ginim, Pietrem atd. V současné literatuře je tento tradiční přístup obohacen o zjišťování rozdílů mezi několika populacemi, lišícími se buď ze sociálně-ekonomického, nebo geografického hlediska, a o kvantitativní vystižení příspěvků jednotlivých subpopulací k nerovnosti příjmů celé heterogenní populace.

Předložený článek je zaměřen na problematiku konstrukce pravděpodobnostního modelu rozdělení příjmů domácností s využitím hustoty známého teoretického rozdělení. Praktická část příspěvku je zaměřena na prezentaci výsledků modelování rozdělení příjmů neagregovaných datových souborů domácností v ČR, pocházejících z Mikrocensu 1996. Důvodem použití uvedeného datového souboru je to, že neagregovaná data umožňují kvalitní odhady parametrů zvoleného teoretického modelu.

1. Pravděpodobnostní modelování rozdělení

Pravděpodobnostní modely rozdělení sledované veličiny umožňují jednoduchou aproximaci často značně komplikovaného výběrového rozdělení. Typ modelu, použitého v praxi, by měl pokud možno vyplývat logicky z charakteru sledovaného znaku nebo z dlouhodobé zkušenosti s jeho chováním. V těch případech, kdy logická kritéria chybějí, používáme k modelování tu teoretickou distribuční funkci, která maximalizuje shodu empirického a teoretického rozdělení četností. Ke konstrukci pravděpodobnostních modelů lze přistupovat dvěma principiálně odlišnými způsoby. Můžeme

- použít některé *známé rozdělení*, jako je rozdělení logaritmicko-normální, Pareto, Weibullovo, gamma apod. (viz Johnson – Kotz – Balakrishnan, 1994),

[#] Článek je zpracován jako jeden z výstupů výzkumného projektu *Získávání a modelování (lékařských) znalostí (ZIMOLEZ)* registrovaného u grantové agentury FRVŠ pod evidenčním číslem 2C/06/019.

* RNDr. Jitka Bartošová, Ph.D. – odborná asistentka; Katedra managementu informací, Fakulta managementu VŠE v Praze, bartosov@fm.vse.cz .

- aproximovat empirické rozdělení některou teoretickou křivkou z Pearsonova, či Johnsonova systému nebo namodelovat jeho tvar vhodnou rychle konvergující řadou (viz Edgeworth, 1907, s. 102–106).

Při aproximaci empirického rozdělení pomocí známých rozdělení (parametrických modelů) si můžeme vybrat jednu ze dvou cest vedoucích k cíli. Můžeme využít hodnoty

- distribuční funkce $F(x)$ (případně její derivace),
- kvantilové funkce $Q(p)$ (popřípadě její derivace).

Používané parametrické modely jsou definovány následovně. Jestliže $X_1, \dots, X_n$ jsou nezávislé, stejně rozdělené náhodné veličiny s touž distribuční funkcí $F = F_\theta$, resp. kvantilovou funkcí $Q = Q_\theta$, pak jednoparametrickým modelem s parametrem θ je myšlen funkcionál

$$F = \{F_\theta | \theta \in \Theta \subseteq R\}, \text{ resp. } Q = \{Q_\theta | \theta \in \Theta \subseteq R\} \quad (1)$$

kde

F_θ	–	distribuční funkce s parametrem θ , kde θ je parametr modelu,
Q_θ	–	kvantilová funkce s parametrem θ ,
Θ	–	otevřená a konvexní množina všech parametrů.

Tuto definici lze jednoduše zobecnit rovněž pro víceparametrické modely s vektorem parametrů $\vec{\theta} = (\theta_1, \dots, \theta_k)$.

1.1 Modelování pomocí generalizovaných systémů

Snaha o optimální aproximaci reálného empirického rozdělení vedla postupně k vytvoření několika teoretických (generalizovaných) systémů, které umožňují aproximaci mnoha empirických rozdělení. Historickou prioritu ve vytvoření „univerzálního systému“ teoretických rozdělení, vhodných k modelování široké třídy empirických rozdělení, lze pravděpodobně přisoudit Pearsonovi (viz Pearson, 1895, s. 343–414). Všechny křivky z Pearsonova systému jsou prvky z množiny kořenů diferenciální rovnice

$$\frac{dy}{dz} = \frac{y(z - b_1)}{b_0 + b_1 z + b_2 z^2} \quad (2)$$

kde

$y = f(z)$	–	hustota rozdělení veličiny $Z = X - \mu$, kde $\mu = E(X)$,
b_0, b_1, b_2	–	konstanty závislé na momentových charakteristikách rozdělení.

Tvar řešení diferenciální rovnice, tj. výsledné křivky, závisí na konstantách b_0, b_1 a b_2 .

Pearsonův systém má některé přednosti, ale i nedostatky. Hlavní předností systému je jeho univerzálnost – velmi často existuje křivka, která zajistí pro praxi dostačující shodu modelu s empirickým rozdělením. Naopak jeho velkou nevýhodou je, že odhad parametrů modelu je produktem málo vydatné a nerobustní momentové metody. Další nevýhodou systému je jeho formálnost – pro parametry jednotlivých křivek bychom jen těžko hledali věcnou interpretaci.

Jiný alternativní přístup k modelování rozdělení je založen na transformaci sledované náhodné veličiny na veličinu s normálním rozdělením. Takovou polynomiální transformaci náhodné proměnné navrhl Edgeworth. Na tomto principu je založen

Johnsonův systém křivek (viz Johnson, 1949, s. 149–176), který tento princip rozvíjí tak, aby vedl k „univerzálnímu“ modelu. Johnsonův systém je tvořen třemi typy křivek – S_L , S_B a S_U . Pro nás zajímavé je to, že křivka typu S_L je shodná s tříparametrickým logaritmicko-normálním rozdělením.

Dalším prostředkem, používaným při modelování empirického rozdělení, jsou rychle konvergující nekonečné řady. Prvním členem řady je vždy tzv. vytvářející funkce, další členy jsou tvořeny jejími derivacemi. Např. *řada Gram-Charlierovu typu A*, která byla poprvé publikována v r. 1905, je tvořena násobky hustoty pravděpodobnosti normovaného normálního rozdělení a jejich derivací. Platí

$$f(u) = a_0\varphi(u) + a_1\varphi'(u) + a_2\varphi''(u) + \dots + a_n\varphi^{(n)}(u) + \dots \quad (3)$$

kde $f(u)$ – hustota rozdělení standardizované veličiny $U = X - \mu$,
kde $\mu = E(X)$,
 $\varphi(u)$ – hustota normovaného normálního rozdělení,
 $\varphi^{(n)}(u)$ – n -tá derivace hustoty normovaného normálního rozdělení $\varphi(u)$,
 $a_n, n=0,1,\dots$ – konstanty modelu.

Při konstrukci modelu se obvykle omezujeme pouze na první tři členy rozvoje. Uvedený model má však jednu velmi nepříjemnou vlastnost – u nesymetrických rozdělení dospíváme na jejich okraji často k záporným výsledkům hustoty pravděpodobnosti. Situace se zlepší, vyjdeme-li z jiného rozdělení, např. z Poissonova rozdělení. Pokud za vytvářející funkci vezmeme hustotu pravděpodobnosti Poissonova rozdělení, dostaneme *Poisson-Charlierovu řadu typu B* (viz Johnson – Kotz – Balakrishnan, 1994). Rozvoje v řady poskytují zvláště vhodný prostředek k vystižení rozdělení četností směsí, kde každá složka má charakteristiky odlišné od ostatních. V takových případech nelze docílit pomocí Pearsonových křivek takové přizpůsobivosti modelu rozdělení, jako pomocí rozvoje.

1.2 Modelování pomocí známých rozdělení

Schopnost citlivě vystihnout empirické rozdělení některým známým parametrickým modelem souvisí s počtem jeho parametrů. Velký počet parametrů však nejenom zvyšuje pružnost přizpůsobení, ale také nestabilitu modelu, takže model je zároveň vysoce citlivý na chyby v datových souborech. Takovým „univerzálním“ pravděpodobnostním modelem empirického rozdělení četností je např. *generalizované lambda rozdělení (GLD)*. Generalizované lambda rozdělení se čtyřmi parametrické (RS GLD), navržené Rambergem a Schmeiserem (viz Ramberg – Schmeiser, 1974, s. 78–82), je dáno kvantilovou funkcí

$$Q(p; \lambda_1, \lambda_2, \lambda_3, \lambda_4) = \lambda_1 + \frac{p^{\lambda_3} - (1-p)^{\lambda_4}}{\lambda_2} \quad \text{pro } p \in (0;1) \quad (4)$$

$$= 0 \quad \text{jinak}$$

kde p – kvantil rozdělení,
 $\lambda_1, \dots, \lambda_4$ – parametry modelu.

Tento způsob konstrukce rozdělení příjmů, využívající kvantilovou funkci, nalezneme např. v práci Sipkové (viz Sipková, 2005, s. 90–164).

Velkou schopnost přizpůsobení má rovněž čtyřparametrická varianta *logaritmicko-normálního modelu*, která je dána hustotou

$$f(x; \mu, \sigma^2, \gamma, \tau) = \frac{1}{\sigma\sqrt{2\pi}} \frac{\tau - \gamma}{(x - \gamma)(\tau - x)} e^{-\frac{\left(\ln \frac{x - \gamma}{\tau - x} \mu\right)^2}{2\sigma^2}} \quad \text{pro } x \in (\gamma, \tau) \quad (5)$$

$$= 0 \quad \text{jinak}$$

kde μ – střední hodnota veličiny $Y = \ln[(X - \gamma)/(\tau - X)]$,
 σ^2 – rozptyl veličiny Y ,
 γ – teoretické minimum veličiny X ,
 τ – teoretické maximum veličiny X .

2. Modelování rozdělení příjmů obyvatelstva

Nejčastěji používaným modelem rozdělení příjmů byl *tříparametrický logaritmicko-normální model*, jehož hustotu rozdělení získáme ze vztahu (5) po dosažení $\tau = 0$. Největším „soupeřem“ logaritmicko-normálního rozdělení při modelování rozdělení příjmů obyvatelstva je Paretovo rozdělení. Logaritmicko-normální model odpovídá empirickému rozdělení příjmů v rozsáhlé centrální oblasti, zatímco na okrajích se někdy značně odchyluje. Naproti tomu Paretova křivka je vhodným modelem rozdělení příjmů na okrajích (viz Johnson – Kotz – Balakrishnan, 1994).

2.1 Logaritmicko-normální model rozdělení příjmů v ČR v roce 1996

K odhadu parametrů μ , σ^2 a γ tříparametrické varianty logaritmicko-normálního modelu rozdělení byla použita metoda maximální věrohodnosti (parametry μ a σ^2) v kombinaci s odhadem teoretického minima (parametr γ) prostřednictvím numerické minimalizace věrohodnostního poměru (viz Bartošová, 2005, s. 150–155). Důvodem této volby byla snaha o maximální kvalitu zkonstruovaných modelů. Odhady byly prováděny iteračně v programu MatLab.

K posouzení kvality zkonstruovaných modelů zde byl použit věrohodnostní poměr v porovnání s 95% kvantilem rozdělení χ^2 s $(k - 4)$ stupni volnosti, kde k je doporučený počet tříd, do nichž jsou data při výpočtu věrohodnostního poměru rozdělena. Důvodem této volby byla skutečnost, že věrohodnostní poměr je založen na stejném principu jako odhady parametrů modelu – na principu maximální věrohodnosti. Zvolená testovací statistika byla použita k získání informace o tom, ve kterých sociálních skupinách je třída logaritmicko-normálních modelů pro vystižení rozdělení příjmů zcela nevhodná (viz Bartošová, 2006, s. 215–222. Pravděpodobnostní model byl v této práci považován za vyhovující, pokud splňoval podmínku „přibližné shody“ s empirickým rozdělením, tj. pokud hodnota věrohodnostního poměru výrazně nepřevyšovala kritickou hodnotu (kvantil $\chi^2_{0,95}(k - 4)$). Dosažené výsledky jsou uvedeny v tabulce 1.

Tab. č. 1: Míra shody tříparametrického logaritmicko-normálního modelu s empirickým rozdělením příjmů domácností v ČR (r. 1996)

Sociální skupina	Rozsah skupiny	Věrohodnostní poměr		Kritická hodnota
		na domác.	na hlavu	
Samostatně hospodařící	131	31,042	22,699	26,296
Družstevní rolníci	195	14,784	11,026	30,144
Ostatní	236	25,315	44,394	31,410
Nezaměstnaní	260	23,024	17,760	32,671
Důchodci s EA členy	1156	21,366	84,804	54,572
Samostatně činní	1748	77,134	55,848	62,830
Zaměstnanci	6 915	109,352	104,240	103,010
Důchodci bez EA členů	8651	3698,544	1519,259	110,898
Dělníci	8856	232,846	163,102	112,022
Všechny domácnosti	28148	3224,410	2534,232	170,807

Zdroj: Mikrocensus 1996 + vlastní výpočty

Závěr

Nejdůležitějším praktickým výsledkem provedeného zkoumání je zjištění, že dříve běžně používaný pravděpodobnostní model rozdělení příjmů (logaritmicko-normální model se třemi parametry) lze ve většině sociálních skupin považovat za vyhovující i po roce 1990. Jako zcela nevhodný se tento model jeví pouze v případě domácností důchodců bez ekonomicky aktivních členů a všech domácností. Odlišnost rozdělení příjmů domácností důchodců bez EA členů lze odůvodnit rozdílností zdroje příjmů – důchody. A protože domácnosti důchodců tvoří prakticky celou jednu třetinu zkoumaného souboru, budou výrazně ovlivňovat i tvar rozdělení všech domácností. Hledání adekvátního modelu rozdělení příjmů domácností v těchto dvou odlišných případech již není náplní tohoto článku.

Literatura

- [1] BARTOŠOVÁ, J., 2005: Logarithmic-Normal Model of Income Distribution in the Czech Republic. *Austrian Journal of Statistics*, 2006, roč. 35, č. 2&3, s. 215–222.
- [2] BARTOŠOVÁ, J., 2005: Maximal Likelihood Estimates of Parameters of Model of Household Income Distribution in the Czech Republic. *Forum Statisticum Slovacum*, 2005, č. 3, s. 150–155.
- [3] EDGEWORTH, F. Y., 1907: On the representation of statistical frequency by a series. *Journal of the Royal Statistical Society, Series A*, 1907, roč. 70, s. 102–106.
- [4] JOHNSON, N. L., 1949: Systems of frequency curves generated by methods of translation. *Biometrika*, 1949, roč. 36, s. 149–176.

- [5] JOHNSON, N. L. – KOTZ, S. – BALAKRISHNAN, N., 1994: *Continuous Univariate Distributions*, vol. 1, 2nd edition. New York, J. Wiley, 1994.
- [6] PEARSON, K., 1895: Contributions to the mathematical theory of evolution. II. Skew variations in homogenous materia. *Philosophical Transactions of the Royal Society of London*, Series A, 1895, roč. 186, s. 343–414.
- [7] RAMBERG, J. – SCHMEISER, B., 1974: An approximate method for generating asymmetric random variables. *Communications of the ACM*, 1974, roč. 17, č. 2, s. 78–82.
- [8] SIPKOVÁ, L., 2005: Modelovanie príjmov domácností zovšeobecneným lambda rozdelením. *Ekonomika a informatika*, 2005, roč. 3, č. 1, s. 90–164.

Pravděpodobnostní model rozdělení příjmů v České republice

Jitka Bartošová

Abstrakt

K pravděpodobnostnímu modelování lze přistupovat několika principiálně odlišnými způsoby. Jednou z možností, vedoucích k cíli, je aproximace empirického rozdělení některým známým rozdělením (parametrickým modelem). Předložený příspěvek se zaměřuje jednak na popis metod, které lze obecně v praxi k modelování využít, a jednak na konstrukci a ověření platnosti jednoho z dosud používaných modelů rozdělení příjmů v České republice – logaritmicke-normálního modelu se třemi parametry.

Klíčová slova: platnost modelu; pravděpodobnostní model; rozdělení příjmů.

Probability Model of Income Distribution in the Czech Republic

Abstract

Probability modeling may be approached in several principally different ways. One of such possibilities to achieve the aim is approximation of empirical distribution by means of an already known distribution (i.e. a parametric model). This paper focuses first on description of methods that may be used for modeling in practice and further on construction and verification of validity of one of present models of income distribution in the Czech Republic – logarithm-normal model with three parameters.

Key words: income distribution; probability model; validity of the model.

JEL classification: G30