# Neparametrický odhad rozdělení doby nezaměstnanosti [#]

***Ivana Malá*** [*]

Nezaměstnanost je problémem, se kterým se potýká řada států. Při analýzách příčin nezaměstnanosti se obvykle zpracovávají agregované údaje, a to buď průřezová data, nebo data ve formě časových řad popisujících míru nezaměstnanosti. Kromě těchto údajů je zřejmě důležité zajímat se také o délku nezaměstnanosti či o šanci na nalezení práce v závislosti na předcházející délce nezaměstnanosti, pohlaví, vzdělání či jiných charakteristikách nezaměstnaného. V dalším textu je konstruován neparametrický odhad rozdělení této délky zvlášť pro ženy a pro muže. Jako ilustrace pozitivního vlivu dosaženého vzdělání na dobu nezaměstnanosti, je pro obě pohlaví uveden také odhad pro nezaměstnané s úplným středním vzděláním. Použitá data pocházejí z Výběrového šetření pracovních sil (dále VŠPS), které organizuje čtvrtletně Český statistický úřad a zahrnují roky 2000–2004. Jsou podrobně popsána v práci [2].

## 1. Úvod

Výběrové šetření pracovních sil pracuje s náhodně vybranými domácnostmi tvořícími rotující panel, ve kterém se každé čtvrtletí pětina domácností obměňuje. Počet domácností s nezaměstnanými, které je možno sledovat po maximální možné období jeden rok (5 návštěv), je poměrně malý. Proto byli do výběru z databáze ČSÚ zařazeni všichni nezaměstnaní vždy z první návštěvy domácnosti a jejich postavení bylo zaznamenáno ještě v nejbližším dalším šetření. Z těchto nezaměstnaných byly dále vybrány pouze osoby, které práci neúspěšně hledají po dobu kratší dvou let a jejich věk v době konání šetření není větší než 65 let. Popsaným způsobem byli do analýzy zahrnuti nezaměstnaní z období 1. čtvrtletí 2000 až 3. čtvrtletí 2004. Údaje 6 141 takto vybraných osob byly (bez ohledu na časovou charakteristiku šetření) použity pro konstrukci odhadu.

V databázi VŠPS není (až na výjimky) uvedeno datum počátku hledání práce. Doba trvání nezaměstnanosti je při každém šetření zaznamenána pomocí kódu představujícího některý z intervalů (v měsících, vždy intervaly uzavřené vpravo) 0 až 1, 1 až 3, 3 až 6, 6 až 12, 12 až 18, 18 až 24, 24 až 48, nad 48, osoby s délkou větší než 2 roky byly ale vyloučeny.

Pokud by u nezaměstnaných byla známa přesná data počátku, případně konce doby hledání zaměstnání, po druhém šetření u všech sledovaných osob byla zaznamenána

[#] Článek byl zpracován jako jeden z výstupů výzkumného projektu Analýza faktorů ovlivňujících dobu do znovuzískání zaměstnání v ČR, IGA VŠE Projekt IG 410043.

[*] RNDr. Ivana Malá, CSc – odborná asistentka; Katedra statistiky a pravděpodobnosti, Fakulta informatiky a statistiky VŠE v Praze, malai@vse.cz.

buď necenzorovaná pozorování doby nezaměstnanosti (v případě, že nezaměstnaný práci během uplynulých 3 měsíců našel) nebo zprava cenzorovaná pozorování, pokud k zaměstnání nedošlo. Vzhledem k tomu, že je uveden jen interval vymezující dobu hledání práce, jde v podstatě o dvojité cenzorování. Po využití všech informací obsažených v databázi o délce nezaměstnanosti byly pro každého nezaměstnaného nalezeny časové meze $(L, R\rangle$ ve kterých k nalezení práce došlo, případně hodnota $L$ charakterizující dobu nezaměstnanosti, pokud k nalezení práce zatím nedošlo.

Ke zkoumání délky nezaměstnanosti lze použít různé statistické postupy. V případě dat pocházejících z VŠPS je možné aplikovat parametrický regresní model, jak je podrobně popsáno v [2]. Tato práce obsahuje analýzu stejného souboru dat, která jsou použita i v tomto článku. Při použití regresního modelu lze využít široce implementované procedury pro odhady modelu, testování hypotéz či regresní diagnostiku. Na druhé straně je třeba učinit předpoklad o pravděpodobnostním rozdělení zkoumané náhodné veličiny a chybný předpoklad může vést k matoucím výsledkům. Proto je pro orientaci a srovnání možno využít i neparametrický odhad. Stručný popis modifikace Kaplanova-Meierova odhadu běžně používaného pro zprava cenzorovaná data na intervalově cenzorovaná data z VŠPS je obsahem následující části.

## 2. Neparametrický odhad funkce přežití

Cenzorovaná data se vyskytují v situacích, kdy se zkoumá doba do určité události. V našem případě je sledovanou událostí znovunalezení práce. Označíme-li tuto dobu $T$, pak její pravděpodobnostní rozdělení můžeme popsat funkcí přežití $S$, která je v čase $t$ definována jako pravděpodobnost, že nezaměstnaný v tomto čase stále zaměstnání nenalezl (tedy $S(t) = P(T > t) = 1 - F(t)$, kde $F$ je distribuční funkce náhodné veličiny $T$). Pokud by pozorované hodnoty byly pouze přesné hodnoty délky nezaměstnanosti (v případě nalezení práce) a data zprava cenzorovaná (v případě, že nezaměstnaný práci dosud nenalezl), bylo by možné použít Kaplanův-Meierův odhad. Tento postup konstruuje odhad funkce přežití v časových bodech ti, ve kterých došlo aspoň k jedné zkoumané události a je definován jako

$$\begin{aligned} S(t) &= 0 && t < t_1 \\ &= \prod_{t_i \le t} \left(1 - \frac{d_i}{Y_i}\right) && t \ge t_1, \end{aligned} \tag{1}$$

kde $t_i$ – uspořádané okamžiky událostí $t_1 \le t_2 \le \ldots$
$t$ – časový okamžik, $t > 0$
$Y_i$ – počet sledovaných osob v čase $t_i$ (počet osob v riziku, tj. těch, kteří ještě práci nenašli a jejich doba hledání zaměstnání dosud nebyla zprava cenzorována),
$d_i$ – počet událostí v čase $t_i$.

Počet osob $n$ ve výběru (čas $t = 0$) je označen $Y_0$. Odhad funkce přežití je po částech konstantní se skoky v bodech $t_i$.

V případě dat intervalově cenzorovaných lze postup (1) modifikovat (viz [4]). Pro každého nezaměstnaného ve výběru budeme uvažovat intervalové cenzorování v intervalu $(L_i, R_i\rangle$, v případě, že práci nalezl a hodnotu zprava cenzorovanou v čase $L_i$,

pokud práci nenalezl. Nechť množina uspořádaných časových okamžiků $\tau_j$ ($j = 0, .., k$) ($0 = \tau_0 < \tau_1 < ... < \tau_k = \infty$) je tvořena všemi časovými okamžiky, které se vyskytují v hodnotách $L_i$ a $R_i$. Odhad funkce přežití se konstruuje konstantní na intervalech definovaných body $\tau_j$. Označíme jako $p_j$ pravděpodobnost, že k nalezení práce dojde v intervalu $(\tau_{j-1}, \tau_j\rangle$. Součet všech k hodnot těchto pravděpodobností se rovná 1 a jejich znalost je ekvivalentní znalosti odhadu funkce $S$. V popsaném případě nejsou v okamžicích $\tau_j$ známy hodnoty $Y$ a $d$ z (1), lze je však aproximovat jako

$$d_j = \sum_{i=1}^{n} \frac{\alpha_{ij} p_j}{\sum_{j=1}^{k} \alpha_{ij} p_j}, \qquad Y_j = \sum_{p=j}^{k} d_p , \tag{2}$$

kde pro každé pozorování $i$ ($i = 1,.,n$) platí pro $j = 1,.,k$

$$\begin{aligned} \alpha_{ij} &= 1 \quad \text{pokud } (\tau_{j-1}, \tau_j\rangle \subseteq (L_i, R_i\rangle \\ &= 0 \quad \text{jinak.} \end{aligned} \tag{3}$$

Ve vzorci (3) klademe pro nezaměstnané, kteří práci nenašli $R_i = \infty$. Na takto upravená „pseudo-data“ lze použít Kaplanův-Maierův odhad $S$ podle (1) a zpětně pak vypočítat opravené hodnoty pravděpodobností $p_i$. Vyjdeme-li z nějakého počátečního odhadu vektoru $\boldsymbol{p}$, je výše popsán iterační postup, který postupně zpřesňuje odhady pravděpodobností jednotlivých intervalů tak dlouho, až se jejich hodnoty ustálí. Dosažené odhady pak definují hledaný odhad funkce přežití.

Odhad nalezený tímto iteračním postupem je totožný s maximálně věrohodným odhadem vektoru $(p_1, p_2, .., p_k)$ (viz [1], [4]) a je tedy možné při zkoumání vlastností využít výsledky známé pro maximálně věrohodné odhady. Oba postupy vyžadují iterační hledání – první nalezení limitních hodnot pravděpodobností, druhý hledání vázaného extrému (maxima) věrohodnostní funkce. K výpočtu odhadů byly použity programy SPlus 6.2 a Excel.

## 3. Aplikace na délku nezaměstnanosti

Analyzovaná data jsou velmi silně cenzorovaná, neboť ze způsobu výběru vyplývá, že ve výběru jsou obsaženi spíše déle nezaměstnaní a velmi málo osob nalezne během tří měsíců mezi dvěma šetřeními práci (tabulka 1). Procento úspěšných při hledání zaměstnání se pohybuje mezi 33 procenty pro muže vysokoškoláky a 12 procenty pro osoby s pouze základním vzděláním (pro muže i pro ženy). Pokud nebudeme uvažovat dělení podle vzdělání, dostáváme 20 procent pro muže a 18 procent pro ženy. Z celkového počtu sledovaných osob nalezlo zaměstnání 1161, z toho do tří měsíců 190 (16 %, z celkového počtu pak 3 %) a do šesti měsíců 627 (54 %, z celkového počtu 10 %). Kromě problémů s cenzorováním (malým počtem úspěšných uchazečů) lze tedy také očekávat, že odhady délky hledání práce jsou nadhodnocené.

**Tab. č. 1: Četnosti skupin nezaměstnaných podle pohlaví a vzdělání**

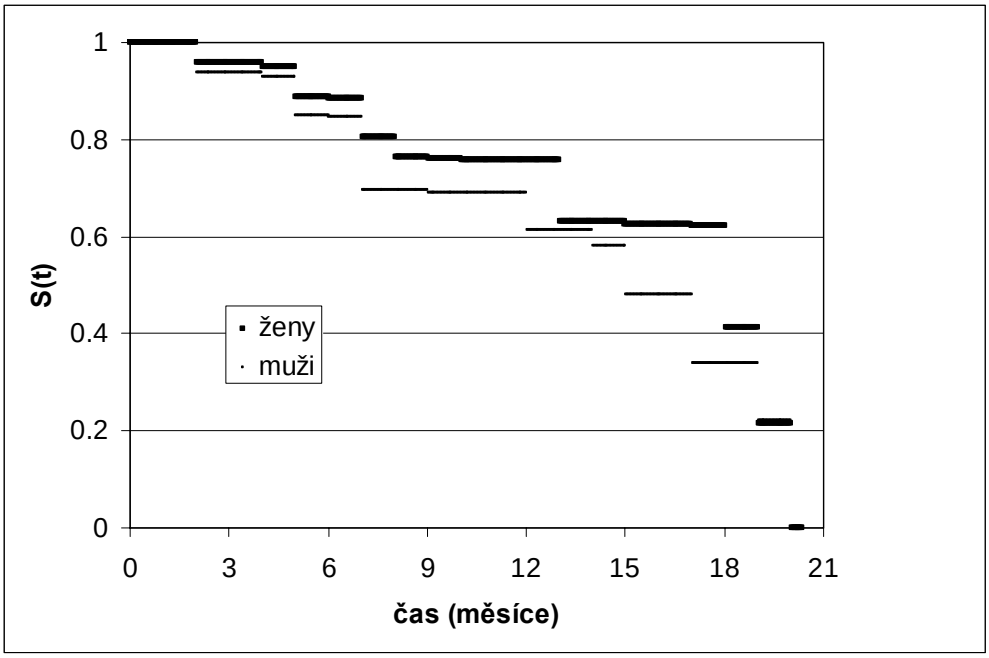
| vzdělání | Počet (% nalezení muži) | Počet (% nalezení ženy) |
|---|---|---|
| **základní** | 558 (12 %) | 681 (12 %) |
| **střední** | 2118 (22 %) | 2557 (19 %) |
| **vysokoškolské** | 114 (33 %) | 113 (31 %) |
| **celkem** | 2790 (20 %) | 3351 (18 %) |

V tabulce 2 jsou uvedeny hodnoty odhadů funkce přežití pro všechny muže a pro všechny ženy. Vzhledem k tomu, že vzdělání příznivě ovlivňuje šance na rychlé znovunalezení práce, v tabulce 2 jsou dále uvedeny hodnoty *S* pro osoby s úplným středoškolským vzděláním. Toto vzdělání bylo voleno jako příklad, neboť skupiny středoškoláků jsou dostatečně obsazeny pozorováními (1122 pro ženy, resp. 574 pro muže) a lze již pozorovat vliv dosaženého vzdělání. Vzhledem k nutnosti popsat funkce přežití pro různé skupiny, je tabulka rozdělena do řádků po jednom měsíci – hodnota *t* znamená levý konec intervalu, ke kterému se řádek vztahuje. První dva sloupce tabulky jsou zobrazeny na grafu 1. Funkce přežití je zakreslena jako po částech lineární. Je vidět, průběh funkce přežití pro muže je stále pod grafem pro ženy, a tedy že ženám trvá hledání zaměstnání déle než mužům. Z grafu lze také velmi přibližně odhadnout medián doby nezaměstnanosti kolem 15 měsíců pro muže a 19 pro ženy.

**Tab. č. 2: Hodnoty funkce přežití *S***

| *t* | muži | ženy | muži – SŠ | ženy – SŠ |
|---|---|---|---|---|
| **0** | 1 | 1 | 1 | 1 |
| **1** | 0,999 | 0,999 | 1 | 0,998 |
| **2** | 0,999 | 0,999 | 0,930 | 0,998 |
| **3** | 0,937 | 0,958 | 0,930 | 0,949 |
| **4** | 0,930 | 0,949 | 0,927 | 0,937 |
| **5** | 0,849 | 0,888 | 0,815 | 0,881 |
| **6** | 0,848 | 0,886 | 0,815 | 0,879 |
| **7** | 0,697 | 0,806 | 0,581 | 0,710 |
| **8** | 0,697 | 0,764 | 0,581 | 0,710 |
| **9** | 0,691 | 0,760 | 0,581 | 0,706 |
| **10** | 0,691 | 0,759 | 0,581 | 0,706 |
| **11** | 0,691 | 0,759 | 0,581 | 0,705 |
| **12** | 0,614 | 0,758 | 0,581 | 0,701 |
| **13** | 0,614 | 0,630 | 0,480 | 0,560 |
| **14** | 0,580 | 0,630 | 0,480 | 0,560 |
| **15** | 0,480 | 0,625 | 0,470 | 0,560 |
| **16** | 0,480 | 0,625 | 0,470 | 0,560 |
| **17** | 0,340 | 0,622 | 0,470 | 0,560 |
| **18** | 0,340 | 0,621 | 0,470 | 0,510 |
| **19** | 0,220 | 0,215 | 0 | 0,220 |
| **20** | 0 | 0 | 0 | 0 |

**Graf č. 1: Graf funkce přežití *S***




# 4. Závěr

Výsledky dosažené metodami obsaženými v této práci jsou srovnatelné s výsledky publikovanými v [2] a dosaženými pomocí regresních metod. Velké směrodatné odchylky pozorované [2] v případě odhadů kvantilů jsou v případě neparametrických odhadů také velké, v některých případech nelze hodnoty odchylek odhadů *S* rozumně vyčíslit. Proto také v této práci nejsou uvedeny (ani způsob jejich konstrukce využívající teorii maximálně věrohodných odhadů) a výsledky obsahují pouze bodové odhady pravděpodobností sledovaných časových intervalů a funkce přežití.

Parametrické modely využívající známé rozdělení sledovaných veličin jsou velmi flexibilní a běžně implementované v statistických programech. Neparametrický odhad nicméně poskytuje představu o rozdělení doby nezaměstnanosti a lze ho použít pro porovnání s odhady dosaženými jinými metodami zvláště v případě, že není spolehlivá představa o pravděpodobnostním rozdělení, které by pro sledovanou dobu do výskytu jevu mělo být použito.

# Literatura

[1] FINKLSTEIN, D. M.: A proportional hazard model for interval-censored failure time data. *Biometrics*, 1986, Vol. 42, č. 4, s. 845–854.

[2] JAROŠOVÁ, E.: Modelování délky trvání nezaměstnanosti. *Statistika*. 2006, roč. 86, č. 3, s. 240–251.

[3] KLEIN J. P. – MOESCHBERGER, M. L.: *Survival Analysis*. Techniques for Censored and Truncated Data, Springer, 1998.

[4] LAWLESS, J. F.: *Statistical models and methods for lifetime data*. Wiley, 2003.

# Neparametrický odhad rozdělení doby nezaměstnanosti

***Ivana Malá***

## Abstrakt

Nezaměstnanost je vážný společenský a hospodářský problémem. Příspěvek se zabývá modelováním délky nezaměstnanosti a konstrukcí neparametrického odhadu jejího rozdělení. Na základě dat pocházejících z Výběrového šetření pracovních sil organizovaného čtvrtletně Českým statistickým úřadem jsou uvedeny odhady rozdělení délky nezaměstnanosti pro ženy a muže. Použitá data zahrnují roky 2000–2004.

**Klíčová slova:** neparametrické metody; cenzorovaná data; maximálně věrohodný odhad.

# Nonparametric estimate of the distribution of the time of unemployment

## Abstract

In the article the length of unemployment in the Czech Republic is treated. A nonparametric estimate of its survival function is constructed using maximum likelihood estimation for the groups of men and women and both groups are compared. The positive influence of education on the length of unemployment is illustrated. Data describing the unemployed were gathered by the labour force sample survey organized by the Czech Statistical Office in years 2000–2004.

**Key words:** nonparametric estimation; censored data; maximum likelihood estimate.

**JEL classification:** C24