Pravděpodobnostní model doby setrvání ministra školství ve funkci

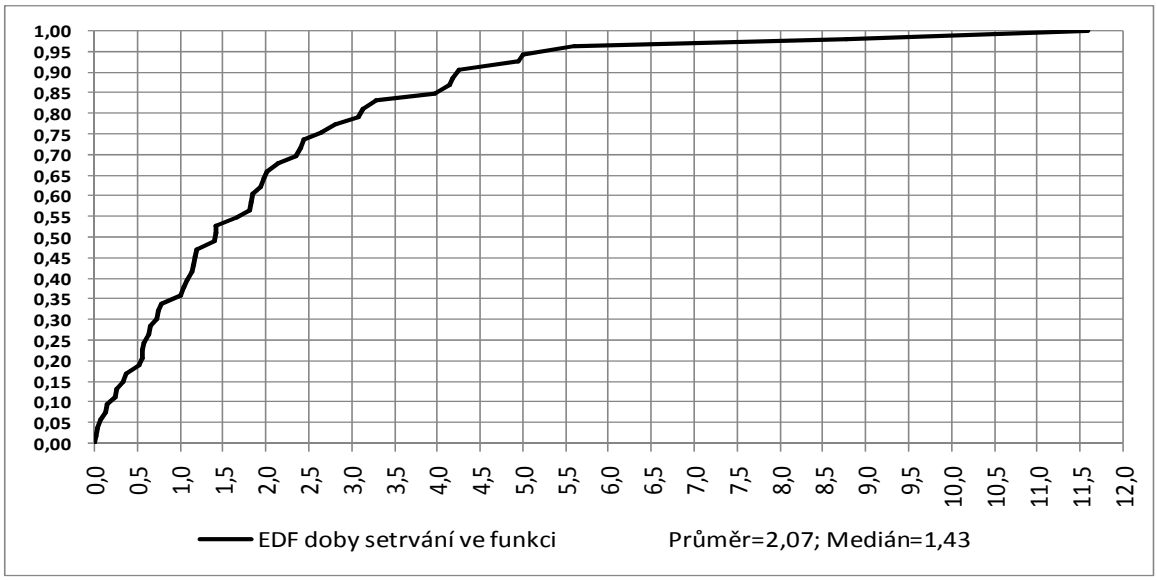
Základní statistická inference

Data

Zdroj: http://www.msmt.cz/ministerstvo/ministri-skolstvi-od-roku-1848. Ke statistickému zpracování byla vzata pozorování od roku (nástupu do funkce) 1900 do 19.7.2010. K datům nelze mít odpovídající statistické námitky. Případné námitky lze mít jen z pohledu historického. Pro přepočet doby trvání v rocích byl použit následující vztah $doba = \frac{T}{365{,}25}$, kde T je počet dnů v postavení ministra bez posledního dne. Případná překrytí (např. den nástupu jednoho je i dnem ukončení funkce druhého, …) byla eliminována (východiskem je den nástupu). Jednotlivé pozorované doby lze chápat jako náhodný výběr. Vzhledem k charakteru dat nelze tento předpoklad ani potvrdit a ani vyvrátit. Celkem bude zpracováváno 53 dob setrvávání ve funkci ministra školství, s uzemní působností v „ZEMÍCH KORUNY ČESKÉ". Data jsou v [1] na listu „Vývoj". Následují základní výběrové statistiky a empirická distribuční funkce.

Výběrová charakteristika	Počet let ve funkci	EDF distribuce X	EDF doby setrvání ve funkci, hodnoty Y
Minimum	0,014	0,000	0,0000
Průměr=	2,066	2,027	0,5000
Medián=	1,426	1,414	0,5000
Maximum	11,578	11,578	1,0000
StD	2,171	2,169	0,2968
25% kvantil	0,630	0,595	0,2500
75% kvantil	2,661	2,610	0,7500
Průměrná abs. odchylka	1,496	1,494	0,2547
Počet	53	54	54

Tabulka 1.: Výběrové charakteristiky dat a empirické distribuční funkce.



Obr.1.: Průběh empirické distribuční funkce. Na vodorovné ose jsou roky setrvání ve funkci.

Cíl zpracování

Cílem zpracování je navrhnout pravděpodobnostní model, který by umožnil v budoucnosti předvídat dobu setrvání ministra ve funkci.

Parametrické modely

Exponenciální model

Protože se jedná o nezápornou náhodnou proměnnou (kterou je možno považovat za spojitou, při měření doby v rocích setrvání ve funkci) mající charakter trvání nabízí se intuitivně klasické exponenciální rozdělení s distribucí a hustotou:

$$F(x) = 1 - e^{-\frac{x}{\tau}} \text{ a } f(x) = \frac{1}{\tau} e^{-\frac{x}{\tau}} \qquad (1)$$

Pro toto rozdělení platí: $E\{x\} = \tau$ a $\sigma^2\{x\} = \tau^2$. Nejlepším (vydatným) nestranným odhadem parametru τ je průměr pozorování, tedy $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$ [2], str. 188, kde t_i je doba setrvání ve funkci *i*-tého ministra ve funkci. Takový odhad dává: $\hat{\tau}$ =2,066. Srovnání modelové distribuční funkce a empirické distribuční funkce je na následujícím obrázku.

Obr.2.: Průběh modelové a empirické distribuční funkce. Na vodorovné ose jsou roky setrvání ve funkci.

Následuje tabulka srovnání výběrových a modelových parametrů. Pro výpočet mediánu a kvantilů byl užit vztah $1 - e^{-\frac{x_p}{\tau}} = p \Rightarrow x_p = -\tau \lg(1-p)$. *p=0,5;0,25;0,75.* Uvedený vztah platí ale pro libovolné $0 \le p < 1$.

Srovnání výběrových a modelových parametrů

Parametr	Výběrový	Modelový	Diference	Diference/Modelový
Minimum	0,0137	---	---	---
Průměr=	2,0657	2,0657	0,0000	0,00%
Medián=	1,4264	1,4318	-0,0054	-0,38%
Maximum	11,5784	---	---	---
StD	2,1713	2,0657	0,1057	5,11%
25% kvantil	0,6297	0,5943	0,0354	5,96%
75% kvantil	2,6612	2,8636	-0,2025	-7,07%
Průměrná abs. odchylka	1,4957	1,5174	-0,0217	-1,43%
Počet	53	---	---	---

Tabulka 2.: Výběrové a modelové charakteristiky[1].

Srovnání u směrodatné odchylky a 25%, 75% kvantilů jsou nepřesvědčivá. Proto je nezbytné uskutečnit nějaký test shody modelu a pozorovaných dat. Zvolíme χ^2 test dobré shody [3]. Výsledek a výpočet takového testu je shrnut v následující tabulce. Požadujeme nízkou hodnotu chyby 1-ho druhu ve výši 1%. Obdobně viz i test, že se jedná o id výběr (viz list „id“, souboru [1]).

	Intervaly kvantování pro a chi-2 test dobré shody					Výpočet kritéria
j		z_j	$F(z_j)$	p_j	n_j	4,9333
1	z_1=	0,2047	0,09434	0,09434	5	0,0000
2	z_2=	0,4319	0,18868	0,09434	4	0,2000
3	z_3=	0,6873	0,28302	0,09434	6	0,2000
4	z_4=	0,9787	0,37736	0,09434	3	0,8000
5	z_5=	1,3181	0,47170	0,09434	7	0,8000
6	z_6=	1,7244	0,56604	0,09434	4	0,2000
7	z_7=	2,2308	0,66038	0,09434	7	0,8000
8	z_8=	2,9030	0,75472	0,09434	5	0,0000
9	z_9=	3,9059	0,84906	0,09434	3	0,8000
10	z_{10}=	5,9320	0,94340	0,09434	7	0,8000
11	z_{11}=	+∞	1,00000	0,05660	2	0,3333
				$\sum_{j=1}^{M} \frac{(n_j - np_j)^2}{np_j}$ =		4,9333
					M-1=	10
					α=	1,0%
					χ^2(M-1)=	23,2093

Tabulka 3.: Výpočet testu shody.

[1] Modelová střední absolutní odchylka se spočte na základě následujícího postupu. Mějme náhodnou veličinu ξ, která má exponenciální rozdělení s parametrem τ. Potom pro náhodnou veličinu $\eta = |\xi - \tau|$ platí

$$F_\eta(x) = P(\eta < x) = P(|\xi - \tau| < x) = P(\tau - x < \xi < \tau + x) = F_\xi(\tau + x) - F_\xi(\tau - x) =$$
$$= 1 - e^{-\frac{\tau + x}{\tau}} - \max\left(0, 1 - e^{-\frac{\tau - x}{\tau}}\right); x \geq 0.$$

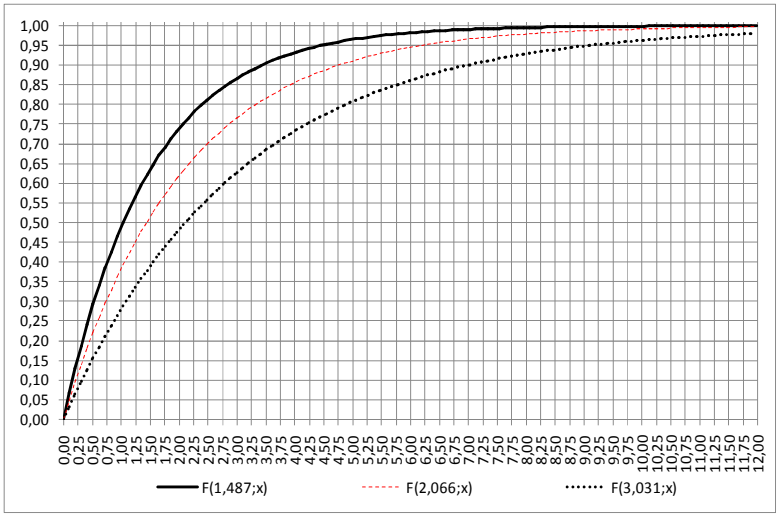
Odtud $E\{\eta\} = \int_0^{+\infty} (1 - F_\eta(x))dx$. Tento integrál lze s přijatelnou přesností spočíst numericky pro dané τ. Viz list „Exponenciální“ souboru [1], sloupce “Distribuce absolutní odchylky a Výpočet střední hodnoty”.

Výpočet je v [1], na listu „Exponenciální“. Hodnota kritéria leží mimo kritický oboru, proto hypotézu o tom, že se doba setrvání řídí rozdělením $F(x)=1-e^{-\frac{x}{\tau}}$, nezamítáme pro odhadnuté τ=2,0657. Proto (hlavně pro výše uvedené pochybnosti „u směrodatné odchylky a 25%, 75% kvantilů“), využijeme zvoleného „tvaru“ rozdělení **avšak ne pro bodový**, ale intervalový odhad parametru τ. Využijeme postup popsaný v [4]. Výpočet je v [1], opět na listu „Exponenciální“. Pro volbu koeficientu spolehlivosti 1-α, použijeme rozklad 1-α=1-(α₁+α₂), kde α=0,01(≡1%); α₁=0,005;α₂=0,005. Výpočet je shrnut v tabulce 4.

Intervalový odhad parametru τ	
$\alpha_1 =$	0,5%
$\alpha_2 =$	0,5%
$n=$	53
$t=$	2,0657
$\tau_d = \frac{nt}{G^{-1}(n,1,1-\alpha_2)} =$	1,4870
$\tau_h = \frac{nt}{G^{-1}(n,1,\alpha_1)}$ [2] $=$	3,0306

Tabulka 4.: Výpočet intervalového odhadu pro parametr τ.

Takovému intervalovému odhadu odpovídá pás distribučních funkcí zobrazený na obr.3. Výpočet je v [1], na listu „Exponenciální, tolerance“.



Obr.3.: Distribuční funkce pro τ_d (černá, plná), $\hat{\tau}$ (červená čárkovaná) a τ_h (černá tečkovaná). Skutečná, ale nedostupná distribuční funkce leží s danou spolehlivostí (99%) mezi oběma černými čarami.

Při tomto pojetí je výrok o budoucí době setrvání ve funkci statistickým výrokem o tolerančním intervalu. Opět užijeme metodiku z [4]. Výpočet je v [1], na listu „Exponenciální“. Výpočet je shrnut v následující tabulce 5.

[2] $G(n,\tau;x) = \int_0^x \frac{z^{n-1}e^{-\frac{z}{\tau}}}{\tau^n (n-1)!} dz$

Intervalový odhad parametru τ		
$\alpha_1 =$	0,5%	
$\alpha_2 =$	0,5%	
$n=$	53	
$t=$	2,0657	
$\tau_d = \frac{nt}{G^{-1}(n,1,1-\alpha_2)} =$	1,4870	
$\tau_h = \frac{nt}{G^{-1}(n,1,\alpha_1)} =$	3,0306	
Toleranční interval pro dobu setrvání ve funkci		
$\beta_1 =$	2,5%	
$\beta_2 =$	2,5%	V měsících
$L(x_1,x_2,...,x_n) = -\left(\sum_{i=1}^{n} x_i\right)\frac{\lg(1-\beta_2)}{G^{-1}(n,1,1-\alpha_2)}$ [rok]=	***0,038***	***0,5***
$U(x_1,x_2,...,x_n) = -\left(\sum_{i=1}^{n} x_i\right)\frac{\lg(\beta_1)}{G^{-1}(n,1,\alpha_1)}$ [rok]=	***11,179***	***134***

$$P_{x_1,x_2,...,x_{53}}\left[P_{doba}\left\{L(x_1,x_2,...,x_{53}) \le \text{doba} \le U(x_1,x_2,...,x_{53})\right\} \ge 0{,}95\right] \ge 0{,}99$$

Tabulka 5.:Tabulka výpočtu toleranční doby setrvání ve funkci.

Shrnuto: S pravděpodobností alespoň 95% setrvá ministr školství ve funkci mezi 0,5 a 134 měsíci. Spolehlivost tohoto výroku (=pravděpodobnost, že bude splněn) je alespoň 99%.

Podkladové soubory a zdroje

[1] ministri skolstvi-studie.xls

[2] Jaroslav Hátle, Jiří Likeš: Základy počtu pravděpodobnosti a matematické statistiky. SNTL Praha 1974.

[3] Přednáška SA1-12n - Neparametrické metody - testy, rozdělení s kategoriálními proměnnými. Str. 2.

[4] Přednáška VSM-3 Intervalové odhady. Metody konstrukce intervalových odhadů.

Příloha 1.: Další empirické důvody pro volbu exponenciálního rozdělení

Výpočetní schéma pro χ^2 test dobré shody využijeme k tomu, že se pokusíme nalézt (metodou pokus-omyl) takové $\hat{\tau}$ (na tři desetinná místa), které minimalizuje hodnotu pravděpodobnosti chyby prvního druhu.

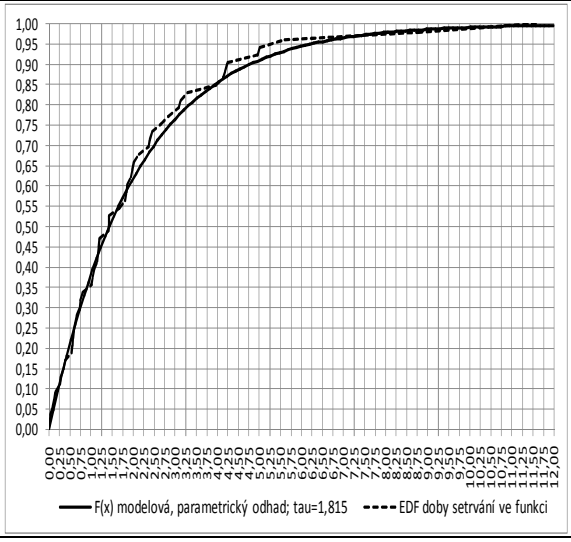
Minimální hodnota kritéria= 0,8000; jí příslušná pravděpodobnost chyby 1-ho druhu= 0,0061%

Minimální $\hat{\tau}$ pro takovou hodnotu= 1,815

Tabulka výpočtu

	Intervaly kvantování pro a chi-2 test dobré shody					Výpočet kritéria
i		z-i	F(z-i)	p-i	n-i	0,8000
1	z-1=	0,1798	0,09434	0,09434	5	0,0000
2	z-2=	0,3795	0,18868	0,09434	4	0,2000
3	z-3=	0,6039	0,28302	0,09434	4	0,2000
4	z-4=	0,8599	0,37736	0,09434	5	0,0000
5	z-5=	1,1581	0,47170	0,09434	5	0,0000
6	z-6=	1,5152	0,56604	0,09434	5	0,0000
7	z-7=	1,9601	0,66038	0,09434	5	0,0000
8	z-8=	2,5507	0,75472	0,09434	6	0,2000
9	z-9=	3,4319	0,84906	0,09434	5	0,0000
10	z-10=	5,2121	0,94340	0,09434	6	0,2000
11	z-11=	∞	1,00000	0,05660	3	0,0000
					Kritérium=	0,8000
					Jeho p-hodnota=	0,006%
					Počet stupňů volnosti=	10
					alfa=	1,000%
					Kritická hodnota=	23,2093

Modelová a empirická distribuční funkce

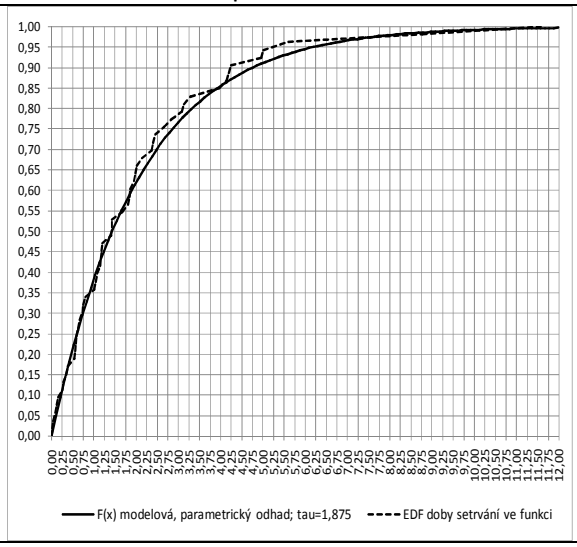


Maximální $\hat{\tau}$ pro takovou hodnotu= 1,853

Tabulka výpočtu

	Intervaly kvantování pro a chi-2 test dobré shody					Výpočet kritéria
i		z-i	F(z-i)	p-i	n-i	0,8000
1	z-1=	0,1836	0,09434	0,09434	5	0,0000
2	z-2=	0,3874	0,18868	0,09434	4	0,2000
3	z-3=	0,6165	0,28302	0,09434	4	0,2000
4	z-4=	0,8779	0,37736	0,09434	5	0,0000
5	z-5=	1,1824	0,47170	0,09434	5	0,0000
6	z-6=	1,5469	0,56604	0,09434	5	0,0000
7	z-7=	2,0011	0,66038	0,09434	6	0,2000
8	z-8=	2,6041	0,75472	0,09434	5	0,0000
9	z-9=	3,5037	0,84906	0,09434	5	0,0000
10	z-10=	5,3212	0,94340	0,09434	6	0,2000
11	z-11=	∞	1,00000	0,05660	3	0,0000
					Kritérium=	0,8000
					Jeho p-hodnota=	0,006%
					Počet stupňů volnosti=	10
					alfa=	1,000%
					Kritická hodnota=	23,2093

Modelová a empirická distribuční funkce



Poznámky

1. Používáme-li bodový odhad, je důležité jakým kritériem nebo vlastností budeme hodnotit jeho „kvalitu“, či k jakému účelu ho budeme dál používat.
2. „Kvalita“ bodového odhadu je při měření jeho *p*-hodnotou dána konstrukcí daného testu.
3. Taková p-hodnota je náhodnou proměnnou (je dána náhodnými pozorováními).
4. Nestrannost a vydatnost parametrického odhadu nemusí být pro některé účely vhodným kritériem pro přijetí takového odhadu.

Výpočet pro popsané empirické testy je v [1] na listu „Exponenciální-ověření“.