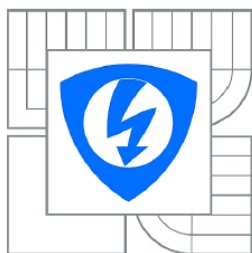


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH
TECHNOLOGIÍ

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF BIOMEDICAL ENGINEERING

PREDIKČNÍ MODELY PRO ANALÝZU PŘEŽITÍ

PREDICTIVE MODELS IN SURVIVAL ANALYSIS

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE

AUTHOR

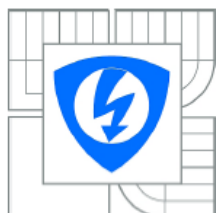
MICHAELA HADWIGEROVÁ

VEDOUCÍ PRÁCE

SUPERVISOR

Ing. HELENA ŠKUTKOVÁ

BRNO 2011



VYSOKÉ UČENÍ
TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

Ústav biomedicínského inženýrství

Bakalářská práce

bakalářský studijní obor
Biomedicínská technika a bioinformatika

Studentka: Michaela Hadwigerová
Ročník: 3

ID: 115092
Akademický rok: 2010/2011

NÁZEV TÉMATU:

Predikční modely pro analýzu přežití

POKYNY PRO VYPRACOVÁNÍ:

Seznamte se s algoritmy pro stanovení odhadu funkce přežití a souvisejících parametrů včetně možností reprezentace výsledků analýzy přežití. Srovnajte parametrické a neparametrické metody pro analýzu přežití. Srovnání dokumentujte na vlastních navržených algoritmech pro odhad funkce přežití realizovaných v prostředí Matlab.

Vytvořte aplikaci umožňující grafickou reprezentaci funkce přežití pomocí těchto základních metod: Kaplan-Meierův odhad, regresní funkce pro lineární a exponenciální modely a metoda nejmenších čtverců. Aplikace bude umožňovat vykreslení křivky přežití s vyznačením a vyčíslením intervalu spolehlivosti a mediánu přežití a bude sloužit jak pro klinické tak populační studie.

DOPORUČENÁ LITERATURA:

- [1] MOTULSKY, H. Intuitive Biostatistics, Oxford: Oxford University Press, 1995.
- [2] ZVÁROVÁ, J. Biomedicínská statistika I - Základy statistiky pro biomedicínské obory, Praha: Karolinum, 2002.

Termín zadání: 7.2.2011

Termín odevzdání: 27.5.2011

Vedoucí práce: Ing. Helena Škutková

prof. Ing. Ivo Provazník, Ph.D.

Předseda oborové rady

UPOZORNĚNÍ:

Autor bakalářské práce nesmí při vytváření bakalářské práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

Abstrakt

Se stále se rozvíjejícími metodami léčby ve zdravotnictví se objevuje i potřeba tyto nové metody nějakým účinným způsobem porovnávat s metodami starými. Je to zejména důležité pro další rozvoj těchto metod. Avšak data, která tyto skutečnosti popisují, nebylo možno obvyklými postupy zpracovávat, a proto došlo v oblasti statistiky k vytvoření nového druhu metod. Jsou známy pod jednotným názvem predikční modely analýzy přežití.

Abstract

With ever-new methods of treatment in health care occurs a requirement of comparing these new methods to the old methods in some effective way. This is particularly important for the further development of these methods. However, data that describe these facts could not be processed by normal procedures and therefore was in the field of statistics to create a new kind of methods. They are known as predictive models of survival analysis.

Klíčová slova:

Analýza, přežití, Kaplan-Meier, Coxův regresivní model, tabulky přežití, funkce přežití, křivka přežití, Statistica, MedClac, cenzorování

Keywords:

Analyzis, survival, Kaplan-Meier, Coxregression model, life-tables, survivalfunction, survivalcurve, Statistica, MedClac, censoring

HADWIGEROVÁ, M. Predikční modely pro analýzu přežití. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2011. 42 s. Vedoucí bakalářské práce Ing. Helena Škutková.

Prohlášení

Prohlašuji, že svou bakalářskou práci na téma Predikční modely pro analýzu přežití jsem vypracoval samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením tohoto projektu jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

V Brně dne

.....
podpis autora

Poděkování

Děkuji vedoucímu bakalářské práce Ing. Heleně Škutkové za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé bakalářské práce.

V Brně dne

.....
podpis autora

Obsah

Úvod	6
1. Analýza přežívání obecně	7
1.1. Cenzorování dat	7
1.2. Funkce přežití	9
2. Odhad funkce přežití	12
2.1. Křivka přežití	12
2.2. Kaplan-Meierův odhad funkce přežití	13
2.3. Odhad pomocí tabulky přežití	15
2.4. Testy shody funkcí přežití	16
2.5. Regresivní model pro analýzu historie událostí	18
3. Programy pro zpracování analýzy přežití	20
3.1. Matlab	20
3.2. Statistica	22
3.3. MedCalc	22
4. Vlastní řešení programu	25
3.1. Lineární regrese	26
3.2. Exponenciální regrese	31
3.3. Srovnání lineární a exponenciální regrese	34
3.4. Kaplan-Meier	35
Závěr	39
Použitá literatura	40
Seznam obrázků	42

ÚVOD

Ve všech oblastech lidského snažení dochází k neustálému a nezadržitelnému pokroku. Jsou vytvářeny nové technologie zpracování surovin, výroby materiálů, staré postupy a technologie jsou nahrazovány novými a lepšími. Tento stále se zrychlující proces s sebou přináší nová rizika. Lidé si proto oprávněně pokládají otázku: Jsou tyto nové technologie opravdu lepší než ty staré? Jsou bezpečnější? A jaká je jejich účinnost? Daly by se tyto nové technologie ještě vylepšit?

Stejná situace nastává i v oblasti medicínského výzkumu a technologií. Neustále jsou objevovány nové látky vhodné pro výrobu léků na nejrůznější nemoci. Jsou vyráběny nové přístroje pro diagnostiku onemocnění i zařízení pro jejich léčbu. A lékaři s pacienty si pokládají podobné otázky: Jsou tyto léky opravdu účinné? Jsou účinnější než léky dosud aplikované na toto onemocnění? Nedochozí po jejich aplikaci ke zhoršení zdravotního stavu pacienta? A právě na tyto otázky se snaží statistické odpovědět použitím analýzy přežití.

Jak se tedy tato statistická metoda přesně uplatňuje? Například při testování účinnosti nového léku jsou účastníci této studie rozděleny do dvou skupin. První skupině je podáván nový lék, zatímco druhé skupině je podáván dosud používaný lék na toto onemocnění. Jsou zaznamenávány počty uzdravených (popřípadě zesnulých) jedinců. Tyto data jsou následně zpracována metodami analýzy přežití. Vypočítané údaje potom můžeme použít na odhad dalšího vývoje událostí nebo na určení střední délky přežití (je to čas kterého se dožije polovina objektů účastnících se výzkumu ve skupině). Je-li tedy hodnota střední délky přežití pro skupinu s novým druhem léku delší než pro skupinu s běžně užívaným lékem, můžeme usoudit, že nový lék má skutečně lepší léčebné účinky než ten starý. Tak dostáváme jasnou představu o účinnosti nového léku, a proto je analýza přežívání skutečným přínosem pro zavádění nových technologií do medicíny a tím i k prodloužení a zkvalitnění našeho života.

Tato práce se věnuje srovnání nejznámějších postupů a metod používaných v analýze přežití. V první části práce se budeme zabývat obecným popisem analýzy dat (funkcí přežití) a jejím cenzorováním. Druhá část bude zaměřena na vlastní metody pro odhad funkce přežití. Ve třetí části se budeme zabývat zpracováním dat pomocí profesionálních počítačových programů. V poslední části je ukázka mnou vytvořeného programu.

1. ANALÝZA PŘEŽÍVÁNÍ OBECNĚ

Analýza přežívání je termín, který se ujal v medicíně pro označení analýzy historie události. Jedná se o statistickou metodu, zabývající se existencí pozorovaných subjektů v čase a změnou jejich stavu. Subjekt začneme pozorovat v určitém čase vstupní události, který označujeme jako tzv. čas počátku (např.: narození, počátek léčby, začátek nemoci, vstup subjektu do studie), do doby než nastane námi sledovaná událost tzv. koncový bod (např.: úmrtí jedince, návrat příznaků nemoci, uzdravení).

S analýzou přežívání se můžeme setkat i v jiných oblastech lidské činnosti, má veliký okruh uplatnění. Například v technických oborech, sociologii, pedagogice a ekonomii. V technických oborech se jedná hlavně o hodnocení životnosti součástek. V sociologii ve studiích týkajících se rozvodovosti, změn zaměstnání, kouření a konzumace drog. V pedagogice se jedná o studie zabývající se průběhem studia a v ekonomii o zjišťování nezaměstnanosti atd.

Metody analýzy přežívání (popř. analýzy časů selhávání u hodnocení životnosti technických součástek) jsou používány ke zvyšování kvality výrobků, k hodnocení účinnosti léčby (pacienti jsou rozděleni do dvou skupin, jedna skupina je léčena standardním postupem, zatímco druhá skupina je léčena novou metodou, zkoumá se, ve které skupině pacienti déle přežívají a jak se mění jejich klinický stav).

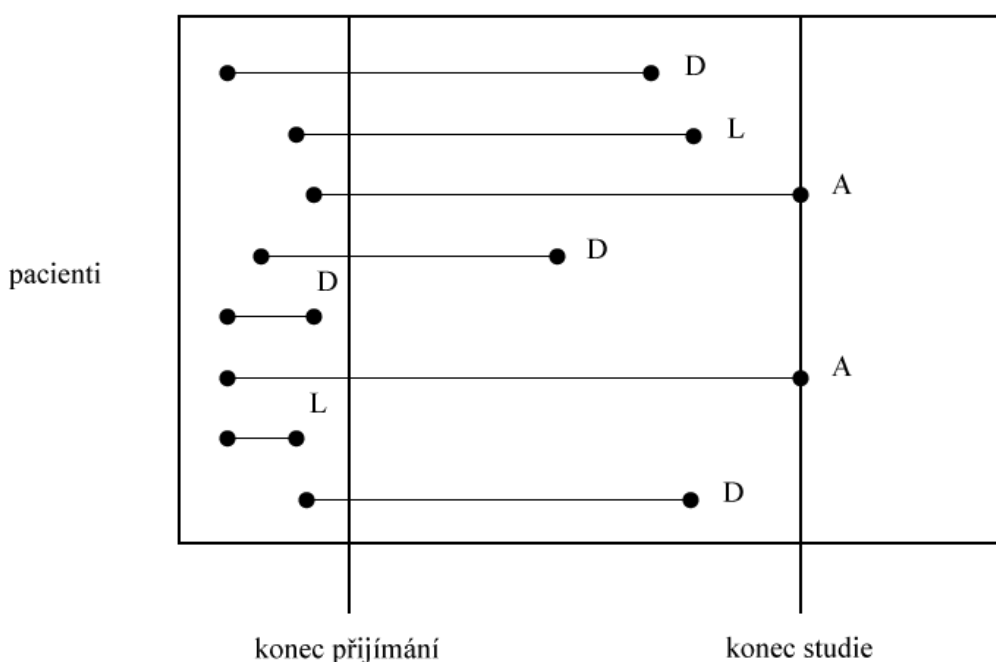
Zpracování dat zabývajících se analýzou přežívání má dva hlavní problémy: data obvykle nemají Gaussovský průběh a nemůžeme začít s analýzou dat pro jednotlivé subjekty, dokud všechny subjekty nedosáhnou koncového bodu (tj. úmrtí). Proto byly vynalezeny speciální metody pro zpracování tohoto typu dat. [6]

1.1.Cenzorování dat

Obecně si pod slovem cenzura nebo **cenzorování** představíme kontrolu a následné omezení nebo okleštění nějakého většího celku dat. Tak je tomu i u analýzy přežívání. Soubor dat je oproštěn od těch, které na jeho zpracování nemají vliv. Nebo přesněji řečeno jsou to data, která obsahují jen část námi požadované informace. Data která dobu pozorování přesahují, doby sledování ani nedosáhnou nebo je z nějakého důvodu přestaneme sledovat a neznáme tudíž jejich přesný vývoj v čase.

Doba přežití je cenzorována v případě, že objekt studie dosáhl koncového bodu mimo pozorování. Jedinec může tohoto bodu dosáhnout před započítáním samotné studie anebo ho dosáhne až po skončení samotné studie. Stejná situace nastává v případě, že s jedincem

ztratíme během studie kontakt a tudíž nevíme, jestli je v době ukončení studie stále na živu. Pokud nastanou sledované události po skončení sledování jedince, jedná se o cenzorování zprava. Skutečná délka života je větší než doba pozorování. U **cenzorování zleva** dochází k úmrtí pacienta, vybraného pro studii, před jejím samotným zahájením. Dalším typem cenzorování je **intervalové cenzorování**. V tomto případě nemůžeme přesně určit skutečnou dobu bez události. Dochází k tomu, když není možné pacienta kontinuálně sledovat. Pacient se účastní studie pouze jednou za určitý časový interval. Pokud v tomto intervalu dojde k události, nejsme schopni přesně odhadnout, v jakém okamžiku mezi poslední prohlídkou a nálezem k této události došlo. Z tohoto plyne, že k vlastní době mezi časem počátku a koncovým bodem musíme navíc přičíst čas, který uběhl mezi koncovým bodem a zjištěním, že námi pozorovaná událost nastala. V praxi se však nejčastěji setkáváme s cenzorováním zprava.



Obrázek 1 - Čas studie pro 8 pacientů [1]

Na obrázku 1 máme příklad, jak může vypadat studie s různou kombinací událostí a počátků pozorování pacienta. Ve studii pozorujeme osm pacientů. Pacienti, které jsme z různých důvodů přestali sledovat, jsme označili písmenem L, pacienti kteří zemřeli (D), ostatní se dožili konce studie (A). Příkladem cenzorování zleva jsou pacienti číslo 5 a 7. Pacient číslo 5 zemřel před započítáním studie a s pacientem číslo 7 jsme z nějakého důvodu

přestali být v kontaktu. V obou těchto případech se jedná o cenzorování zleva. Příklad cenzorování zprava máme u pacientů číslo 2, 3 a 6. S pacientem číslo 2 byl v průběhu studie přerušen kontakt, tudíž nevíme, jestli se konce studie dožil. U pacientů číslo 3 a 6 víme, že se dožili konce studie, ke sledované události v průběhu studie nedošlo. Jen u pacientů číslo 1, 4 a 8 jsme sledovanou událost mohli pozorovat v průběhu studie. U těchto pacientů k žádné formě cenzorování dat nedochází.

V praxi používáme čtyři druhy nezávislého cenzorování:

1. jednoduchý typ I – všichni jedinci jsou sledováni pevnou dobu;
2. progresivní typ I – všichni jedinci jsou cenzorováni ve stejný časový okamžik;
3. typ II – studie trvá do uskutečnění n událostí;
4. náhodné pozorování – časy cenzorování jsou nezávislé na době události. [1]

K označení délky pozorování (do sledované události nebo do cenzorování) používáme náhodnou proměnnou T . δ je indikátor události. V případě že událost nastala je δ rovno 1, pokud byl jedinec cenzorován je δ rovno 0. Pro n pozorovaných subjektů dostáváme množinu dvojic $\{(t_j, \delta_j), j = 1 \dots n\}$. Tato pozorování na sobě musí být nezávislá. K popisu těchto dat používáme nejčastěji čtyřpolní tabulku. Zvolí se určitá časová hranice, nacházející se v období pozorování. Po té se jedinci rozdělí do čtyř skupin, podle toho jestli této časové hranice dosáhli bez události, anebo jí bez události nedosáhli. Nejedná se však o nejdokonalejší techniku. Dochází ke směšování jedinců, u kterých událost nenastala 40 dnů s jedinci, u kterých nenastala 90 dnů. Nedomáhá se tak k využití veškerých informací v datech. [1][6]

1.2. Funkce přežití

Funkce přežití udává pravděpodobnost přežití jedince. Značí se $S(t)$ a zapisuje se ve tvaru $S(t) = P(T \geq t)$, kde T je délka pozorování a t je konkrétní časový okamžik. Funkce nabývá pouze kladných hodnot a má exponenciální průběh (křivka klesá s tím, jak se pravděpodobnost přežití jedince zmenšuje), proto je maximum na počátku měření (v čase $t = 0$ je maximum rovno 1), kdy subjekty vstupují do studie, a můžeme si být jisti, že jsou stále na živu. Data o přežívání jsou popisována mediánem časů přežití. Je to čas, kterého se dožila polovina pozorovaných jedinců (50% jedinců se dožilo času menšího než je hodnota mediánu a 50% se dožilo času většího nebo rovného mediánu). Mohli bychom také říci, že se u mediánu jedná o střední hodnotu délky přežití.

Analýza historie událostí:

- a) v diskrétním čase – popis dat je jednodušší, časy zařazujeme do intervalů, jednodušší matematické řešení
- b) ve spojitém čase – využívají z dat více informací, jsou vhodné jen pro malé soubory dat

Funkci přežití s parametry t a T , kde t je aktuální přežití jedince a T je nezáporná náhodná veličina, označující délku pozorování do sledované události. V analýze přežívání používáme nejčastěji pro popis rozložení pravděpodobností pro náhodnou veličinu T distribuční funkci, hustotu pravděpodobnosti, funkci přežití a rizikovou funkci.

- a) *Distribuční funkce* – značí se $F(t)$ a je pro náhodnou proměnnou T dána vztahem

$$F(t) = P(T < t)$$

Tato funkce nám říká, s jakou pravděpodobností dosáhl jedinec v čase t sledované události, tj. jestli je sledovaný jedinec v čase t mrtev. Jestli je doba přežití T menší než doba t .

- b) *Hustota pravděpodobnosti* $f(t)$ – značí se $F(t)$ a určuje rozdělení spojitě náhodné veličiny T , je to po částech spojitá funkce, kterou si můžeme vyjádřit vzorcem

$$F(t) = \int_0^t f(u) du$$

Tuto funkci si může vyjádřit také ve tvaru

$$f(t) = \frac{dF(t)}{dt}$$

- c) *Funkce přežití* $S(t)$ - se definuje jako pravděpodobnost, že doba přežití jedince T je větší nebo rovna t . Když tohle vyjádříme vztahem

$$S(t) = P(T \geq t) = 1 - F(t)$$

Z tohoto plyne, že funkce přežití udává s jakou pravděpodobností je sledovaný subjekt v čase t stále na živu, sledovaná událost ještě nenastala.

U funkce přežití je jedná o nerostoucí funkci (může klesat nebo být konstantní). Jak už jsme si dříve říkali, platí pro ni že:

$$S(0) = 1,$$

$$\lim_{t \rightarrow \infty} S(t) = 0.$$

Hodnota funkce v okamžiku $t = 0$ se rovná 1 (maximum), s ubíhajícím časem t ($t \rightarrow \infty$) se hodnoty pravděpodobnosti blíží k nule.

Mezi hustotou pravděpodobnosti a funkcí přežití platí následující vztah:

$$f(t) = -\frac{dS(t)}{dt}$$

Pokud se u funkce přežití jedná o diskrétní rozložení (náhodná veličina T nabývá pouze diskrétních hodnot), platí pro funkci přežití tento vzorec

$$S(t) = \sum_{t_i \geq t} p(t_i)$$

kde $p(t_i)$ je hodnota pravděpodobnosti pro jednotlivé časy t . Můžeme napsat ve formě $p(t_i) = P(T = t_i)$, $i = 1, 2, \dots, n$, kde n je počet diskrétních hodnot. [2]

Střední hodnota a rozptyl náhodné veličiny T :

- *Střední hodnota* – ET

$$ET = \int_0^{\infty} u f(u) du = \int_0^{\infty} S(u) du$$

- *Rozptyl* – DT

$$DT = 2 \cdot \int_0^{\infty} u S(u) du - \left(\int_0^{\infty} S(u) du \right)^2$$

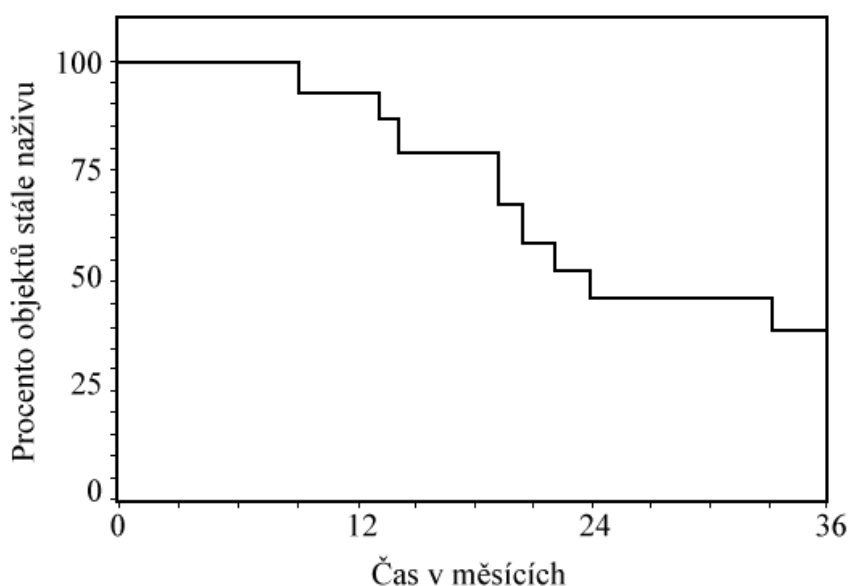
2. ODHAD FUNKCE PŘEŽITÍ

Nejčastěji se pro odhad funkce přežití používá **Kaplan-Meierova metoda**. Jedná se o **neparametrickou metodu**, která nevyžaduje znalost pravděpodobnostního rozložení doby přežití u jedinců zkoumaných se studií.

Soubor dat, který budeme zpracovávat, můžeme prezentovat několika způsoby. Například numerickým způsobem ve formě tabulky nebo formou grafu (křivka přežití).

2.1. Křivka přežití

Křivky přežívání se používají nejčastěji pro grafické zobrazení průběhu dat u analýzy přežívání. Na Obrázku 2 máme znázorněnu jednoduchou křivku přežití. Ve studii bylo pozorováno patnáct pacientů po dobu 36 měsíců. Devět pacientů zemřelo během studie, jejich čas úmrtí tedy známe. Šest pacientů bylo po ukončení studie stále na živu.



Obrázek 2 - Křivka přežití [8]

Čas nula je okamžik, kdy byli pacienti zařazeni do studie (nejedná se však o určité kalendářní datum). V čase nula jsou všichni pacienti na živu, z toho plyne, že $P(T = 0) = 1$ neboli 100%. Kdykoli dojde k úmrtí pacienta, procento přeživších pacientů poklesne (v grafu se to projevuje skokem dolů, v devatenáctém měsíci došlo k úmrtí dvou pacientů, skok je tedy dvakrát tak velký). Pokud by byla studie prodloužena na dostatečně dlouhou dobu,

pravděpodobnost by dosáhla 0. Tato studie skončila po 36 měsících s 40% přeživších pacientů.

Slovo „přežití“ v názvu této křivky může být špatně interpretováno. U konce studie se totiž nemusí vždy jednat o úmrtí pacienta. Toto slovo může reprezentovat jakýkoli předem definovaný konečný bod studie. Například to může být okamžik výskytu metastáz, obnova funkcí určitého orgánu nebo propuštění z nemocnice. Musí se však jednat o události uskutečňující se v jednom čase. Křivky přežití jsou nevhodné pro analýzu opakujících se událostí. [6]

2.2. Kaplan-Meierův odhad funkce přežití

U odhadu funkce přežití podle Kaplana a Meiera se jedná o metodu neparametrickou, nevyžadují se tedy znalosti pravděpodobnostního rozložení, kterou se řídí doba přežití u jednotlivých jedinců.

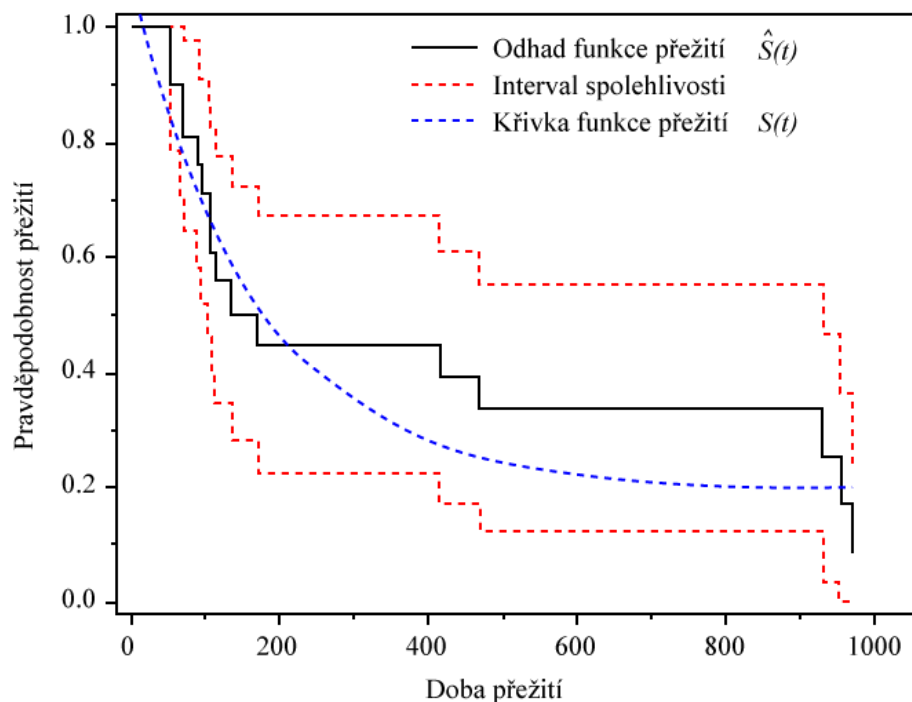
Kaplan-Meierova metoda nám poskytuje odhad funkce přežití v každém okamžiku, ve kterém došlo ke sledované události (smrt, návrat příznaků nemoci, uzdravení jedince).

Před započítáním samotné Kaplan-Meierovi metody je nejprve nutné rozdělit dobu, po kterou jedince pozorujeme, na jednotlivé časové intervaly. V každém tomto intervalu se nachází jedna pozorovaná událost (smrt, návrat příznaků nemoci, uzdravení jedince). Čas kdy k této události dochází, je zároveň počátkem jednotlivých intervalů.

Dále zjistíme časy, kdy pro jednotlivé časové intervaly nastala námi pozorovaná událost $(\delta_j = 1)t_1 < t_2 < \dots < t_k$, počty (d_j) pacientů s událostí pro každý z těchto časů a počty n_j pacientů pozorovaných právě do doby t_j . Pokud se ve studii vyskytuje skupina, ve které nedošlo k žádné události, označujeme ji jako skupinu v riziku. Odhad funkce přežití $S(t)$ v bodě t_j dle Kaplan-Meiera je

$$\hat{S}(t_j) = \hat{S}(t_{j-1})(1 - d_j/n_j) = \hat{S}(t_{j-1})\hat{p}(t_j),$$

kde $\hat{p}(t_j) = 1 - \frac{d_j}{n_j}$ je odhad podmíněné pravděpodobnosti, že pacient, který se dožil doby t_{j-1} , se dožije i doby t_j . Funkce je konstantní v intervalu t_j až t_{j+1} , včetně t_j . Obecně lze provést odhad $\hat{p}(t_j)$ v X. týdnu, kdy nastalo x pozorovaných událostí a y pozorování bylo cenzorovaných:



Obrázek 3 - Odhad funkce přežití s intervalem spolehlivosti

$$\hat{p}(X) = \frac{\text{skupina v riziku pro } X. \text{ týden} - \text{počet událostí v } X. \text{ týdnu}}{\text{skupina v riziku pro } X. \text{ týden}}$$

Pokud se v dané skupině nachází cenzorovaný jedinec, započítává se do skupiny v riziku, ale nezapočítává se do počtu pacientů s událostí. Odhad funkce přežití je cenzorováním ovlivňován. Cenzorovaní pacienti se postupně odstraňují ze skupiny pacientů v riziku pro příslušné výpočty. V tabulkách, které používáme pro výpočet odhadu funkce přežití, uvádíme také směrodatnou chybu SE odhadu křivky přežití v daném bodě. Nejčastěji používaným vzorcem pro výpočet směrodatné chyby je **Greenwoodův vzorec**. Pro jednotlivé časy t z intervalu t_j až t_{j+1} , včetně t_j se rozptyl odhadu vypočítá vzorcem:

$$Var(\hat{S}(t)) = \hat{S}(t) \sum_{i=1}^j \frac{d_i}{n_i(n_i - d_i)}$$

Interval spolehlivosti pro křivku přežití, jejíž spolehlivost dosahuje hodnot $100(1 - \alpha)$ pro daný čas t , má tvar

$$(\hat{S}(t) - z_{\alpha/2}SE[\hat{S}(t)]; \hat{S}(t) + z_{\alpha/2}SE[\hat{S}(t)]),$$

kde se $SE[\hat{S}(t)]$ vypočítá Greenwoodovým vzorcem. Tento interval se dokresluje k odhadnuté křivce přežití. Na grafech se vyznačují přerušovanými nebo tečkovanými čarami, které se nachází nad a pod křivkou přežití. [1][2]

Tabulka 1 - Příklad výpočtu odhadu funkce přežití

Doba do relapsu	Cenzorovaný údaj	Úmrtí	Skupina v riziku	$\hat{p}(t_j)$	Odhad $\hat{S}(t_j)$	Směrodatná chyba odhadu
0	-	-	21	21/21 = 1,000	1,000	-
6	1	3	21	18/21 = 0,857	0,857	0,006
7	-	1	17	16/17 = 0,941	0,807	0,008
10	1	1	15	14/15 = 0,933	0,753	0,009
13	-	1	12	11/12 = 0,917	0,690	0,011
16	-	1	11	10/11 = 0,909	0,628	0,013
22	-	1	7	6/7 = 0,857	0,538	0,016
23	-	1	6	5/6 = 0,833	0,448	0,018

2.3.Odhad pomocí tabulky přežití

Této metodě se také někdy říká **aktérský odhad**. Při použití této metody jsou časy přežití rozděleny do časových intervalů, kdy událost nastala (obvykle jsou tyto intervaly stejné délky). Nejčastěji se používají týdenní, měsíční, roční nebo pětiroční intervaly. Stejně jako při Kaplan-Meierově metodě pracujeme s cenzorovanými údaji, jestliže byl pacient v daném časovém intervalu přestán sledován.

Odhad křivky přežití se provede pomocí součinu odhadu pravděpodobnosti přežití v jednotlivých intervalech do uvažované doby t

$$\hat{S}(t) = \prod_{j:t_j \leq t} \hat{p}(t_j),$$

kde se

$$\hat{p}(j) = 1 - \frac{d_j}{n_j - c_j/2},$$

kde je d_j počet událostí v j -tém intervalu, n_j počáteční počet pacientů v rizikové skupině a c_j je počet cenzorovaných jedinců. Pro tyto cenzorované jedince však musí platit, že ve sledovaném intervalu zůstávají v průměru polovinu doby.

Tabulka 2 - Příklad tabulky přežití (roční intervaly)

Interval	Čas	Počet událostí d_j	Počet cenzorovaných jedinců c_j	Počet jedinců v riziku n_j	$1 - \frac{d_j}{n_j - c_j/2}$	$\hat{S}(t)$
1	0-	16	4	48	0,652	0,652
2	12-	10	4	28	0,615	0,401
3	24-	1	0	14	0,929	0,373
4	36-	3	1	13	0,760	0,283
5	48-	2	2	9	0,750	0,212
6	60-	4	1	5	0,111	0,024

Tabulková metoda je vhodná zejména v případech, kdy nevíme, v jakém přesném okamžiku ke sledované události došlo. Známe jen časový interval, kdy k tomuto došlo. Pokud výsledky této metody zobrazíme graficky, dostáváme schodovou funkci, která má konstantní hodnotu v intervalech $(t_j; t_{j+1})$.

Tuto metodu používáme, pouze pokud známe jen časové intervaly, kdy k událostem dochází. Pokud známe časy, kdy sledovaná událost nastala, přesně je vhodnější použít Kaplan-Meierovu metodu. [1]

2.4. Testy shody funkcí přežití

Tyto testy slouží k porovnávání dvou křivek přežití (z dvou výběrů). Těchto testů existuje celá řada druhů, každý z nich má optimální vlastnosti pro jinou situaci.

K nejznámějším testům patří asymptoticky platný neparametrický **Coxův-Mantelův test**. Někdy se pro něj používá i označení **log rank test**. K předpokladům u tohoto testu se počítá, že proces cenzorování je nezávislý na procesu, jenž vede ke vzniku události. Coxův-Mantelův test byl zobecněn i pro hodnocení více než dvou testovaných skupin.

Při použití Coxova-Mantelova testu zohledňujeme každý čas t_j , ve kterém došlo k pozorované události. Abychom tento test mohli použít, musíme si nejdříve zjistit počet jedinců v každé rizikové skupině, která sledované události předcházela. Dále musíme zjistit počet událostí v tomto okamžiku. Dále si stanovíme nulovou hypotézu rovnosti křivek přežití ($H_0: S_1(t) = S_2(t)$) a vypočítáme si, kolik pozorovaných událostí bychom očekávali, kdyby tato nulová hypotéza platila. Toto si můžeme ukázat na příkladu dvou skupin A a B. Pokud má skupina A v čase t 6 jedinců a skupina B 3 jedince, tak podle námi stanovené nulové hypotézy by měl mezi oběma skupinami platit poměr 2:1. Podle tohoto poměru můžeme očekávat, že pokud dále nastanou tři události, budou tyto rozděleny mezi skupinou A a B tak, že ve skupině A nastanou události dvě, zatímco ve skupině B jen jedna.

V testové statistice poté pokračujeme zvlášť výpočtem součtů počtu pozorovaných událostí (O_A, O_B) a součtem odhadnutých počtů událostí ve všech okamžicích, kdy došlo aspoň k jedné události (E_A, E_B). Tyto součty následně porovnáme.

Tabulka 3 - Schéma pomocných tabulek, které konstruuje pro jednotlivé časy při testu funkcí přežití

	Počet událostí	Počet bez události	Počet v riziku
Skupina 1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
Skupina 2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Součet	d_j	$n_j - d_j$	n_j

Obecný výpočet pro skupinu x E_x

$$E_x = d \cdot n_x / n ,$$

kde d je součet počtu událostí všech pozorovaných skupin, n_x počet jedinců v riziku a n součet počtů jedinců v riziku pro všechny skupiny.

Následně sečteme všechny počty pozorovaných četností (O_x) a počty odhadnutých četností (E_x) pro všechny časy, ve kterých došlo k události. Ty následně dosadíme do vzorce testové statistiky.

Testová statistika pro skupiny A a B

$$\chi^2 = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B}$$

Vypočítaná testovací statistika χ^2 se srovná s kritickou mezí χ^2 -rozdělení s jedním stupněm volnosti.

Vypočítané hodnoty O_A, O_B a E_A, E_B se následně využijí pro výpočet *podílu šancí OR*, který kvalifikuje rozdíl křivek přežití:

$$OR = \frac{O_A/E_A}{O_B/E_B}$$

Při sestřování intervalů spolehlivosti vycházíme ze znalosti rozptylu logaritmu hodnoty OR .

$$Var[\ln(OR)] = (1/O_A + 1/O_B)$$

Hodnoty intervalu dostaneme odlogaritmováním těchto hodnot.

Coxův-Mantelův test předpokládá, že hodnoty funkce rizika (viz dále) mají pro obě skupiny konstantní poměr pro všechny hodnoty časů události t . [1]

2.5. Regresivní model pro analýzu historie událostí

Dříve popsané neparametrické metody (křivka přežití, Kaplan-Meierův odhad funkce přežití, tabulka přežití) jsou vhodné jen k analýze dat s jedním výběrem časů událostí nebo při porovnávání dvou nebo více skupin. Obvyklé však je, že o jedincích máme více informací, které mohou časový průběh ovlivnit. Například ve studii kde porovnáváme časy událostí dvou skupin pacientů, u nichž známe pohlaví, věk nebo hodnoty dalších fyziologických charakteristik pacienta, které by mohli mít vliv na průběh nemoci (jako je např.: kouření, dieta, věk, terapie apod.). Tyto hodnoty považujeme za ovlivňující nebo nezávislé proměnné. Použitím regresivních modelů se snažíme zachytit vliv těchto faktorů na průběh křivek přežití.

Nejnámějším regresivním modelem je **Coxova regrese**. Tento model je založen na funkci rizika. Funkci rizika značíme $h(t)$ nám udává, jak se v čase mění pravděpodobnost, že dojde k pozorované události. U funkce rizika se v podstatě jedná o podmíněnou pravděpodobnost $P(x < t < y | t > x)$, že dojde k události v daném časovém intervalu $(x; y)$ za předpokladu, že k ní nedošlo před tímto intervalem. Funkci rizika můžeme ke křivce přežití $S(t)$ najít jako limitu, kdy se délka intervalu $(x; y)$ blíží k nule pro danou hodnotu x . mezi těmito dvěma křivkami platí vztah

$$h(t) = S'(t)/S(t) = [\ln(S(t))]' = -f(t)/S(t)$$

kde $f(t)$ je hustota rozdělení časů přežití a čárka značí derivování podle času.

Coxova regrese vychází při popisu vztahu mezi nezávislými proměnnými a funkcí rizika touto rovnicí

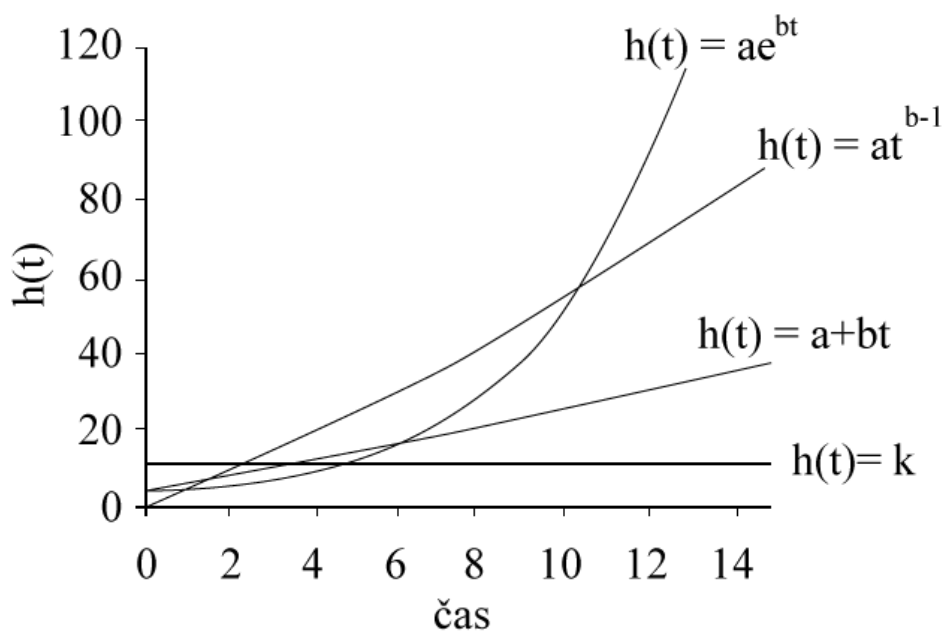
$$h(t, x_i) = h_0(t) \exp\left(\sum_j \beta_j x_{ij}\right)$$

kde x_{ij} ($j = 1, 2, \dots, k$) jsou hodnoty nezávislých proměnných pro jedince i .

Logaritmický tvar této rovnice

$$\ln[h(t, x_i)] = \ln[h_0(t)] + \sum_j \beta_j x_{ij}$$

$h_0(t)$ se nazývá bázová funkce rizika a je společná pro všechny jedince. Pro jednotlivé jedince se mění v závislosti na hodnotách nezávisle proměnných. Touto funkcí může být jakákoli nezáporná funkce.



Obrázek 4 – Příklad průběhu funkce rizika [10]

Coxův model předpokládá, že poměr funkcí rizika nezávisí na t a pro dva různé jedince má tvar

$$\exp\left(\sum \beta_j (x_j - y_j)\right)$$

Pokud je hodnota poměru větší než jedna, je riziko události větší pro jedince x než pro jedince y . Proto představuje Coxův model model relativního rizika. Protože se hodnoty parametru β_j odhadují nezávisle na funkci rizika je Coxův model semiparametrickým regresivním modelem.

Při použití Coxovy metody je důležité nejprve ověřit úměrnost funkcí rizika pro všechny skupiny jedinců s danými hodnotami nezávisle proměnných. V případě že jsou hodnoty nezávisle proměnných různé, způsobují posun funkce rizika k vyšším nebo nižším hodnotám. Změnou velikosti koeficientů β_j zjišťujeme, které nezávisle proměnné mají vliv na průběh funkce rizika a křivku přežití. [1] [3]

3. PROGRAMY PRO ZPRACOVÁNÍ ANALÝZY PŘEŽITÍ

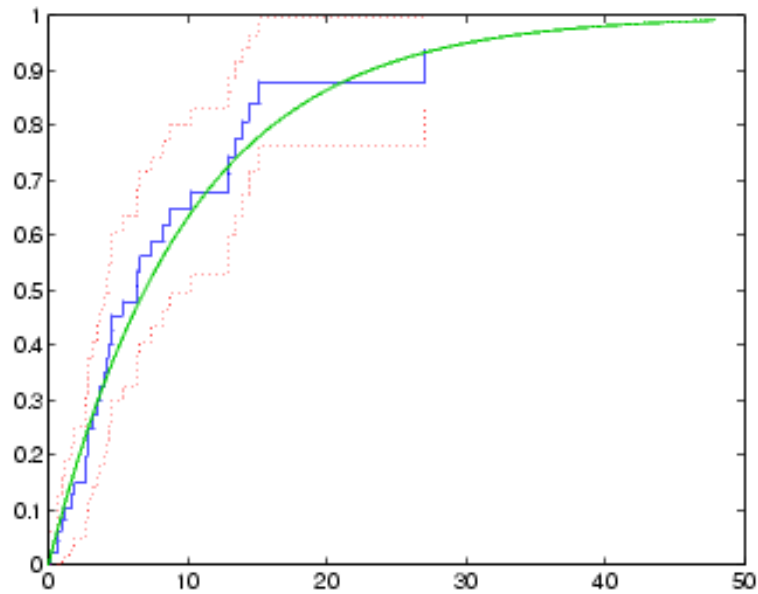
V praxi existuje mnoho programů jejich součástí nebo hlavním zaměřením je zpracovávání dat k analýze přežití. K takovým patří například program MedCalc, který je zaměřen pouze na analýzu medicínských dat. Další možností jak tyto data zpracovat je program Statistica. Jak už vyplívá z názvu, je tento program zaměřen na všechny možné druhy statistické analýzy nejen na analýzu přežití. A nakonec program Matlab u kterého je statistická část jen malou součástí celku. Pomocí tohoto programu můžeme zpracovávat data zcela odlišná od statistických.

3.1. Matlab

V Matlabu existuje speciální toolbox pro zpracovávání dat k analýze přežití. Zpracované data můžeme reprezentovat například ve formě čtyř grafů. Jedná se o graf hustoty pravděpodobnosti (probability densityfunction - PDF), graf funkce přežití (1-CDF = 1-cumulative distributionfunction = 1-distribuční funkce), graf hazardní funkce (hazard ratefunction) a graf pravděpodobnostní funkce (probability plot). Příkazy které pro tyto grafy používáme jsou následující:

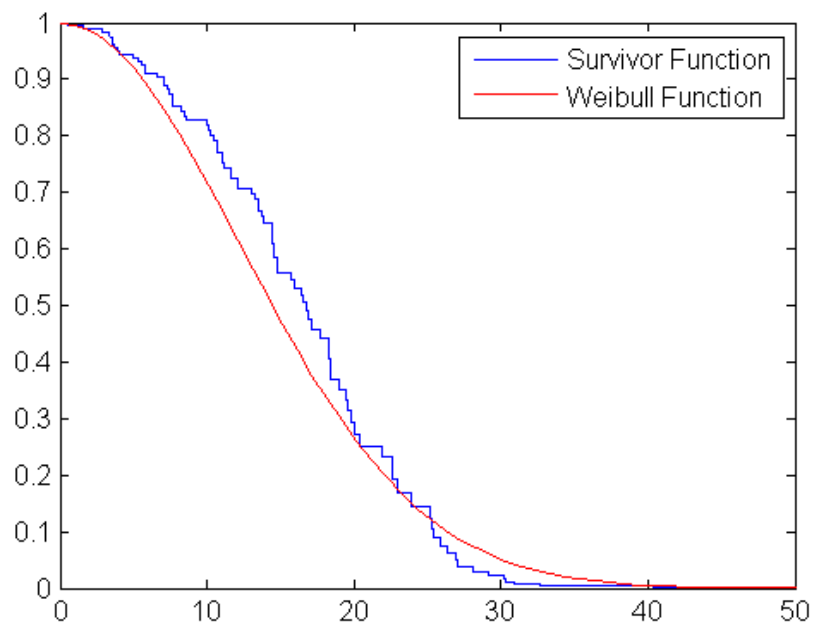
1. PDF – příkaz: `wblpdf`
2. 1-CDF – příkaz: `1-wblcdf`
3. Hazard rate fiction – příkaz: `wblhaz`

Další možností je použít příkaz $[x, y] = \text{ecdf}(y)$ pro odhad křivky přežití podle Kaplan-Meiera.



Obrázek 5 - Graf Kaplan-Meier (Matlab)

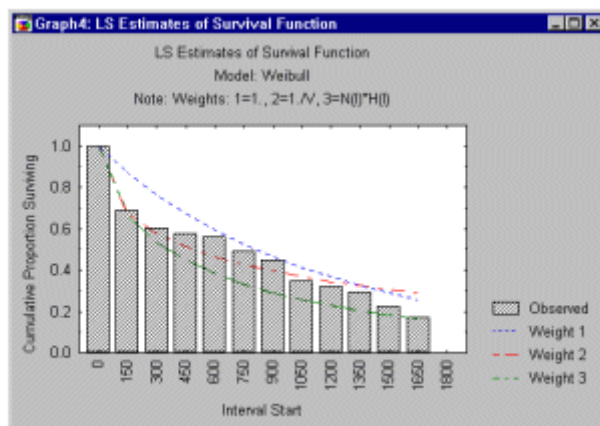
Dále v Matlabu existuje příkaz pro výpočet regrese podle Coxe: $b = \text{coxphfit}(X, y)$.



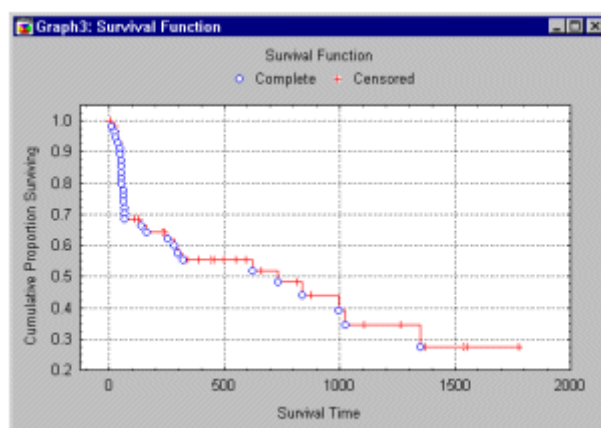
Obrázek 6 - Coxův regresivní model (Matlab)

3.2.Statistica

Program Statistica nám nabízí široký výběr statistických testů a postupů. Je možné v něm provést všechny metody výpočtu odhadu analýzy přežití, které jsme si ukázali ve druhé kapitole. To zahrnuje tabulky přežití, odhad funkce přežití podle Kaplan-Meiera, testy shody funkcí a regresivní modely podle Coxe.



Obrázek 7 – Ukázka odhadu funkce přežití (Statistica)

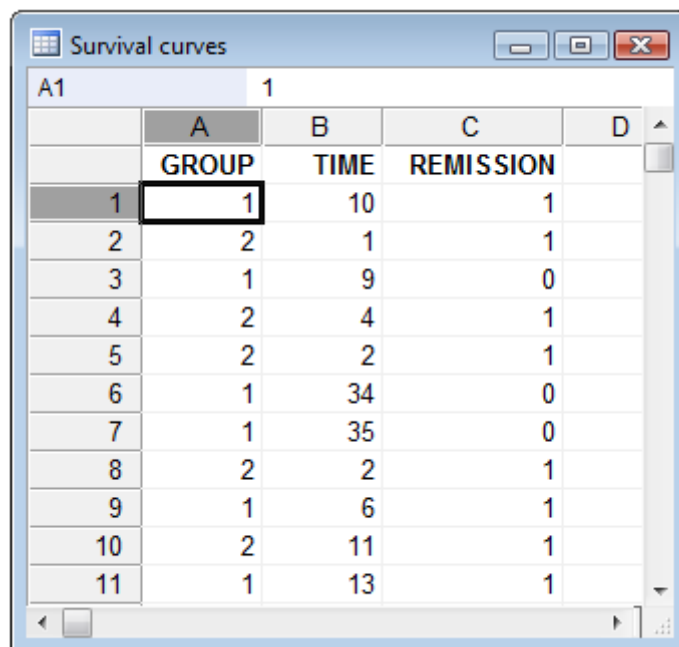


Obrázek 8 - Ukázka funkce přežití (Statistica)

3.3.MedCalc

V tomto profesionálním programu pro zdravotnická zařízení a zdravotnický výzkum je možno provést odhad funkce přežití podle Kaplan-Meiera a Coxův regresivní model.

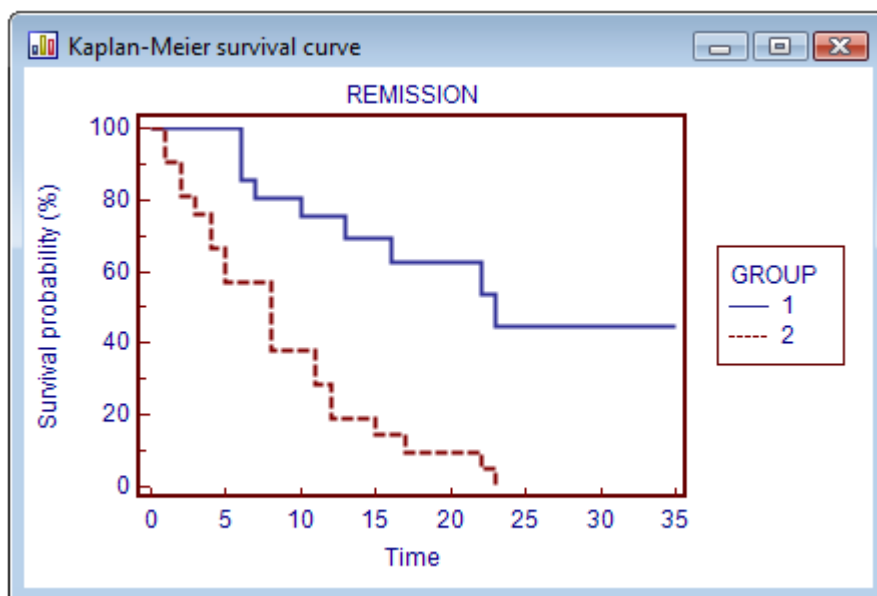
Pokud se rozhodneme pro Kaplan-Meierovu metodu, je nám nabídnuta tabulka se třemi sloupci. Do prvního sloupce napíšeme číslo skupiny, do druhého čas přežití a do třetího sloupce jestli se jedná o data, která dosáhla sledované události nebo jestli se jedná o data cenzorovaná.



	A	B	C	D
	GROUP	TIME	REMISSION	
1	1	10	1	
2	2	1	1	
3	1	9	0	
4	2	4	1	
5	2	2	1	
6	1	34	0	
7	1	35	0	
8	2	2	1	
9	1	6	1	
10	2	11	1	
11	1	13	1	

Obrázek 9 - Tabulka pro Kaplan Meiera [7]

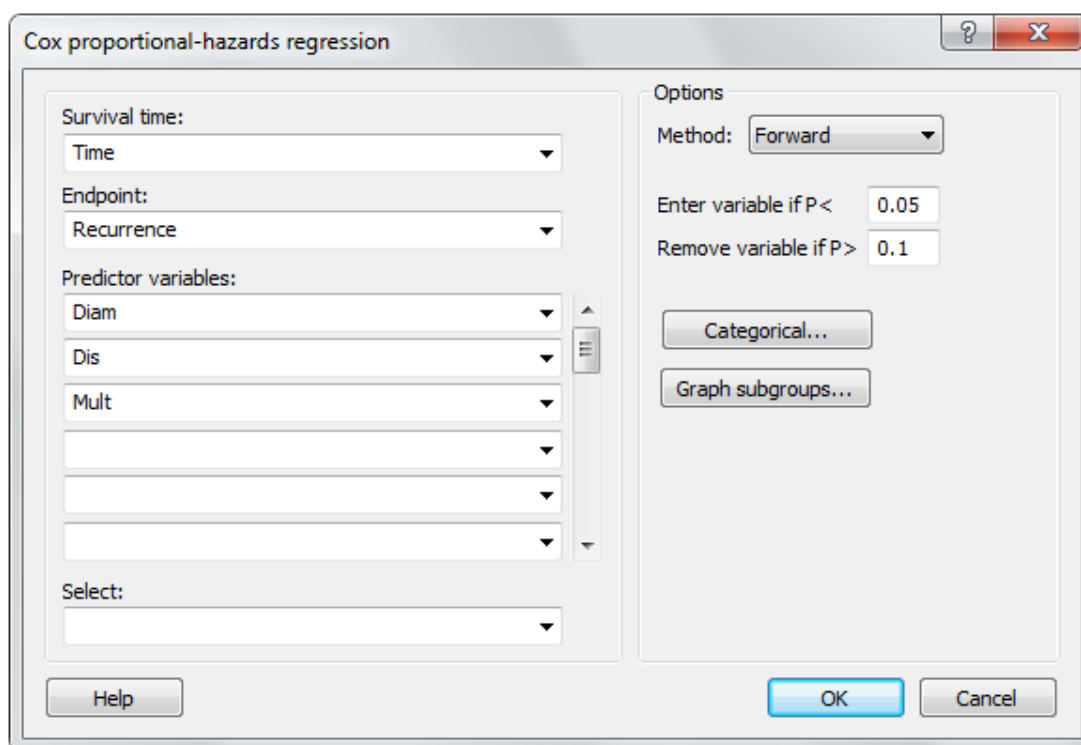
Následně musíme vyplnit tabulku, která se týká možnosti zobrazení grafu z těchto dat. Poté se nám zobrazí samotný graf.



Obrázek 10 - Graf Kaplan-Meier MedCalc [7]

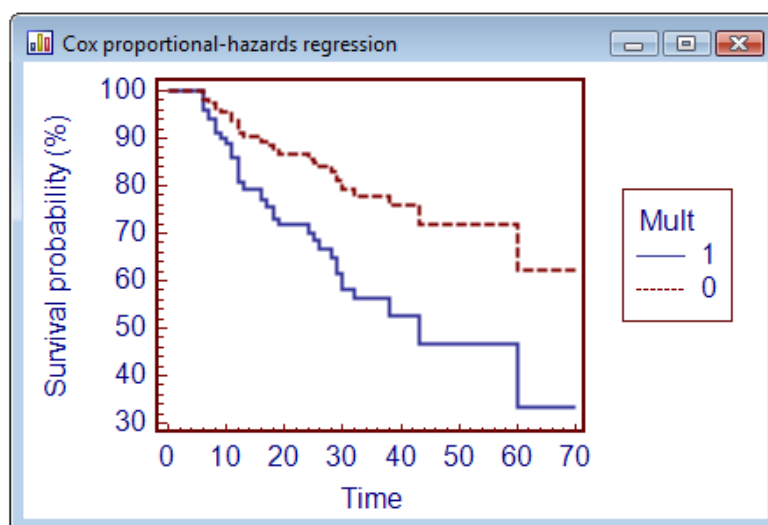
Výsledky celé studie se následně zobrazí i ve formě tabulky, vypisující všechny důležité údaje (kolik se účastnilo studie vzorků, velikost mediánu).

V případě že Kaplan-Meierovu metodu nemůžeme použít, využijeme Coxův regresivní model. Obvykle u objektů s větším počtem faktorů, které je ovlivňují. Před zahájením této metody je také nutné vyplnit několik údajů. Nejdůležitější je vyplnit jaké proměnné budou výslední čas přežití ovlivňovat.



Obrázek 11 - Tabulka pro Coxovu regresivní analýzu MedCalc [7]

Výsledný graf vypadá podobně jako u metody podle Kaplan-Meiera.

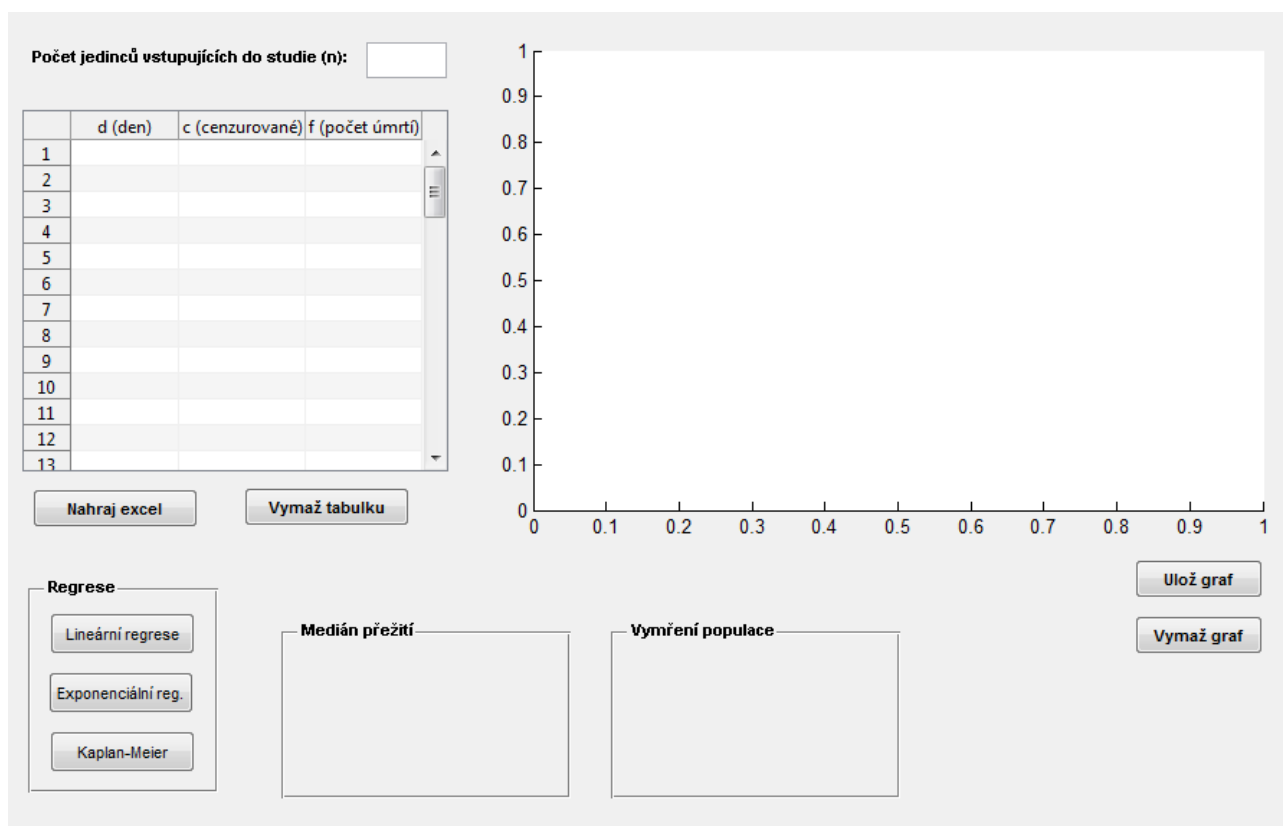


Obrázek 12 - Graf pro Coxovu regresivní analýzu

4. VLASTNÍ ŘEŠENÍ PROGRAMU

Jak už bylo naznačeno v předchozí kapitole, existuje mnoho programů pro zpracování analýzy přežití. Některé z nich jsou zaměřeny pouze na tento specifický druh problému, zatímco u jiných programů se jedná jen o jednu z mnoha možností využití. Program `survival_guide.fig` je mnou vytvořený algoritmus v programovacím prostředí Matlab. Je vytvořen v grafickém rozhraní GUIDE.

Pomocí tohoto programu je možné vytvářet regresivní analýzu přežití využitím tří různých způsobů, které si později podrobněji vysvětlíme (oddíly 4.1 až 4.3). Jedná se o lineární regresi využívající metodu nejmenších čtverců, exponenciální regresi a Kaplan-Meierův odhad. Všechny tyto metody jsou běžně v praxi využívány a jsou součástí většiny profesionálních programů určených pro analýzu přežívání (jako je například program Statistica od firmy StatSoft nebo i Microsoft Office Excel). Výsledky získané mnou navrženým programem a výsledky těchto programů budou uvedeny a porovnány v jednotlivých oddílech 4.1 až 4.3. Na Obrázku 13 vidíme čelní panel mnou navrženého programu.



Obrázek 13 - Čelní panel programu

K zadávání dat slouží v levé části umístěná tabulka s textovým polem nad ní. Do textového pole zadáváme počet jedinců, kteří do studie vstupují. Do prvního sloupce tabulky zadáváme čísla dnů studie, kdy došlo k pozorované události (úmrtí pacienta nebo k cenzurování). Druhý sloupec slouží k zapsání počtu cenzurovaných jedinců (pacienti, kteří ze studie odešli nebo není znám jejich současný zdravotní stav – ztratili jsme s nimi kontakt). Do třetího sloupce tabulky zapisujeme počet úmrtí, ke kterým došlo v daný den. Tabulku můžeme vyplnit buď ručně anebo použitím tlačítka *Nahraj excel* umístěným pod tabulkou. Toto tlačítko využívá k nahrání souboru matlabovský příkaz `xlsread`.

Tlačítka provádějící samotné regrese se nachází pod tabulkou v panelu s názvem *Regrese*. Po stisknutí tlačítka dochází k výpočtu regrese a zobrazení příslušného grafu v poli axes, umístěném na pravé straně čelního panelu. Současně se pod grafem objeví název uživatelem zvolené regrese.

Dále se na čelním panelu nacházejí panely s názvy *Medián přežití* a *Vymření populace*. V těchto panelech umístěná textová pole zobrazují, v jaký den dojde v populaci k poklesu na padesát procent a v jaký den dojde k poklesu počtu populace na nulu. Tyto údaje jsou pro predikci přežití velice důležité.

Pro ulehčení práce s programem jsou pro uživatele na čelním panelu umístěna další tři tlačítka. Tlačítko *Vymaž tabulku*, umístěné pod vlastní tabulkou, slouží jednak ke smazání údajů v tabulce a dále také ke smazání čísla v textovém poli nad ní. Další tlačítka jsou umístěna pod grafem. Jedná se o tlačítko *Ulož graf* a *Vymaž graf*. Z jejich názvů je patrné k čemu tato tlačítka slouží. Po stisknutí tlačítka *Ulož graf* dochází k uložení grafu pod názvem `survival.png`. Jak vyplývá z názvů, obrázek se uloží ve formátu png (Portable Network Graphics). Toto tlačítko využívá nejdříve matlabovský příkaz `getframe(gca)`, který zachytí graf formou snímku. Poté následuje příkaz `imwrite`, který uloží snímek do námi zvoleného formátu. Samotný obrázek je umístěn do složky adresované Matlabem.

3.1.Lineární regrese

Jedná se o nejsnadnější druh aproximace množiny bodů polynomem. Při použití lineární regrese se jedná o polynom, jehož hodnoty se nacházejí v přímce – polynom prvního řádu. Pro výpočet hodnot polynomu se využívá metody nejmenších čtverců. Tato metoda je založena na sumaci druhých mocnin nejmenších odchylek bodů, které chceme aproximovat, od výsledné přímky.

Rovnice přímky je dána ve formě:

$$y = x \cdot A + b$$

Proměnná A udává směrnici přímky a b posunutí přímky po ose y. Proměnné A a b získáme vyřešením soustavy rovnic:

$$A = \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i} \cdot \frac{\sum_{i=1}^n x_i}{n}$$

$$b = \frac{\sum_{i=1}^n y_i \cdot c}{\sum_{i=1}^n y_i}$$

n – počet prvků v x

Dále je nutné vypočítat interval spolehlivosti pro danou metodu. Interval spolehlivosti udává v jakém rozmezí se v daný čas regresní křivka (v tomto případě přímka) nachází s určitou pravděpodobností (v našem případě s pravděpodobností 95%). Neudává však, v jakém rozmezí se nachází pravá křivka přežití (které skutečný průběh neznáme), ta se může nacházet i mimo interval spolehlivosti pro lineární regresi. Interval spolehlivosti určíme výpočtem směrodatné odchylky $s_{y,x}$. Směrodatnou odchylku vypočítáme z rovnice [11]:

$$s_{y,x} = \sqrt{\frac{\sum (y_i - Y_i)^2}{n - 2}}$$

Kde Y_i je hodnota vypočítaná z regresivní rovnice pro odpovídající x_i . Z této směrodatné odchylky můžeme vypočítat odchylky pro koeficienty A a b.

$$s_a = s_{y,x} \sqrt{\frac{\sum x_i^2}{n \cdot \sum x_i^2 - (\sum x_i)^2}}$$

$$s_b = \frac{s_{y,x}}{\sqrt{\sum x_i^2 - (\sum x_i)^2/n}}$$

Vlastní interval spolehlivosti spočítáme ze vztahů:

$$L(a)_{1,2} = a \pm s_a \cdot t_\alpha$$

$$L(b)_{1,2} = b \pm s_b \cdot t_\alpha$$

t_α udává hladinu významnosti. V našem případě se $t_\alpha = 2,093$. Hodnota je udávána v tabulkách pro 95% pravděpodobnost výskytu u dvaceti měření.

Výpočet pro konkrétní hodnoty pro $n = 100$:

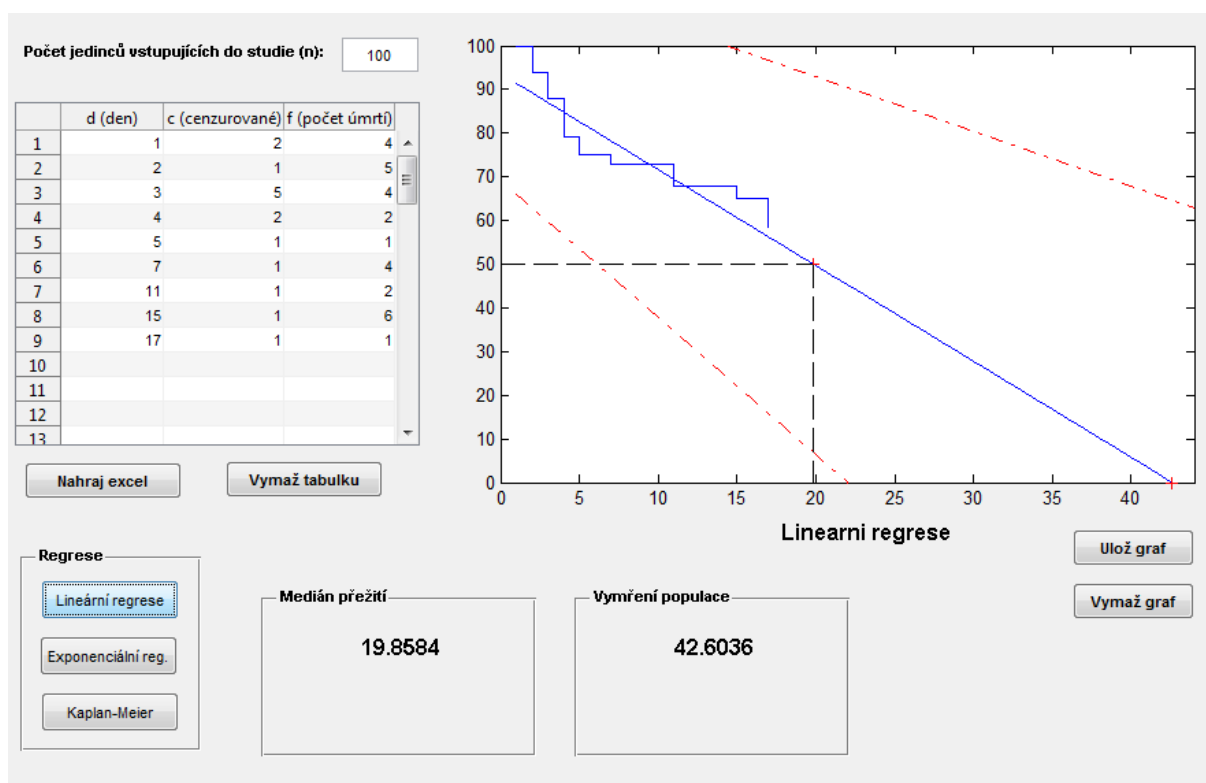
Tabulka 4 - Tabulka zpracovávaných hodnot

d (den)	c (cenzurované)	f (počet úmrtí)
1	2	4
2	1	5
3	5	4
4	2	2
5	1	1
7	1	4
11	1	2
15	1	6

Na Obrázku 14 vidíme čelní panel se zadanými hodnotami spolu s grafem a vypočtenými hodnotami pro medián přežití a vymření populace.

Tabulka 5 - Hodnoty mediánu přežití a vymření populace pro lineární regresi

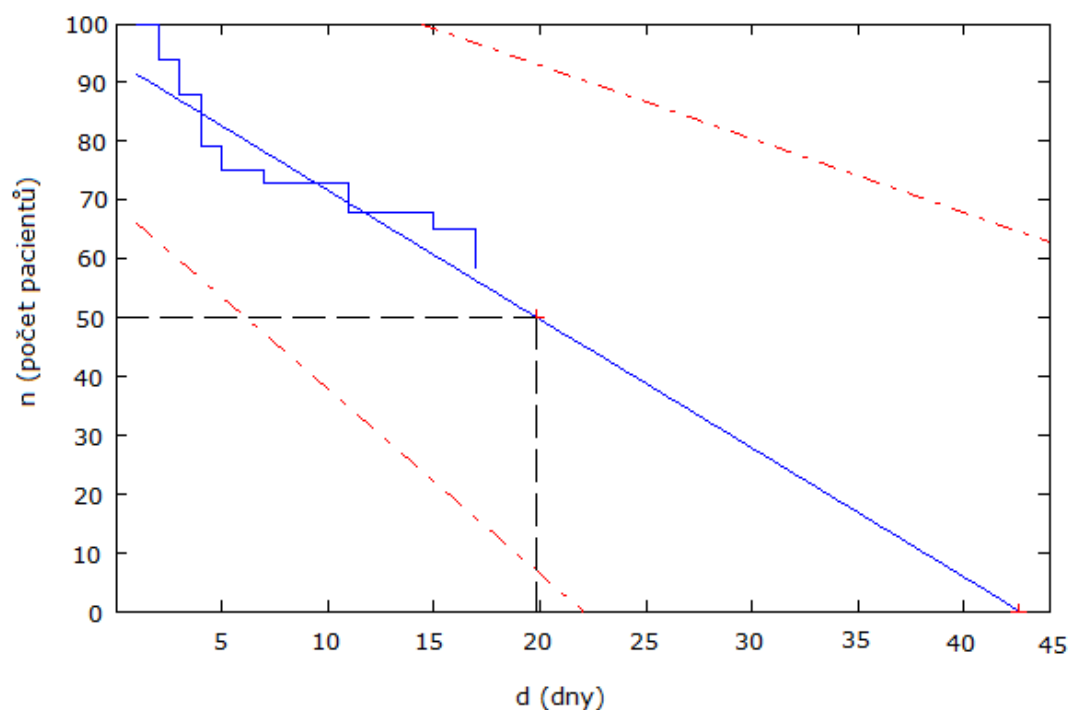
Medián přežití	Vymření populace
19,858	42,604



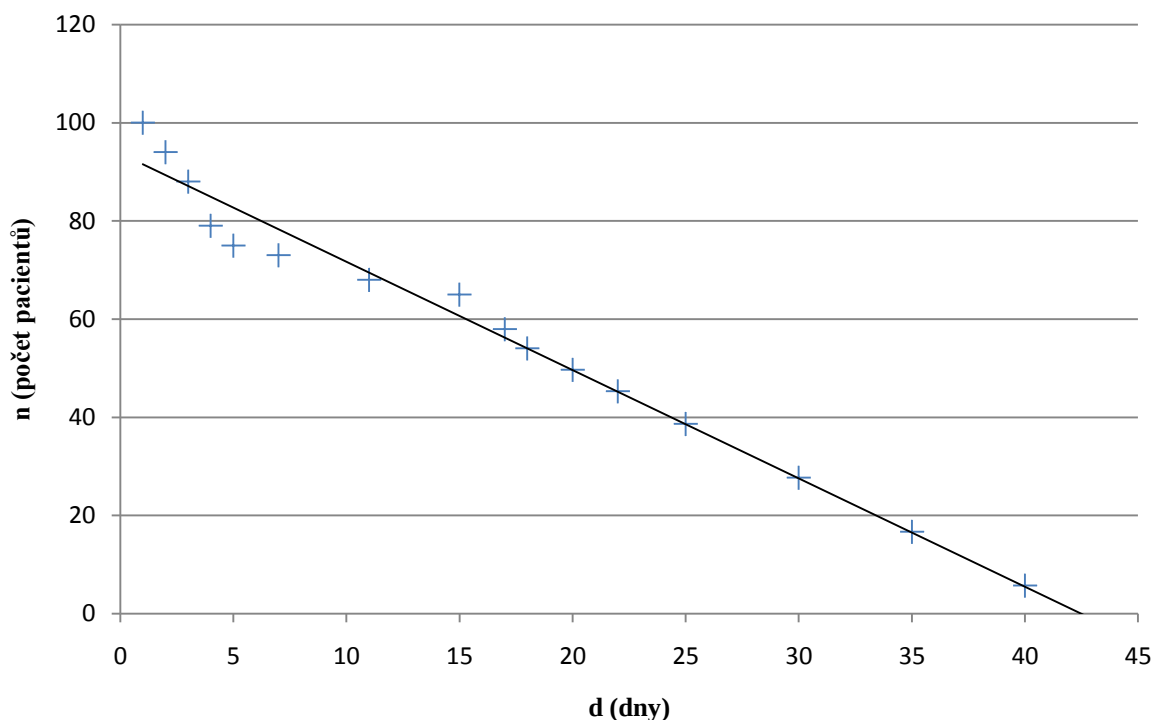
Obrázek 14 - Lineární regrese

Pro ověření správnosti výpočtů prováděných mým programem jsem si vybrala program Excel, protože je ze statistických programů nejdostupnější uživateli (v porovnání na příklad s programem MedCalc).

Porovnání výsledků programu s výsledky Excelu:



Obrázek 15 - Graf z mnou vytvořeného programu (lineární regrese)



Obrázek 16 - Graf vytvořený v Excelu (lineární regrese)

K vytvoření regresivního modelu v Excelu byl použit tento příkaz:

`=SUMA(LINREGRESE(pole_y; pole_x)*{a;1})`

pole_y – množina čísel udávající počet pacientů

pole_x – množina čísel udávající dny

a – číslo dne, pro který chceme vypočítat počet pacientů

Pro ověření správnosti výpočtu mediánu a vymření populace jsme do vzorce z Excelu dosadili hodnoty vypočítané mnou vytvořeným programem (Tabulka 5):

`=SUMA(LINREGRESE(E1:E9;A1:A9)*{19,8584;1})`

Pro zadanou hodnotu nám vyšlo: 50,00006.

`=SUMA(LINREGRESE(E1:E9;A1:A9)*{42,6036;1})`

Pro zadanou hodnotu nám vyšlo: $4,94641 \cdot 10^{-7}$. Toto číslo se velmi blíží nule.

Po porovnání přímek na Obrázku 15 a 16 a čísla pro medián a vymření populace vypočítaného v Excelu, můžeme říci, že obě přímky mají téměř shodný průběh. Z tohoto můžeme usoudit, že mnou navržený algoritmus pracuje pro lineární regresi správně. Navíc znázorňuje i interval spolehlivosti a vyznačuje medián přežití i vymření populace.

3.2.Exponenciální regrese

Stejně jako u lineární regrese se v tomto případě využívá pro výpočet hodnot metoda nejmenších čtverců (metoda je však vyššího řádu). Hodnoty polynomu popisující exponenciální regrese se nacházejí na exponenciále – funkce popisující exponenciálu je prostá, monotónní, nemá minimum ani maximum, svým průběhem se blíží ose x . Exponenciální regrese se používá v případě, že se čas přežití řídí exponenciálním rozdělením. Jedná se o nelineární křivky přežití.

Exponenciální funkce je zadána ve tvaru:

$$y = a \cdot e^{bx}$$

Abychom tuto rovnici mohli vyřešit, musíme si ji převést na lineární tvar. Nejdříve exponenciální rovnici zlogaritmujeme

$$\ln y = \ln a + b \cdot x,$$

následně provedeme substituci za $\ln y$ a $\ln a$

$$Y = \ln x$$

$$A = \ln a$$

$$Y = A + b \cdot x$$

Vzniklou rovnici vyřešíme stejně jako u lineární regrese.

$$b = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i n}$$
$$A = \frac{\sum_{i=1}^n y_i \cdot c}{\sum_{i=1}^n y_i}$$

n – počet prvků v x

Při dalším řešení rovnice je nutné výsledné A odlogaritmovat, abychom z rovnice dostali výsledné číslo a . Následně všechny proměnné dosadíme do původní rovnice. Za proměnnou x dosazujeme množinu dnů, u kterých chceme vypočítat odhadovanou hodnotu počtu pacientů.

Stejně jako u lineární regrese musíme vypočítat interval spolehlivosti. Vypočítáme ho z rovnice

$$IS = p \pm 1,96 \cdot \sqrt{(1 - p)/N},$$

kde IS je interval spolehlivosti, p je odhadovaný počet pacientů a N je celkový počet pacientů na počátku studie.

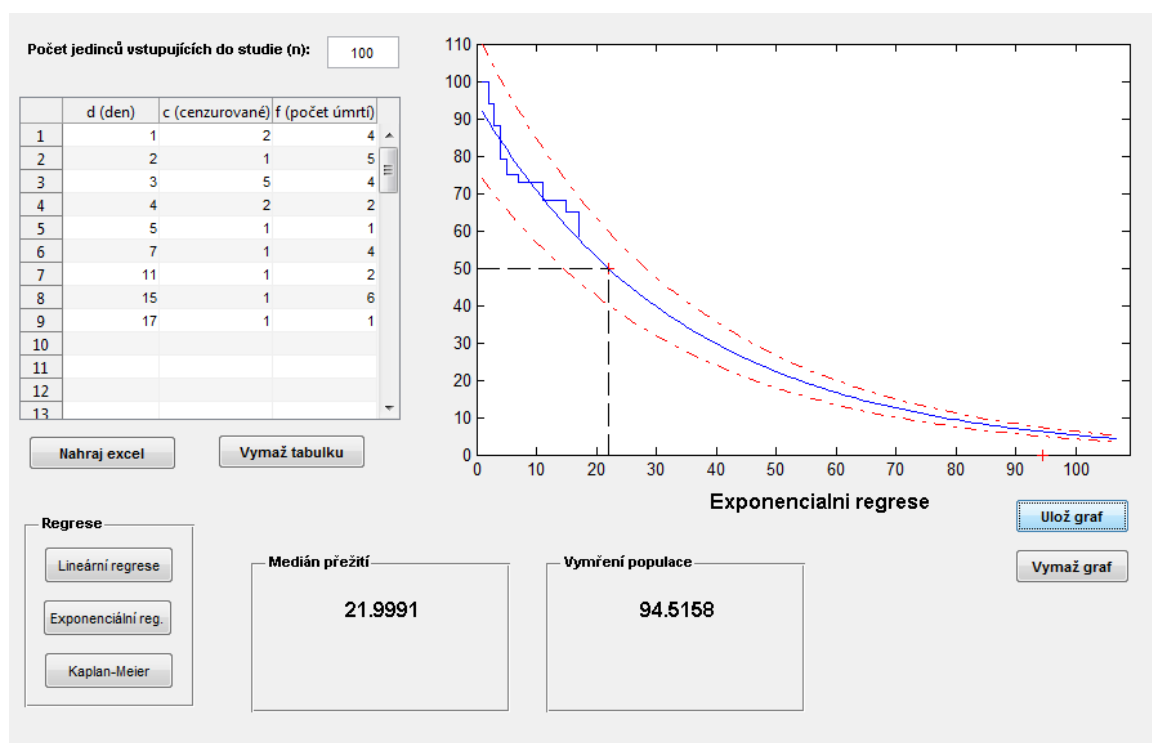
K výpočtu exponenciální regrese využijeme stejné hodnoty jako u regrese lineární, viz Tabulka 4.

Výsledné hodnoty mediánu přežití a vymření populace vidíme v Tabulce 6.

Tabulka 6 - Hodnoty mediánu přežití a vymření populace pro exponenciální regresi

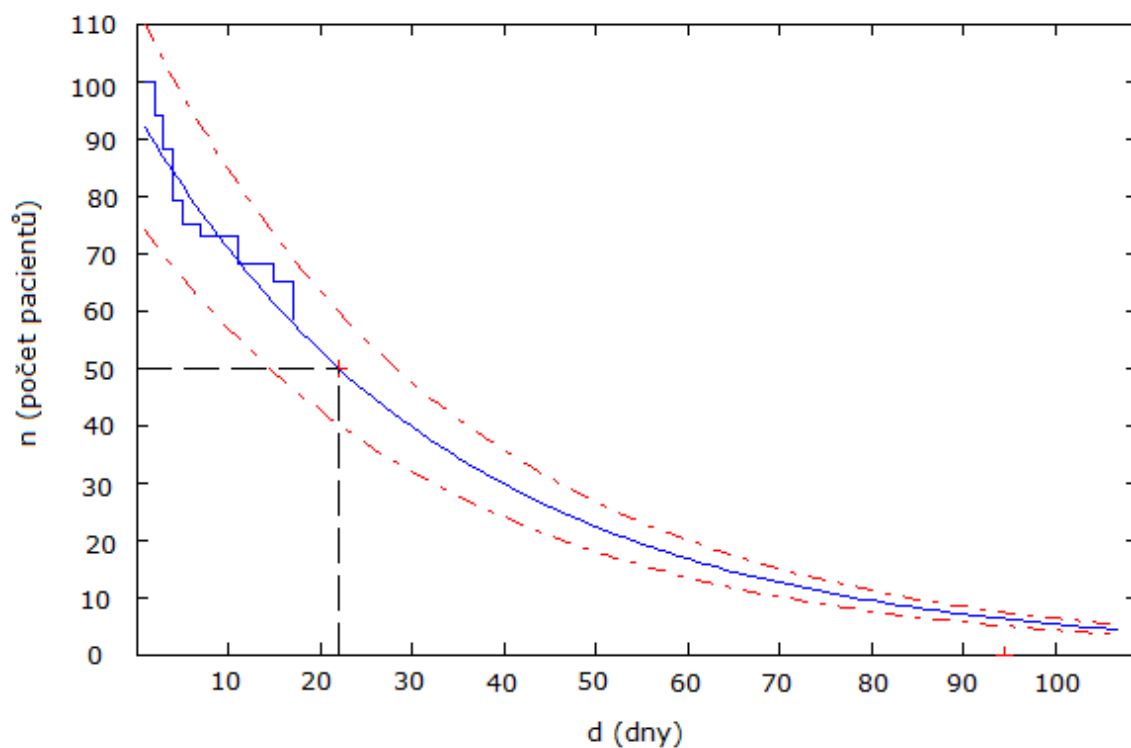
Medián přežití	Vymření populace
21,999	94,516

Na Obrázku 14 vidíme čelní panel se zadanými hodnotami spolu s grafem a vypočtenými hodnotami pro medián přežití a vymření populace.

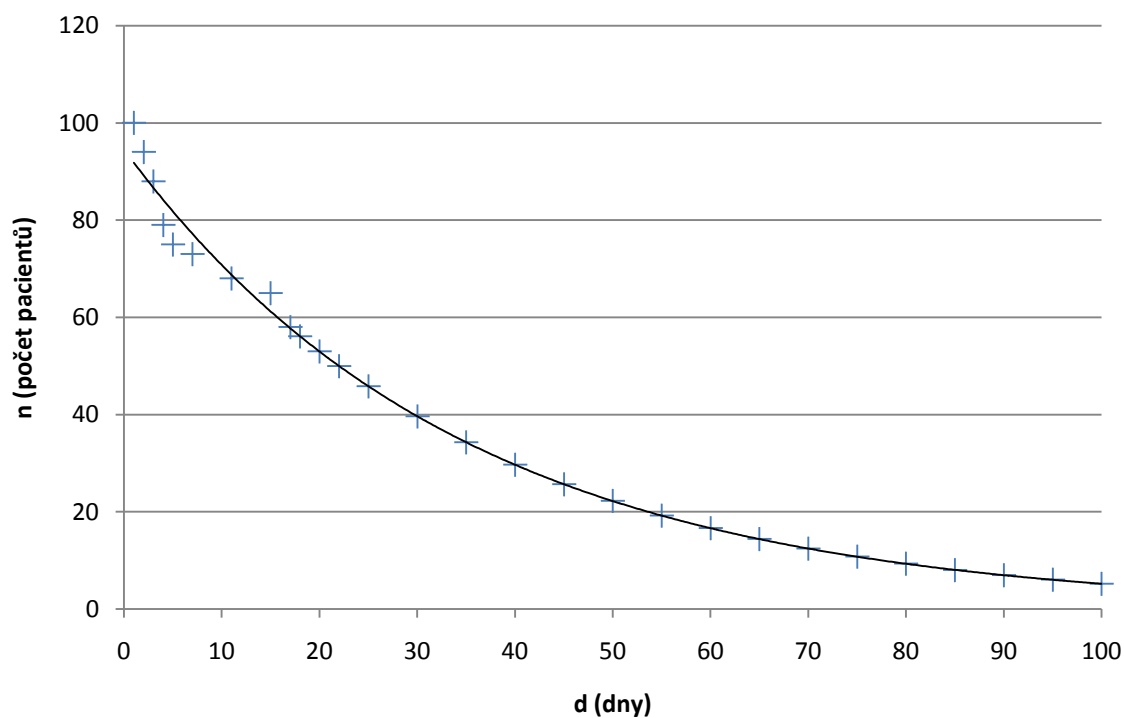


Obrázek 17 - Exponenciální regrese

Porovnání výsledků programu s výsledky Excelu:



Obrázek 18 - Graf z mnou vytvořeného programu (exponenciální regrese)



Obrázek 19 - Graf vytvořený v Excelu (exponenciální regrese)

Pro vytvoření grafu v Excelu byl použit příkaz:

=LOGLINTREND(pole_y;pole_x;nová_x)

kde mají proměnné tento význam:

pole_y – počet pacientů

pole_x – množina čísel udávajících den

nová_x – čísla dnů, pro která se má exponenciální odhad provést

Pro ověření správnosti výpočtu mediánu a vymřeni populace jsme do vzorce z Excelu dosadili hodnoty vypočítané mnou vytvořeným programem (Tabulka 6):

=LOGLINTREND(E1:E9;A1:A9;D20)

D20 = 21,9991

Pro zadanou hodnotu nám vyšlo: 49,99995. Hodnota se blíží číslu 50.

=LOGLINTREND(E1:E9;A1:A9;D23)

D23 = 94,5158

Pro zadanou hodnotu nám vyšlo: 6,12934.

Porovnáním průběhů křivek na Obrázku 18 a Obrázku 19 a čísla pro medián a vymřeni populace vypočítaného v Excelu zjišťujeme, že jsou oba průběhy téměř shodné. Číslo udávající vymřeni populace se nerovná nule, protože exponenciální funkce se jen blíží číslu nula, ale nikdy ho nedosáhne (neprotne osu x v žádném bodě). Z těchto údajů můžeme usoudit, že výpočty provedené mým programem jsou s největší pravděpodobností správné. Stejně jako algoritmus pro výpočet lineární regrese i tento znázorňuje interval spolehlivosti spolu s mediánem přežití. Vymřeni populace není na grafu znázorněno, protože nabylo příliš vysoké hodnoty.

3.3.Srovnání lineární a exponenciální regrese

Obě metody můžeme porovnat, jelikož jsou původní zpracovávaná data totožná (Tabulka 4). Budeme srovnávat grafy na Obrázku 15 a Obrázku 18.

Lineární regrese se od exponenciální liší strmostí svého poklesu. Přímka lineární regrese klesá v celém svém průběhu konstantně, zatímco křivka exponenciální regrese klesá nejdříve strmě a poté teprve dochází k jejímu pozvolnému poklesu.

Z průběhu obou křivek vidíme, že se k modelování odhadovaného průběhu zpracovávaných dat více hodí exponenciální regrese. Křivka lépe kopíruje schodovitý průběh reprezentující zpracovávaná data. Výsledný medián přežití se u obou těchto metod příliš neliší. Co se ovšem výrazně liší je hodnota vymřeni populace. U exponenciální regrese je tato

hodnota téměř dvojnásobek hodnoty lineární regrese. Je to způsobeno faktem, že exponenciální křivka se nikdy neprotne s osou x, a proto musíme vyměření populace počítat jako limitu této křivky.

Z předchozích údajů můžeme odvodit, že se lineární regrese hodí spíše na modelování průběhu nemoci s konstantním úbytkem pacientů za časovou jednotku. Zatímco exponenciální regrese se bude spíše hodit na onemocnění, na která pacienti umírají v prvních dnech onemocnění, s postupem času však počet pacientů klesá pomaleji a po překročení určité časové hranice počet pacientů téměř neklesá.

3.4.Kaplan-Meier

Jedná se o nejjednodušší ze zpracovávaných metod (lineární a exponenciální regrese) pro výpočet křivek přežití. Jak už bylo uvedeno dříve je odhad přežití pacienta přepočítán vždy znovu po úmrtí pacienta. Kvůli své jednoduchosti je tato metoda preferována pro studie s malým počtem subjektů. Počet přeživších pacientů je počítán jako podíl počtu pacientů, kteří jsou naživu na konci dne, počtem pacientů, kteří se dožili následujícího dne. [6][15]

Konkrétně v našem případě jsou hodnoty počítány těmito vzorci:

$$n_{i+1} = n_i - (c_i + f_i)$$

Pro $i = 1$: n_1 - počet pacientů na počátku studie

c_i – cenzorovaní pacienti

f_i – úmrtí pacienta

$$S_i = 1 - (f_i/n_i)$$

n_i – počty pacientů během studie

Dále musíme vypočítat kumulaci výsledků. Pro jednotlivé dny studie vypočítáme kumulaci jako násobek všech pravděpodobností, které předcházeli tomuto dnu. [15]

Interval spolehlivosti pro Kaplan-Meiera vypočítáme tímto vzorcem:

Pro horní mez

$$S(t)^{\exp \left(\frac{1}{1.96} \cdot SE_{Tr} \right)},$$

pro dolní mez

$$S(t)^{\exp \left(-\frac{1}{1.96} \cdot SE_{Tr} \right)}.$$

Kde SE_{Tr} je vypočítáno

$$SE_{Tr}[S(t)] = \frac{\left[\sum_{j=0}^{t-1} \frac{d_j}{n_j(n_j - d_j)} \right]^{1/2}}{\left[- \sum_{j=0}^{t-1} \log \left(\frac{n_j - d_j}{n_j} \right) \right]}$$

K výpočtu použijeme následující data (Tabulka 7), počet pacientů vstupujících do studie $n = 36$.

Tabulka 7 - Hodnoty pro Kaplan-Meiera

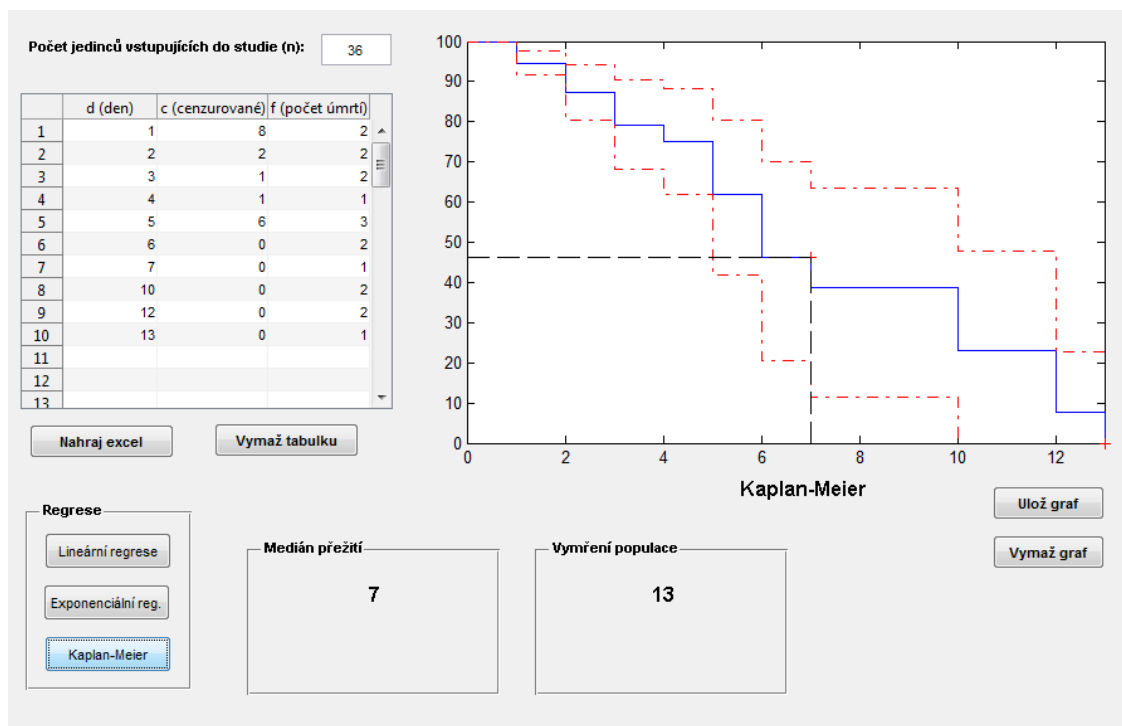
d (den)	c (cenzorovaní)	f (počet úmrtí)
1	8	2
2	2	2
3	1	2
4	1	1
5	6	3
6	0	2
7	0	1
10	0	2
12	0	2
13	0	1

Tabulka 8 - Hodnoty mediánu přežití a vymření populace pro Kaplan-Meiera

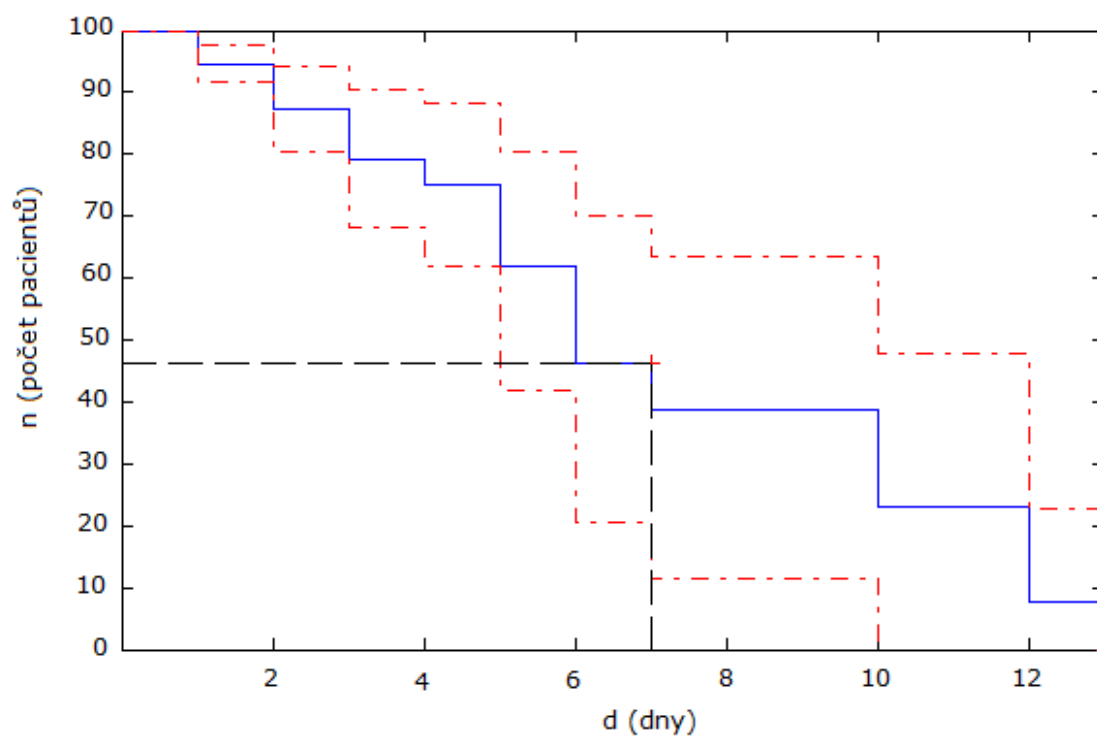
Medián přežití	Vymření populace
7	13

Kaplan-Meierův odhad je vhodný na porovnávání dvou a více skupin pacientů a jejich pravděpodobností přežití. Zadávaná data musí ovšem splňovat podmínku, že na konci měření je populace rovna nule.

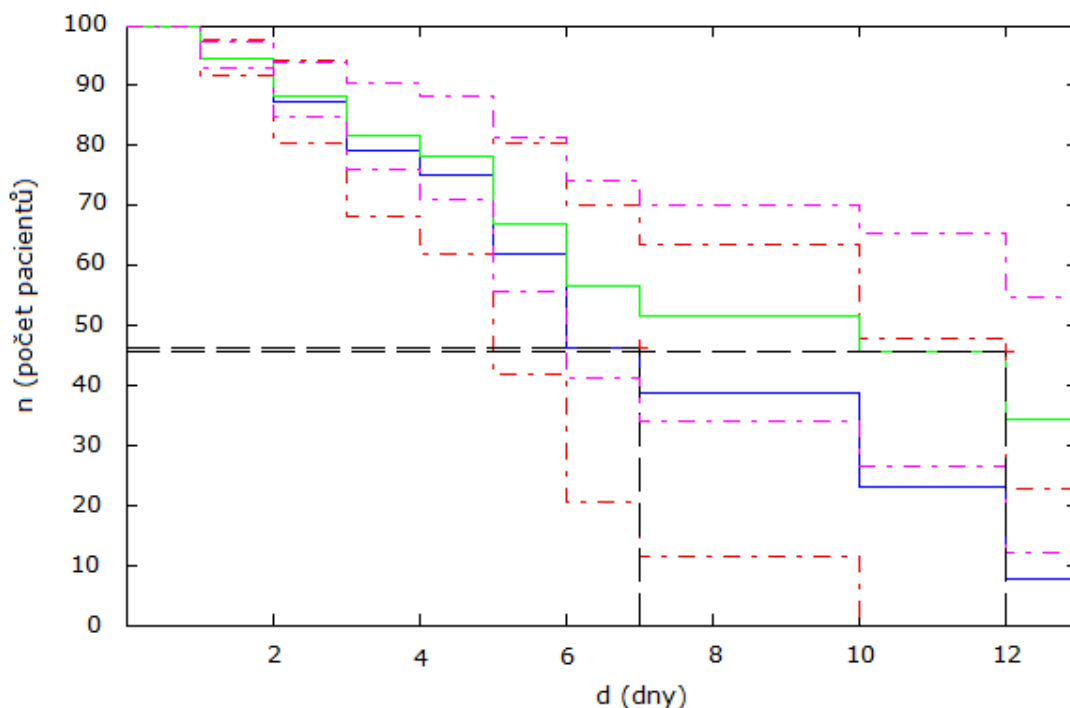
Na Obrázku 20 vidíme čelní panel se zadanými hodnotami spolu s grafem a vypočtenými hodnotami pro medián přežití a vymření populace.



Obrázek 20 - Kaplan-Meier



Obrázek 21 - Graf z mnou vytvořeného programu (Kaplan-Meier)



Obrázek 22 - Kaplan-Meier srovnání dvou skupin

Na Obrázku 22 vidíme srovnání dvou druhů léčby pro dvě skupiny pacientů. První skupina je znázorněna modrou barvou s červeným intervalem spolehlivosti. Druhá skupina je znázorněna zelenou barvou s fialovým intervalem spolehlivosti.

Tabulka 9 - Hodnoty mediánu přežití a vymření populace pro srovnávané skupiny

Skupiny	Medián přežití	Vymření populace
Modrá	7	13
Zelená	12	13

Z Obrázku 22 a Tabulky 9 můžeme usoudit, že účinnější je způsob léčby vyznačený zelenou barvou. Medián přežití je v tomto případě dvanáct dní, zatímco u léčby znázorněné modrou barvou je medián přežití roven sedmi.

ZÁVĚR

Možností jak zpracovávat statistická data týkající se analýzy přežívání je velké množství. Pro laika však není důležité, aby všechny tyto způsoby znal dopodrobna. Stačí mu jen základní znalosti o této problematice a sofistikovaný počítačový program, který tyto data zpracuje za něj.

Na trhu existuje velké množství tohoto druhu programů, ale pokud by měl člověk zcela specifické požadavky na program, který bude pro analýzu přežití používat, může si jej, s dostatečnou znalostí této problematiky, sám navrhnout v nějakém určitém, jemu vyhovujícím, programovacím jazyce. Také jsem se pokusila takovýto program vytvořit v programovacím prostředí Matlab. Tento program je vcelku jednoduchý na používání. Stačí když uživatel načte zpracovávaná data pomocí tlačítka *Nahrej Excel*, doplní počet pacientů účastnících se studie do políčka nad tabulkou a stiskne tlačítko pro jím zvolený způsob analýzy. Po vyhodnocení dat je uživateli poskytnut graf, který uživateli ukazuje závislost počtu objektů studie na čase, dále pak interval spolehlivosti pro daný graf, hodnoty mediánu přežití (střední hodnota délky přežití, polovina pozorovaných objektů se dožila času kratšího nebo rovného mediánu, druhá polovina se dožila času delšího než je hodnota mediánu) a vyměření populace (čas kdy se velikost populace se rovná nule), které jsou na grafu znázorněny červenými křížky. Tyto body jsou pro názornost spojeny s osami černými čárkovanými přímkami. Výsledný graf si může uložit pomocí tlačítka *Ulož graf*.

Porovnáním výsledků analýz mého programu a programu Excel, které vycházeli jen s minimální odchylkou, můžu říci, že mnou navržený program pracuje spolehlivě přesně. Odchyly hodnot mohou být způsobeny zaokrouhlováním hodnot při výpočtech.

Podklady a zdroje informací týkajících se analýzy přežití jsou ve většině případů napsány anglicky. České knihy o statistice obsahovali obvykle jen jednu nebo pár kapitol o analýze přežívání. Proto slouží moje práce i jako literární rešerše pro základní proniknutí do moderní problematiky analýzy přežívání, které se česká literatura zatím příliš nevěnuje.

POUŽITÁ LITERATURA

- [1] HENDL, Jan. *Přehled statistických metod zpracování dat: analýza a metaanalýza dat*. Praha : Portál, 2006. 583 s. ISBN 80-7367-123-9
- [2] ILLOVÁ, Kateřina. *Kaplanův - Meierův odhad funkce přežití*. Brno, 2006. 43 s. Bakalářská práce. Masarykova univerzita.
- [3] REISNEROVÁ, Soňa. *Analýza přežití a Coxův model pro diskrétní čas*. 2004. Praha : Robust, 2004. 8 s.
- [4] ILLOVÁ, Kateřina. *Neparametrické metody v analýze přežití*. Brno, 2009. 63 s. Diplomová práce. Masarykova univerzita.
- [5] KAHOUNOVÁ, Jana. *Odhad funkce přežití*. Praha, 2007. 9 s. Oborová práce. Acta OeconomicaPragensia.
- [6] MOTULSKY, Harvey. *Intuitive Biostatistics*. New York : Oxford University Press, 1995. 404 s. ISBN 0-19-508606-6
- [7] *MedCalc* [online]. 1993-2010 [cit. 2010-12-30]. MedCalc. Dostupné z WWW: <<http://www.medcalc.org/manual/survival-analysis.php>>.
- [8] *Survival Analysis - Kaplan-Meier* [online]. 2010 [cit. 2010-12-30]. OriginLab. Dostupné z WWW: <<http://www.originlab.com/index.aspx?go=Products/Origin/Statistics/SurvivalAnalysis>>.
- [9] *StatSoft* [online]. 2010 [cit. 2010-12-30]. Statistica. Dostupné z WWW: <<http://www.statsoft.com/textbook/survival-failure-time-analysis/#dgeneral>>.
- [10] *Emerald* [online]. 1984 [cit. 2011-01-03]. International Journal of Quality and Reliability Management. Dostupné z WWW: <<http://www.emeraldinsight.com/journals.htm?articleid=1463555&show=html>>.
- [11] ECKSCHLAGER, K.; HORSÁK, I.; KODEJŠ, Z. *Vyhodnocování analytických výsledků a metod*. Praha : SNTL, 1980. 223 s.
- [12] *GraphPad* [online]. 2010 [cit. 2011-05-15]. Quick Answers Database. Dostupné z WWW: < <http://www.graphpad.com/faq/viewfaq.cfm?faq=1499>>
- [13] MACHIN, David; CHEUNG, Yin Bun; PARMAR, Mahesh K.B. *Survival Analysis : A practical approach*. druhé. Chichester, England : John Willey & Sons Ltd, 2006. 278 s. ISBN 978-0470870402.

- [14] HEBÁK, Petr. *Regrese : I. část*. Praha : Ediční oddělení VŠE Praha, 1998. 138 s. ISBN 80-7079-909-9.
- [15] *Children's Mercy* [online]. 2000 [cit. 2011-05-17]. Kaplan Meier. Dostupné z WWW: <<http://www.childrensmercy.org/stats/model/kaplan.asp>>.
- [16] GARSON, David. *Quantitative Research in Public Administration* [online]. North Carolina : 2011, 21.4.2011 [cit. 2011-05-19]. Event History Analysis (Survival Analysis)). Dostupné z WWW: <<http://faculty.chass.ncsu.edu/garson/PA765/event.htm>>
- [17] DUNN, Steve. *Cancer Guide* [online]. 1995, July 2008 [cit. 2011-05-19]. Survival Curves: Accrual and The Kaplan-Meier Estiamate. Dostupné z WWW: <http://cancerguide.org/scurve_km.html>

SEZNAM OBRÁZKŮ

Obrázek 1 - Čas studie pro 8 pacientů [1]	8
Obrázek 2 - Křivka přežití [8]	12
Obrázek 3 - Odhad funkce přežití s intervalem spolehlivosti	14
Obrázek 4 – Příklad průběhu funkce rizika [10]	19
Obrázek 5 - Graf Kaplan-Meier (Matlab)	21
Obrázek 6 - Coxův regresivní model (Matlab)	21
Obrázek 7 – Ukázka odhadu funkce přežití (Statistica)	22
Obrázek 8 - Ukázka funkce přežití (Statistica)	22
Obrázek 9 - Tabulka pro Kaplan Meiera [7]	23
Obrázek 10 - Graf Kaplan-Meier MedCalc [7]	23
Obrázek 11 - Tabulka pro Coxovu regresivní analýzu MedCalc [7]	24
Obrázek 12 - Graf pro Coxovu regresivní analýzu	24
Obrázek 13 - Čelní panel programu	25
Obrázek 14 - Lineární regrese	29
Obrázek 15 - Graf z mnou vytvořeného programu (lineární regrese)	29
Obrázek 16 - Graf vytvořený v Excelu (lineární regrese)	30
Obrázek 17 - Exponenciální regrese	32
Obrázek 18 - Graf z mnou vytvořeného programu (exponenciální regrese)	33
Obrázek 19 - Graf vytvořený v Excelu (exponenciální regrese)	33
Obrázek 20 - Kaplan-Meier	37
Obrázek 21 - Graf z mnou vytvořeného programu (Kaplan-Meier)	37
Obrázek 22 - Kaplan-Meier srovnání dvou skupin	38