Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE

Zuzana Kaderjáková

Analýza přežití s programem STATISTICA

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Šárka Hudecová Ph.D.

Studijní program: Matematika

Studijní obor: Finanční matematika

Praha 2012

Na tomto mieste by som sa chcela poďakovať RNDr. Šárke Hudecovej Ph.D., vedúcej mojej bakalárskej práce, za jej ochotu, cenné rady a pripomienky. Zároveň ďakujem spoločnosti StatSoft CR s.r.o. za poskytnutie licencie k používanému softvéru.

Prehlasujem, že som túto bakalársku prácu vypracovala samostatne a výhradne s použitím citovaných prameňov, literatúry a ďalších odborných zdrojov.

Beriem na vedomie, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., autorského zákona v platnom znení, predovšetkým skutočnosť, že Univerzita Karlova v Praze má právo na uzavretie licenčnej zmluvy o použití tejto práce ako školského diela podľa §60 odst. 1 autorského zákona.

V Prahe dňa 25.5.2012 Zuzana Kaderjáková

Názov práce: Analýza přežití s programem STATISTICA

Autor: Zuzana Kaderjáková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedúci bakalárskej práce: RNDr. Šárka Hudecová Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Analýza prežitia je samostatnou štatistickou oblasťou, ktorá našla uplatnenie v mnohých odboroch. Táto bakalárska práca sa zaoberá výkladom základných pojmov, princípov a ďalej metód, ktoré sa používajú, a ktoré sú implementované v softvére *STATISTICA*. Venuje sa cenzorovaniu a možnostiam charakterizácie rozdelenia času prežitia. Uvádza Kaplan-Meierovu metódu odhadu funkcie prežitia a taktiež metódu tabuliek úmrtnosti. Neskôr pojednáva o základných možnostiach porovnania rozdelení času prežitia v dvoch skupinách a ich vhodnosti pre rôzne situácie. V práci sa ďalej zaoberáme možnosťami aplikácie metód analýzy prežitia vo finančnom sektore, kde uvádzame Coxov model proporcionálnych rizík. V závere práce aplikujeme teoretické vedomosti na reálnu skupinu dát.

Kľúčové slová: čas prežitia, riziková funkcia, Kaplan-Meierov odhad, logrankový test, finančné aplikácie

Title: Survival analysis with STATISTICA

Author: Zuzana Kaderjáková

Department: Department of probability and mathematical statistics

Supervisor: RNDr. Šárka Hudecová Ph.D., Department of probability and mathematical statistics

Abstract: Survival analysis is a separate statistical area. This paper discusses the interpretation of basic concepts, principles and methods used and implemented in the software *STATISTICA*. First, we introduce censoring and ways of characterizing a distribution of survival time. We present Kaplan-Meier estimate of a survival function and also a method of mortality tables. Later, we discuss basic methods of comparison of the survival time distribution in two groups and their suitability for different situations. The paper also deals with application of the survival analysis methods in the financial sector, where we introduce Cox proportional hazards model. Finally, we apply theoretical knowledge to a real data set.

Keywords: survival time, hazard function, Kaplan-Meier estimation, logrank test, financial applications

Obsah

Úvod

Analýza prežitia je súborom štatistických metód, ktoré skúmajú dobu do výskytu sledovanej udalosti, prípadne čas zotrvania v určitom stave. Metódy, ktoré sa s touto oblasťou štatistiky spájajú, našli svoje uplatnenie v rôznych vedných odboroch - v medicíne, farmácii a neskôr i v technických vedách a vo finančnom sektore. Metódy a spôsoby, ktoré popisujeme v nasledujúcich kapitolách boli vyberané na základe ponuky softvéru *STATISTICA*.

Cieľom tejto práce je vyložiť základnú problematiku, koncepty a používané metódy. Teoretické poznatky budeme priebežne ilustrovať na príkladoch z praxe a v závere práce zhrnieme väčšinu získaných poznatkov v štúdií na rozsiahlejšej skupine dát.

Práca je členená do piatich kapitol. Prvá kapitola pokrýva základné definície, pojmy a vzťahy medzi nimi. Predstavuje koncept cenzorovania odlišujúci analýzu prežitia od iných štatististických oblastí. Definuje tri základné funkcie charakterizujúce rozdelenie času prežitia a vysvetľuje vzťahy medzi nimi. Pre názornosť tieto vzťahy ilustrujeme na základných parametrických modeloch.

V druhej kapitole sa venujeme možnostiam odhadu rozdelenia času prežitia prostredníctvom funkcie prežitia. Medzi hlavné metódy, ktoré sa pre tento účel využívajú sú Kaplan-Meierova metóda a odhad pomocou tabuliek úmrtnosti.

Tretia kapitola je zameraná na základnú skupinu testov, ktoré sa v analýze prežitia najčastejšie používajú pre porovnanie rozdelenia času prežitia v skupinách dát.

Štvrtá kapitola venuje pozornosť aplikácií metód analýzy prežitia vo finančnom sektore. Primárnou oblasťou vzniku a rozvoja analýzy prežitia bola síce bioštatistika, no koncepty, ktoré boli navrhnuté, našli uplatnenie aj vo financiách. Spomenieme niekoľko príkladov z tejto oblasti, kde sa prvky objavujú a uvedieme známy Coxov model proporcionálnych rizík ako metódu, ktorá našla najväčšie uplatnenie v oblasti bankovníctva.

Záverečná piata kapitola obsahuje praktické spracovanie reálnych dát o prežití. Zahŕňa praktické návody ako postupovať pri práci so softvérom a poukazuje na rôzne ďalšie možnosti, ktoré sú v ponuke.

Obrázky a grafy použité v tejto práci, boli zostrojené v software *Mathematica* a *STATISTICA*.

1. Základné pojmy

Dáta o prežití majú určité charakteristické vlastnosti, ktoré je potrebné zohľadniť pri ich analyzovaní. Je preto nutné tieto vlastnosti popísať a zadefinovať základné pojmy, s ktorými sa v analýze prežitia pracuje.

Pre analýzu prežitia nie je podstatná samotná nastávajúca udalosť (smrť, porucha, bankrot), ale čas, ktorý do tejto udalosti uplynul. Preto ako prvý zadefinujeme pojem čas prežitia.

Definícia 1.1. *Čas prežitia* je doba, ktorá uplynula do výskytu pozorovanej udalosti. Značíme ju T.

Poznámka 1.1. *Veličina T nutne nadobúda len nezáporné hodnoty vzhľadom na to, že sa jedná o čas.*

Pre jednoznačné určenie času prežitia je potrebné jasne určiť tri základné elementy, ktorými sú počiatočný čas, nastávajúca udalosť a mierka pre prislúchajúci časový interval. Pre väčšinu udalostí možno zmysluplne určiť viacero počiatočných časov, viď príklad 1.1.

Príklad 1.1. Skúmame dobu do úmrtia. Počiatočný čas môžme definovať ako čas narodenia, čas vypuknutia choroby, čas začiatku liečby, prípadne čas hospitalizácie.

Čas prežitia často nazývame *odozvou*. Toto označenie zdôrazňuje fakt, že čas prežitia je veličina zväčša závislá na viacerých faktoroch, viď príklad 1.2.

Príklad 1.2. Dĺžka doby nezamestnatnosti je ovplyvnená úrovňou vzdelania, vekom danej osoby, prípadne geograficko-politickými faktormi. Dĺžka doby do úmrtia po transplantácií kostnej drene je ovplyvnená vekom pacienta, vekom darcu, hodnotami rôznych medicínskych ukazovateľov oboch osôb, prípadne doba do bankrotu na kreditnej karte je ovplyvnená makroekonomickými ukazovateľmi.

Existujú dva dôvody prečo dáta o prežití nie je vhodné analyzovať štandardnými analytickými metódami (viď zdroj [2]).

Po prvé, tieto dáta sú vo väčšine prípadov rozdelené nesymetricky, prevláda kladné zošikmenie. Tento rys je spôsobený tým, že odozva T môže nadobúdať len nezáporné hodnoty. Nie je teda vhodné využívať analytické nástroje založené na predpoklade normality rozdelenia.

Po druhé, dáta bývajú veľmi často cenzorované. Znamená to, že čas T je pozorovaný len čiastočne. U niektorých subjektov zo súboru počas štúdie nemuselo dôjsť ku zisteniu skúmanej udalosti, prípadne nemožno určiť, kedy presne udalosť nastala. Ak sa v pozorovanom súbore nenachádzajú cenzorované dáta, hovoríme o tzv. *kompletnom súbore časov T*.

1.1 Cenzorovanie a krátenie dát

Analýza prežitia sa od ostatných štatistických metód líši práve prítomnosťou cenzorovaných a krátených dát. Rozlišujeme niekoľko typov cenzorovania (viď zdroje [2], [6]). Najčastejšie sa vyskytuje cenzorovanie sprava.

Definícia 1.2. Nech $X_1, X_2, ..., X_n$ sú nezávislé, rovnako rozdelené doby prežitia a nech $C_1, C_2, ..., C_n$ sú nezávislé, rovnako rozdelené doby cenzorovania. Doba prežitia X_i príslušná i-temu subjektu bude známa práve vtedy, keď $X_i < C_i$. Ak $C_i < X_i$ potom bude čas do výskytu sledovanej udalosti u i-teho subjektu *cenzorovaný sprava v* C_i.

Je teda vhodné reprezentovať informáciu o prežití pomocou dvojice náhodných veličín (T_i, δ_i), kde $T_i = \min(X_i, C_i)$ a $\delta_i = \mathrm{I}(X_i < C_i)$ a I je indikátor výskytu udalosti, $\delta_i = 1$ ak udalosť nastala, $\delta_i = 0$ v prípade cenzorovania.

Najčastejšie dôvody pre cenzorovanie sprava:

- Subjekt nemôže byť ďalej sledovaný (strata záujmu o štúdiu, presťahovanie sa, strata kontaktu so subjektom).
- Pri ukončení štúdie u niektorých subjektov nenastala sledovaná udalosť.
- U sledovaného subjektu nastala iná udalosť, ktorá znemožnila ďalšie sledovanie (autonehoda, smrť z iných dôvodov).
- Potreba vyradenia sledovaného subjektu zo štúdie z rôznych dôvodov (neplnenie požiadaviek, zistenie dodatočných relevantných informácií).

Špeciálne typy cenzorovania sprava:

Typ I Za predpokladu, že nedošlo k žiadnym náhodným stratám subjektov, majú všetky cenzorované pozorovania rovnakú dĺžku a to čas trvania štúdie.

Typ II Pri cenzorovaní typu II štúdia končí, keď napozorujeme prvých N časov, ktoré nás zaujímajú. Všetky cenzorované pozorovania sú potom rovné najdlhšiemu necenzorovanému pozorovaniu, opäť za predpokladu, že nedošlo k žiadnym náhodným stratám.

Typ III Pri cenzorovaní typu III majú všetky cenzorované pozorovania rôzne hodnoty, ktoré závisia na povahe subjektov. Cenzorovaniu typu III sa tiež často hovorí náhodné.

Ďalšími, menej sa vyskytujúcimi typmi cenzorovania, sú cenzorovanie zľava a intervalové cenzorovanie.

Definícia 1.3. Nech $X_1, X_2, ..., X_n$ sú nezávislé, rovnako rozdelené doby prežitia a nech $C_1, C_2, ..., C_n$ sú nezávislé, rovnako rozdelené doby cenzorovania. Doba prežitia X_i príslušná i-temu subjektu bude známa práve vtedy, keď $X_i > C_i$. Ak $C_i > X_i$ potom bude čas do výskytu sledovanej udalosti u i-teho subjektu *cenzorovaný zľava v* C_i.

Opäť informáciu o prežití reprezentujeme pomocou dvojice náhodných veličín (T_i, δ_i), kde tentokrát $T_i = \max(X_i, C_i)$ a $\delta_i = \mathrm{I}(X_i > C_i)$ a I je indikátor výskytu udalosti, $\delta_i = 1$ ak udalosť nastala, $\delta_i = 0$ v prípade cenzorovania.

Definícia 1.4. Nech $C_1, C_2, ..., C_n$ sú nezávislé, rovnako rozdelené doby cenzorovania. Skutočný čas T bude *intervalovo cenzorovaný* ak platí, že $L < T \leq U$. Kde za L určíme najväčšie také C_i, pri ktorom nebolo zistené nastanie sledovanej udalosti a za U stanovíme najmenšie také C_j, pri ktorom bol prvýkrát zaznamenaný výskyt sledovanej udalosti, pre $i, j = 1, ..., n$.

Poznámka 1.2. *Je chybou určovať T ako ktorúkoľvek hodnotu z intervalu $(L, U]$.*

Pre aplikovanie štatických metód na dáta prežitia je ďalej nutné rozlišovať či sa jedná o cenzorovanie informatívne alebo neinformatívne.

Definícia 1.5. Povieme, že cenzorovanie je *neinformatívne*,ak hodnoty C_i sú nezávislé na T_i. Cenzorovanie považujeme za *informatívne*, ak distribúcia C_i obsahuje akúkoľvek informáciu o rozdelení T_i.

Väčšina štatistických metód vyžaduje neinformatívne cenzorovanie. V dobre navrhnutých štúdiách možno informatívne cenzorovanie takmer úplne eliminovať. Naďalej predpokladáme, že pracujeme s neinformatívnym cenzorovaním.

Ďalším typickým javom pre analýzu prežitia je *krátenie dát* (*truncation*). Krátenie dát sa vyskytuje v prípadoch, keď sú do štúdie zahrnuté len tie jednotky, ktorých udalosť nastala v danom časovom intervale . Podobne ako pri cenzorovaní, rozlišujeme krátenie zľava a sprava. Pri *krátení zľava* je $T_R = \infty$, pri *krátení sprava* je $T_L = 0$. O jednotkách, ktorých udalosť nastala mimo interval (T_L, T_R), nemáme žiadne informácie. Rozdiel medzi cenzorováním a krátením dát je ten, že v prípade cenzorovania máme o každej jednotke aspoň čiastočné informácie.

1.2 Rozdelenie času prežitia

Čas prežitia je náhodná veličina, ktorá môže nadobúdať len nezáporné hodnoty. Jej rozdelenie môže byť diskrétne ako aj absolútne spojité. Rozdelenie času prežitia zvyčajne charakterizujú tri funkcie: hustota pravdepodobnosti, funkcia prežitia a riziková funkcia. V praxi každá z týchto funkcií ilustruje rôzne vlastnosti dát.

Definícia 1.6. Nech T má diskrétne rozdelenie a nadobúda hodnoty $a_1, a_2, ..., a_n$ s pravdepodobnosťami $p_1,\ p_2,\ ...,\ p_n$, pre ktoré platí $\sum_{i=1}^{n} p_i = 1$. Potom *hustotu* definujeme ako

$$f(t) = \begin{cases} p_i & t = a_i, i = 1, 2, ..., n, \\ 0, & inak. \end{cases} \tag{1.1}$$

Pre spojitú náhodnú veličinu definujeme hustotu ako

$$f(t) = \lim_{\Delta t \to 0_+} \frac{P[t \leq T \leq t + \Delta t]}{\Delta t}. \tag{1.2}$$

Takto definovanú hustotu chápame v klasickom význame hustoty náhodnej veličiny (viď zdroj [7]). Naviac platí,že $f(t) \geq 0$ pre $t \geq 0$ a $f(t) = 0$ pre $t < 0$.

Definícia 1.7. *Funkcia prežitia* (*survival function*) náhodnej veličiny T je definovaná ako

$$S(t) = P[T \geq t]. \tag{1.3}$$

Funkcia prežitia teda udáva pravdepodobnosť, že individuálny subjekt prežije dlhšie ako t. Pre diskrétnu náhodnú veličinu T má funkcia prežitia predpis

$$S(t) = \sum_{a_i \geq t} f_i. \tag{1.4}$$

Pre T spojitú náhodnú veličinu získame zintegrovaním hustoty $f(t)$ v príslušných medziach

$$S(t) = \int_t^\infty f(u)\, du. \tag{1.5}$$

Ak označíme $F(t)$ distribučnú funkciu času T tak funkciu prežitia môžeme vyjadriť ako

$$S(t) = 1 - F(t). \tag{1.6}$$

Takto definovaná funkcia $S(t)$ je nerastúca a platí:

$$S(0) = 1,$$

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

Graf funkcie $S(t)$ nazývame krivkou prežitia. Strmá krivka reprezentuje krátky čas prežitia, naopak, pozvoľne klesajúca krivka reprezentuje dlhší čas prežitia.

Z funkcie prežitia môžeme určiť základné charakteristiky dát ako strednú hodnotu a medián. Tieto hodnoty potom môžu slúžiť na zachytenie trendu dát v skupine alebo na porovnávanie rôznych skupín dát. Použitie mediánu ako referenčnej hodnoty je preferované, pretože medián na rozdiel od strednej hodnoty nie je ovplyvnený extrémnymi pozorovaniami, a teda podáva neskreslenú informáciu o dátach.

Čas prežitia možno charakterizovať aj z iného pohľadu. Dôležitú informáciu o jeho rozdelení poskytujú *riziková funkcia* a *kumulatívna riziková funkcia*.

Definícia 1.8. *Riziková funkcia* (*hazard function*) udáva mieru pravdepodobnosti výskytu sledovanej udalosti v budúcom okamihu za predpokladu, že udalosť doposiaľ nenastala. Pre diskrétnu náhodnú veličinu T rizikovú funkciu definujeme ako podmienenú pravdepodobnosť

$$h_i = P[T = a_i \mid T \geq a_i]. \tag{1.7}$$

Pre spojitú náhodnú veličinu T rizikovú funkciu definujeme ako

$$h(t) = \lim_{\Delta t \to 0_+} \frac{P[t \leq T \leq t + \Delta t \mid t \leq T]}{\Delta t}. \tag{1.8}$$

Riziková funkcia môže byť rastúca, klesajúca, konštantná alebo môže indikovať zložitejší proces.

Definícia 1.9. Pre diskrétnu náhodnú veličinu T *kumulatívnu rizikovú funkciu* (*cumulative hazard function*) definujeme ako

$$H_i = \sum_{i:a_i<t} h_i. \tag{1.9}$$

Pre spojitú náhodnú veličinu T *kumulatívnu rizikovú funkciu* definujeme ako

$$H(t) = \int_0^t h(u)\, du. \tag{1.10}$$

Hustota, funkcia prežitia a riziková funkcia sú navzájom matematicky ekvivalentné, a to v prípade diskrétneho aj spojitého rozdelenia. Vo väčšine aplikácií (viď kapitoly 2, 3) sa však pracuje len s rozdelením spojitým, preto sa diskrétnymi prípadmi nebudeme ďalej zaoberať.

Veta 1.1. *Pre absolútne spojitú náhodnú veličinu T platia medzi vyššie definovanými funkciami nasledujúce vzťahy:*

1. $h(t) = \dfrac{f(t)}{S(t)}$,
2. $f(t) = -S'(t)$,
3. $h(t) = -[logS(t)]'$,
4. $S(t) = \exp\left\{-\int_0^t h(x)\, dx\right\}$,
5. $f(t) = h(t)\ \exp\left\{-\int_0^t h(x)\, dx\right\}$.

Dôkaz. Vzťah číslo 1 možno odvodiť zo vzorcov (1.8), (1.2) a znalosti podmienenej pravdepodobnosti nasledovne

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0_+} \frac{P[t \leq T \leq t + \Delta t \mid t \leq T]}{\Delta t} \\ &= \lim_{\Delta t \to 0_+} \frac{P[(t \leq T \leq t + \Delta t) \bigcap t \leq T]}{\Delta t P[t \leq T]} \\ &= \frac{1}{S(t)} \lim_{\Delta t \to 0_+} \frac{P[(t \leq T \leq t + \Delta t) \bigcap t \leq T]}{\Delta t} = \frac{f(t)}{S(t)}. \end{aligned}$$

Pri odvodzovaní vzťahu číslo 2 vychádzame zo vzťahu hustoty a distribučnej funkcie a vzorca (1.6)

$$f(t) = \frac{d\,F(t)}{d\,t} = \frac{d\,(1 - S(t))}{d\,t} = -S'(t).$$

Substitúciou vzťahu 1 do 2 dostávame vzťah 3

$$h(t) = -\frac{S'(t)}{S(t)} = -[logS(t)]'.$$

Vzťah číslo 4 možno dostať úpravami vzťahu 3. Ak využijeme vzťah rizikovej funkcie a kumulatívnej rizikovej funkcie (1.10), môžeme tiež vyjadriť $S(t)$ ako $S(t) = \exp\{-H(t)\}$. Substitúciou vzťahu 4 do 1 a úpravami získavame vzťah číslo 5

$$f(t) = h(t)\,S(t) = h(t) \exp\left\{-\int_0^t h(x)\,dx\right\}.$$

Hustotu taktiež možno vyjadriť pomocou kumulatívnej rizikovej funkcie ako

$$f(t) \;=\; h(t)\;\exp\{-H(t)\}\,.$$

□

1.3 Parametrické modely

Dáta analýzy prežitia zväčša dobre aproximujú exponenciálne, Weibullovo, lognormálne, prípadne gamma rozdelenie. Z nich exponenciálne a Weibullovo rozdelenie program *STATISTICA* ponúka pri modelovaní dát (viď kapitola 5). Preto sa zameriame výhradne na tieto dve rozdelenia.

1.3.1 Exponenciálne rozdelenie

Uvažujme, že T má exponenciálne rozdelenie s parametrom $\lambda > 0$. Hustota exponenciálneho rozdelenia je definovaná ako

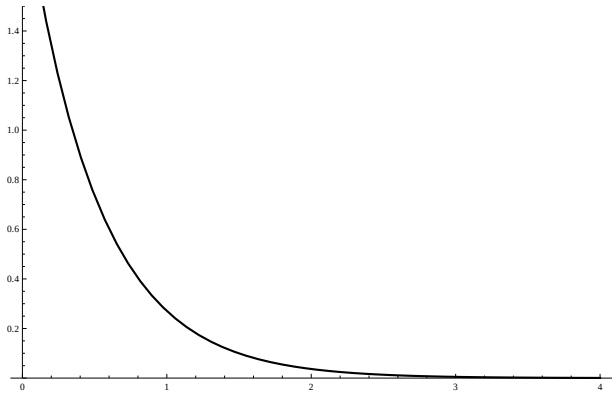
$$f(t) = \begin{cases} \lambda \exp\{-\lambda t\} & pre\ \ t \geq 0, \\ 0 & pre\ \ t < 0. \end{cases}$$

Pomocou vzťahov 1-5 z vety 1.1 odvodíme funkciu prežitia a rizikovú funkciu

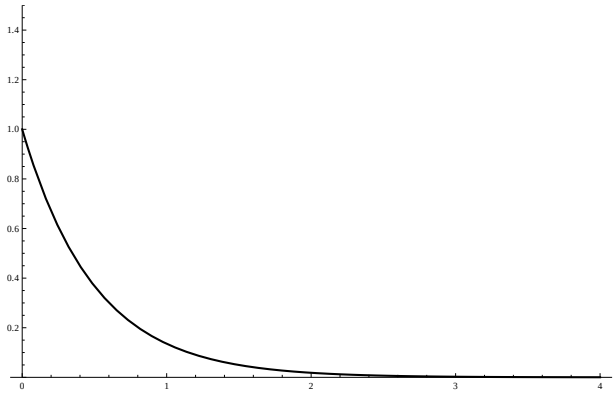
$$S(t) = \exp\{-\lambda t\},$$

$$h(t) = \lambda.$$

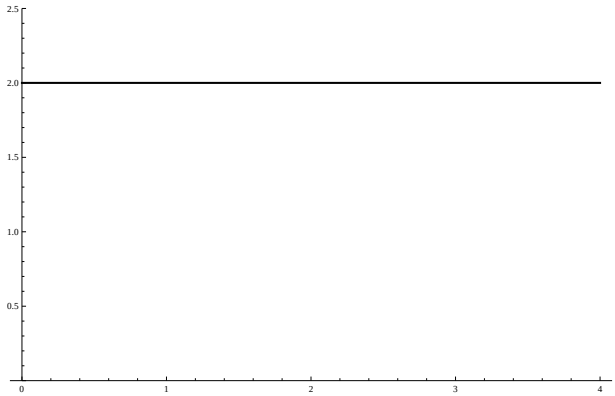
Na nasledujúcich troch grafoch sú ilustrované dané funkcie pre konkrétnu hodnotu parametra $\lambda = 2$. Všimneme si, že pri exponenciálnom rozdelení je riziková funkcia konštantná.



Obr. 1.1: Hustota exponenciálneho rozdelenia



Obr. 1.2: Funkcia prežitia exponenciálneho rozdelenia



Obr. 1.3: Riziková funkcia exponenciálneho rozdelenia

1.3.2 Weibullovo rozdelenie

Uvažujme, že T má Weibullovo rozdelenie s parametrami $\gamma > 0, \lambda > 0$. Weibullovo rozdelenie je zovšeobecnením exponenciálneho rozdelenia. V porovnaní s ním je však flexibilnejšie, preto má širšie využitie. Hustota Weibullovho rozdelenia je definovaná ako

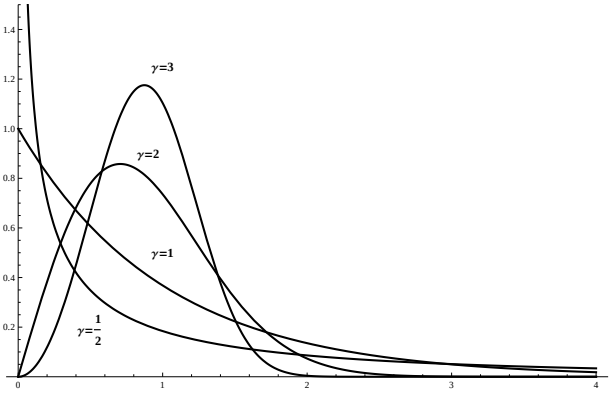
$$f(t) = \begin{cases} \lambda\gamma(\lambda t)^{\gamma-1} \exp\{-(\lambda t)^{\gamma}\} & pre\ t \geq 0, \\ 0 & pre\ t < 0. \end{cases}$$

Pomocou vzťahov 1-5 z vety 1.1 odvodíme funciu prežitia a rizikovú funkciu

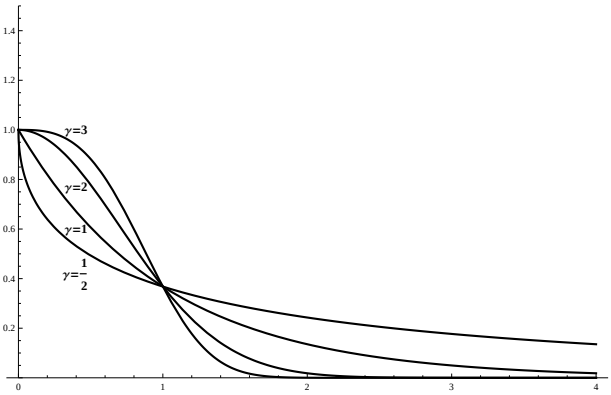
$$S(t) = \exp\left\{-(\lambda t)^{\gamma}\right\},$$

$$h(t) = \lambda\gamma(\lambda t)^{\gamma-1}.$$

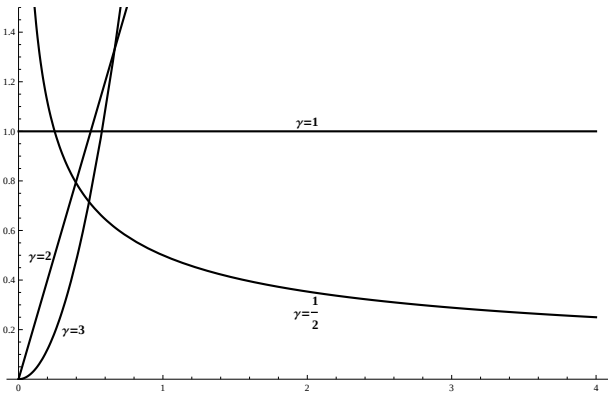
Parameter λ určuje škálovanie funkcií Weibullovho rozdelenia, parameter γ určuje ich tvar. Na nasledujúcich troch grafoch vidíme, ako sa pri zafixovanej hodnote parametra $\lambda = 1$ mení tvar funkcií pre rôzne hodnoty parametra γ.



Obr. 1.4: Hustota Weibullovho rozdelenia



Obr. 1.5: Funkcia prežitia Weibullovho rozdelenia



Obr. 1.6: Riziková funkcia Weibullovho rozdelenia

2. Odhad rozdelení

Jedným zo základných problémov analýzy prežitia je odhad rozdelenia doby prežitia T. Pri uvedených metódach odhadov budeme predpokladať, že T je absolútne spojitá náhodná veličina a pri uvažovaní cenzorovaných dát pracujeme len s neinformatívnym cenzorovaním sprava.

Rozdelenie doby prežitia T charakterizujú tri ekvivalentné funkcie. Najčastejšie sa využíva funkcia prežitia, respektíve jej grafická reprezentácia krivka prežitia. Preto sa zameriame predovšetkým na metódy odhadu tejto funkcie.

2.1 Empirický odhad

Empirický odhad funkcie prežitia možno aplikovať len na *dáta s kompletným súborom časov* T.

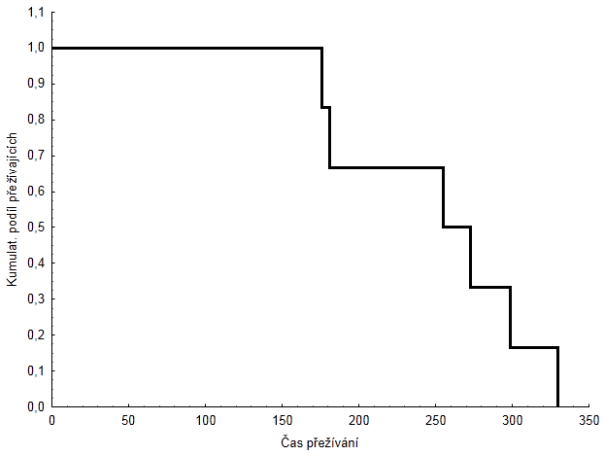
Definícia 2.1. Predpokladáme, že máme n pozorovaní $t_1, ..., t_n$. *Empirickú funkciu prežitia* definujeme ako

$$\widehat{S}(t) = \frac{\sum_{i=1}^{n} \mathrm{I}(t_i \geq t)}{n}, \tag{2.1}$$

kde I je identifikátor.

Takto definovaná empirická funkcia prežitia je skokovitá funkcia. Ak všetky pozorovania $t_1, ..., t_n$ nadobúdajú rôzne hodnoty, tak má $\widehat{S}(t)$ skok veľkosti $\frac{1}{n}$ v každom bode t_i, $i = 1, ..., n$. Ak označíme $\widehat{F}(t)$ *empirickú distribučnú funkciu času prežitia*, tak *empirickú funkciu prežitia* zo vzťahu (2.1) môžeme vyjadriť ako

$$\widehat{S}(t) = 1 - \widehat{F}(t).$$



Obr. 2.1: Empirický odhad funkcie prežitia zostrojený z ukončených pozorovaní z príkladu 2.1

Empirický odhad funkcie prežitia je najjednoduchším spôsobom odhadu, avšak pre väčšinu dát prežitia nepostačujúci. Nasledujúce dva uvedené spôsoby odhadu, ktoré sú implementované aj v softvéri *STATISTICA*, umožnujú prácu aj so skupinami dát s cenzorovanými pozorovaniami.

2.2 Kaplan-Meierov odhad

Ak sa v skupine dát vyskytujú cenzorované pozorovania, funkciu prežitia možno odhadnúť pomocou Kaplan-Meierovej metódy, známej tiež ako *product limit estimate*. Jedná sa o metódu založenú na idei podmienenej pravdepodobnosti, ktorú ako prví predstavili E. L. Kaplan a P. Meier v roku 1958.

Definícia 2.2. Označme $t_{(1)} < ... < t_{(m)}$ zoradené časy, v ktorých pre pozorované udalosti platí, že $\delta_i = 1$, n_j počet jednotiek, u ktorých nastala pozorovaná udalosť v čase t_j alebo neskôr, d_j počet jednotiek, u ktorých nastala pozorovaná udalosť v čase t_j. Potom *Kaplanov-Meierov odhad funkcie prežitia* je definovaný ako

$$\widehat{S}(t) = \widehat{P}[T > t] = \prod_{j:t_j \leq t} \frac{n_j - d_j}{n_j} = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \tag{2.2}$$

Poznámka 2.1. *Podiel* $\frac{d_j}{n_j}$ *odhaduje riziko udalosti v čase* t_j.

Takto definovaný odhad funkcie prežitia nájdeme v zdroji [1]. Existuje niekoľko alternatívnych zápisov odhadu (viď zdroje [2], [3]).

Funkcia $\widehat{S}(t)$ získaná pomocou Kaplan-Meierovej metódy je taktiež skokovitá funkcia, ktorá má skok v každom bode t_i, v ktorom je $\delta_i = 1$. Cenzorované pozorovania skok nespôsobujú, avšak navyšujú menovateľa n_j, a teda tiež prispievajú informáciou pre odhad funkcie prežitia. Ak by sa v skupine dát nevyskytovali žiadne cenzorované pozorovania, tak Kaplan-Meierov odhad funkcie prežitia je totožný s empirickou funkciou prežitia.

Príklad 2.1. Počas časového intervalu jeden rok skúmame vplyv ťažkých kovov na život rastlín. Desať kusov rastlín vystavíme silnému vplyvu týchto kovov a výsledky zaznamenávame do tabuľky 2.1, čas prežitia uvádzame v dňoch. Niektoré rastliny pod vplyvom ťažkých kovov umreli, niektoré prežili a o niektorých sme stratili informáciu (napr. umreli z dôvodu prudkého slnečného žiarenia). Pre prvú skupinu bude $\delta_i = 1$, pre zvyšné dve skupiny je $\delta_i = 0$.

číslo rastliny	čas prežitia	cenzorovacia premenná δ_i
1	365	0
2	181	1
3	176	1
4	204	0
5	245	0
6	299	1
7	255	0
8	273	1
9	330	1
10	255	1

Tabuľka 2.1: Výsledky výskumu

Pre Kaplan-Meierov odhad funkcie prežitia najskôr jednotlivé pozorovania usporiadame vzostupne podľa času prežitia.

číslo rastliny	čas prežitia	cenzorovacia premenná δ_i
3	176	1
2	181	1
⋮	⋮	⋮
9	330	1
1	365	0

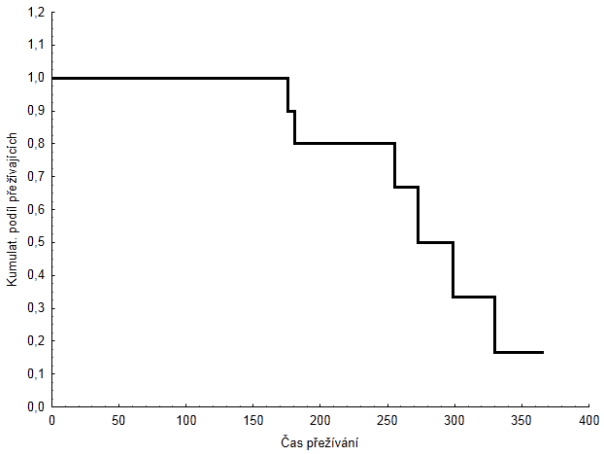
Tabuľka 2.2: Vzostupne usporiadané pozorovania

Odhad funkcie prežitia potom počítame postupne podľa vzorca (2.2).

čas t_i	n_i	d_i	$\widehat{S}(t_i)$
0	10	0	1
176	10	1	0,9
181	9	1	0,8
204	8	0	0,8
245	7	0	0,8
255	6	1	0,666
273	4	1	0,5
299	3	1	0,333
330	2	1	0,166
365	1	0	0,166

Tabuľka 2.3: Hodnoty funkcie prežitia

Získané hodnoty potom zakreslíme do grafu.



Obr. 2.2: Kaplan-Meierov odhad funkcie prežitia

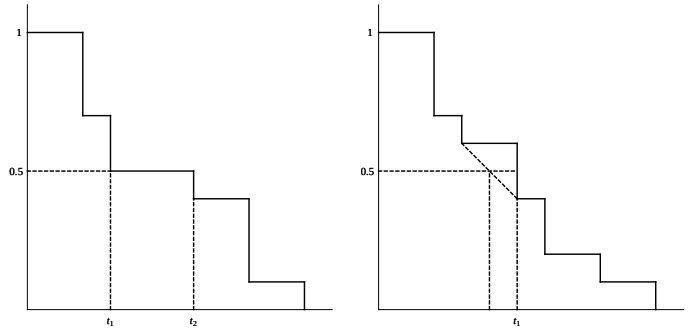
Poznámka 2.2. *Na obrázku 2.2 si možno všimnúť jednu z vlastností Kaplan-Meierovho odhadu a síce, že v skupine zoradených pozorovaní* $t_{(1)} \leq ... \leq t_{(n)}$,

z ktorých bol odhad zostrojený, bolo $t_{(n)}$ cenzorované, preto odhadnutá funkcia prežitia nedosiahla 0. Ak by $t_{(n)}$ bolo necenzorované, tak Kaplan-Meierov odhad v tomto bode by bol rovný 0.

Kaplan-Meierov odhad je veľmi rozšírená metóda odhadu funkcie $S(t)$. Jedná sa o jej neparametrický maximálne vierohodný odhad. Breslow, Crowley a Meier tiež dokázali, že za určitých podmienok je tento odhad tiež konzistentný a asymptoticky normálny (viď zdroj [2]).

Medián

Ako sme už uviedli, najčastejšie používaným ukazovateľom polohy je medián a nie stredná hodnota. Medián[1] možno určiť z Kaplan-Meierovho odhadu funkcie prežitia ako čas t, pre ktorý platí $\widehat{S}(t) = 0,5$. Avšak kvantily odhadu nie sú určené jednoznačne. Na obrázku 2.3 vľavo je riešením rovnice $\widehat{S}(t) = 0,5$ celý interval (t_1, t_2). V praxi sa potom za medián najčastejšie berie artitmetický, prípadne vážený priemer časov t_1, t_2. Na obrázku 2.3 vpravo uvádzame situáciu, kedy pre riešenie t_1 rovnica $\widehat{S}(t) = 0,5$ nadobúda aj iných hodnôt okrem $0,5$. V takomto prípade spojíme konce intervalu a následne lokalizujeme medián.



Obr. 2.3: Kaplan-Meierov odhad mediánu

2.3 Tabuľky úmrtnosti

Metóda tabuliek úmrtnosti, ktorá sa často označuje ako aktuárska metóda, je jednou z najstarších techník analýzy prežitia. Je založená na podobnom princípe ako Kaplan-Meierova metóda. Používa sa zväčša pri veľkom množstve dát, ktoré sú rozdelené do skupín. Cieľom je opäť odhadnúť funkciu prežitia, avšak tentokrát je situácia skomplikovaná faktom, že nevieme, kedy presne situácia nastala v danom časovom intervale. Táto skutočnosť nás vedie k úprave hodnoty n_j, ktorá naďalej reprezentuje počet jednotiek, u ktorých nastala pozorovaná udalosť v čase t_j alebo neskôr. Vzhľadom k tomu, že jednotky nemusia byť vystavené riziku počas celého časového intervalu $I_j = (t_{j-1}, t_j)$ tak prirodzeným spôsobom upravíme hodnotu n_j na novú

$$n_j^{'} = n_j - \frac{c_j}{2},$$

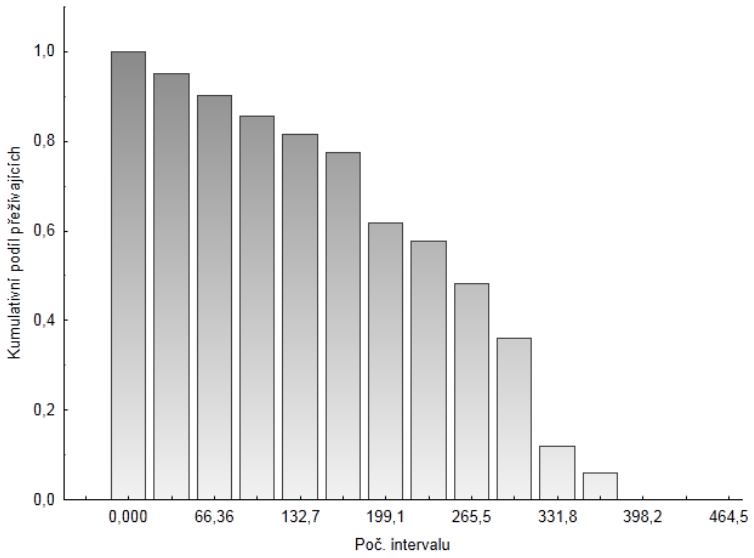
[1] Analogicky ostatné kvantily.

kde c_j je počet cenzorovaných jednotiek v intervale I_j.

Definícia 2.3. *Aktuársky odhad funkcie prežitia* za interval I_j je definovaný ako

$$\widehat{S}(t_j) = \prod_{k=1}^{j} \left(1 - \frac{d_k}{n_k^{'}}\right) \tag{2.3}$$

Poznámka 2.3. *Podobne ako pri Kaplan-Meierovom odhade zlomok* $\frac{d_j}{n_j^{'}}$ *odhaduje riziko udalosti v intervale* I_j.



Obr. 2.4: Aktuársky odhad funkcie prežitia zostrojený z dát z príkladu 2.1

3. Porovnanie rozdelení

V predchádzajúcej kapitole sme uviedli niekoľko možností ako odhadnúť distribúciu časov T. Úlohou dvojvýberových a viacvýberových testov je určiť či sa táto distribúcia líši medzi dvomi prípadne viacerými skupinami pozorovaní. V analýze prežitia sa táto úloha formuluje ako porovnávanie funkcií prežitia v jednotlivých skupinách. Testujeme hypotézu

$$H_0 : \ S_1 = S_2 = ... = S_k,$$

kde S_i je funkcia prežitia v i-tej skupine pre $i = 1, 2, ..., k$.

Keď sa v skupinách pozorovaní nevyskytujú žiadne cenzorované pozorovania, na porovnanie distribúcií možeme využiť klasické neparametrické testy, napríklad Wicoxonov test, jeho upravenú podobu Mannov-Whitneyov test, Kolmogorovov-Smirnov test na porovnanie dvoch nezávislých skupín, prípadne Kruskallov-Wallisov test na porovnanie viacerých skupín. Každý z testov sa líši prístupom k danej problematike a teda je optimálny pre rôzne skupiny dát.

Pre skupiny dát obsahujúcich cenzorované pozorovania, boli špeciálne vyvinuté testy, ktoré si s nimi poradia. Možno ich samozrejme aplikovať aj na skupiny bez cenzorovaných pozorovaní, ktoré by v tomto prípade boli považované za špeciálny prípad.

V nasledujúcom texte popisujeme niektoré vybrané testy pre 2 skupiny pozorovaní. Testujeme teda hypotézu $H_0 : \ S_1 = S_2$ proti $H_1 : \ S_1 \neq S_2$.

3.1 Logrankový test

Gehanov-Wilcoxonov test, Peto-Petoov Wilcoxonov test a Cox-Mantelov test (všetky implementované v programe *STATISTICA*) patria do skupiny testov, ktoré vychádzajú z jednotného základu. Je ním tzv. logrankový test patriaci medzi najznámejšie a najrozšírenejšie testy. Tento vychádza z Cochran-Mantel-Haenszelovho testu nezávislosti v stratifikovaných kontingenčných tabuľkách. Označme n_1, n_2 počet pozorovaní v jednotlivých skupinách, $n = n_1 + n_2$ celkový rozsah výberu, $t_{(1)} \leq t_{(2)} \leq ... \leq t_{(M)}$ zotriedené necenzorované časy z oboch skupín, $n_{j,k}$ počet pozorovaní z k-tej skupiny, pre ktoré nastala udalosť v čase $t_{(j)}$ alebo neskôr, $d_{j,k}$ počet jednotiek z k-tej skupiny, pre ktoré nastala udalosť v čase $t_{(j)}$. Pre každý z časov $t_{(1)} \leq t_{(2)} \leq ... \leq t_{(M)}$ máme kontingenčnú tabuľku

	udalosť		
skupina	nastala	nenastala	celkom
1	$d_{j,1}$	$n_{j,1} - d_{j,1}$	$n_{j,1}$
2	$d_{j,2}$	$n_{j,2} - d_{j,2}$	$n_{j,2}$
celkom	d_j	$n_j - d_j$	n_j

Pri platnosti hypotézy H_0 je

$$\mathrm{E}(d_{j,1}) = \frac{n_{j,1} d_j}{n_j}$$

a rozptyl (určený na základe hypergeometrického modelu) je

$$\mathrm{Var}(d_{j,1}) = \frac{n_{j,1} n_{j,2} (n_j - d_j) d_j}{n_j^2 (n_j - 1)}.$$

Označíme štatistiku S

$$S = \sum_{j=1}^{M} w_i [d_{j,1} - E(d_{j,1})],$$

kde w_j sú rôzne určené váhy (viď tabuľka 3.1). Potom pri platnosti hypotézy H_0 má S strednú hodnotu

$$\mathrm{E}(S) = 0$$

a rozptyl

$$\mathrm{Var}(S) = \sum_{j=1}^{M} w_j^2 \mathrm{Var}(d_{j,1}) = \sum_{j=1}^{M} \frac{w_j^2 n_{j,1} n_{j,2} (n_j - d_j) d_j}{n_j^2 (n_j - 1)}.$$

Štandardizovaná testová štatistika

$$L = \frac{S}{\sqrt{\mathrm{Var}(S)}} \tag{3.1}$$

má za platnosti hypotézy H_0 asymptoticky normované normálne rozdelenie a L^2 má χ^2 rozdelenie s jedným stupňom voľnosti. Hypotézu H_0 teda zamietame pre veľké hodnoty testovej štatistiky L^2.

Rôznou voľbou váh w_j v (3.1) potom získavame vyššie zmienené testy, pričom rozdelenie testovej štatistiky L ostáva nezmenené.

Test	Váhy w_j
Cox-Mantelov test	$w_j = 1$
Gehanov-Wilcoxonov test	$w_j = n_j$
Peto-Petoov Wilcoxonov test	$w_j = \begin{cases} \widehat{S}(t+) + \widehat{S}(t-) - 1, & pre\ \delta_j = 1 \\ \widehat{S}(t) - 1, & pre\ \delta_j = 0 \end{cases}$

Tabuľka 3.1: Rôzne voľby váh w_j

3.2 Poznámky k jednotlivým testom

Gehanov-Wilcoxonov, Peto-Petoov Wilcoxonov a Cox-Mantelov test sú založené na rovnakom princípe prideľovania váh jednotlivým pozorovaniam, avšak citilivé na rôzne typy odlišností v rozdeleniach. Ak porovnávame silu týchto testov

spolu s Coxovým F-testom[1] zistíme, že pri malom rozsahu skupín ($n_1, n_2 \leq 50$) je Coxov F-test silnejší ako Gehanov-Wilcoxonov test, ak pozorovania pochádzajú z exponenciálneho alebo Weibullovho rozdelenia. Pre pozorovania pochádzajúce z exponciálneho rozdelenia ďalej platí, že Cox-Mantelov test je silnejší ako Gehanov-Wilcoxonov aj Peto-Petoov Wilcoxonov test. Medzi je zovšeobecnenými Wilcoxonovými testami je len malý rozdiel v sile. Tieto testy ďalej dávajú väčšiu váhu skorším „úmrtiam“ než neskorším, preto na rozdiel od Cox-Mantelovho testu, ktorý dáva všetkým „úmrtiam“ rovnakú váhu, detekujú skoršie odlišnosti v distribúcii. Vo všeobecnosti možno povedať, že logrankové testy nie sú veľmi účinné v prípade križujúcich sa funkcií prežitia či rizikových funkciách. V tomto prípade je potrebné zvážiť použitie iných testov.

Tieto a ďalšie podrobnosti o používaní a vlastnostiach zmienených testov možno nájť v zdroji [2] kapitola 5, prípadne v zdroji [1] kapitola 3.

Poznámka 3.1. *V tejto kapitole sme sa zamerali predovšetkým na dvojvýberové problémy. Všetky spomenuté testy možno rozšíriť aj pre viacvýberové problémy. Softvér STATISTICA opäť ponúka niekoľko možností riešenia danej problematiky.*

[1] Coxov F-test nájdeme popísaný v zdroji [2] sekcia 5.1 alebo v zdroji [1] sekcia 3.2.

4. Analýza prežitia vo financiách

Počiatky analýzy prežitia siahajú do 18. storočia, kedy boli zaznamenané prvé analýzy ľudskej úmrtnosti. V čase 2. svetovej vojny bolo primárnou oblasťou záujmu strojárenstvo a zbrojný priemysel, analýza prežitia bola teda zameraná na dobu životnosti vojenskej výzbroje. V technických odboroch sa však častejšie ako s pojmom analýza prežitia stretávame s pojmom analýza spoľahlivosti, ide však o synonymické označenie. Po skončení 2. svetovej vojny sa oblasťou záujmu opäť stalo zdravotníctvo, farmaceutický priemysel a ekonómia.

V tejto kapitole si ukážeme niekoľko oblastí z finančnej sféry, kde možno pozorovať prvky analýzy prežitia a uvedieme metódu, ktorá našla v tejto oblasti najväčšie uplatnenie.

S nastupujúcim rozvojom ekonómie sa analýza prežitia začala využívať v rade finančných modelov, ktoré sledujú dobu trvania určitého javu, typicky dobu do transakcie cenného papieru na burze od predchádzajúcej transakcie, dobu splácania úveru, dobu do likvidácie poistnej udalosti prípadne dobu nezamestnanosti. Opäť sa môžeme stretnúť s novým označením pre analýzu prežitia, ktoré lepšie vystihuje podstatu ekonomických problémov, a síce analýza durácie (doby trvania).

Ako prvý bod, ktorým sa analýza prežitia vo financiách líši od ostatných oblastí je *cenzorovanie a krátenie dát*. K cenzorovaniu a kráteniu dochádza zväčša z časových dôvodov no nemusí to tak byť vždy. Vo svete financií a poisťovníctva môže byť cenzorovanou premennou napríklad výška platov, ktoré firma nechce zverejniť nad určitou hranicou, povedzme 100 000 Kč (*cenzorovanie sprava*). Prípadne rodiny, ktoré by nekúpili nejaký výrobok ako príliš drahý vzhľadom k danému cenovému limitu, sposobujú v ankete *cenzorovanie zľava*. V poisťovníctve sa pri navrhovaní modelov pre rozdelenie výšky škôd využívajú rozdelenia, ktoré predpokladajú hornú hranicu pre výšku vzniknutej škody (*cenzorovanie sprava*). *Krátenie dát* môžeme pozorovať na príklade investičnej spoločnosti, ktorá svojím klientom ponúkne nový produkt a skúma výšku investícií jednotlivých klientov. Avšak mnoho klientov o daný produkt neprejaví záujem, preto ich spoločnosť zo svojej monitorovanej skupiny vypustí.

Ďalej uvádzame *tabuľky úmrtnosti*, ktoré sa využívali už v minulosti s cieľom monitorovať vývoj ľudskej populácie. V súčasnej dobe sa v takejto klasickej podobe využívajú napríklad v poisťovníctve. Životné poisťovne ponúkajú poistenie pre prípad smrti, dožitia, príbadne zmiešané poistenie. Pri zostavovaní tarifných skupín, určovaní výšky poistného a iných vecí s poistením súvisiacich, využívajú tabuľky úmrtnosti ako podklad prvého rádu.

K úlohám, ktoré analýza prežitia rieši, patrí aj skúmanie vplyvu rôznych faktorov na rozdelenie času prežitia. Štatisitcké aparáty, ktoré sa zaoberajú týmto problémom, nazývame regresné modely.

Jeden z modelov veľmi dobre prispôsobených pre prácu s dátami o prežití je *Coxov model proporcionálnych rizík*. Tento model vychádza zo zložitejšej teórie, ktorej sa nebudeme venovať. Preto len načrtneme ako daný model vyzerá a zameriame sa najmä na jeho využitie vo finančnom sektore.

Definícia 4.1. *Coxov model proporcionálych rizík* je model, v ktorom je vplyv faktorov určený pomocou rizikovej funkcie

$$h(t) = h_0(t) \exp\left\{\beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k\right\},$$

kde $X_1, ..., X_k$ sú vplývajúce faktory, $\beta_1, ..., \beta_k$ sú hľadané regresné parametre a h_0 je základná riziková funkcia. Funkcia h_0 odráža riziko, ktoré prislúcha jednotke, kde sú všetky faktory rovné nule.

Podrobnejšie informácie o Coxovom modeli možno nájsť v zdroji [2] kapitola 12 alebo v zdroji [1] kapitola 4.

Banky, ako sprostredkovatelia na finančnom trhu, zhromažďujú voľné finančné prostriedky a následne ich poskytujú formou pôžičiek, úverov a iných produktov svojim klientom. Počas tohto procesu sa vystavujú tzv. kreditnému riziku (riziko zlyhania zmluvnej protistrany). Musia teda vyvinúť aparáty na ohodnocovanie svojich klientov, aby minimalizovali straty plynúce z tohto typu rizika. Štandardne sa pre tieto účely používa model logistickej regresie. *Coxov model proporcionálnych rizík* má oproti logistickej regresii tú výhodu, že okrem odpovede na otázku či dôjde k danej situácií (neschopnosť klienta dodržať svoje záväzky) pridáva aj informáciu o tom, kedy k nej dôjde.

Pomocou *Coxovho modelu* teda možno skúmať vplyv rôznych makroekonomických ukazovateľov, akými sú napr. úrokové miery a index nezamestnanosti, na pravdepodobnosť bankrotu na klientských kreditných kartách, prípadne ich vplyv na neschopnosť klienta splácať pôžičku (*credit-scoringové modely*). Tieto ukazovatele sú však spravidla závislé na čase (čo je rozpor so základnými predpokladmi modelu), preto sa používa upravený *model proporcionálneho rizika s časovo závislými kovariantmi*.

Podrobnejšie informácie o využití Coxovho modelu vo financiách možno nájsť v článkoch [4], [5].

5. Ukážka analýzy dát

V tejto kapitole uvádzame postup pri analýze dát v programe *STATISTICA*. Balíček *analýza prežitia* sa nachádza v menu medzi štatistikami ako jeden z pokročilých lineárnych/nelineárnych modelov. Zo základného výberu si možno zvoliť:

- úmrtnostné tabuľky & rozdelenia,
- Kaplan-Meierova metóda,
- porovnanie 2 vzoriek,
- porovnanie viacerých vzoriek,
- regresné modely,
- časovo závislé kovarianty.

Na konkrétnych dátach demonštrujeme teoretické vedomosti z predchádzajúcich kapitol a poukážeme na rôzne možnosti, ktoré program ponúka. Zameriavame sa predovšetkým na Kaplan-Meierov odhad funkcie prežitia, odhad funkcie prežitia pomocou tabuliek úmrtnosti a na porovnanie rozdelenia času prežitia v dvoch skupinách dát.

V predchádzajúcej kapitole sme sa venovali rôznym aplikáciám vo finančnej sfére, avšak väčšina dát z tohto prostredia nie je verejne prítupná, preto sme pre demonštráciu zvolili dáta z oblasti, kde sa analýza prežitia začala využívať pôvodne, a to oblasti medicíny.

5.1 Dáta a ich reprezentácia

Máme údaje o 863 pacientoch, ktorí podstúpili transplantáciu obličiek. V softvéri *STATISTICA* sú dáta uložené v prehľadnej tabuľke. Ak dáta priamo netvoríme v softvéri možno ich importovať z iných zdrojov (napr. Excel). V našom prípade má tabuľka 863 riadkov a 5 stĺpcov. Počet riadkov odpovedá počtu pacientov a počet stĺpcov počtu údajov prisluchájúcich jednotlivým pacientom. O každom pacientovi máme v premennej

time zaznamenaný čas prežitia v dňoch,

indicator indikátor udalosti: $\delta_i = 1$ *pre smrť a* $\delta_i = 0$ *pre život,*

gender pohlavie pacienta: *1 = muž, 2 = žena*,

race rasu pacienta: *1 = biela, 2 = čierna*,

age vek pacienta v čase transplantácie.

Obrázok 5.1 predstavuje ukážku dát uložených v tabuľke. Kompletné dáta možno nájsť v zdroji [8].

	1 time	2 indicator	3 gender	4 race	5 age
1	1	0	1	1	46
2	5	0	1	1	51
3	7	1	1	1	55
4	9	0	1	1	57
5	13	0	1	1	45
6	13	0	1	1	43
7	17	1	1	1	47
8	20	0	1	1	65
9	26	1	1	1	55
10	26	1	1	1	44
11	28	1	1	1	49
12	32	0	1	1	52
13	32	0	1	1	31
14	43	0	1	1	63
15	43	1	1	1	55
16	44	1	1	1	50

Obr. 5.1: Ukážka dát

5.2 Kaplan-Meierov odhad

Ak v programe *STATISTICA* používame na odhad funkcie prežitia Kaplan-Meierovu metódu, najskôr vyberieme požadované premenné: časy prežitia a cenzorovanú premennú. V našom prípade *time* a *indicator*. Následne nastavíme správne kódy pre cenzorovanie[1]. Ako výsledok Kaplan-Meierovej analýzy sa ukáže tabuľka skladajúca sa z dvoch častí. V hornej časti znázornenej na obrázku 5.2 vidíme základný sumár informácií o pozorovaniach.

```
Proměnná: time
Proměnná s indikátorem cenzor.:    indicator

Celkový počet platných pozorování : 863
    necenzor. :    140 ( 16,22%)    cenzor. :  723 ( 83,78%)
```

Obr. 5.2: Sumár informácií

V dolnej časti tabuľky sa nachádzajú ďalšie možnosti výberu:

- **základné výsledky**: analýza prežívania (obrázok 5.3),
- **detaily**: kvantily funkcie prežívania (obrázok 5.5),
- **Kaplan-Meierove grafy**: rôzne podoby grafov (obrázok 5.4).

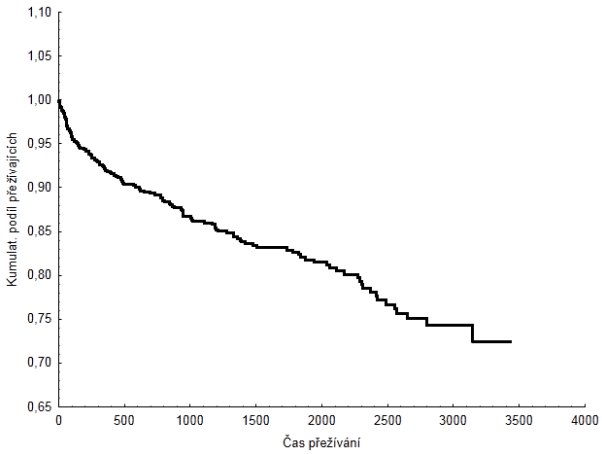
Obrázkok 5.3 je ukážkou výstupu základnej Kaplan-Meierovej analýzy. V prvom stĺpci sa nachádzajú čísla pozorovaní spolu s informáciou o cenzorovaní. Softvér *STATISTICA* pre označenie cenzorovaných pozorovaní používa symbol +. Nasleduje stĺpec s hodnotami časov prežitia. V stĺpci označenom kumulatívne podiely prežívajúcich sú hodnoty odhadnutej funkcie prežitia a v ďalšom stĺpci smerodajné odchýlky. Pozorovania, v ktorých hodnota odhadnutých funkčných hodnôt ostáva nemenná sú prázdne, podobne ako pri smerodajných odchýlkach.

[1]Nie vždy musí v dátach platiť, že 1 je kód pre ukončené a 0 pre cenzorované pozorovanie.

Číslo případu	Čas	Kumulat. přežív.	Směrod. chyba
6+	5,000		
5+	5,000		
8	7,000	0,996513	0,002010
7	7,000	0,995349	0,002320
10+	9,000		
9+	9,000		
12	10,000	0,994182	0,002594
11	10,000	0,993015	0,002842
13+	13,000		
14+	13,000		
15+	14,000		
16	17,000	0,991844	0,003070

Obr. 5.3: Ukážka výstupu Kaplan-Meierovej analýzy

Obrázok 5.4 ukazuje jednu z podôb, v akej sa Kaplan-Meierov graf dá vykresliť.



Obr. 5.4: Kaplan-Meierov graf

Znázornená funkcia prežitia na začiatku prudko klesne a potom zvoľní a klesá pomalšie. Tento fakt vysvetľujeme skutočnosťou, že riziko úmrtia býva najvyššie hneď po transplantácii. Najnižšia úroveň, na ktorú sa hodnoty funkcie dostanú, je približne 0,72. To znamená, že 9 rokov po transplantácii je nažive 72 % pacientov.

Faktom, že funkcia prežitia dosahuje hodnoty nad hranicou 0,72 vysvetľujeme aj prázdne miesta pri hodnotách mediánu a dolného kvartilu znázornených na obrázku 5.5.

Kvantily	Čas přežív.
25. kvantil (dolní kvartil)	2662,330
50. kvantil (medián)	
75. kvantil (horní kvartil)	

Obr. 5.5: Kvantily Kaplan-Meierovho odhadu

5.3 Tabuľky úmrtnosti

Pri používaní úmrtnostných tabuliek opäť najskôr nastavíme požadované premenné, kódy pre cenzorovanie a naviac zvolíme podobu tabuľky z dvoch možností. Voľba „počet intervalov“ vytvorí daný počet intervalov, kde šírka intervalu je určená automaticky, voľba „krok“ vytvorí intervaly so šírkou daného kroku a ich počet je daný automaticky. Pre naše dáta bude logické zvoliť možnosť „krok“ a nastaviť ju na hodnotu 365, tj. dáta zoskupíme do jednotlivých rokov. Ako výsledok tejto analýzy je tabuľka zložená z dvoch častí. V hornej časti sa nachádza opäť sumár základných informácií a v dolnej časti sú možnosti výberu:

- **základné výsledky**: úmrtnostná tabuľka (obrázok 5.6),
- **grafy funkcií**:
  - graf funkcie prežitia (obrázok 5.7),
  - graf intenzity zlyhania (riziková funkcia) (obrázok 5.8),
  - graf hustoty (obrázok 5.9),
- **detaily**:
  - odhad funkcie prežitia,
  - odhad intenzity zlyhania,
  - odhad hustoty,
  - odhady parametrov.

Úmrtnostná tabuľka poskytuje rozsiahle informácie o vzniknutých intervaloch, počtoch pozorovaní spadajúcich do jednotlivých intervalov, o odhadoch funkcie prežitia, rizikovej funkcie a hustoty spolu s ich smerodajnými odchýlkami. Ukážku tabuľky vidíme na nasledujúcom obrázku 5.6.

Interval	Interval počátek	Střed	Interval šířka	Počet vstupuj.	Počet vyřazen.	Počet exponov.	Počet ukončen.	Podíl ukončen.	Podíl přežil.	KumPodíl přežil.	Hustota pravděp.	Míra selhání	Sm.chyba kum.přež	Sm.chyba hustoty	Sm.chyba m.selh.	Medián oč.živ.	Sm.chyba oč.živ.
Int.č.1	0,000	182,500	365,0000	863	131	797,5000	65	0,081505	0,918495	1,000000	0,000223	0,000233	0,000000	0,000027	0,000029	3285,000	0,00
Int.č.2	365,000	547,500	365,0000	667	85	624,5000	18	0,028823	0,971177	0,918495	0,000073	0,000080	0,009689	0,000017	0,000019	2920,000	0,00
Int.č.3	730,000	912,500	365,0000	564	66	531,0000	19	0,035782	0,964218	0,892022	0,000087	0,000100	0,011241	0,000020	0,000023	2555,000	0,00
Int.č.4	1095,000	1277,500	365,0000	479	81	438,5000	13	0,029647	0,970353	0,860104	0,000070	0,000082	0,013007	0,000019	0,000023	2190,000	0,00
Int.č.5	1460,000	1642,500	365,0000	385	86	342,0000	5	0,014620	0,985380	0,834604	0,000033	0,000040	0,014416	0,000015	0,000018	1825,000	0,00
Int.č.6	1825,000	2007,500	365,0000	294	67	260,5000	7	0,026871	0,973129	0,822403	0,000061	0,000075	0,015203	0,000023	0,000028	1460,000	0,00
Int.č.7	2190,000	2372,500	365,0000	220	65	187,5000	8	0,042667	0,957333	0,800303	0,000094	0,000119	0,016934	0,000032	0,000042	1095,000	0,00
Int.č.8	2555,000	2737,500	365,0000	147	63	115,5000	4	0,034632	0,965368	0,766157	0,000073	0,000097	0,020059	0,000036	0,000048	730,000	0,00
Int.č.9	2920,000	3102,500	365,0000	80	61	49,5000	1	0,020202	0,979798	0,739624	0,000041	0,000056	0,023343	0,000041	0,000056	365,000	0,00
Int.č.10	3285,000			18	18	9,0000	0	0,055556	0,944444	0,724682			0,027237				

Obr. 5.6: Tabuľka úmrtnosti

Odhad funkcie prežitia je znázornený na obrázku 5.7. Z neho vyplýva, že najväčší počet osôb umrie hneď po transplantácií a v ďalších rokoch tento ukazovateľ klesá pomaly. Z daných dát konštatujeme, že transplantácia obličiek je pomerne úspešným zákrokom.

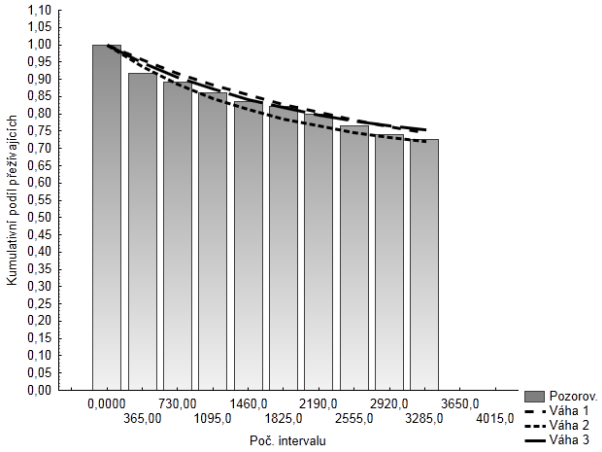
Poznámka 5.1. *STATISTICA ponúka možnosť aproximácie funkcie prežitia, rizikovej funkcie a hustoty pomocou parametrických modelov. V ponuke sú exponenciálny, Weibullov, Gompertzov a lineárny model. Tieto odhady STATISTICA*

konštruuje na základe metódy vážených najmenších štvorcov. Pri odhade používa tri varianty voľby váh:

1. $w_i = 1$, *klasická metóda najmenších štvorcov,*
2. $w_i = \dfrac{1}{v_i}$,
3. $w_i = n_i h_i$,

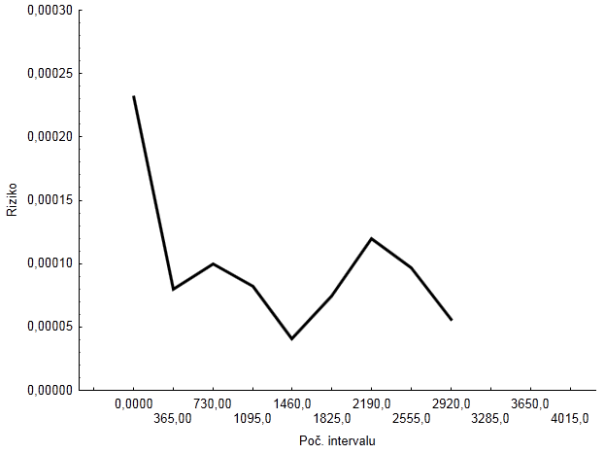
kde v_i je rozptyl odhadu rizikovej funkcie, h_i je šírka intervalu a n_i je počet pozorovaní vystavených riziku v i-tom intervale. Konkrétne hodnoty váh pre jednotlivé intervaly a modely možno nájsť v tabuľkách v sekcii ***detaily****.*

Okrem klasického odhadu sú na obrázku 5.7 znázornené aj odhady funkcie prežitia pomocou Gompertzovho rozdelenia, ktoré sme spomedzi ponúkaných parametrických modelov vybrali ako najvhodnejšie pre voľby váh 1 a 3.



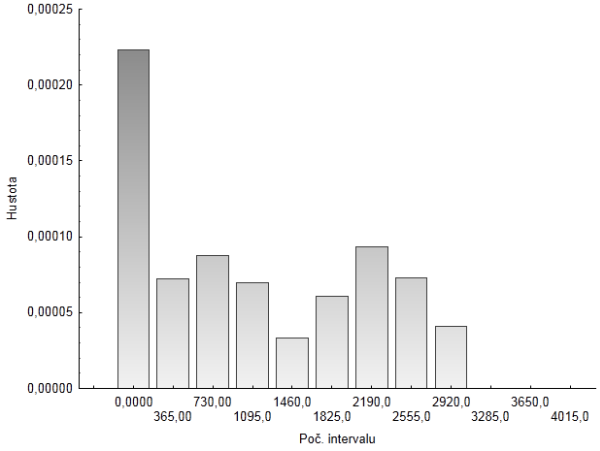
Obr. 5.7: Odhad funkcie prežívania pomocou tabuliek úmrtnosti

Pre podrobnejšie informácie o rozdelení času prežitia môžeme zostrojiť aj odhad hustoty a rizikovej funkcie.



Obr. 5.8: Odhad rizikovej funkcie zostrojený z dát

Riziková funkcia pre naše dáta, znázornená na obrázku 5.8, dosahuje najvyššiu hodnotu blízko času 0 a potom klesá. Funkcia však nie je monotónna pre celý skúmaný časový úsek. Tento fakt môže spôsobiť vplyv rôznych faktorov, ktorý by sme mohli testovať pomocou regresných modelov. *STATISTICA* opäť ponúka možnosť aproximácie pomocou parametrických modelov, avšak žiadny z ponúkaných modelov nie je pra naše dáta vhodný vzhľadom k zložitému tvaru rizikovej funkcie. Z rovnakých dôvodov parametricky neaproximujeme ani hustotu, znázornenú na obrázku 5.9.



Obr. 5.9: Odhad hustoty zostrojený z dát

5.4 Dvojvýberové testy

V programe *STATISTICA* dvojvýberové testy nájdeme v možnosti „porovnanie dvoch vzoriek“. Pomocou dvojvýberových testov na našich dátach môžeme skúmať napríklad rozdiely v rozdelení času prežitia pre ženy a mužov. Po zadaní požadovaných premenných, ktorými sú časy prežitia *time*, cenzorovaná premenná *indicator* (nastavenie správnych kódov) a premenná s kódmi skupín *gender* sa opäť objaví dvojdielna tabuľka. V hornej časti, znázornenej na obrázku 5.10, je sumár informácií o pozorovaniach v oboch skupínach.

```
Proměnná: time
Proměnná s indikátorem cenzorov. : indicator
Grupov. proměnná : gender
Celkový počet platných pozorování : 863
    necenzor. : 140 ( 16,22%)       cenzor. : 723 ( 83,78%)
Platná pozorování:   Skup. 1: 524              Skup. 2: 339
        Necenzor. :     87 ( 16,60%)              53 ( 15,63%)
          Cenzor. :    437 ( 83,40%)             286 ( 84,37%)
```

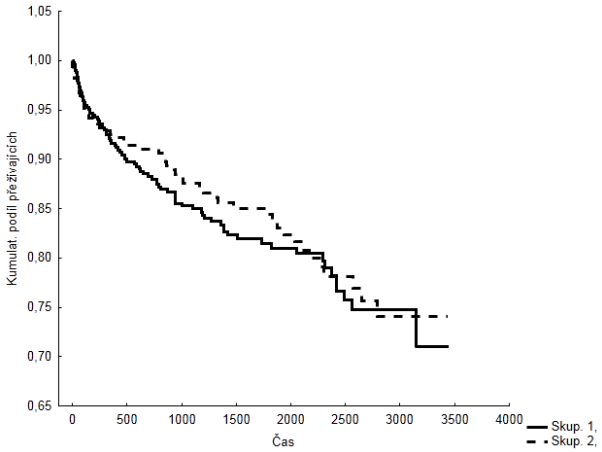
Obr. 5.10: Sumár informácií pre obe skupiny pozorovaní

V dolnej časti sú opäť možnosti ďalšieho výberu:

- **grafy funkcií**: Kaplan-Meierove grafy pre obe skupiny (obrázok 5.11),
- **dvojvýberové testy**:

– Gehanov-Wilcoxonov test,
– Peto-Petoov Wilcoxonov test,
– Coxov F-test,
– Ln-poradový test,
– Cox-Mantelov test.

Najskôr pre obe skupiny vygenerujeme Kaplan-Meierove grafy.



Obr. 5.11: Muži vs. ženy

Z pohľadu na grafy možno usúdiť, že rozdelenie času prežitia medzi mužmi a ženami je na pohľad podobné. Pre potvrdenie nášeho predpokladu zvolíme základný Cox-Mantelov test. Výsledkom testu je tabuľka obsahujúca rôzne hodnoty. V záhlaví tabuľky (obrázok 5.12) vyčítame hodnotu testovej štatistiky a p-hodnotu, ktorá je rovná 0,57749. Test sme realizovali na hladine 95 %, teda na základe p-hodnoty hypotézu H_0 nezamietame. Test potvrdil náš predpoklad, že rozdelenie časov prežitia medzi mužmi a ženami sa príliš nelíši.

Cox-Mantelův test (oblicky)
I = 33,58538 U = -3,22829
Test. statist. = -,557054 p = ,57749

Obr. 5.12: P-hodnota pre Cox-Mantelov test

test	p-hodnota
Gehanov-Wilcoxonov test	0,52169
Peto-Petoov Wilcoxonov test	0,57044
Coxov F-test	0,34133
Ln-poradový test	0,57613

Tabuľka 5.1: P-hodnoty pre rôzne testy

Pre testovanie hypotézy rovnosti rozdelenia časov prežitia medzi mužmi a ženami, môžeme samozrejme použiť aj ostatné testy, ktoré sú v ponuke. Všeobecne

testy volíme s prihliadnutím na ich vlastnosti opísané v podkapitole 3.2. V tabuľke 5.1 znázorňujeme p-hodnoty pre iné testy.

Pri testovaní hypotéz sa vždy riadime výsledkom jednoho, predom vybraného testu. Ak by sme si pre testovanie zvolili hociktorý z ponúkaných testov, dospejeme k rovnakému výsledku, pretože p-hodnoty všetkých testov sa pohybujú nad potrebnou hranicou 0,05.

5.5 Ostatná ponuka softvéru

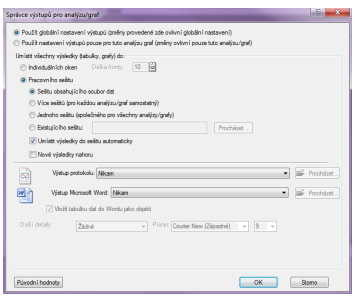
Okrem uvedených postupov *STATISTICA* ďalej ponúka možnosti testovania viacvýberových problémov a modelovanie dát na základe regresných modelov. Touto problematikou sme sa nezaoberali v teoretickej časti, preto len spomenieme príklady, ako by sa dané metódy dali aplikovať na naše dáta obohatené o ďalšie údaje.

Predpokladajme, že súbor informácií o pacientoch, ktoré máme k dispozícií, pochádza z údajov poskytnutých viacerými nemocnicami. V tomto prípade by sme mohli použiť viacvýberové testy a testovať, či sa rozdelenie času prežitia líši medzi rôznymi nemocnicami a nemocnice potom zoradiť podľa percenta úspešnosti zákroku.

Čas prežitia u jednotlivých pacientov isto závisí aj na iných faktoroch ako je miesto operácie. Typicky sú to vek, pohlavie, predchádzajúci zdravotný stav a životospráva. Ak by sme mali k dispozícií tieto (alebo rôzne ďalšie) údaje o pacientoch, mohli by sme pomocou regresných modelov skúmať ich vplyv na rozdelenie času prežitia.

5.6 Možnosti výstupov

STATISTICA ponúka rôzne možnosti formátovania a úpravy obrázkov, prípadq-ne tabuliek podľa individuálnej potreby užívateľov. Obrázky a grafy je možné ukladať v štandardne používaných formátoch. Pri zhotovovaní protokolov je ďalej možné nastavenie výstupu z analýz v užívateľom zvolenom formáte. Túto možnosť nájdeme v základnej tabuľke výstupu z analýz voľbou „možnosti“ a následne „výstup“. vygeneruje sa tabuľka 5.13, v ktorej nastavíme požadované voľby.



Obr. 5.13: Správca výstupov

Záver

Analýza prežitia patrí k štatistickým metódam, ktoré našli uplatnenie v mnohých odboroch. Spolu s rozvojom jednotlivých postupov prebieha ich implementácia do štatistických programov. Jedným z nich je aj *STATISTICA*.

V úvodných kapitolách sme sa venovali teoretickému rozboru základných pojmov, konceptov a metód, ktoré sa používajú a sú zároveň súčasťou používaného programu. Popísali sme rôzne možnosti, ako sa dá vyjadriť rozdelenie času prežitia. Neskôr sme si ukázali možnosti odhadu funkcie prežitia, ako najvýznamnejšej z nich. Opísali sme základné testy pre porovnávanie rozdelenia času prežitia v dvoch skupinách a teoretickú časť sme zakončili výkladom o uplatnení metód analýzy prežitia vo finančnom sektore.

Teoretické vedomosti sme v závere práce aplikovali na reálne dáta pochádzajúce z medicínskeho prostredia. Ukážka spracovaných dát slúži ako demonštrácia rôznych možností, ktoré program ponúka a boli spracované v teoretickej časti práce a zároveň poukazuje na ďalšie metódy, ktoré sme v práci neobsiahli, ale sú súčasťou programu.

Zoznam použitej literatúry

[1] Le, Chap T. *Applied Survival Analysis*. Wiley series in probability and statistics. Wiley, 1997.

[2] Lee, Elisa T., Wang, John Wenyu. *Statistical Methods for Survival Data Analysis*. Wiley series in probability and statistics. Wiley, 2004.

[3] StatSoft, Inc. *Electronic Statistics Textbook*. 2011.
WEB: $http://www.statsoft.com/textbook/$.

[4] Bellotti, Tony, Crook, Jonathan. *Credit Scoring With Macroeconomic Variables Using Survival Analysis*. Management School and Economics University of Edinburgh, 2007.
WEB: $http://fic.wharton.upenn.edu/fic/papers/07/0715.pdf$.

[5] Stepanova, Maria, Thomas, Lyn. *Survival Analysis Methods For Personal Loan Data*. 1999.
WEB: $http://teaching.ust.hk/\ ismt253w/cox.pdf$.

[6] Ibrahim, Joseph G. *Applied Survival Analysis*. University of North Carolina at Chapel Hill, 2005.
WEB: $http://www.amstat.org/chapters/northeasternillinois/pastevents/presentations/summer05\_Ibrahim\_J.pdf$.

[7] Dupač, Václav, Hušková, Marie. *Pravděpodobnost a matematická statistika*. Nakladatelství Karolinum, 1999.

[8] Medical College of Wisconsin.
WEB: $http://www.mcw.edu/FileLibrary/Groups/Biostatistics/Publicfiles/DataFromSection/DataFromSectionTXT/Data\_from\_section\_1.16.txt$.