Janez Stare[1]

Krivulje preživetja

Survival Curves

IZVLEČEK

KLJUČNE BESEDE: krivulje preživetja, okrnjeni podatki

Krivulje preživetja uporabljamo za prikaz spreminjanja deleža preživelih v času. Del podatkov o preživetju je praviloma okrnjen, zaradi česar potrebujemo za ocenjevanje deleža preživelih posebne metode. V članku je opisana metoda Kaplan-Meierja, ki se najpogosteje uporablja. Zaradi lažjega razumevanja pa je pojasnjena tudi povezanost krivulj preživetja s histogramom.

ABSTRACT

KEY WORDS: survival curves, censored data

Survival curves are used to describe the proportion of subjects alive through time. Since some of the data on survival are usually censored, special methods are needed to estimate the survival proportion. In this paper the commonly used method of Kaplan and Meier is described. For better understanding, the connection between the survival curves and the histogram is explained.

[1] Prof. dr. Janez Stare, dipl. inž. mat., Inštitut za biomedicinsko informatiko, Medicinska fakulteta, Vrazov trg 2, 1104 Ljubljana.

UVOD

Čas, ki mine med dvema dogodkoma, je v medicini pogosto mera uspešnosti zdravljenja.

Največkrat gre za čas preživetja, npr. od diagnoze ali začetka zdravljenja naprej, zato je področje statistike, ki se ukvarja z analizo takšnih podatkov, dobilo ime **analiza preživetja** (angl. *survival analysis*)*. Seveda je čas, ki nas zanima, lahko tudi kaj drugega, npr. trajanje remisije, čas med dvema infarktoma ipd. Druga področja znanosti zanimajo drugačni časi, na primer čas nezaposlenosti, trajanje sodnih procesov, čas učenja, čas trajanja elektronskih komponent, čas med dvema aretacijama itn. Zunaj medicine se analizi preživetja včasih reče **analiza poteka dogodkov** (angl. *event history analysis*). V tem članku bomo času med dvema dogodkoma, ne glede na vrsto dogodkov, rekli **čas preživetja**.

Če smo zbrali podatke o preživetju za skupino ljudi, nas najprej zanima, kako so ljudje umirali v času oz. koliko jih preživi določeno obdobje. Tabeli, ki za vsak čas podaja delež preživelih, rečemo **tabela preživetja** (angl. *life table*), če te podatke narišemo, pa dobimo **krivuljo preživetja** (angl. *survival curve*).

Osnovno sporočilo krivulje preživetja, delež preživelih v določenem obdobju, je torej povsem enostavno. Prav tako se zdi enostavno izračunavanje takšnih krivulj, saj je treba le prešteti število živih v vsakem času. Žal se v praksi stvari zapletejo zaradi težav z merjenjem časa preživetja. V večini študij preživetja se namreč dogaja, da nekaterih časov ne poznamo, vemo le, da so bolniki preživeli določeno obdobje, končnega dogodka pa še ni bilo (ni bilo smrti, ni bilo recidiva ...). Temu pojavu rečemo **krnjenje** (angl. *censoring*). Razlogi za krnjenje so lahko zelo različni. Najpogosteje gre za to, da študijo končamo, nekateri opazovanci pa so še živi. Krnitve pa lahko povzročijo tudi vzroki, kot so: bolnik se je preselil v oddaljen kraj in spremljanje ni več možno; bolnik je umrl zaradi vzroka, ki ni povezan z boleznijo, katere potek raziskujemo (na primer prometna nesreča); način zdravljenja bolnika se je pomembno spremenil (transplantacija, prekinitev zdravljenja zaradi stranskih učinkov in podobno).

V članku skušam na čimbolj razumljiv način razložiti računanje krivulj preživetja za okrnjene podatke. Ker pa se po mojih izkušnjah krivulj preživetja drži svojevrstna avreola mističnosti, skušam najprej to odstraniti z opisom njihove povezanosti z bolj običajnim prikazom podatkov, s histogramom.

HISTOGRAM IN KRIVULJE PREŽIVETJA

Čas je seveda numerična spremenljivka, največkrat jo merimo v dnevih in v principu ni nič posebnega. Za prikaz porazdelitve numeričnih spremenljivk imamo na voljo različne možnosti. Zakaj potem teh ne uporabimo tudi za prikaz porazdelitve časa preživetja? Razlog ni v naravi spremenljivke, ampak v že omenjenih težavah z merjenjem časa preživetja, se pravi s krnjenjem.

A za zdaj na krnjenje pozabimo in si oglejmo primer, ko so vsi časi preživetja znani.

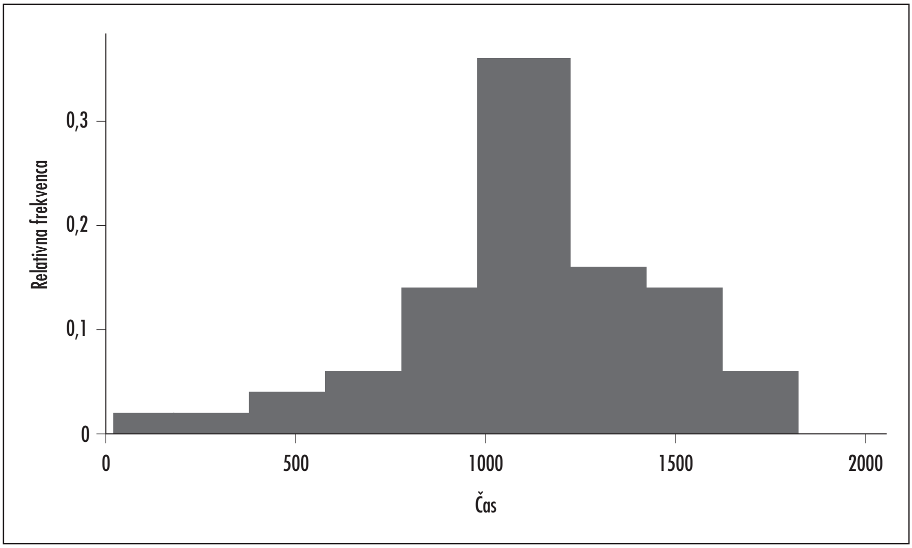
Primer 1. Vzemimo, da imamo 50 neokrnjenih časov. Za tiste, ki bi želeli slediti članku s papirjem in svinčnikom v roki (ali, še bolje, z računalnikom), tukaj navajam čase, urejene po velikosti: 123, 310, 466, 553, 681, 753, 766, 852, 882, 920, 921, 940, 957, 973, 1021, 1037, 1039, 1040, 1053, 1062, 1065, 1077, 1107, 1134, 1147, 1148, 1162, 1167, 1172, 1192, 1196, 1198, 1202, 1254, 1277, 1286, 1328, 1348, 1358, 1389, 1435, 1494, 1495, 1537, 1541, 1563, 1570, 1603, 1646, 1667.

Kako bi se običajno lotili prikazovanja takšnih podatkov? Med različnimi možnostmi se najbrž najpogosteje uporablja histogram. Pa ga dajmo še mi! Če razdelimo časovno obdobje na podobdobja z dolžino 200 dni, dobimo histogram, ki ga kaže slika 1.

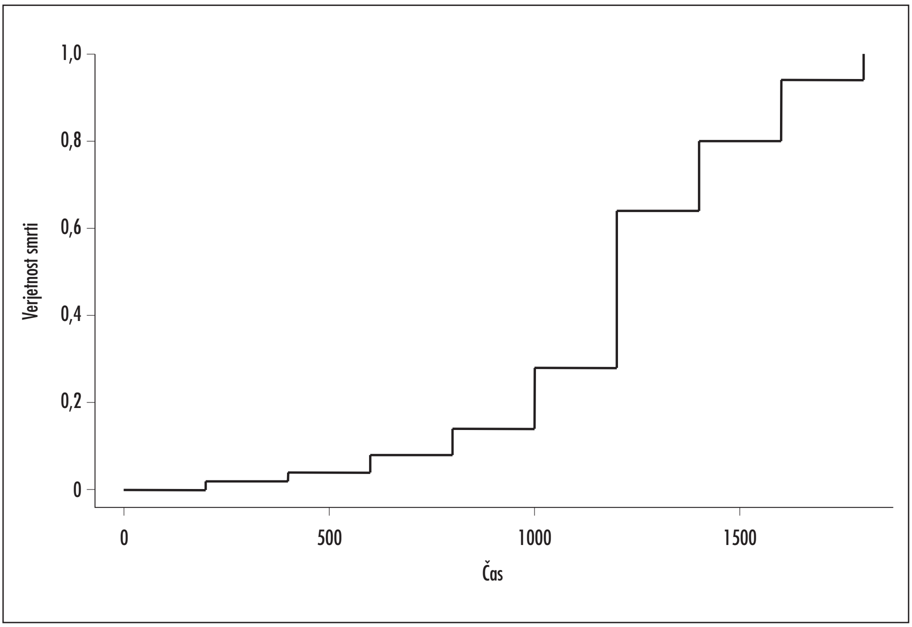
S histograma preberemo, kolikšen je delež časov v določenem podobdobju. Povedano drugače, histogram nam pove, kolikšen delež ljudi je umrl v posameznih 200-dnevnih podobdobjih.

Iz histograma lahko hitro dobimo takoimenovano **kumulativno porazdelitev**. Ta nam pove, kolikšen delež časov je manjši od neke vrednosti oz. kolikšen delež ljudi je umrl do določenega obdobja. Na primer, če sta

* Za večino statističnih izrazov s področja, ki ga obravnava članek, ni dogovorjenih slovenskih izrazov. V članku na to opozarjam z navedbo angleških izrazov.

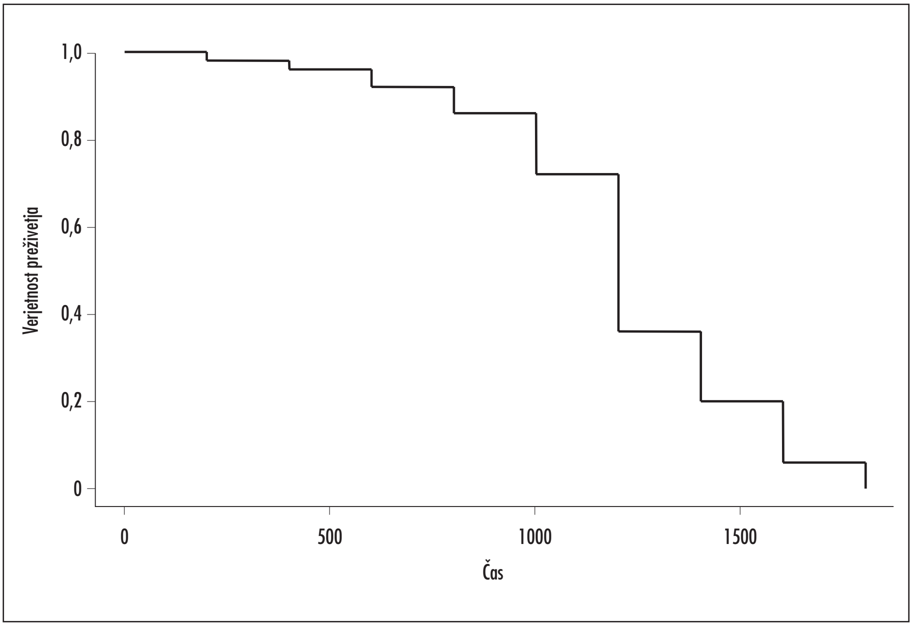


Slika 1. *Histogram časov preživetja iz primera 1.*



Slika 2. *Krivulja umrljivosti za podatke iz primera 1.*

v prvih 200 dneh umrla 2 % ljudi in v naslednjih 200 dneh še 2 %, so v prvih 400 dneh umrli 4 % ljudi. Za naš primer nam kumulativno porazdelitev kaže slika 2. Iz nje npr. preberemo, da je v prvih 800 dneh umrlo približno 14 %, v 1200 dneh pa že 64 % ljudi. Povejmo še enkrat, da smo sliko 2 dobili iz slike 1 s preprostim seštevanjem deležev na sliki 1. Dobili smo krivuljo, ki bi ji lahko rekli **krivulja umrljivosti**. Ta nam za vsak trenutek pove, kolikšen delež ljudi je do takrat umrl.



Slika 3. *Krivulja preživetja za podatke iz primera 1.*

Kako bomo zdaj prišli do krivulje preživetja, je najbrž že jasno. Če namreč vemo, kolikšen delež ljudi je do nekega trenutka umrl, naj bo to na primer *p*, potem je delež preživelih $1 - p$. Krivuljo na sliki 2 moramo torej spremeniti le toliko, da vsako vrednost na njej odštejemo od 1. Tako dobimo sliko 3, to pa je že **krivulja preživetja**. Te praviloma uporabljamo v analizi preživetja, pač zato, ker raje govorimo o živih kot o mrtvih. Z njih odčitamo, kolikšen delež ljudi preživi določeno obdobje.

Krivulje preživetja so torej samo na glavo obrnjene kumulativne porazdelitve, se pravi nič novega ali specifičnega. Morda velja vseeno poudariti, da slik, kot sta naši sliki 2 in 3, ne vidimo prav pogosto v zvezi z običajnimi numeričnimi spremenljivkami. Mnogo bolj pogosti so histogrami, čeprav so vsi prikazi povsem enakovredni. Iz vsakega namreč lahko dobimo druga dva s preprostim seštevanjem ali odštevanjem. Histogrami imajo prednost predvsem zato, ker nam olajšujejo ocenjevanje vrste porazdelitve, predvsem normalnosti, 'na oko'.

Če bi bil naš cilj le krivulja preživetja, za njeno računanje seveda ne bi potrebovali histograma. Takšno pot sem uporabil le zaradi njene poučnosti (upam!). Seveda bi do enakih rezultatov prišli, če bi kar naravnost izračunali delež preživelih v določenem obdobju. Na primer: časov, ki so daljši od 800 dni, je 43, se pravi, da je delež preživelih po 800 dneh $43/50 = 0{,}86$ (spomnimo se, da je bil delež umrlih v tem času 0,14!).

Zakaj torej krivulje preživetja? Rekel sem že, da je razlog v krnjenju, nisem pa tega podrobneje pojasnil. Stvar je preprosta: če nek čas ni popoln, se pravi, da smrti ni bilo, dokler smo bolnika opazovali, potem ne vemo, koliko bi bila prava vrednost in takšnega časa ne moremo uvrstiti v neko 200-dnevno podobdobje. Primer: če je nekdo preživel 300 dni, lahko živi še 1000, ali pa samo 5 dni. S takšnimi podatki ne moremo računati niti povprečja! Izkaže pa se, da krivulje preživetja, torej delež preživelih v določenih časih, vseeno lahko ocenimo. O tem podrobneje v naslednjem razdelku.

RAČUNANJE KRIVULJ PREŽIVETJA ZA OKRNJENE PODATKE

Če eno leto preživi polovica ljudi in naslednje leto še polovica preostalih, potem dve leti preživi četrtina ljudi. Z računom: $0{,}5 \times 0{,}5 = 0{,}25$. V splošnem: če neko obdobje razdelimo na dva dela in je delež preživelih v prvem obdobju p_1, delež preživelih v drugem obdobju (od še živečih na začetku drugega obdobja) pa p_2, je delež preživelih na koncu celotnega obdobja $p_1 \times p_2$. Na tem preprostem dejstvu temelji metoda za izračunavanje krivulj (in tabel) preživetja. Najlažje jo bo razložiti na primeru.

Primer 2. Naj bodo nekateri časi iz primera 1 okrnjeni: 123, 144+, 238+, 310, 346+, 357+, 532+, 550+, 554+, 681, 753, 766, 828+, 852, 873+, 882, 920, 921, 940, 951+, 957, 964+, 973, 993+, 1021, 1028+ 1037, 1039+ 1053, 1065, 1077, 1107, 1147, 1148, 1167, 1172, 1192, 1196, 1198, 1254, 1301+ 1348, 1494, 1495, 1537, 1541, 1563, 1603, 1646, 1667. Gre torej za iste čase kot v primeru 1, le da smo nekatere okrnili. Okrnjene čase smo označili s +. To so torej tisti časi, za katere pravih vrednosti ne poznamo, vemo le, da so daljši od navedenih. Recimo, vrednost 144+ pomeni, da je bil pacient po 144 dneh še živ, ne vemo pa, kaj se je pozneje z njim dogajalo. Če bi iz teh podatkov izračunali povprečje, bi pravo vrednost seveda bistveno podcenili. Prav tako ne moremo narisati histograma, saj ne vemo, kam naj damo vrednosti, kot je 144+. Kako pa je z oceno deleža preživelih? Pojdimo lepo po vrsti: prvih 123 dni je preživelo 49 ljudi, torej je preživetje 49/50. Od 123. do 144. dneva je bilo 49 ljudi izpostavljenih tveganju, umrl ni nobeden, torej je delež preživelih v tem (!!) času $49/49 = 1$. Če pomnožimo oba deleža, dobimo $\frac{49}{50} \times \frac{49}{49} = \frac{49}{50}$, kar je v skladu z dejstvom, da je v 144 dneh umrla le ena oseba. Na 144. dan je pri enem bolniku prišlo do krnitve, v nadaljevanju jih torej opazujemo le še 48. Do 238. dneva se nič ne zgodi, na ta dan pa zopet prenehamo opazovati enega bolnika. Preživetje do 238 dne je seveda $\frac{49}{50} \times \frac{49}{49} \times \frac{48}{48} = \frac{49}{50}$. V obdobju po 238. dnevu opazujemo še 47 ljudi. Na 310. dan je en bolnik umrl, torej je preživetje v obdobju med 238. in 310. dnem 46/47. Skupaj pa potem $\frac{49}{50} \times \frac{49}{49} \times \frac{48}{48} \times \frac{46}{47} = \frac{49}{50} \times \frac{46}{47} =$ 0,959. Najbrž je jasno, kako nadaljujemo. Bistvo metode je torej v tem, da razdelimo celotno opazovalno obdobje na podobdobja, za katera lahko ocenimo preživetje, in te ocene med seboj množimo. Začetke in konce podobdobij določajo dnevi, v katerih je prišlo do smrti ali krnjenja. Pomembno pri tem je, da se bolniki, katerih podatki so okrnjeni, pojavljajo v izračunih (preko dejstva, da so med izpostavljenimi tveganju) vse do časa krnjenja.

Pravkar opisano metodo ponavadi imenujemo **Kaplan-Meierjeva metoda** *(1)*, po avtorjih, ki sta jo prva predlagala. V literaturi naletimo tudi na **metodo življenjskih tabel**, ki se od opisane ne razlikuje dosti, je pa primernejša za ročno računanje. Pri tej metodi razdelimo celotno opazovalno obdobje na le nekaj podobdobij, pogosto uporabljena dolžina je eno leto. Problem okrnjenih opazovanj rešimo tako, da privzamemo, da je do krnjenja prihajalo slučajno in enakomerno znotraj obdobij. To pomeni, da so bili okrnjeni bolniki znotraj podobdobja opazovani v povprečju polovico podobdobja. Za toliko potem zmanjšamo število izpostavljenih tveganju na začetku podobdobja. Na primer: v prvem letu smo imeli v naših podatkih dve smrti (123. in 310. dan) ter štiri krnitve (144., 238., 346. in 357. dan). Začetno število izpostavljenih tveganju zmanjšamo za 2 (vsaka krnitev prispeva polovico!), se pravi, da imamo 48 izpostavljenih tveganju in ker sta dva umrla, je preživetje po enem letu $46/48 = 0{,}958$. Na začetku naslednjega leta bi bilo število izpostavljenih tveganju 46, zmanjšano za polovico krnitev v tem letu, ki so bile tri (532., 550. in 554. dan), torej bi bilo dejansko število izpostavljenih tveganju na začetku drugega leta 44,5. In tako naprej!

Metoda življenjskih tabel se v praksi vse manj uporablja, ker nam pri izračunih pomagajo računalniki. Seveda podatke pogosto **prikazujemo** v obliki takšnih tabel, a izračuni lahko vseeno temeljijo na bolj natančni Kaplan-Meierjevi metodi. Primer si bralec lahko ogleda v predzadnjem razdelku.

ČESA NE SMEMO POČETI

Morda bi koga zamikalo, da bi pri izračunavanju preživetja na okrnjene podatke preprosto pozabil, jih torej ne bi uporabil. Da je s tem nekaj narobe, nam hitro pokaže tale preprost premislek.

Vzemimo, da bolnike vključujemo v študijo 5 let in jih po koncu tega obdobja spremljamo še eno leto. Tisti, ki vstopijo v študijo v začetku, so torej lahko opazovani do 6 let, tisti, ki vstopijo proti koncu pa kvečjemu do-

Tabela 1. *Primerjava različnih metod izračunavanja preživetja*

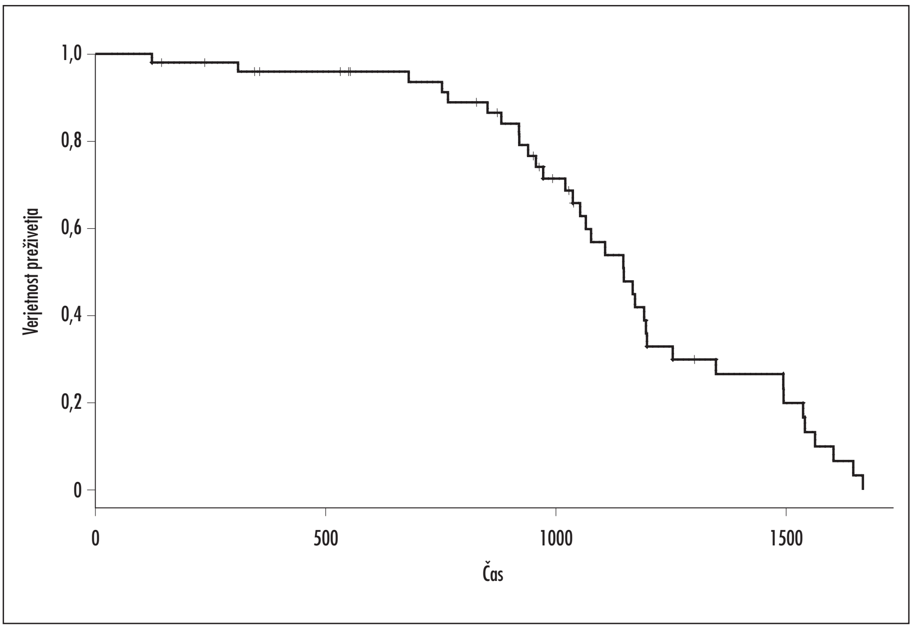
Čas	Kaplan-Meier (upoštevano krnjenje) (a)	Preživetje, če zanemarimo okrnjene podatke (b)	Preživetje, če privzamemo, da so vsi okrnjeni umrli na dan krnjenja (c)	Preživetje, če privzamemo, da so okrnjeni preživeli celotno opazovalno obdobje (č)	Preživetje – prave vrednosti (d)
852	0,8649	0,8286	0,72	0,88	0,84
973	0,7141	0,6571	0,54	0,76	0,72
1147	0,5085	0,4571	0,34	0,62	0,50
1348	0,2659	0,2286	0,16	0,46	0,24

bro leto. Če je nekdo, ki je vstopil v začetku, živel 5 let in je bilo opazovanje okrnjeno, bi ga morali iz analize izpustiti, nekoga, ki bi vstopil proti koncu in bi umrl po treh mesecih, bi pa upoštevali. Očitno bi torej smrtim dajali preveliko težo in umrljivost precenili, preživetje pa podcenili.

Prav tako bi bilo seveda napak, če bi privzeli, da so vsi okrnjeni umrli na dan krnitve, ali pa da so vsi preživeli celotno opazovalno obdobje.

V tabeli 1 so za naš primer z okrnjenimi podatki prikazane vrednosti preživetja za nekaj časov. Vrednosti v stolpcu (a) so izračunane po Kaplan-Meierjevi metodi, v stolpcu (b) po metodi, ki zanemari vse krnitve, v stolpcu (c) po metodi, ki privzame, da so vsi okrnjeni umrli na dan krnjenja, in v stolpcu (č) po metodi, ki privzame, da so vsi okrnjeni živeli dlje kot vsi mrtvi. Za primerjavo so v stolpcu (d) prikazane še vrednosti, ki jih dobimo iz neokrnjenih podatkov, prave vrednosti torej. V praksi vrednosti v stolpcu (d) seveda ne moremo izračunati, ker pravih vrednosti preživetja ne poznamo za vse. A ker smo naše podatke simulirali, teh težav tukaj seveda nimamo.

Vrednosti v stolpcu (a) so očitno najbliže pravim vrednostim, nekako slučajno skačejo okrog pravih vrednosti. Statistiki rečemo, da je metoda dosledna, kar pomeni, da se z na-



Slika 4. *Krivulja preživetja po Kaplan-Meierju za primer 2.*

Tabela 2. *Preživetje po Kaplan-Meierju za podatke iz primera 2.*

Čas v dnevih	Izpostavljenih tveganju	Preživetje	Čas v dnevih	Izpostavljenih tveganju	Preživetje
123	50	0,9800	1147	18	0,5085
310	47	0,9591	1148	17	0,4786
681	41	0,9358	1167	16	0,4487
753	40	0,9124	1172	15	0,4187
766	39	0,8890	1192	14	0,3888
852	37	0,8649	1196	13	0,3589
882	35	0,8402	1198	12	0,3290
920	34	0,8155	1254	11	0,2991
921	33	0,7908	1348	9	0,2659
940	32	0,7661	1494	8	0,2326
957	30	0,7406	1495	7	0,1994
973	28	0,7141	1537	6	0,1662
1021	26	0,6866	1541	5	0,1329
1037	24	0,6580	1563	4	0,0997
1053	22	0,6281	1603	3	0,0665
1065	21	0,5982	1646	2	0,0332
1077	20	0,5683	1667	1	0,0000
1107	19	0,5384			

raščanjem velikosti vzorca, ob enakem deležu krnitev, vrednosti po Kaplan-Meierju vse bolj bližajo pravim vrednostim. Stolpca (b) in (c) preživetje podcenita (stolpec (c) bolj!), (č) pa preceni. Bralcu svetujem, da sam premisli, zakaj. Napake so seveda tem večje, čim večji je delež krnitev.

ŠE NEKAJ PRIMEROV

V tabeli 2 najprej podajam podrobne rezultate Kaplan-Meierjeve metode za naše podatke iz primera 2. Slika 4 prikazuje ustrezno krivuljo preživetja. Časi krnitev so na krivulji prikazani z navpičnimi črticami. Prikazovanje teh seveda ni nujno, odločamo se od primera do primera. Krivulja je bolj natančna kot tista na sliki 3, ker smo tam čase grupirali v razrede.

Metodo Kaplan-Meierja pozna danes že vsak resnejši statistični program, od nedavnega tudi Epi Info 2000, ki je na voljo brezplačno in je vsakomur dosegljiv na medmrežju. Tako si lahko bralec tabelo 2 in sliko 4 pripravi sam, kar vsekakor priporočam.

Primer 3. Naredimo iz tabele 2, ki gotovo ni primerna za objavo v kakšnem članku o preživetju bolnikov, življenjsko tabelo, ki naj vsebuje podatke o preživetju po letih. Iz tabele 2 vidimo, da se preživetje med 310. in

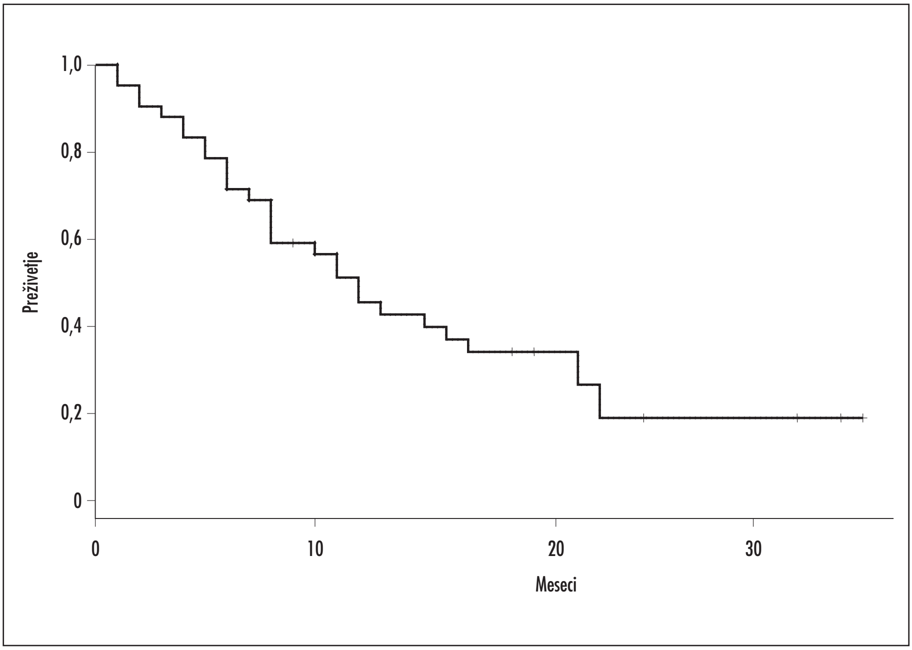
Tabela 3. *Tabela preživetja po posameznih letih za podatke iz primera 2.*

Leto	Preživetje
1	0,9591
2	0,9124
3	0,5683
4	0,2659
5	0

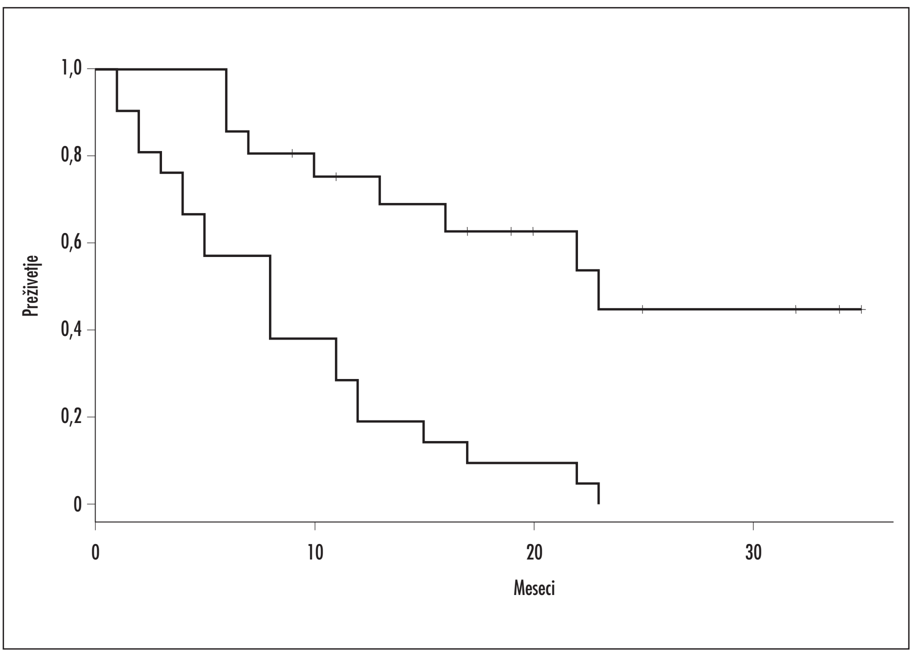
680. dnevom ne spreminja, po prvem letu je torej enako kot preživetje po 310. dnevu, se pravi 0,9591. Po dveh letih oz. 730 dnevih je enako kot 681. dan, torej 0,9124. In tako naprej, dokler nimamo tabele 3.

Primer 4. Slika 5 prikazuje krivuljo preživetja za bolnike z levkemijo *(2)*. Primer je poučen zaradi dejstva, da krivulja preživetja ne doseže ničle. To se pogosto dogaja in seveda pomeni, da so najdaljši časi krnjeni.

Primer 5. V praksi največkrat želimo primerjati krivulje preživetja med različnimi skupinami. Tako so bili na primer pacienti iz primera 4 zdravljeni na dva različna načina in slika 6 prikazuje krivulji preživetja za obe skupini. V eni skupini preživetje torej pade na ničlo (ker so v njej vsi bolniki umrli), v drugi pa preživetje ostaja nad 40 odstotki do konca opazovanj. Ob takšni sliki je seveda na-



Slika 5. *Krivulja preživetja bolnikov z levkemijo*



Slika 6. *Primerjava krivulj preživetja dveh skupin bolnikov z levkemijo (2).*

ravno vprašanje, ali je razlika med krivuljama statistično značilna. Za takšno analizo imamo na voljo več testov, najpogostejši je test **log-rank** *(3)*, a to že presega namen tega članka.

OPOZORILO IN NASVET

Opisana metoda za izračunavanje krivulj preživetja sodi med neparametrične metode, saj nam ni bilo potrebno sprejeti nobenih predpostavk o obliki ali parametrih porazdelitve. Pa vendar čisto brez predpostavk le ni šlo. Privzeli smo namreč (čeprav tega nismo posebej poudarjali), da so izračunane verjetnosti preživetja v posameznih obdobjih dobra ocena za prave verjetnosti. To pravzaprav pomeni, da predpostavimo, da se stvari ne bi bistveno spremenile, če ne bi bilo okrnjenih opazovanj. To je sicer ponavadi res, a prav je, da se predpostavke zavedamo. Lahko bi npr. do krnitev prihajalo predvsem pri bolnikih z dobro prognozo, kar bi pomenilo, da bi bila naša ocena preživetja prenizka. Zato *opozorilo*: krnitve morajo biti slučajne. To je vedno res, če npr. do njih prihaja zaradi končanja študije.

Seveda je na vzorcu izračunana krivulja preživetja le ena od možnih krivulj, ki bi jih dobili, če bi lahko vzorčili večkrat. To dejstvo se v računalniških izpisih izraža s standardno napako preživetja, krivulje pa imajo včasih vrisane tudi meje zaupanja. Kot vedno, večji vzorci pomenijo ožje meje in obratno. *Nasvet*: ker gre pri krivuljah preživetja za deleže, ocene teh pa zahtevajo velike vzorce, ni smotrno načrtovati študij preživetja, če bomo lahko spremljali le nekaj deset bolnikov. Bržčas bomo le zapravljali čas in denar.

LITERATURA

1. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Ass 1958*; 53: 457–81.
2. Freireich EO, Gehan E, Frei E, Schroeder LR, Wolman IJ, Anbari R, Burgert EO, Mills SD, Pinkel D, Selawry OS, Moon JH, Gendel BR, Spurr CL, Storrs R, Haurani F, Hoogstraten B, Lee S. The effect of 6-mercaptopmine on the duration of steroid induced remission in acute leukemia. *Blood 1963*; 21: 699–716.
3. Peto R, Pike MC, Armitage P, Breslow NE et al. Design and analysis of randomised clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *Br J Cancer 1977*; 35: 1–39.

Prispelo 4. 8. 2001