Statistička analiza preživljavanja i primjene

Antolković, Mateja

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO-MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Mateja Antolković

STATISTIČKA ANALIZA PREŽIVLJAVANJA I PRIMJENE

Diplomski rad

Voditelj rada:
Prof.dr.sc. Siniša Slijepčević

Zagreb, rujan 2015.

Ovaj diplomski rad obranjen je dana _____ pred nastavničkim povjerenstvom u sastavu:

1. _____, **predsjednik**
2. _____, **član**
3. _____, **član**

Povjerenstvo je rad ocijenilo ocjenom _____ .

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

Analiza preživljavanja je skup tehnika kojima se procjenjuje i opisuje vrijeme potrebno do pojave jednog ili više određenih događaja. Naziv *'Analiza preživljavanja*' dolazi od prvih istraživanja gdje je događaj od interesa uglavnom bila smrt. Danas se ona primjenjuje i u mnogim drugim širim područjima, primjerice, kod procjene vremena potrebnog za razvijanje nekih bolesti, pada burze, kvara raznih strojeva, pojave potresa, razvoda braka, začeća, prestanka pušenja, financijama, itd.

Nije poznato kada je točno ova grana statistike nastala, no vjeruje se da je to bilo prije nekoliko stoljeća. Nakon Drugog svjetskog rata njezina primjena postaje sve veća, i to ponajviše u testiranju pouzdanosti vojne opreme. Potom se sve više koristi i u industriji, jer su ljudi počeli zahtijevati što sigurnije i pouzdanije proizvode. Posljednji napredak analize preživljavanja je posljedica razvoja softverskih paketa i visokih performansi računala koji su sada u mogućnosti veoma efikasno izvršiti ove zahtjevne algoritme.

Kroz prvo poglavlje ovog rada ćemo se upoznati s osnovnim veličinama analize preživljavanja. Definirat ćemo funkciju preživljavanja, funkciju rizika, te očekivano trajanja preostalog života. Potom ćemo opisati problem višestrukog rizika. Nadalje, objasnit ćemo razliku između cenzuriranih i odrezanih podataka te dati primjer za svaki.

U drugom poglavlju dajemo neparametarske procjene za veličine definirane u prvom poglavlju, na temelju desno cenzuriranih podataka, te pokazujemo kako se one mogu koristiti i u slučaju lijevo odrezanih podataka. U zadnjem poglavlju primjenjujemo tehnike iz drugog poglavlja na konkretne podatke vezane za oporavak nakon presađivanja koštane srži. Za obradu podataka korišten je programski jezik R.

1. Osnovni pojmovi analize preživljavanja

1.1. Osnovni pojmovi i definicije statistike i vjerojatnosti

Neka nam Ω predstavlja prostor elementarnih događaja, a $\mathcal{P}(\Omega)$ partitivni skup od Ω.

Definicija 1.1.1. Familija $\mathcal{F}$ podskupova od Ω je $\boldsymbol{\sigma}$**−algebra skupova** ako vrijedi sljedeće:

i) $\emptyset \in \mathcal{F}$

ii) $A \in \mathcal{F} \Rightarrow A^C \in \mathcal{F}$

iii) $A_i \in \mathcal{F}, i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Uređeni par $(\Omega, \mathcal{F})$ se zove **izmjeriv prostor**.

Definicija 1.1.2. Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor. Funkcija $\mathbb{P}: \mathcal{F} \to \mathbb{R}$ je **vjerojatnost** ako ispunjava sljedeće uvjete:

i) $\mathbb{P}(A) \geq 0, A \in \mathcal{F}$

ii) $\mathbb{P}(\Omega) = 1$

iii) $A_i \in \mathcal{F}, i \in \mathbb{N}$, A_i međusobno disjunktni ⇒

$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Uređena trojka $(\Omega, \mathcal{F}, \mathbb{P})$ se zove **vjerojatnosni prostor**.

Definicija 1.1.3. Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor, te neka su $A, B \in \mathcal{F}$ takvi da je $\mathbb{P}(B) > 0$. Uvjetna vjerojatnost od A uz uvjet B definira se kao

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Definicija 1.1.4. Neka je $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ vjerojatnosni prostor. **Slučajna varijabla** je proizvoljna realna funkcija definirana na Ω.

Ukoliko imamo više slučajnih varijabli na istom vjerojatnosnom prostoru, možemo na njih gledati kao na komponente nekog vektora. Takav vektor zovemo slučajni vektor.

Definicija 1.1.5. Neka je X neprekidna slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$. Funkcija $f: \mathbb{R} \to [0, +\infty\rangle$ je **funkcija gustoće** slučajne varijable X ako vrijedi:

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)dt,\ x \in \mathbb{R}.$$

U slučaju da je X diskretna slučajna varijabla, funkcija gustoće je dana s

$$f(x) = \mathbb{P}(X = x), x \in \mathbb{R}.$$

Funkcija gustoće je nenegativna funkcija sa površinom ispod krivulje jednakom jedan u slučaju neprekidne varijable, odnosno vrijedi $\sum_{x_j} f(x_j) = 1$, gdje je x_j vrijednost koju X može poprimiti, u slučaju diskretne slučajne varijable.

Definicija 1.1.6. Funkcija distribucije slučajne varijable je definirana s

$$F(x) := \mathbb{P}(X \leq x), x \in \mathbb{R}.$$

U slučaju neprekidne slučajne varijable vrijedi $F(x) = \int_{-\infty}^{x} f(t)dt$, a u slučaju diskretne slučajne varijable $F(x) = \sum_{x_j \leq x} f(x_j)$, gdje su $x_j, j = 1, 2, \ldots$, vrijednosti koje X može poprimiti. Vrijedi da je $F(-\infty) = 0$ i $F(+\infty) = 1$.

Definicija 1.1.7. Matematičko očekivanje slučajne varijable X, $\mathbb{E}[X]$, je definirano s

$\mathbb{E}[X] = \sum_x xf(x),\ \ x \in Im\, X$, u diskretnom slučaju, te

$\mathbb{E}[X] = \int_{-\infty}^{+\infty} xf(x)dx$, u neprekidnom slučaju.

Definicija 1.1.8. Varijanca slučajne varijable X je definirana formulom

$$Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Standardna devijacija slučajne varijable X je korijen iz njezine varijance.

Definicija 1.1.9. Gustoća diskretnog slučajnog vektora (X, Y) je funkcija $f_{X,Y}: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ definirana s

$$f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y).$$

Gustoća neprekidnog slučajnog vektora (X, Y) je funkcija $f_{X,Y}: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ za koju vrijedi:

$\mathbb{P}(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y)\, dy\, dx$, za sve $a < b$, $c < d$.

Definicija 1.1.10. Za diskretan slučajni vektor (X, Y) s funkcijom gustoće $f_{X,Y}$, **marginalna gustoća** od X je dana s

$$f_X(x) = \sum_y f_{X,Y}(x, y),\ y \in Im\, Y.$$

Analogno se definira marginalna gustoća od Y.

Za neprekidni slučajni vektor (X, Y) s funkcijom gustoće $f_{X,Y}$, marginalna gustoća od X je dana s

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\, dy.$$

Analogno se definira marginalna gustoća od Y.

Definicija 1.1.11. Neka je (X, Y) slučajni vektor sa funkcijom gustoće $f_{X,Y}$ i marginalnim gustoćama f_X i f_Y. Ako je $f_Y(y) > 0$ za sve $y \in Im\, Y$, tada je **uvjetna gustoća** od X za dano $Y = y$ funkcija $x \to f_{X|Y}(x|y)$ definirana s:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Definicija1.1.12. Neka je (X, Y) slučajni vektor. **Uvjetno očekivanje od Y uz dano X** u diskretnom slučaju je

$$\mathbb{E}[Y|X = x] = \sum_y y f_{Y|X}(y|x),\ y \in Im\, Y.$$

U neprekidnom slučaju je

$$\mathbb{E}[Y|X=x] = \int_{-\infty}^{+\infty} y\, f_{Y|X}(y|x)dy.$$

U oba slučaja mora vrijediti da je $f_X(x) > 0$.

Teorem 1.1.13. (Centralni granični teorem) Neka je $(X_n, n \in \mathbb{N})$ niz nezavisnih jednako distribuiranih slučajnih varijabli s očekivanjem m i varijancom σ^2, $0 < \sigma^2 < \infty$ i neka je $S_n = \sum_{k=1}^{n} X_k$. Tada vrijedi

$$\frac{S_n - \mathbb{E}S_n}{\sigma\sqrt{n}} \to N(0,1).$$

Teorem 1.1.14. (Delta metoda) Neka je X normalna slučajna varijabla s očekivanjem m i varijancom σ^2, te neka je g funkcija $g: \mathbb{R} \to \mathbb{R}$. Tada je $g(Y)$ aproksimativno normalno distibuirana s očekivanjem $g(m)$ i varijancom $[g'(m)]^2\sigma^2$.

1.2. Funkcija preživljavanja

Definicija 1.2.1. Funkcija preživljavanja slučajne varijable X je definirana s:

$$S(x) := P(X > x). \qquad (1.2.1.)$$

Funkcija preživljavanja je vjerojatnost da slučajna varijabla X postigne vrijednost veću od x. Na primjer, to može biti vjerojatnost da osoba živi duže od x godina, ili da iskusi neki drugi događaj od interesa tek nakon trenutka x.

Ako je X neprekidna slučajna varijabla, tada je i $S(X)$ neprekidna, strogo padajuća funkcija, te vrijedi $S(x) = 1 - F(x)$, odnosno, $S(x)$ je inverzna funkcija distribucije. Dodatno, vrijedi i sljedeće:

$$S(x) = P(X > x) = \int_x^{\infty} f(t)dt. \qquad (1.2.2.)$$

Stoga vrijedi

$$f(x) = -\frac{dS(x)}{dx}. \tag{1.2.3.}$$

Vrijednost $f(x)dx$ možemo shvatiti kao približnu vjerojatnost da će se događaj pojaviti u trenutku x.

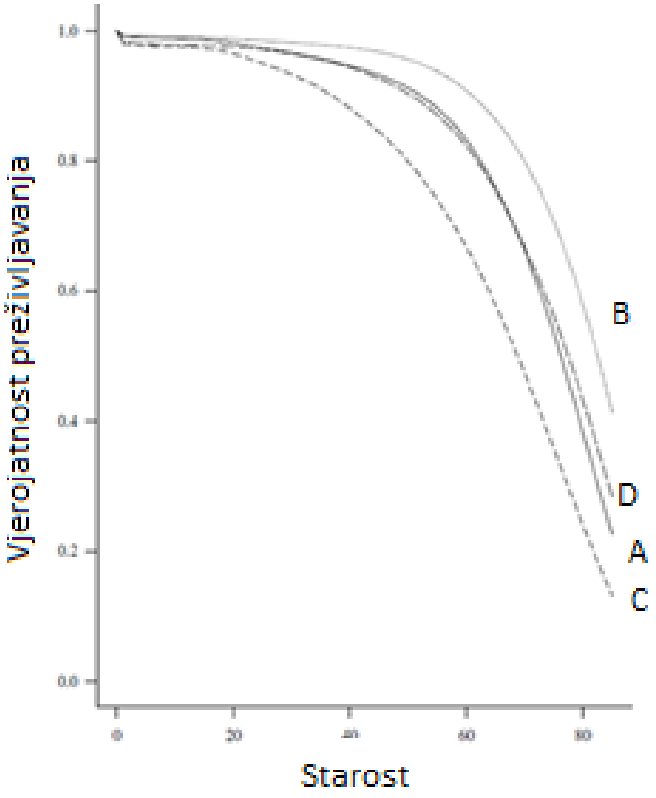
Zajedničko svojstvo svim funkcijama preživljavanja je da su monotone i nerastuće, te u nuli postižu vrijednost jedan, a za $x \to \infty$, postižu vrijednost nula. Gledajući samo graf funkcije preživljavanja možda i ne možemo puno zaključiti, no može biti korisno i zanimljivo uspoređivati graf s dvije, pa i više funkcija.

Funkcija preživljavanja za diskretnu slučajnu varijablu je dana sljedećom formulom:

$$S(x) = P(X > x) = \sum_{x_j > x} f(x_j), \tag{1.2.4.}$$

gdje su $x_j, j = 1,2, \ldots$ vrijednosti koje X može poprimiti, $x_1 < x_2 < \cdots$, a $f(x_j)$ gustoća diskretne slučajne varijable X. Diskretna slučajna varijabla može biti posljedica zaokruživanja rezultata istraživanja, grupiranja trenutaka u kojima je zabilježen neuspjeh u intervale i slično. U ovom slučaju $S(x)$ je također nerastuća funkcija.

Primjer 1.2.2. Promatramo funkcije preživljavanja vezane za događaj smrti, među stanovnicima SAD-a, dobivenih na temelju podataka iz 1990. godine, prema primjeru iz [2, str. 23-25]. Stanovnici su podijeljeni u četiri skupine, ovisno o rasi i spolu. U tablici 1.2.2. su prikazane vjerojatnosti preživljavanja za svaku skupinu, a graf 1.2.1. prikazuje krivulje preživljavanja svake skupine. Uspoređujući krivulje vidimo da su žene bijele rase (B) imale najveće vjerojatnosti preživljavanja, muškarci bijele rase (A) i žene crne rase (D) podjednake, a muškarci crne rase (C) najgore vjerojatnosti preživljavanja.



Slika 1.2.1. Krivulje preživljavanja za stanovnike SAD-a iz primjera 1.2.2., izvor [2, str.25]

STAROST	A	B	C	D	STAROST	A	B	C	D
0	1.00000	1.00000	1.00000	1.00000	43	0.93771	0.97016	0.85917	0.93361
1	0.99092	0.99285	0.97996	0.98283	44	0.93477	0.96862	0.85163	0.92998
2	0.99024	0.99232	0.97881	0.98193	45	0.93161	0.96694	0.84377	0.92612
3	0.98975	0.99192	0.97792	0.98119	46	0.92820	0.96511	0.83559	0.92202
4	0.98937	0.99160	0.97722	0.98059	47	0.92450	0.96311	0.82707	0.91765
5	0.98905	0.99134	0.97664	0.98011	48	0.92050	0.96091	0.81814	0.91300
6	0.98877	0.99111	0.97615	0.97972	49	0.91617	0.95847	0.80871	0.90804
7	0.98850	0.99091	0.97571	0.97941	50	0.91148	0.95575	0.79870	0.90275
8	0.98825	0.99073	0.97532	0.97915	51	0.90639	0.95273	0.78808	0.89709
9	0.98802	0.99056	0.97499	0.97892	52	0.90086	0.94938	0.77685	0.89103
10	0.98782	0.99041	0.97472	0.97870	53	0.89480	0.94568	0.76503	0.88453
11	0.98765	0.99028	0.97449	0.97847	54	0.88810	0.94161	0.75268	0.87754
12	0.98748	0.99015	0.97425	0.97823	55	0.88068	0.93713	0.73983	0.87000
13	0.98724	0.98999	0.97392	0.97796	56	0.87250	0.93222	0.72649	0.86190
14	0.98686	0.98977	0.97339	0.97767	57	0.86352	0.92684	0.71262	0.85321
15	0.98628	0.98948	0.97258	0.97735	58	0.85370	0.92096	0.69817	0.84381
16	0.98547	0.98909	0.97145	0.97699	59	0.84299	0.91455	0.68308	0.83358
17	0.98445	0.98862	0.97002	0.97658	60	0.83135	0.90756	0.66730	0.82243
18	0.98326	0.98809	0.96829	0.97612	61	0.81873	0.89995	0.65083	0.81029
19	0.98197	0.98755	0.96628	0.97559	62	0.80511	0.89169	0.63368	0.79719
20	0.98063	0.98703	0.96403	0.97498	63	0.79052	0.88275	0.61584	0.78323
21	0.97924	0.98654	0.96151	0.97429	64	0.77501	0.87312	0.59732	0.76858
22	0.97780	0.98607	0.95873	0.97352	65	0.75860	0.86278	0.57813	0.75330
23	0.97633	0.98561	0.95575	0.97267	66	0.74131	0.85169	0.55829	0.73748
24	0.97483	0.98514	0.95267	0.97174	67	0.72309	0.83980	0.53783	0.72104
25	0.97332	0.98466	0.94954	0.97074	68	0.70383	0.82702	0.51679	0.70393
26	0.97181	0.98416	0.94639	0.96967	69	0.68339	0.81324	0.49520	0.68604
27	0.97029	0.98365	0.94319	0.96852	70	0.66166	0.79839	0.47312	0.66730
28	0.96876	0.98312	0.93989	0.96728	71	0.63865	0.78420	0.45058	0.64769
29	0.96719	0.98257	0.93642	0.96594	72	0.61441	0.76522	0.42765	0.62723
30	0.96557	0.98199	0.93273	0.96448	73	0.58897	0.74682	0.40442	0.60591
31	0.96390	0.98138	0.92881	0.96289	74	0.56238	0.72716	0.38100	0.58375
32	0.96217	0.98073	0.92466	0.96118	75	0.53470	0.70619	0.35749	0.56074
33	0.96038	0.98005	0.92024	0.95934	76	0.50601	0.68387	0.33397	0.53689
34	0.95852	0.97933	0.91551	0.95740	77	0.47641	0.66014	0.31050	0.51219
35	0.95659	0.97858	0.91044	0.95336	78	0.44604	0.63494	0.28713	0.48663
36	0.95457	0.97779	0.90501	0.95321	79	0.41503	0.60822	0.26391	0.46020
37	0.95245	0.97696	0.89922	0.95095	80	0.38355	0.57991	0.24091	0.43291
38	0.95024	0.97607	0.89312	0.94855	81	0.35178	0.54997	0.21819	0.40475
39	0.94794	0.97510	0.88677	0.94598	82	0.31991	0.51835	0.19583	0.37573
40	0.94555	0.97404	0.88021	0.94321	83	0.28816	0.48502	0.17392	0.34588
41	0.94307	0.97287	0.87344	0.94023	84	0.25677	0.44993	0.15257	0.31522
42	0.94047	0.97158	0.86643	0.93703	85	0.22599	0.41306	0.13191	0.28378

Tablica 1.2.2. Vrijednosti funkcije preživljavanja za stanovnike SAD-a iz primjera 1.2.2., izvor [2, str. 24]

1.3. Funkcija rizika

Sljedeća funkcija koju ćemo spomenuti također ima veoma važnu ulogu u analizi preživljavanja, a to je funkcija rizika, koju ćemo označavati s $h(x)$.

Definicija 1.3.1. Funkcija rizika definirana je sljedećom formulom:

$$h(x) := \lim_{\Delta x \to 0} \frac{\mathbb{P}[x \leq X < x + \Delta x | X \geq x]}{\Delta x}. \tag{1.3.1.}$$

U slučaju da je X neprekidna slučajna varijabla vrijedi:

$$\mathbb{P}[x \leq X < x + \Delta x | X \geq x] = \frac{\mathbb{P}[x \leq X < x + \Delta x]}{\mathbb{P}[X \geq x]} = \frac{F(x + \Delta x) - F(x)}{S(x)}$$

i

$$f(x) = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x}.$$

Dakle, funkciju rizika možemo zapisati i u sljedećem obliku:

$$h(x) = f(x)/S(x) = -d \ln[\mathrm{S(x)}]/dx. \tag{1.3.2.}$$

Definicija 1.3.2. Funkcija kumulativnog rizika, $H(x)$, je definirana sljedećom formulom:

$$H(x) := \int_0^x h(u)du = -\ln[S(x)]\,. \tag{1.3.3.}$$

Iz gornje formule slijedi:

$$S(x) = \exp[-H(x)] = \exp\left[-\int_0^x h(u)du\right]. \tag{1.3.4.}$$

Iz (1.3.1) vidimo da se $h(x)\Delta x$ može shvatiti kao aproksimativna vjerojatnost da osoba koja ima x godina, umre u nadolazećem trenutku. Pomoću te funkcije možemo vidjeti kako se vjerojatnost da osoba umre, odnosno iskusi događaj od interesa, mijenja kroz vrijeme. Funkcija rizika može poprimit razne oblike, ovisno o vrsti problema koju opisuje. Jedino zajedničko svojstvo svim funkcijama rizika je da su nenegativne.

Ako je X diskretna slučajna varijabla, funkcija rizika je dana sljedećom formulom:

$$h(x_j) = P(X = x_j | X \geq x_j) = \frac{p(x_j)}{S(x_{j-1})}, j = 1, 2, \ldots \quad (1.3.5.)$$

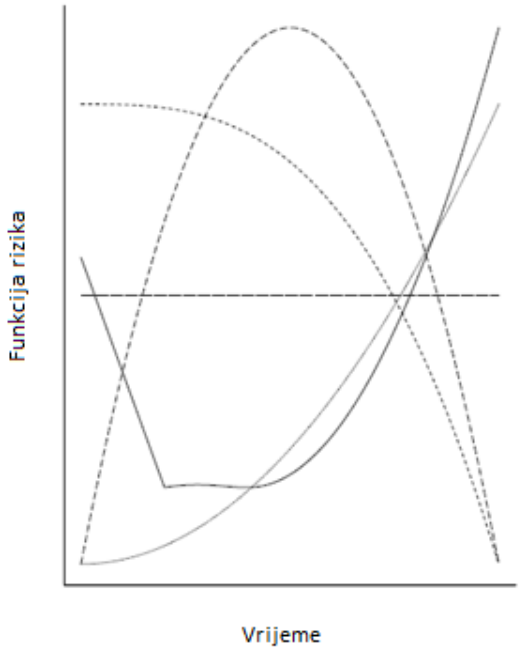
gdje je $S(x_0) = 1$. Zbog $p(x_j) = S(x_{j-1}) - S(x_j)$ vrijedi sljedeća jednakost $h(x_j) = 1 - S(x_j)/S(x_{j-1}),\ j = 1, 2, \ldots.$

Primijetimo da se funkcija preživljavanja može napisati i kao produkt uvjetnih vjerojatnosti preživljavanja, odnosno:

$$S(x) = \prod_{x_j \leq x} S(x_j)/S(x_{j-1}). \quad (1.3.6.)$$

Iz toga lako izvedemo relaciju koja povezuje funkciju rizika i funkciju preživljavanja za diskretnu slučajnu varijablu:

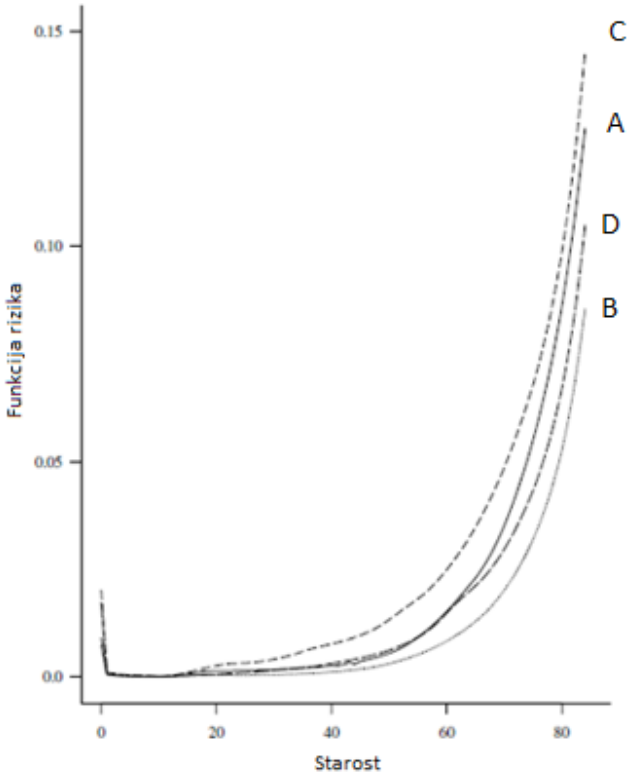
$$S(x) = \prod_{x_j \leq x} [1 - h(x_j)]. \quad (1.3.7.)$$



Slika 1.3.1. Razni oblici funkcija rizika, izvor [2, str. 28]

Slika 1.3.1. sadrži nekoliko oblika funkcija rizika. Rastuća funkcija rizika bi mogla biti posljedica prirodnog starenja, odnosno što su ljudi stariji, veća je stopa umiranja. Padajuća funkcija rizika se mnogo rjeđe pojavljuje, i to kod slučajeva kad je odmah na početku najveća vjerojatnost neuspjeha, kao kod elektroničkih uređaja i transplantacija organa. Funkcija u obliku '*kade*' je specifična za živa bića, koja imaju veću stopu smrtnosti kod rođenja, te u ranim fazama života zbog raznih dječjih bolesti, te kasnije kada nastupi starost. Razni strojevi se također mogu odmah na početku upotrebe pokvariti zbog neispravnih dijelova, a kasnije opet postati neispravni zbog dotrajalosti. Krivulja funkcije rizika u obliku 'grbe' je specifična za operacije koje su uspješno završene, no s vremenom se povećava vjerojatnost infekcije, hemoragije ili neke druge komplikacije, koja opet s vremenom pada kako se pacijent oporavlja.

Primjer 1.3.3. Na slici 1.3.2. su prikazane četiri krivulje funkcija rizika vezane primjer 1.2.2. o smrtnosti među stanovnicima SAD-a. Kod sve četiri krivulje vidimo nisku stopu rizika na početku, nakon kojeg slijedi period konstantne stope, a potom rast, no za svaku grupu u drugom trenutku. [2, str. 30].



Slika 1.3.2. Funkcije rizika za stanovnike SAD-a iz primjera 1.3.3., [2, str. 30]

1.4. Očekivano trajanje života

Sljedeći pojam koji spominjemo je očekivanog trajanja ostatka života, u oznaci $mrl(x)$, *eng.mean residual life*. Ono nam govori koliko je očekivano trajanje ostatka života osobe koja ima x godina.

Definicija 1.4.1. Očekivano trajanje ostatka života u trenutku x, $mrl(x)$, je:

$$mrl(x) := \mathbb{E}(X - x | X > x). \tag{1.4.1.}$$

Ako je X neprekidna slučajna varijabla vrijedi sljedeće:

$$mrl(x) = \frac{\int_x^\infty (t-x) f(t) dt}{S(x)}. \tag{1.4.2.}$$

Nadalje, zbog $f(t)dt = -dS(t)$ imamo:

$$\mathbb{E}(X - x|X > x)S(x) = \int_x^\infty (t-x)f(t)dt$$

$$= (parcijalna\ integracija) =$$

$$= -(t-x)\,S(x)|_x^\infty + \int_x^\infty S(t)dt$$

$$= 0 + \int_x^\infty S(t)dt. \qquad (1.4.3.)$$

Tada možemo pisati:

$$mrl(x) = \frac{\int_x^\infty (t-x)f(t)dt}{S(x)} = \frac{\int_x^\infty S(t)dt}{S(x)}. \qquad (1.4.4.)$$

Uvrštavanjem $x = 0$ i $S(0) = 1$ u gornju jednadžbu dobivamo formulu za očekivano ukupno trajanje života, $\mu := mrl(0)$:

$$\mu = \mathbb{E}(X) = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt. \qquad (1.4.5.)$$

Iz gornjih jednakosti vidimo da je očekivano trajanje preostalog života jednako površini ispod krivulje funkcije preživljavanja za vrijednosti veće od x, podijeljenoj sa $S(x)$. Za očekivano trajanje života, μ, pak vrijedi da je ono jednako ukupnoj površini ispod krivulje funkcije preživljavanja.

Sada lako možemo uočiti povezanost između varijance i funkcije preživljavanja:

$$\mathbb{E}[X^2] = \int_0^\infty t^2 f(t)dt = -\,t^2 S(x)|_x^\infty + \int_x^\infty 2tS(t)dt = 0 + 2\int_0^\infty tf(t)dt,$$

$$Var(X) = 2\int_0^\infty tS(t)dt - \left[\int_0^\infty S(t)dt\right]^2. \qquad (1.4.6.)$$

Definicija 1.4.2. *p*-ti kvantil distribucije *X* je najmanji x_p takav da zadovoljava:

$$S(x_p) \leq 1 - p, \quad \text{tj.} \quad x_p = \inf\{t: S(t) \leq 1 - p\}. \tag{1.4.7.}$$

Ako je X neprekidna slučajna varijabla, tada je *p*-kvantil rješenje jednadžbe $S(x_p) = 1 - p$.

Definicija 1.4.3. Medijan životnog vijeka je 0.5-kvantil distribucije slučajne varijable *X*.

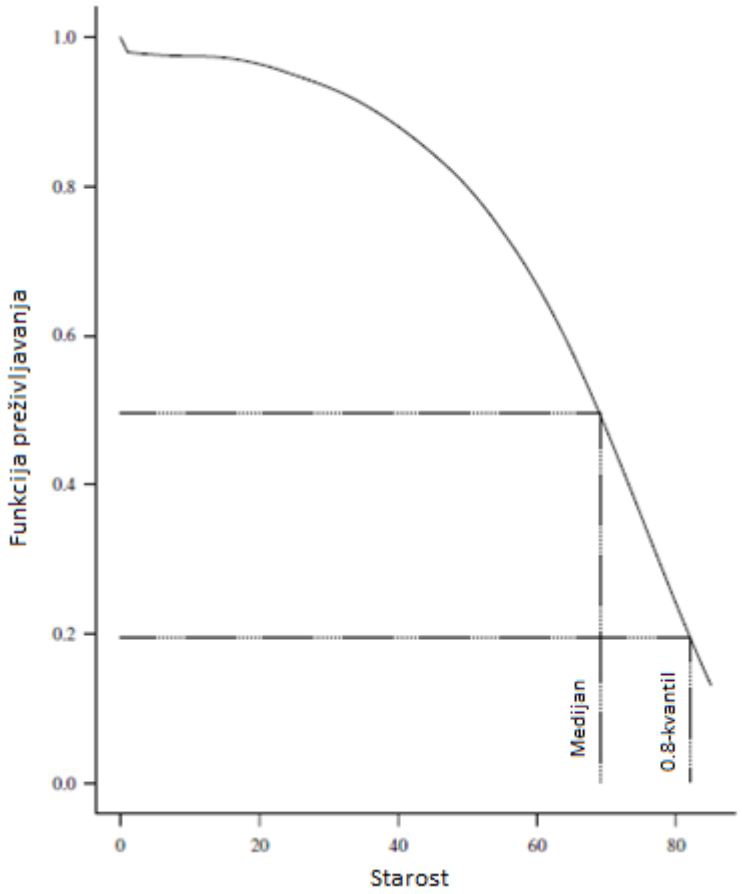
Stoga, za neprekidne varijable X vrijedi da medijan životnog vijeka zadovoljava jednadžbu

$$S(x_{0.5}) = 0.5.$$

Primjer 1.4.4. Medijan životnog vijeka, no i ostali kvantili, lako se mogu iščitati iz grafova, . Pokazat ćemo to na primjeru smrtnosti u SAD-u, vidi [2, str. 33-34], koristeći sliku 1.4.1. Prvo se pronađe odgovarajuća vjerojatnost preživljavanja, te se tada pogleda za koji x se ona postiže. Na gornjem grafu je pokazano kako očitati medijan i 0.8-kvantil. Približni rezulati su 69, odnosno 82 godine. Mnogo točniji rezultat možemo dobiti linearnom interpolacijom koristeći podatke iz tablice 1.2.2.. Vidimo da je $S(68) = 0.51679 > 0.5$ i $S(69) = 0.49520 < 0.5$, stoga medijan sigurno poprima neku vrijednost između 68 i 69 godina. Linearnom interpolacijom dobijemo:

$$x_{0.5} = 68 + \frac{S(68) - 0.5}{S(68) - S(69)} = 68.78 \text{ godina}.$$

Na isti način dobijemo da je $x_{0.8} = 81.81$ godina.



Slika 1.4.1. Određivanje medijana funkcije preživljavanja sa grafa, izvor [2, str. 34]

1.5. Problem višestrukog rizika

U nekim istraživanjima, često medicinskim, imamo problem višestrukog rizika. Takav problem se javlja kada osobe koje sudjeluju u istraživanju su izložene većem broju rizika K $(K \geq 2)$. Na primjer, osoba može dobiti status da je neuspješno izliječena ukoliko joj se bolest vrati ili nastupi smrt. Ovdje su povratak bolesti i smrt višestruki rizici kojima je osoba izložena. Nadalje, ako je događaj od interesa smrt, ona može također nastupiti zbog više razloga, poput srčanog infarkta, tumora, prometnih nesreća itd.

Neka su $X_i, i = 1, \dots, K$ potencijalna vremena pojave *i-tog* rizika. Ono što mi možemo u istraživanju saznati je vrijeme kada je osoba umrla zbog nekog od

mogućih rizika, $T = \min(X_1, \dots, X_p)$, i indikator δ, koji nam govori koji od *K* mogućih rizika je uzrokovao smrt.

Definicija 1.5. Uzrok-marginalna stopa rizika je definirana s

$$h_i(t) := \lim_{\Delta t \to 0} \frac{P[t \le T < t + \Delta t, \delta = i | T \ge t]}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{P\left[t \le X_i < t + \Delta t, \delta = i \middle| X_j \ge t, \ \ j = 1, \dots, K\right]}{\Delta t}. \qquad (1.5.1.)$$

Ona nam govori po kojoj stopi osobe koje su još uvijek žive, tj. nisu nastradale od niti jednog mogućeg rizika, umru zbog *i*-tog rizika u trenutku t. Ukupna stopa rizika u trenutku *T* je suma uzrok-marginalnih stopa rizika, tj.

$$h_T(t) = \sum_{i=1}^{K} h_i(t). \qquad (1.5.2.)$$

Neka je $S(t_1, .., t_K) = P[X_1 > t_1, \dots, X_K > t_K]$ zajednička funkcija preživljavanja svih *K* mogućih rizika. Uzročno specifična stopa rizika se može izvesti iz zajedničke funkcije preživljavanja na sljedeći način

$$h_i(t) = \frac{-\partial S(t_1, \dots, t_K)/\partial t_i|_{t_1 = \dots = t_K = t}}{S(t, \dots, t)}. \qquad (1.5.3.)$$

U modelu s više mogućih rizika, ponekad nas zanima samo vjerojatnost pojavljivanja nekih određenih događaja. U tu svrhu ćemo definirati sljedeće tri vjerojatnosti: grubu, neto i djelomično grubu vjerojatnost. **Gruba vjerojatnost** je vjerojatnost da će osoba umrijeti od nekog određenog uzroka, u realnom svijetu gdje je mogla stradati i od bilo kojeg drugog uzroka smrti. Na primjer, želimo izračunati kolika je vjerojatnost da će osoba umrijeti od neke srčane bolesti u pedesetoj godini života, ako znamo da je izložena i ostalim uzročnicima smrti.

Neto vjerojatnost je vjerojatnost da će osoba umrijeti od nekog određenog uzroka, ukoliko znamo da je to zapravo jedini uzrok zbog kojeg ona može umrijeti. Na primjer, to je vjerojatnost da osoba umre od neke srčane bolesti, ako živi u hipotetičkom svijetu gdje se može umrijeti jedino zbog te bolesti.

Djelomično gruba vjerojatnost je vjerojatnost da će osoba umrijeti u hipotetičkom svijetu gdje su neki rizici eliminirani. Na primjer, to je vjerojatnost da osoba umre od srčanog infarkta u svijetu gdje se ne može umrijeti od raka.

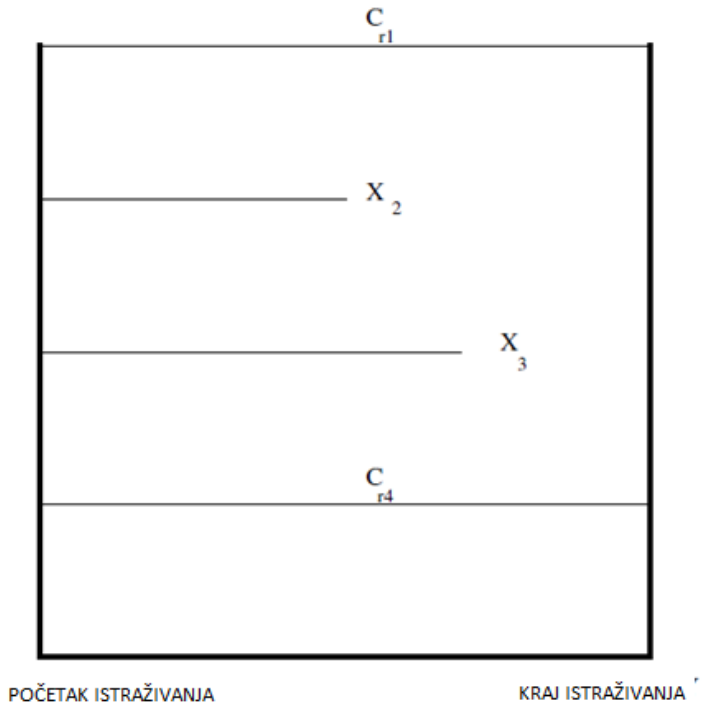
1.6. Cenzurirani i odrezani podaci

1.6.1. Desno cenzuriranje

Prvi način cenzuriranja koji ćemo opisati je *Type I Censoring.* Kod ovog tipa prikupljanja podataka je određen trenutak do kojeg se vrši promatranje, te ukoliko do tog trenutka osoba nije iskusila događaj od interesa, smatrat ćemo to cenzuriranim promatranjem, jer je sasvim moguće da se događaj pojavio nakon završetka promatranja. To je i motivacija za naziv ovog načina prikupljanja podataka, jer su te osobe cenzurirano promatrane s desne strane, odnosno nisu promatrane nakon završetka istraživanja. Ovaj način promatranja se koristi kada zbog nedostatka vremena ili velikih troškova, istraživanje mora završiti prije nego što sve osobe iskuse događaj od interesa ili napuste promatranje zbog nekog razloga. Vrijeme promatranja nije isto za sve osobe, i ovisi o tome da li su iskusile događaj od interesa ili napustile promatranje prijevremeno.

Sa C_r ćemo označiti trenutak u kojem prestaje promatranje, dok nam *X* označava varijablu životnog vijeka osobe, s funkcijom gustoće *f(x)* i funkcijom preživljavanja *S(x)*. Točan životni vijek pojedine osobe saznat ćemo, ako i samo ako ona umre za vrijeme promatranja, i tada će *X* biti manji od C_r. Ako je *X* veći od C_r, to znači da je osoba preživjela razdoblje promatranja, i umrla nakon trenutka C_r. Podatke s istraživanja prezentiramo u obliku uređenih parova (T, δ), gdje nam

T označava trenutak kada je nastupila smrt, tj. događaj od interesa, ili trenutak u kojem promatranje prestaje za određenu osobu. Nadalje, δ je indikator koji ukazuje na to da li je osoba iskusila događaj od interesa za vrijeme promatranja, i tada postiže vrijednost jedan, a u suprotnom slučaju postiže vrijednost nula.

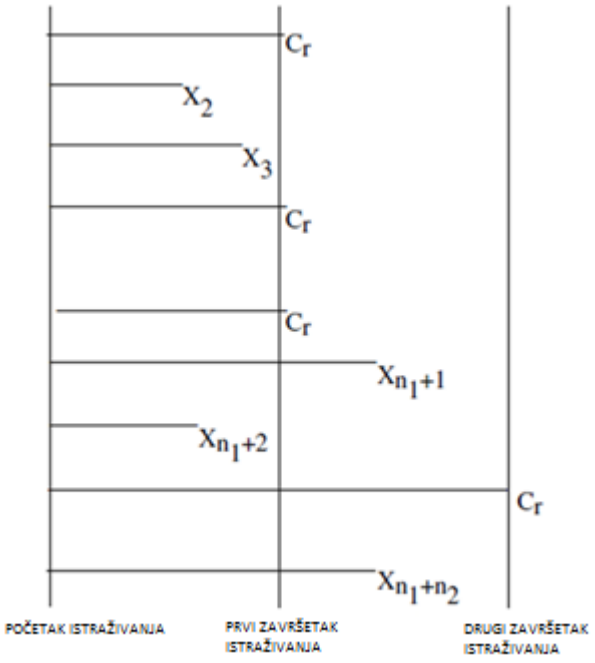


Slika 1.6.1.1. *Type I Censoring,* izvor [2, str. 65]

Primjer 1.6.1.1. Promatramo istraživanje gdje je miševima bila davana određena doza kancerogene hrane, iz [2, str. 65]. Svrha istraživanja bila je utvrditi štetnost kancerogenih tvari. Miševi su bili promatrani od početka istraživanja pa do smrti, odnosno do trenutka prestanka istraživanja. Ovaj primjer prikazan je na slici 1.6.1.1. Miševi kojima pripadaju varijable X_2 i X_3 su umrli tokom promatranja, dok su ostala dva doživjela kraj istraživanja.

Progressive type I censoring je zapravo *Type I Censoring* ali s dva vremena završetka istraživanja, odnosno, za dio promatranih objekata istraživanje prestaje u nekom unaprijed određenom trenutku, dok za ostatak prestaje u nekom kasnijem unaprijed određenom trenutku.

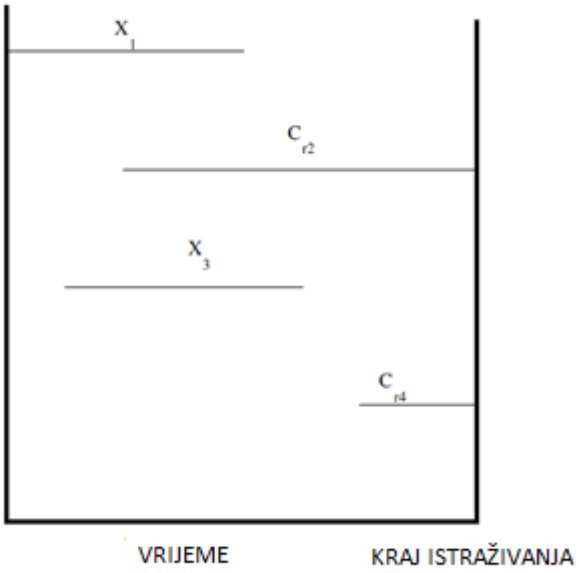
Primjer 1.6.1.2. Na slici 1.6.1.2. vidimo skicu istraživanja s dva vremena završetka, vidi [2, str. 66-67]. Vidimo da su neki objekti ostali pod promatranjem i nakon prvog završetka istraživanja, dok je jedan objekt ostao na promatranju i do drugog završetka istraživanja. Ovaj način je pogodan jer se njime reduciraju troškovi istraživanja, a opet znamo što se događa sa nekim objektima koji su ostali na promatranju.



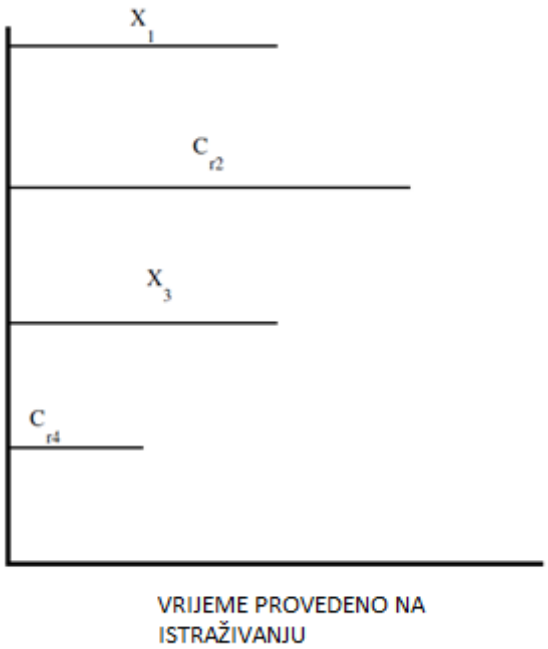
Slika 1.6.1.2. *Progressive type I censoring,* izvor [2, str. 66]

Sljedeći način prikupljanja podataka je *generalized type I censoring*. U ovom slučaju osobe ulaze u istraživanje u različitim trenucima, no istraživanje prestaje za sve osobe u istom, unaprijed određenom trenutku. Stoga svaka osoba ima svoje vlastito vrijeme tokom kojeg je bila pod promatranjem. Ovaj način prikupljanja podataka je prikazan na slici 1.6.1.3. Ove podatke možemo grafički još prikazati i tako da pomaknemo svako vrijeme početka u nulu, kao što se vidi na slici 1.6.1.4.. Treći način prezentiranja je Lexis dijagram. Kod njega se na horizontalnoj osi nalazi vrijeme ulaska u istraživanje, a linija pod kutom od $45°$

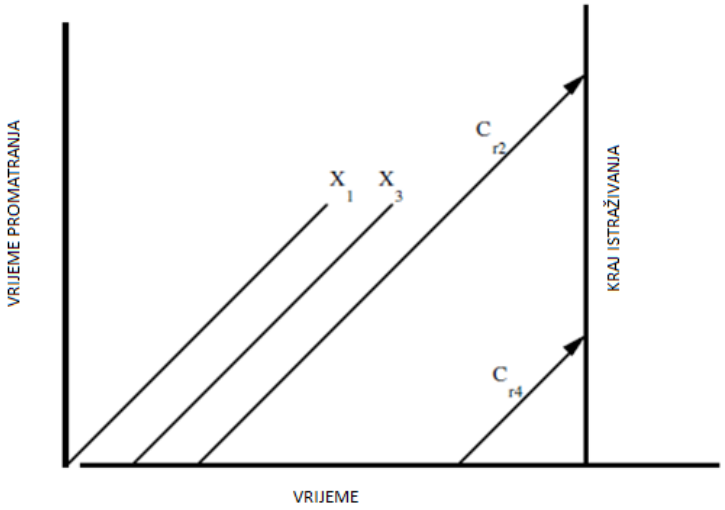
označava duljinu života. Lexis dijagram imamo na slici 1.6.1.5. Na slikama (1.6.1.3.-1.6.1.5.) objekti 1 i 3 su iskusili događaj od interesa prije, a objekti 2 i 4 nakon kraja istraživanja.



Slika 1.6.1.3. *Generalized type I censoring,* izvor [2, str. 66]



Slika 1.6.1.4. *Generalized type I censoring,* izvor [2, str. 68]



Slika 1.6.1.5. Lexis dijagram, izvor [2, str. 68]

U drugom tipu desnog cenzuriranja, *Type II Censoringu*, istraživanje traje sve dok se ne dogodi prvih r neuspjeha, gdje je r neki prirodni broj unaprijed određen. Ova metoda se koristi kod testiranja raznih alata i oprema, a njezine prednosti su manji trošak vremena i novaca. Dakle, imat ćemo r neuspjeha, $n-r$ cenzuriranih promatranja, i te su brojke unaprijed određene, a $T_{(r)}$, trenutak u kojem promatranje završava je slučajno.

Složenija verzija gornje metode je *progressive Type II Censoring.* Početak je isti kao kod *Type II Censoringa*, unaprijed odredimo prirodni broj r_1, i čekamo da se dogodi r_1 neuspjeha u uzorku od n jedinki. Tada od preostalih $n-r_1$ jedinki, njih n_1-r_1 maknemo iz promatranja, a na ostalim $n-n_1$ ponovimo gornji postupak za neki, unaprijed određen prirodni broj r_2. Također, unaprijed je određeno koliko puta će se izvršiti serija ponavljanja.

1.6.2. Lijevo i intervalno cenzuriranje

Kažemo da je podatak lijevo cenzuriran ako se događaj od interesa za promatranu osobu dogodio prije njenog ulaska u istraživanje. Za takav slučaj znamo da se događaj od interesa već pojavio nekad prije trenutka C_l, gdje je C_l početak istraživanja, no ne znamo točno kad. Točno vrijeme pojave događaja od interesa znamo samo za one osobe koje su ga iskusile nakon trenutka C_l. Podatke

lijevo cenzuriranog uzorka prikazujemo u obliku uređenih parova slučajnih varijabli oblika (T, ε) gdje je $T = \max(X, C_l)$, a ε je indikator koji ukazuje na to da li se događaj od interesa dogodio za vrijeme istraživanja $(\varepsilon = 1)$, ili nije $(\varepsilon = 0)$.

Primjer 1.6.2.1. Za primjer ovakve vrste podataka ćemo spomenuti jedno istraživanje, vidi [2, str. 70-71], kojem je cilj bio odrediti funkciju distribucije prvog konzumiranja marihuane među srednjoškolcima u Kaliforniji. Pitanje koje im je bilo postavljano je glasilo "Kada ste prvi put konzumirali marihuanu?". Neki odgovori su glasili ovako "Konzumirao sam je, no ne sjećam se kada je sam to učinio prvi put.". Odgovori poput ovog su primjer lijevo cenzuriranih podataka, zbog toga što znamo da se događaj od interesa, prvo konzumiranje marihuane, dogodio nekad prije, no ne i točno kad.

Često se može dogoditi kombinacija lijevo i desno cenzuriranih podataka, stoga imamo duplo cenzurirane podatke. Takve podatke prikazujemo uređenim parom varijabla (T, δ), gdje je $T = \max[\min(X, C_r), C_l]$, a δ je indikator koji poprima vrijednost 1 ako je *T* trenutak smrti, vrijednost 0 ako je *T* desno cenzurirano, i vrijednost -1 ako je *T* lijevo cenzurirano. Stoga točan trenutak smrti znamo samo ako se ona dogodila unutar intervala $[C_l, C_r]$.

Mogući odgovor na pitanje "Kada ste prvi put konzumirali marihuanu?" bi mogao biti "Nisam je nikad konzumirao". Takav odgovor daje desno cenzuriran podatak. Ukoliko imamo i lijevo i desno cenzuriranih podataka, tada je naš uzorak zapravo duplo cenzuriran.

Sljedeći tip cenzuriranja se pojavljuje u slučajevima gdje se promatranja ponavljaju nekoliko puta u različitim intervalima. Ovaj način se pojavljuje u nekim medicinskim istraživanjima, gdje pacijenti dolaze na promatranja u nekoliko perioda, te se samo na promatranju može ispostaviti da je nastupio događaj od interesa. Ako se događaj od interesa pojavio između dva promatranja kažemo da je intervalno cenzuriran. Također, ovaj način promatranja je čest i kod ispitivanja raznih proizvoda, gdje inspekcija dolazi u periodima kako bi ispitala kvalitetu i valjanost samih proizvoda.

1.7. Odrezani podaci

Sljedeće vrsta podataka koja se pojavljuje među podacima o preživljavanju su odrezani podaci (*eng.truncation data*). Kažemo da je uzorak podataka odrezan ukoliko se u njemu nalaze samo podaci za one osobe koje su iskusile događaj od interesa točno u unaprijed određenom intervalu (Y_l, Y_r). Osobe koje nisu iskusile događaj od interesa u navedenom intervalu nisu promatrane, i istražitelji nemaju nikakve informacije o njima, one su 'odrezane' od istraživanja. To je i razlika između odsječenih i cenzuriranih podataka, jer smo kod cenzuriranih podataka imali barem djelomičnu informaciju i o tim osobama. Veoma je važno da znamo da podaci s kojima radimo odrezani, jer se za procjene osnovnih parametara analize preživljavanja koriste druge tehnike.

Kada je Y_r beskonačan, imamo lijevo odsijecanje. Ovdje dolaze u obzir sve one osobe koje događaj od interesa iskuse nakon trenutka Y_l. Ovakvu vrstu podataka možemo imati kada, na primjer, želimo procijeniti distribuciju veličina nekih sitnih čestica. Tada pomoću mikroskopa mjerimo promjer tih čestica, no, neke čestice su možda toliko malene da neće biti moguće niti pomoću mikroskopa izmjeriti njihov promjer, te ćemo njih zanemariti, odnosno nećemo uzeti nikakve podatke o njima. U ovom slučaju odrezali smo sve podatke vezane za čestice koje imaju promjer manji od onog kojeg može zabilježiti mikroskop. Još jedan primjer lijevog odsijecanja može biti istraživanje nad stanovnicima umirovljeničkog doma. Da bi osoba mogla biti primljena u umirovljenički dom, ona mora imati dovoljan broj godina, stoga sve osobe koje su umrle ranije ne mogu biti dio istraživanja, i zato ove podatke smatramo lijevo odrezanima.

Desno odrezane podatke imamo kada je $Y_l = 0$. U ovom slučaju bilježimo podatke samo o onim osobama koje su događaj od interesa iskusile prije trenutka Y_r. Pretpostavimo da želimo procijeniti distribuciju udaljenosti zvijezda od zemlje. Neke zvijezde su toliko daleke da neće biti moguće izmjeriti njihovu udaljenost od zemlje, stoga njih nećemo uzeti u obzir, odnosno podaci o njima će biti odrezani. Desno odrezani podaci se pojavljuju i u sljedećem istraživanju kojim se želi

procijeniti distribucija vremena potrebnog za razvijanje bolesti HIV-a, ukoliko se osoba njime zarazila transfuzijom. Recimo da imamo fiksiran datum kada ćemo uzimati podatke za ovo istraživanje. U obzir nećemo uzimati podatke za one osobe kojima je razdoblje potrebno za razvijanje bolesti veće od razdoblja koje je prošlo od dana transfuzije do datuma kada se prikupljaju podaci. Stoga, podaci za osobe kojima se bolest nije razvila do dana uzimanja podataka su desno odrezani.

2. Neparametarska procjena osnovnih veličina analize preživljavanja za desno cenzurirane i lijevo odrezane podatke

U ovom odjeljku ćemo pokazati kako se neparametarski mogu procijeniti funkcija preživljavanja, $S(x) = P(X > x)$, i funkcija kumulativnog rizika, $H(x) = \int_0^x h(u)du$, ukoliko je u podacima prisutno desno cenzuriranje ili su lijevo odrezani. Glavna pretpostavka kroz cijelo ovo poglavlje je da su potencijalno vrijeme cenzuriranja i potencijalno vrijeme pojave događaja od interesa nezavisni.

2.1. Procjena funkcije preživljavanja i kumulativnog rizika za desno cenzurirane podatke

Primjer 2.1.1. *(Procjena $S(t)$ na temelju podataka koji ne sadrže cenzuriranje)* Promatramo dvadeset i jednu osobu koje boluju od leukemije, te primaju neki određeni lijek, za više vidi [9, str. 44-46]. Događaj od interesa je ozdravljenje. Sljedeće brojke predstavljaju broj tjedana potrebnih za ozdravljenje svakog od bolesnika:

$$1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23.$$

Vidimo da su dva pacijenta ozdravila za jedan tjedan, stoga je $\hat{S}(1) = 19/21 = 0.905$. Broj pacijenata koji su ozdravili nakon drugog tjedna je 17, stoga je $\hat{S}(2) = 17/21 = 0.810$. Na isti način dobijemo da je $\hat{S}(3) = 16/21 = 0.762$, jer je šesnaest pacijenata ozdravilo nakon trećeg tjedna. Nastavimo li ovaj postupak sve do $\hat{S}(21)$ konstruirat ćemo procijenjenu funkciju preživljavanja za dani primjer. Općenita formula glasi:

$$\hat{S}(t) = \frac{broj\ pojedinaca\ s\ X>t}{veličina\ uzorka}. \qquad (2.1.1.)$$

t	$\hat{S}(t)$
$t < 1$	$21/21 = 1.000$
$1 \leq t < 2$	$19/21 = 0.905$
$2 \leq t < 3$	$17/21 = 0.810$
$3 \leq t < 4$	$16/21 = 0.762$
$4 \leq t < 5$	$14/21 = 0.667$
$5 \leq t < 8$	$12/21 = 0.571$
$8 \leq t < 11$	$8/21 = 0.381$
$11 \leq t < 12$	$6/21 = 0.286$
$12 \leq t < 15$	$4/21 = 0.190$
$15 \leq t < 17$	$3/21 = 0.143$
$17 \leq t < 22$	$2/21 = 0.095$
$22 \leq t < 23$	$1/21 = 0.048$

Tablica 2.1.1. Procjena funkcije preživljavanja

Pretpostavimo sada da su nam podaci bili sljedećeg oblika:

$$6^+, 6, 6, 6, 7, 9^+, 10^+, 10, 11^+, 13, 16, 17^+, 19^+, 20^+, 22, 23, 25^+, 32^+, 32^+, 34^+, 35^+,$$

gdje indeks $'+'$ označava desno cenzuriranje. Vidimo da su u šestom tjednu ozdravila tri pacijenta, te je za jednog pacijenta zabilježeno cenzurirano promatranje. No, ne možemo reći da je $\hat{S}(6) = 17/21$ jer ne znamo što se dogodilo s pacijentom koji je cenzuriran u šestom tjednu. 1958. godine. Kaplan i Meier su predložili rješenje ovog problema, odnosno, neparametarsku procjenu funkcije

preživljavanja za slučajeve gdje je prisutno desno cenzuriranje, pod nazivom Produkt-Limit procjenitelj.

2.1.1. Produkt-Limit procjenitelj

Pretpostavimo da se događaj od interesa može pojaviti u $D, D \in \mathbb{N}$ različitih trenutaka $t_1 < t_2 < \cdots < t_D, t_i \epsilon \mathbb{R}, i = 1, 2, \dots, D$. Neka je $t \in \mathbb{R}$ realan broj za koji vrijedi $t_k \leq t < t_{k+1}, k \in \{1, 2, \dots, D-1\}$. Tada imamo:

$$S(t) = \mathbb{P}(X \geq t_{k+1})$$

$$= \mathbb{P}(X \geq t_1, X \geq t_2, \dots, X \geq t_{k+1})$$

$$= \mathbb{P}(X \geq t_1) \prod_{j=1}^{k} \mathbb{P}(X \geq t_{j+1} | X \geq t_j)$$

$$= \prod_{j=1}^{k} \left[1 - \mathbb{P}\left(X = t_j \middle| X \geq t_j\right)\right]. \qquad (2.1.1.1.)$$

Označimo sa d_i broj događaja koji su se pojavili u trenutku t_i, s sa Y_i broj osoba koje su bile izložene riziku u trenutku t_i, tj. to je broj osoba koje su bile žive u trenutku t_i ili su umrle upravo u tom trenutku. Vremenski raspon istraživanja možemo podijeliti i u D intervala, pa nam d_i može predstavljati broj događaja koji su se pojavili u tom intervalu. Vjerojatnost umiranja u trenutku $t_{i,}$ λ_i, možemo procijeniti metodom maksimalne vjerodostojnosti. Funkcija vjerodostojnosti, $L(\lambda)$, je sljedećeg oblika:

$$L(\lambda) = \prod_{i=1}^{D} \lambda_i^{d_i} (1 - \lambda_i)^{Y_i - d_i}. \qquad (2.1.1.2.)$$

Procjenitelj za λ_i je tada $\hat{\lambda}_i = d_i / Y_i$.

Gornji rezultati opravdaju sljedeći oblik procjene funkcije preživljavanja:

$$\hat{S}(t) := \begin{cases} 1, t < t_1 \\ \prod_{t_i \leq t} \left[1 - \frac{d_i}{Y_i}\right], t_1 \leq t \end{cases}. \qquad (2.1.1.3.)$$

Ovaj procjenitelj je predložen od strane Kaplana i Meiera 1958. godine, i naziva se Produkt-Limit procjenitelj. To je skok funkcija sa skokovima u vrijednostima $t_1, t_2, \dots, t_D$. Veličina skoka ovisi o broju događaja koji su zabilježeni u pojedinom trenutku, no i o broju osoba koje su cenzurirane prije tog trenutka. Vidimo da za $t > t_D$ Produkt-Limit procjenitelj nije dobro definiran. Ako je zadnji zabilježeni događaj smrt, tada procjena postiže vrijednost jednaku nuli za sve $t > t_D$. No ako je u tom zadnjem trenutku zabilježeno cenzuriranje, ne možemo odrediti procjenu za $t > t_D$, jer ne znamo što se dalje događa s tim pojedincem. Imamo dva prijedloga rješenja ovog problema. Efron je 1967. godine predložio da $\hat{S}(t)$ poprima vrijednost nula za sve $t > t_D$, što se temelji na pretpostavci da će ta osoba koja je cenzurirana u trenutku t_D ubrzo umrijeti, stoga kažemo da je ova procjena negativno pristrana. Gial je 1980.godine predložio da $\hat{S}(t)$ poprima vrijednost $\hat{S}(t_D)$ za sve $t > t_D$, što se temelji na pretpostavci da će osoba cenzurirana u trenutku t_D živjeti beskonačno dugo, pa kažemo da je ova procjena pozitivno pristrana. Iako obje verzije na velikim uzorcima konvergiraju k pravoj vrijednosti funkcije preživljavanja, Klein je 1991. godine pokazao se Gillov prijedlog bolji na manjim uzorcima, kako se navodi u [2, str. 99-100].

Sada želimo procijeniti varijancu Produkt-Limit procjenitelja. Prisjetimo se, sa $\hat{\lambda}_i$ smo označavali procijenjenu vjerojatnost umiranja u trenutku/intervalu t_i. $\hat{\lambda}_i$ možemo shvatiti kao parametar (proporciju) binomne razdiobe, stoga slijedi da je $\hat{\lambda}_i$ aproksimativno normalna s očekivanjem λ_i i varijancom $\frac{\hat{\lambda}_i(1-\hat{\lambda}_i)}{Y_i}$.

Za slučajnu varijablu *Y*, s očekivanjem μ i varijancom σ^2, primjenom delta metode vrijedi sljedeće:

$$Z = \log(Y) \rightarrow Z \sim N\left[\log(\mu), \left(\frac{1}{\mu}\right)^2 \sigma^2\right] \qquad (2.1.1.4.)$$

i

$$Z = \exp(Y) \rightarrow Z \sim N[e^{\mu}, [e^{\mu}]^2 \sigma^2]. \qquad (2.1.1.5.)$$

Pogledajmo logaritam Produkt-Limit procjenitelja:

$$\log\left[\hat{S}(t)\right] = \sum_{t_i \le t} \log(1 - \frac{d_i}{Y_i}). \qquad (2.1.1.6.)$$

Tada, zbog nezavisnosti od $\frac{d_i}{Y_i}$ i gornjih rezultata imamo sljedeće:

$$var\left(\log\left[\hat{S}(t)\right]\right) = \sum_{t_i \le t} var\left[\log\left(1 - \frac{d_i}{Y_i}\right)\right]$$

$$= \sum_{t_i \le t} \left(\frac{1}{1 - \frac{d_i}{Y_i}}\right)^2 var\left(\frac{d_i}{Y_i}\right)$$

$$= \sum_{t_i \le t} \left(\frac{1}{1 - \frac{d_i}{Y_i}}\right)^2 \frac{d_i}{Y_i}(1 - \frac{d_i}{Y_i})/Y_i$$

$$= \sum_{t_i \le t} \frac{\frac{d_i}{Y_i}}{(1 - \frac{d_i}{Y_i})Y_i}$$

$$= \sum_{t_i \le t} \frac{d_i}{(Y_i - d_i)Y_i}. \qquad (2.1.1.7.)$$

Nadalje, $\hat{S}(t) = \exp\left[\log\left[\hat{S}(t)\right]\right]$, te zbog (2.1.1.5.) slijedi:

$$var\left(\hat{S}(t)\right) = \left[\hat{S}(t)\right]^2 var\left[\log\left[\hat{S}(t)\right]\right]. \qquad (2.1.1.8.)$$

Sada smo izveli Greenwood-ovu formulu za procjenu varijance Produkt-Limit procjenitelja:

$$var\left(\hat{S}(t)\right) = [\hat{S}(t)]^2 \sum_{t_i \le t} \frac{d_i}{Y_i(Y_i - d_i)}. \qquad (2.1.1.9.)$$

Standardna devijacija Produkt-Limit procjenitelja je dana s $\{var[\hat{S}(t)]\}^{1/2}$.

Zbog jednakosti $H(t) = -\ln[S(t)]$, Produkt-Limit procjenitelj se može upotrijebiti i za procjenu kumulativne funkcije rizika. Njezin procjenitelj je tada $\hat{H}(t) = -\ln[\hat{S}(t)]$.

2.1.2. Nelson-Aalen procjenitelj

Želimo procijeniti funkciju kumulativnog rizika, $H(x) = \int_0^x h(u)du$. Pretpostavimo da je razdoblje promatranja podijeljeno u sitne intervale, tako da se u jednom intervalu može pojaviti najviše jedan događaj, smrt, cenzuriranje, ili niti jedno od toga. Pretpostavimo da imamo podjelu vremena kao na slici 2.1.2.1., gdje *D* označava smrt, a *C* cenzuriranje.

Slika 2.1.2.1. Podjela razdoblja promatranja na intervale, izvor [9, str.97]

$H(x)$ možemo aproksimirati sljedećom sumom:

$$\tilde{H}(t) = \sum_i h_i \, \Delta, \qquad (2.1.2.1.)$$

gdje suma ide po intervalima, h_i je vrijednost funkcije rizika u *i*-tom intervalu, a Δ je širina intervala. Kako je $h_i\Delta$ procjena vjerojatnosti umiranja u i-tom intervalu, funkciju kumulativnog rizika možemo procijeniti na sljedeći način:

$$\tilde{H}(t) = \begin{cases} 0, t < t_1 \\ \sum_{t_1 \le t} \frac{d_i}{Y_i}, t_1 \le t. \end{cases} \qquad (2.1.2.2.)$$

$\tilde{H}(t)$ se naziva Nelson-Aalen procjenitelj. Ovaj procjenitelj je prvo bio predložen od strane Nelsona 1972. godine, a Aalen ga je 1978.godine ponovo izveo pomoću procesa prebrojavanja.

Procjena varijance je prema [8, str. 4] dana s:

$$\sigma_H^2 = \sum_{t_i \le t} \frac{d_i}{Y_i^2}. \qquad (2.1.2.3.)$$

Procjenu funkcije preživljavanja možemo izvesti i iz Nelson-Aalenovog procjenitelja, te tada ona glasi $\tilde{S}(t) = \exp\left[-\tilde{H}(t)\right]$.

2.2. Pouzdani intervali za funkciju preživljavanja

Standardna devijacija nam pruža informacije vezane za točnost procjene, tj. pomoću nje se mogu konstruirati pouzdani intervali za funkciju preživljavanja u točci t_0. Pouzdani intervali na razini značajnosti $(1-\alpha)$, nam govore da je $(1-\alpha)\times 100\%$ vjerovatno da će prava vrijednost funkcije preživljavanja biti upravo neka vrijednost iz tog intervala.

Najčešće korišteni $(1-\alpha)\times 100\%$ pouzdani interval za funkciju preživljavanja u trenutku t_0 je linearni pouzdani interval, definiran sa:

$$\hat{S}(t_0) - Z_{1-\frac{\alpha}{2}}\sigma_S(t_0)\hat{S}(t_0), \hat{S}(t_0) + Z_{1-\frac{\alpha}{2}}\sigma_S(t_0)\hat{S}(t_0), \qquad (2.2.1.)$$

gdje je $Z_{1-\alpha/2}$ kritična vrijednost za standardnu normalnu distribuciju, odnosno, $\Phi\left(Z_{1-\frac{\alpha}{2}}\right) = 1-\frac{\alpha}{2}$, a Φ je funkcija distribucije standardne normalne varijable, te je $\sigma_S^2 = Var(\hat{S}(t))/\hat{S}^2(t)$. Ovako konstruiran interval može poprimiti vrijednosti izvan intervala $[0,1]$, a znamo da funkcija preživljavanja može poprimiti vrijednosti samo iz tog intervala. Taj problem se može riješiti raznim transformacijama funkcije preživljavanja. Dodatno, tako konstruirani intervali su se pokazali boljima od klasičnog linearnog intervala.

Neka je g transformacija koju ćemo primijeniti na funkciju preživljavanja $S(t)$. Standardna devijacija transformacije $g(S(t))$, $sd(g(S(t)))$, je procijenjena delta metodom i jednaka je:

$$sd(g\left(\hat{S}(t)\right) = g'\left(\hat{S}(t)\right)\widehat{sd}\left(\hat{S}(t)\right). \qquad (2.2.2.)$$

$(1-\alpha)\times 100\%$ pouzdani interval je tada dan sa:

$$g^{-1}(g\left(\hat{S}(t)\right)\pm Z_{1-\frac{\alpha}{2}}g'\left(\hat{S}(t)\right)\widehat{sd}\left(\hat{S}(t)\right)), \qquad (2.2.3.)$$

gdje je g^{-1} inverzna funkcija od g.

Prva transformacija koju ćemo spomenuti je log-log transformacija. Kod ove transformacije prvo radimo log transformaciju kumulativne funkcije rizika, $\log\left(\widehat{H}(t)\right)=\ln(-\ln\left(\hat{S}(t)\right)$. Uvrštavajući u 2.2.2. vidimo da je $Var(\ln\left(-\ln\left(\hat{S}(t)\right)\right))=\frac{Var(\hat{S}(t))}{[\hat{S}(t)\ln\left(\hat{S}(t)\right)]^2}$. $(1-\alpha)\times 100\%$ pouzdani interval za $S(t)$ je tada dan sa:

$$[\widehat{\boldsymbol{S}}(\boldsymbol{t_0})]^{\exp\left(Z_{1-\frac{\alpha}{2}}\frac{\sqrt{\boldsymbol{Var}[\widehat{\boldsymbol{S}}(\boldsymbol{t_0})]}}{\widehat{\boldsymbol{S}}(\boldsymbol{t_0})\ln\left(\widehat{\boldsymbol{S}}(\boldsymbol{t_0})\right)}\right)},[\widehat{\boldsymbol{S}}(\boldsymbol{t_0})]^{\exp\left(-Z_{1-\frac{\alpha}{2}}\frac{\sqrt{\boldsymbol{Var}[\widehat{\boldsymbol{S}}(\boldsymbol{t_0})]}}{\widehat{\boldsymbol{S}}(\boldsymbol{t_0})\ln\left(\widehat{\boldsymbol{S}}(\boldsymbol{t_0})\right)}\right)}. \qquad (2.2.4)$$

Primijetimo da ovaj interval nije simetričan u odnosu na procijenjenu vrijednost $S(t_0)$.

Druga transformacija koju ćemo spomenuti je arcsin-korijen transformacija. Procjena varijance je $Var\left(\arcsin\left(\sqrt{\hat{S}(t_0)}\right)\right)=\frac{Var[\hat{S}(t_0)]}{4\hat{S}(t_0)[1-\hat{S}(t_0)]}$. $(1-\alpha)\times 100\%$ pouzdani interval za $S(t)$ je dan sa

$$sin^2\left\{\max\left[0,sin^{-1}\sqrt{\hat{S}(t_0)}-\frac{1}{2}Z_{1-\frac{\alpha}{2}}\sqrt{\frac{Var[\hat{S}(t_0)]}{\hat{S}(t_0)\left(1-\hat{S}(t_0)\right)}}\right]\right\},$$

$$sin^2\left\{\min\left[\frac{\pi}{2},sin^{-1}\sqrt{\hat{S}(t_0)}+\frac{1}{2}Z_{1-\frac{\alpha}{2}}\sqrt{\frac{Var[\hat{S}(t_0)]}{\hat{S}(t_0)\left(1-\hat{S}(t_0)\right)}}\right]\right\}. \qquad (2.2.5.)$$

Konstrukcija pouzdanih intervala slijedi iz konvergencije Produkt-Limit procjenitelja k Gaussovom procesu, što znači da za fiksni t, procjenitelj ima

normalnu distribuciju. Dokaz konvergencije Produkt-Limit procjenitelja Gaussovom procesu se može naći u [6, str.89-104].

2.3. Pouzdano područje za funkciju preživljavanja

Ponekad nam je od interesa pronaći pouzdano područje koje garantira s nekom razinom pouzdanosti, da funkcija preživljavanja poprima vrijednosti unutar tog područja za svaki t iz nekog intervala. U tu svrhu cilj nam je pronaći dvije funkcije, $L(t)$ i $U(t)$, takve da vrijedi $1-\alpha = \mathbb{P}[L(t) \leq S(t) \leq U(t), t_L \leq t \leq t_U]$. $[L(t), U(t)]$ zovemo $(1-\alpha)\times 100\%$ pouzano područje (*eng. confidence bands*) za $S(t)$. Više o pouzdanim područjima se može naći u [2, str. 109-117]

2.3.1. EP područje

Jedan način konstruiranja pouzdanih područja, predložen od Naira 1948. godine, daje pouzdano područje koje je proporcionalno pouzdanim intervalima iz prošle točke. Ovako dobiveno područje se zove *equal probability* ili EP područje. Prvo izaberemo t_L i t_D takve da je $t_L < t_D$, te da je t_L veći ili jednak najmanjem t za koji imamo podatke, a t_D manji ili jednak najvećem t za kojeg imamo podatke. Potrebno je još definirati sljedeće veličine:

$$a_L = \frac{n\sigma_S^2(t_L)}{1+n\sigma_S^2(t_L)} \qquad (2.3.1.1.)$$

i

$$a_U = \frac{n\sigma_S^2(t_U)}{1+n\sigma_S^2(t_U)}, \qquad (2.3.1.2.)$$

gdje je n veličina uzorka. Potom trebamo pronaći koeficijent $c_\alpha(a_L, a_U)$ iz pripadajuće tablice, tablice s ovim koeficijentima se mogu naći u [2, str. 459-467]. Kao i u slučaju pouzdanih intervala pokazat ćemo tri oblika pouzdanih područja, linearno, log i arcsin- korijen transformirano, izvor [2, str.109-110].

Linearno:

$$\hat{S}(t) - c_\alpha(a_L, a_U)\sigma_S(t)\hat{S}(t), \hat{S}(t) + c_\alpha(a_L, a_U)\sigma_S(t)\hat{S}(t). \quad (2.3.1.3.)$$

Log-transformirano:

$$\left[\hat{S}(t_0)\right]^{\exp\left(c_\alpha(a_L,a_U)\frac{\sqrt{Var[\hat{S}(t_0)]}}{\hat{S}(t_0)\ln\left(\hat{S}(t_0)\right)}\right)}, \left[\hat{S}(t_0)\right]^{\exp\left(-c_\alpha(a_L,a_U)\frac{\sqrt{Var[\hat{S}(t_0)]}}{\hat{S}(t_0)\ln\left(\hat{S}(t_0)\right)}\right)}. \quad (2.3.1.4.)$$

Arcsin-Korijen transformacija:

$$sin^2\left\{\max\left[0, sin^{-1}\sqrt{\hat{S}(t_0)} - \frac{1}{2}c_\alpha(a_L, a_U)\sqrt{\frac{Var\left[\hat{S}(t_0)\right]}{\hat{S}(t_0)\left(1-\hat{S}(t_0)\right)}}\right]\right\},$$

$$sin^2\left\{\min\left[\frac{\pi}{2}, sin^{-1}\sqrt{\hat{S}(t_0)} + \frac{1}{2}c_\alpha(a_L, a_U)\sqrt{\frac{Var\left[\hat{S}(t_0)\right]}{\hat{S}(t_0)\left(1-\hat{S}(t_0)\right)}}\right]\right\}. \quad (2.3.1.5.)$$

2.3.2. Hall Wellner pouzdano područje

Drugi način konstruiranja pouzdanih područja su predložili Hall i Wellner 1980. godine. U ovom slučaju dopušteno je da t_L poprimi vrijednost nula. Za konstruiranje pouzdanog područja nad $[t_L, t_U]$ trebamo pronaći odgovarajući koeficijent pouzdanosti $k_\alpha(a_L, a_U)$, tablice s koeficijentima se nalaze u [2, str. 468-476]. Ponovo, imamo tri oblika pouzdanih područja, izvor [2, str.110-112].

Linearno:

$$\hat{S}(t) - \frac{k_\alpha(a_L, a_U)[1 + n\sigma_S^2(t)]}{n^{1/2}}\hat{S}(t), \hat{S}(t) + \frac{k_\alpha(a_L, a_U)[1 + n\sigma_S^2(t)]}{n^{1/2}}\hat{S}(t).$$

(2.3.2.1.)

Log-transformirano:

$$[\hat{S}(t_0)]^{\exp(\frac{k_\alpha(a_L,a_U)[1+n\sigma_S^2(t)]}{n^{1/2}\ln[\hat{S}(t)]})}, [\hat{S}(t_0)]^{\exp(-\frac{k_\alpha(a_L,a_U)[1+n\sigma_S^2(t)]}{n^{1/2}\ln[\hat{S}(t)]})}. \quad (2.3.2.2.)$$

Arcsin-Korijen transformacija:

$$sin^2\left\{\max\left[0, sin^{-1}\sqrt{\hat{S}(t_0)} - 0.5\frac{k_\alpha(a_L,a_U)[1+n\sigma_S^2(t)]}{n^{1/2}}\sqrt{\frac{\hat{S}(t_0)}{\left(1-\hat{S}(t_0)\right)}}\right]\right\},$$

$$sin^2\left\{\min\left[\frac{\pi}{2}, sin^{-1}\sqrt{\hat{S}(t_0)} + 0.5\frac{k_\alpha(a_L,a_U)[1+n\sigma_S^2(t)]}{n^{1/2}}\sqrt{\frac{\hat{S}(t_0)}{\left(1-\hat{S}(t_0)\right)}}\right]\right\}.$$

(2.3.2.3.)

Konstrukcija pouzdanih područja također slijedi iz konvergencije Produkt-Limit procjenitelja k Gaussovom procesu. Za više, vidi [2, str. 109.-117.].

2.4. Procjena očekivanog trajanja života i kvantila

U formuli 1.4.5. smo pokazali da je očekivano trajanje života dano formulom $\mu = \int_0^\tau S(t)dt$. Procjenu očekivanog trajanja života lako možemo dobiti tako da $S(t)$ supstituiramo Produkt-Limit procjeniteljom. Ova procjena je prikladna samo u slučaju kad je u t_D zabilježena smrt, jer inače, u slučaju cenzuriranja Produkt-Limit procjenitelj nije definiran za $t > t_D$. Ovaj problem se može riješiti Efronovom korekcijom repa koja zadnji zabilježeni događaj promijeni u smrt, ukoliko je bilo cenzuriranje, te se napravi restrikcija intervala na $[0, t_{max}]$. Drugi način rješenja ovog problema je restrikcija na interval $[0, \tau]$, gdje je τ izabran od strane istraživača, te bi trebao predstavljati najduže moguće trajanje života. U svakom slučaju, procjena očekivanog trajanja života na intervalu $[0, \tau]$, gdje τ predstavlja

zadnji trenutak za koji imamo podatke, ili trenutak određen od strane istraživača, je dana s:

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)dt. \tag{2.4.1.}$$

Prema [2, str. 118] , varijanca je dana s:

$$\hat{V}[\hat{\mu}_\tau] = \sum_{i=1}^{D} [\int_{t_i}^{\tau} \hat{S}(t)dt]^2 \frac{d_i}{Y_i(Y_i - d_i)}, \tag{2.4.2.}$$

a $100 \times (1-\alpha)\%$ pouzdani interval za očekivano trajanje života je:

$$\hat{\mu}_\tau \pm Z_{1-\alpha/2}\sqrt{\hat{V}[\hat{\mu}_\tau]}. \tag{2.4.3.}$$

Produkt-Limit procjenitelj se koristi i kod procjene kvantila. Prisjetimo se, p-ti kvantil je definiran sa $x_p = \inf\{t: S(t) \leq 1-p\}$, to jest, x_p je najmanje vrijeme u kojem $S(t)$ postiže vrijednost manju ili jednaku $1-p$. Za procjenu od x_p uzimamo najmanji $\hat{x}_p$ za koji je Produkt-Limit procjenitelj manji ili jednak $1-p$, $\hat{x}_p = \inf\{t: \hat{S}(t) \leq 1-p\}$. Standardnu devijaciju od $\hat{x}_p$ je teško izračunati jer zahtijeva procjenu funkcije gustoće od X. Brookmeyer i Crowley su 1982.godine konstruirali pouzdani interval za $\hat{x}_p$ koji ne zahtijeva procjenu funkcije gustoće od X. Prema njima $100 \times (1-\alpha)\%$ pouzdani interval za $\hat{x}_p$, baziran na linearnom pouzdanom intervalu, je skup svih t-ova koji zadovoljavaju sljedeću jednadžbu:

$$-Z_{1-\alpha/2} \leq \frac{\hat{S}(t) - (1-p)}{\hat{V}^{1/2}[\hat{S}(t)]} \leq Z_{1-\alpha/2}. \tag{2.4.4.}$$

$100 \times (1-\alpha)\%$ pouzdani interval za $\hat{x}_p$, baziran na log transformiranom pouzdanom intervalu, je skup svih t-ova koji zadovoljavaju sljedeću jednadžbu:

$$-Z_{1-\alpha/2} \leq \frac{[\ln\{-\ln[\hat{S}(t)]\} - \ln\{-\ln[1-p]\}]\hat{S}(t)\ln[\hat{S}(t)]}{\hat{V}^{1/2}[\hat{S}(t)]} \leq Z_{1-\alpha/2}. \tag{2.4.5.}$$

$100 \times (1-\alpha)\%$ pouzdani interval za $\hat{x}_p$, baziran na arcsin-korijen transformaciji pouzdanom intervalu, je skup svih t-ova koji zadovoljavaju sljedeću jednadžbu:

$$-Z_{1-\alpha/2} \leq \frac{2\left\{\arcsin\left[\sqrt{\hat{S}(t)}\right] - \arcsin\left[\sqrt{(1-p)}\right]\right\}[\hat{S}(t)(1-\hat{S}(t))]^{1/2}}{\hat{V}^{1/2}[\hat{S}(t)]} \leq Z_{1-\alpha/2}.$$

(2.4.6.)

2.5. Procjena funkcije preživljavanja za lijevo odrezane i desno cenzurirane podatke

Sve procjene koje smo do sada prezentirali su bile bazirane ne temelju desno cenzuriranih podataka. Sada ćemo pokazati kako se te procjene mogu prilagoditi za uzorak s desno odrezanim i lijevo cenzuriranim podacima. Sa L_j ćemo označiti broj godina koje je pojedinac imao kada je ušao u istraživanje, a s T_j broj godina koje je imao kada je ili umro ili bio cenzuriran. Ponovo sa $t_1 < t_2 < \cdots < t_D$ označimo trenutke u kojima se dogodila smrt, te sa d_i broj pojedinaca koji su umrli u trenutku t_i. Ono što se mijenja u odnosu na do sad je Y_i, broj pojedinaca koji su izloženi riziku u t_i. Sada je Y_i broj pojedinaca koji su ušli u istraživanje prije t_i, te u njemu bili najmanje do trenutka t_i, odnosno Y_i je broj pojedinaca za koje vrijedi $L_j < t_i \leq T_j$.

Koristeći ovako definiran Y_i, sve procjene koje smo do sada definirali su valjane. Ipak, treba biti oprezan u interpretaciji ovih statistika. Na primjer, procjenitelj funkcije preživljavanja je sada vjerojatnost preživljavanja trenutka t, uz uvjet da je doživljena L-ta godina, gdje je L najmanji broj godina s kojih je netko ušao u istraživanje.

Također, integral u funkciji rizika sada ide od L do t.

2.6. Višestruki rizici

U poglavljima (2.1.-2.5.) nam je pretpostavka bila da su vremena pojave događaja i cenzuriranja nezavisna. U slučaju višestrukih rizika to ne mora biti tako. U ovom odjeljku ćemo se upoznati s tri tehnike obrade podataka u modelu

višestrukih rizika. Kako bismo lakše vidjeli razliku između te tri tehnike, razmotrit ćemo primjer presađivanje koštane srži. Događaj od interesa kojeg ovdje promatramo je neuspješno presađivanje. Presađivanje smatramo neuspješnim ako je pacijent umro, ili mu se bolest vratila. Ovdje su smrt i povratak bolesti višestruki rizici. Nas zanima kako se vjerojatnost pojave ovih događaja mijenja kroz vrijeme. Pojava jednog od ova dva događaja isključuje mogućnost pojave drugog.

Vjerojatnost pojave jednog od ovih događaja se može procijeniti komplementom Produkt-Limit procjenitelja. Pojavu drugog događaja tretiramo kao cenzurirano promatranje. Na primjer, procjena vjerojatnosti za povratak bolesti prije trenutka t je jedan minus Produkt-Limit procjenitelj za povratak bolesti, koji je dobiven tako da se povratak bolesti smatrao događajem od interesa, a pojava smrti prije povratka bolesti se tretirala kao cenzurirano promatranje. Ovo se može interpretirati kao vjerojatnost povratka bolesti, u hipotetičkom svijetu gdje je isključen rizik od smrti koja može nastupiti prije povratka bolesti.

Druga procjena koju ćemo prezentirati je funkcija kumulativnih događaja. Neka nam $t_1 < t_2 < \cdots < t_D$ budu vremena kada su zabilježene pojave nekog od mogućih događaja. Neka Y_i bude broj pojedinaca koji su izloženi rizicima u trenutku t_i, r_i neka bude broj pojedinaca koji su iskusili događaj od interesa u trenutku t_i, a d_i broj pojedinaca koji su iskusili neki drugi događaj kojem su bili izloženi u trenutku t_i. Primjetimo da je $r_i + d_i$ broj pojedinaca koji su iskusili bilo koji od mogućih događaja u trenutku t_i. U ovom slučaju se cenzuriranje ne smatra jednim od mogućih događaja, no vidljivo je u promjeni vrijednosti Y_i. Funkcija kumulativnog događaja je definirana sljedećom formulom:

$$CI(t) = \begin{cases} 0, t < t_1 \\ \sum_{t_i \le t} \{ \prod_{j=1}^{i-1} \frac{1 - [d_j + r_j]}{Y_j} \} \frac{r_i}{Y_i}, t_1 \le t \end{cases} . \qquad (2.6.1.)$$

Primijetimo da za $t \geq t_1$ vrijedi $CI(t) = \hat{S}(t_i-)\,\frac{r_i}{Y_i}$, gdje je $\hat{S}(t_i-)$ vrijednost Produkt-Limit izračunata netom prije trenutka t_i, koji je dobiven tako da se bilo koji od mogućih događaja tretirao kao događaj od interesa. $CI(t)$ je procjena vjerojatnosti da se događaj od interesa pojavio prije trenutka t, i prije svih drugih mogućih događaja. To je procjena vjerojatnosti pojave događaja od interesa u svijetu gdje je moguća pojava i ostalih događaja.

Treća procjena je funkcija uvjetne vjerojatnosti za višestruke rizike. Za određeni rizik K, sa $CI_K(t)$ i $CI_{K^c}(t)$ označimo funkciju kumulativnih događaja za rizik K, odnosno za sve preostale rizike skupljene zajedno. Funkcija uvjetne vjerojatnosti je tada definirana sa:

$$CP_K(t) = \frac{CI_K(t)}{1 - CI_{K^c}(t)}\ . \qquad (2.6.2.)$$

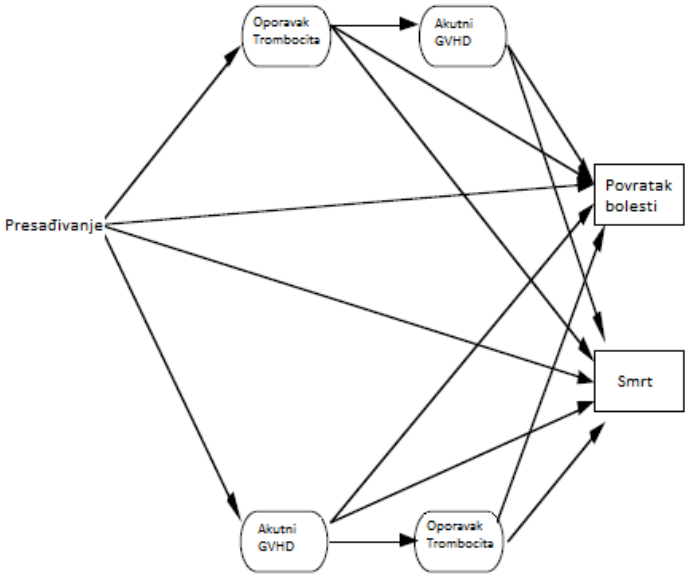
3. Primjena analize preživljavanja kod oporavka nakon presađivanje koštane srži

3.1. Uvod

U ovom poglavlju ćemo primijeniti tehniku procjenjivanja iz drugog poglavlja na podatke vezane za oporavak nakon presađivanja koštane srži. Podaci su preuzeti iz [2, str.3-133, 483-489]. Također, tamo se mogu naći još neke dodatne informacije.

Presađivanje koštane srži je standardni tretman koji se provodi nad oboljelima od akutne leukemije. Oporavak nakon presađivanja ovisi o više faktora, poput starosti i spola pacijenta, no i donora, stupnju bolesti, te o vremenu koje je prošlo od dijagnoze do presađivanja. Nakon presađivanja može doći do razvoja akutne i kronične bolesti transplantanta protiv domaćina, eng. graft-versus host (GVHD), te infekcija, no i do povratka trombocita i granulocita na normalnu razinu. Presađivanje koštane srži se smatra neuspješnim ako se bolest vratila, ili pacijent umro.

Slika 3.1.1. pokazuje pojednostavljeni dijagram procesa oporavka pacijenta. Imamo dva moguća posredna događaja koji se mogu pojaviti tokom oporavka, razvoj akutnog GVHD-a i povratak broja trombocita na normalnu razinu. Odmah nakon presađivanja, pacijent ima smanjen broj trombocita, te nema akutni GVHD. Povratak trombocita na normalnu razinu, te razvoj akutnog GVHD mogu utjecati na to da nastupi smrt ili se bolest vrati. Slika 3.1.1. pokazuje redoslijed kojim se mogu pojaviti ovi događaji.



Slika 3.1.1. Proces oporavka nakon presađivanja koštane srži, izvor [2, str. 4]

Mi ćemo uzeti podatke dobivene promatranjem pacijenata koji idu na presađivanje, a boluju od akutne mijeloidne leukemije (AML) ili akutne limfoblastične leukemije (ALL). Na promatranju je bilo ukupno 137 pacijenata, 99 njih koji su bolovali od AML i 38 koji su bolovali od ALL. Promatrani pacijenti su se liječili u jednoj od četiri bolnice koje su se nalazile u Columbusu, Philadelphiji, Sydneyu i Melbourneu, u razdoblju od početka ožujka 1984. godine, pa do kraja lipnja 1989. godine. Najduže promatranje nekog pacijenta je trajalo sedam godina. Nakon presađivanja bolest se vratila kod 42 pacijenta, dok je 41 pacijent umro. Kod 26 pacijenata se razvio akutni GVHD, dok je 17 pacijenata umrlo bez da im se broj trombocita vratio na normalnu razinu.

Pacijenti su bili podijeljeni u sljedeće kategorije ovisno o statusu bolesti u trenutku presađivanja: ALL(38 pacijenata), AML-niski rizik (54 pacijenta), i AML-visok rizik (45 pacijenata). Kod presađivanja su još zabilježeni sljedeći podaci: spol primatelja i donora koštane srži (80, odnosno 88 muškog spola), status citomegalovirusa primatelja i donora (68, odnosno 58 njih je bilo pozitivno), starost primatelja i donora (bila je u rasponu od 7-52 godine za primatelje, te 2-56 godina za donore), duljina razdoblja između dijagnoze i presađivanja (u rasponu od 0.8 do 87.2 mjeseca), te FAB klasifikaciju (44 njih je su pripadali grupi M4 ili M5).

Pacijentima u bolnicama u Australiji je dana kombinacija metotreksada, ciklosporina i metilprednizolona kao preventiva GVHD-u, dok je pacijentima u ostalim bolnicama dana ta kombinacija ali bez metotreksada.

3.2. Obrada podataka pomoću programskog jezika R

Sada ćemo pomoću Produkt-Limit i Nelson-Aalen procjenitelja aproksimirati funkciju preživljavanja i funkciju kumulativnog rizika. Podaci s ovog promatranja se nalaze u tablici A.1. u Dodatku A. Procijenit ćemo vjerojatnost da se pacijentu koji je bio podvrgnut presađivanju koštane srži bolest vrati ili da umre tek nakon trenutka t. Indikator će nam biti δ_3, $\delta_3 = \max(\delta_1, \delta_2)$, gdje je δ_1 indikator pojave smrti, a δ_2 indikator povratka bolesti. Za broj dana provedenih na promatranju smo uzimali onaj koji je bio potreban da dođe do povratka bolesti ili smrti, ovisno o tome koji događaj je nastupio prije. U tablicama 3.2.1-3.2.3. su izračunate vrijednosti Produkt-Limit i Nelson Aalen procjenitelja za sve tri skupine ALL, AML - nizak rizik, AML - visok rizik. Za računanje je korišten programski jezik R, a kod ćemo prikazati kasnije.

t_i	d_i	Y_i	$\hat{S}(t_i)$	$\widetilde{H}(t_i)$
1	1	38	0.9736842	0.9736842
55	1	37	0.9473684	0.9473684
74	1	36	0.9210526	0.9210526
86	1	35	0.8947368	0.8947368
104	1	34	0.8684211	0.8684211
107	1	33	0.8421053	0.8421053
109	1	32	0.8157895	0.8157895
110	1	31	0.7894737	0.7894737
122	2	30	0.7368421	0.7368421
129	1	28	0.7105263	0.7105263
172	1	27	0.6842105	0.6842105
192	1	26	0.6578947	0.6578947
194	1	25	0.6315789	0.6315789
230	1	23	0.6041190	0.6041190
276	1	22	0.5766590	0.5766590
332	1	21	0.5491991	0.5491991
383	1	20	0.5217391	0.5217391
418	1	19	0.4942792	0.4942792
466	1	18	0.4668192	0.4668192
487	1	17	0.4393593	0.4393593
526	1	16	0.4118993	0.4118993
609	1	14	0.3824779	0.3824779
662	1	13	0.3530566	0.3530566

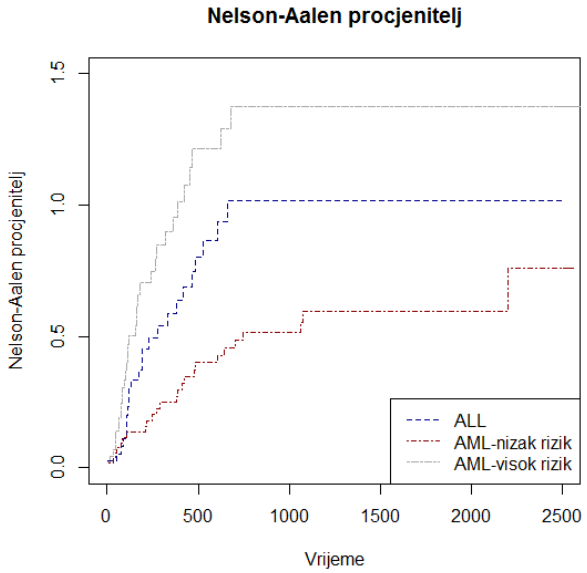
Tablica 3.2.1. Procijenjene vrijednosti funkcije preživljavanja i funkcije rizika za skupinu ALL

t_i	d_i	Y_i	$\hat{S}(t_i)$	$\tilde{H}(t_i)$
10	1	54	0.9814815	0.01851852
35	1	53	0.9629630	0.03738644
48	1	52	0.9444444	0.05661721
53	1	51	0.9259259	0.07622506
79	1	50	0.9074074	0.09622506
80	1	49	0.8888889	0.11663322
105	1	48	0.8703704	0.13746655
211	1	47	0.8518519	0.15874315
219	1	46	0.8333333	0.18048228
248	1	45	0.8148148	0.20270450
272	1	44	0.7962963	0.22543177
288	1	43	0.7777778	0.24868759
381	1	42	0.7592593	0.27249711
390	1	41	0.7407407	0.29688735
414	1	40	0.7222222	0.32188735
421	1	39	0.7037037	0.34752838
481	1	38	0.6851852	0.37384417
486	1	37	0.6666667	0.40087120
606	1	36	0.6481481	0.42864897
641	1	35	0.6296296	0.45722040
704	1	34	0.6111111	0.48663217
748	1	33	0.5925926	0.51693520
1063	1	26	0.5698006	0.55539674
1074	1	25	0.5470085	0.59539674
2204	1	6	0.4558405	0.76206340

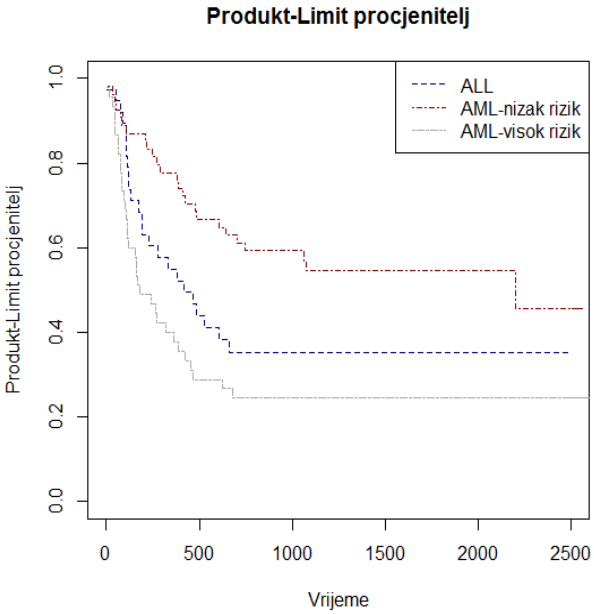
Tablica 3.2.2. Procijenjene vrijednosti funkcije preživljavanja i funkcije rizika za skupinu AML-nizak rizik

t_i	d_i	Y_i	$\hat{S}(t_i)$	$\tilde{H}(t_i)$
2	1	45	0.9777778	0.02222222
16	1	44	0.9555556	0.04494949
32	1	43	0.9333333	0.06820531
47	2	42	0.8888889	0.11582436
48	1	40	0.8666667	0.14082436
63	1	39	0.8444444	0.16646538
64	1	38	0.8222222	0.19278117
74	1	37	0.8000000	0.21980820
76	1	36	0.7777778	0.24758598
80	1	35	0.7555556	0.27615741
84	1	34	0.7333333	0.30556917
93	1	33	0.7111111	0.33587220
100	1	32	0.6888889	0.36712220
105	1	31	0.6666667	0.39938026
113	1	30	0.6444444	0.43271360
115	1	29	0.6222222	0.46719636
120	1	28	0.6000000	0.50291064
157	1	27	0.5777778	0.53994768
162	1	26	0.5555556	0.57840922
164	1	25	0.5333333	0.61840922
168	1	24	0.5111111	0.66007588
183	1	23	0.4888889	0.70355415
242	1	22	0.4666667	0.74900869
268	1	21	0.4444444	0.79662774
273	1	20	0.4222222	0.84662774
318	1	19	0.4000000	0.89925932
363	1	18	0.3777778	0.95481487
390	1	17	0.3555556	1.01363840
422	1	16	0.3333333	1.07613840
456	1	15	0.3111111	1.14280507
467	1	14	0.2888889	1.21423364
625	1	13	0.2666667	1.29115672
677	1	12	0.2444444	1.37449005

Tablica 3.2.3. Procijenjene vrijednosti funkcije preživljavanja i funkcije rizika za skupinu AML-visok rizik



Graf 3.2.4. Procjena kumulativne funkcije rizika



Graf 3.2.5. Procjena funkcije preživljavanja

Na grafu 3.2.5. vidimo funkciju preživljavanja za sve tri skupine. Najbolju prognozu imaju pacijenti iz skupine AML-nizak rizik, dok najgoru prognozu imaju pacijenti iz skupine AML-visok rizik. Na primjer, vjerojatnost da je pacijent još uvijek zdrav nakon tri godine je 0.3531 za ALL skupinu, 0.5470 za skupinu AML-nizak rizik, te 0.2444 za skupinu AML-visok rizik. Graf 3.2.4. predstavlja procjenu funkcije kumulativnog rizika za sve tri skupine. Iz grafa se vidi da skupina AML-visok rizik ima najveću stopu rizika od povratka bolesti i smrti, dok AML-nizak rizik ima najmanju.

Sada ćemo prikazati kod kojim smo izračunali vrijednosti iz tablica (3.2.1.-3.2.3.) i nacrtali grafove (3.2.4.-3.2.5.) pomoću programskog jezika R.

Kako bismo mogli koristiti neke gotove funkcije trebamo instalirati paket *OIsurv.*

```
>install.packages("OIsurv")
>library(OIsurv)
>data(bmt)
>attach(bmt)
```

Funkcija za obradu podataka glasi *Surv(time, indicator)* ili *Surv(time1, time2, indicator)* za desno cenzurirane, odnosno desno cenzurirane i lijevo odrezane podatke.

```
>podaci<-Surv(t2[group==1], d3[group==1])
>podaci
[1] 2081+ 1602+ 1496+ 1462+ 1433+ 1377+ 1330+  996+  226+ 1199+ 1111+
530+
[13] 1182+ 1167+  418   383   276   104   609   172   487   662   194   230
[25]  526   122   129    74   122    86   466   192   109    55     1   107
[37]  110   332

>survfit(podaci ~ 1)
```

```
Call: survfit(formula = podaci ~ 1)

records   n.max n.start  events  median 0.95LCL 0.95UCL
     38      38      38      24     418     194      NA
```

Na sljedeći način možemo ispisati vektore vremena, broja događaja, broja pojedinaca izloženih riziku, funkcije preživljavanja i standardne greške.

```
>procjena <- survfit(podaci~1)

>summary(procjena)$time
>summary(procjena)$n.event
>summary(procjena)$n.risk
>summary(procjena)$surv
>summary(procjena)$std.err
```

Konstruiranje Nelson-Aalenovog procjenitelja:

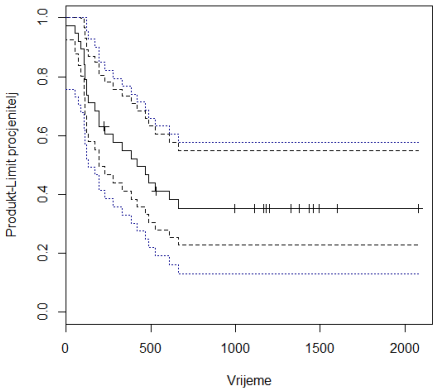
```
>na.pomocni<-procjena$n.event/procjena$n.risk
>nelson_aalen<-cumsum(na.pomocni)
```

Crtanje grafova Produkt-Limit i Nelson-Aalen procjenitelja:

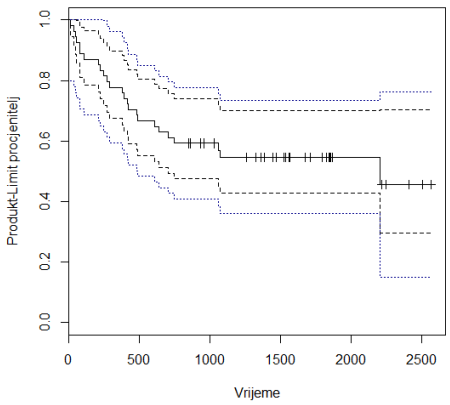
```
>plot(c(procjena$time, 2500), c(procjena$surv,tail(procjena$surv,1)),
+type="s", ylim=c(0,1),lty=2, col="darkblue", xlab="Vrijeme",
+ylab="Produkt-Limit procjenitelj", main="Produkt-Limit procjenitelj")

>plot(c(procjena$time, 2500), c(nelson_aalen,tail(nelson_aalen,1)),
+ylim=c(0,1.5),type="s",lty=2, col="darkblue", xlab="Vrijeme",
+ylab="Nelson-Aalen procjenitelj", main="Nelson-Aalen procjenitelj")

>
```

Crtanje pouzdanih intervala i područja:

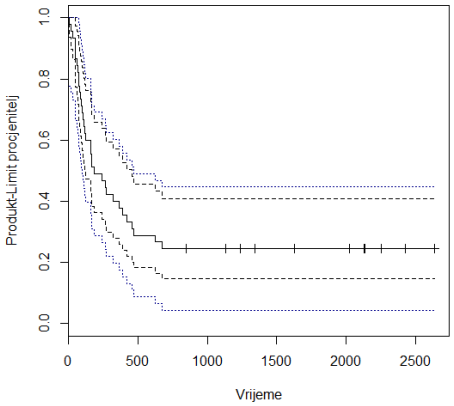
```
>plot(survfit(podaci ~ 1),xlab="Vrijeme",ylab="Produkt-Limit procjenitelj")
>granice<-confBands(podaci,confLevel=0.95,type="hall")
>lines(granice$time,granice$lower,lty=3,type="s", col="darkblue")
>lines(granice$time,granice$upper,lty=3,type="s", col="darkblue")
```



Graf 3.2.6. Pouzdani intervali i područje za funkciju preživljavanja za skupinu ALL



Graf 3.2.7. Pouzdani intervali i područje za funkciju preživljavanja za skupinu AML-nizak rizik



Graf 3.2.7. Pouzdani intervali i područje za funkciju preživljavanja za skupinu AML-visok rizik

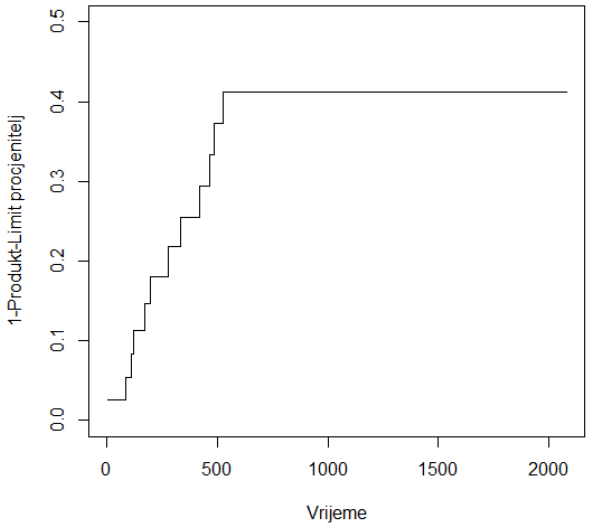
Medijan:

```
> print(survfit(podaci ~ 1),print.rmean=TRUE)
Call: survfit(formula = podaci ~ 1)

  records     n.max   n.start    events    *rmean *se(rmean)    median
       38        38        38        24       899       146       418
  0.95LCL   0.95UCL
      194        NA
   * restricted mean with upper limit =  2081
>
> print(survfit(my.surv.object2 ~ 1),print.rmean=TRUE)
Call: survfit(formula = my.surv.object2 ~ 1)

  records     n.max   n.start    events    *rmean *se(rmean)    median
       54        54        54        25      1549       151      2204
  0.95LCL   0.95UCL
      704        NA
   * restricted mean with upper limit =  2569
> print(survfit(my.surv.object3 ~ 1),print.rmean=TRUE)
Call: survfit(formula = my.surv.object3 ~ 1)

  records     n.max   n.start    events    *rmean *se(rmean)    median
       45        45        45        34       792       158       183
  0.95LCL   0.95UCL
      115       456
   * restricted mean with upper limit =  2640
>
```

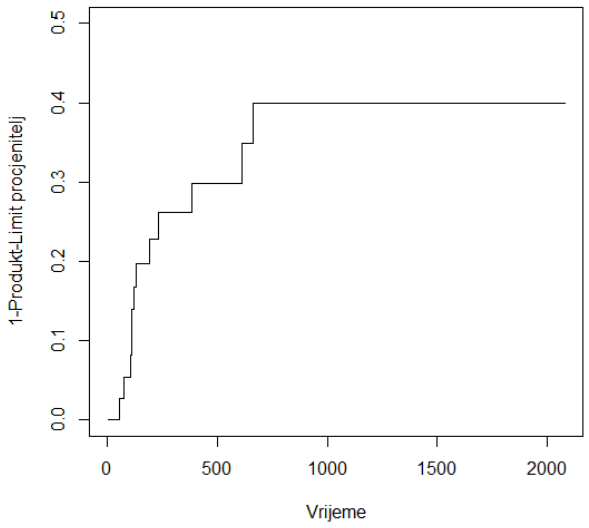
Računanje grube vjerojatnosti za povratak bolesti, u hipotetičkom svijetu gdje se ne može umrijeti i obratno:

```
>com.risk<-Surv(t2[group==1],d2[group==1])
>survfit(com.risk~1)
>proc.cr.1<-survfit(com.risk~1)
>cr.relapse.1<-1-summary(proc.cr.1)$surv
> cr.relapse.1
 [1] 0.02702703 0.05405405 0.08187599 0.11056737 0.13925874 0.16795012
 [7] 0.19766619 0.22852518 0.26206756 0.29896418 0.34903817 0.39911216

>d=d1[group==1]-d2[group==1]
>com.risk2<-Surv(t2[group==1],d)
```

```
>survfit(com.risk2~1)
>proc.cr.2<-survfit(com.risk2~1)
>cr.death.1<-1-summary(proc.cr.2)$surv
> cr.death.1
 [1] 0.02631579 0.05413534 0.08279790 0.11337131 0.14620941 0.18036103
 [7] 0.21761735 0.25487366 0.29409084 0.33330802 0.37252519 0.41174237
>
```



Graf 3.2.8. Gruba vjerojatnost za smrt (u hipotetičkom svijetu gdje je povratak bolesti nemoguć)



Graf 3.2.9. Gruba vjerojatnost za povratak bolesti (u hipotetičkom svijetu gdje je smrt nemoguća)

Dodatak A

Tablica A.1. Podaci o oporavku nakon presađivanja koštane srži, izvor [2, str. 483-488]:

grupa	t1	t2	d1	d2	d3
1	2081	2081	0	0	0
1	1602	1602	0	0	0
1	1496	1496	0	0	0
1	1462	1462	0	0	0
1	1433	1433	0	0	0
1	1377	1377	0	0	0
1	1330	1330	0	0	0
1	996	996	0	0	0
1	226	226	0	0	0
1	1199	1199	0	0	0
1	1111	1111	0	0	0
1	530	530	0	0	0
1	1182	1182	0	0	0
1	1167	1167	0	0	0
1	418	418	1	0	1
1	417	383	1	1	1
1	276	276	1	0	1
1	156	104	1	1	1
1	781	609	1	1	1
1	172	172	1	0	1
1	487	487	1	0	1
1	716	662	1	1	1
1	194	194	1	0	1
1	371	230	1	1	1
1	526	526	1	0	1
1	122	122	1	0	1
1	1279	129	1	1	1
1	110	74	1	1	1
1	243	122	1	1	1
1	86	86	1	0	1
1	466	466	1	0	1
1	262	192	1	1	1
1	162	109	1	1	1
1	262	55	1	1	1
1	1	1	1	0	1
1	107	107	1	0	1

1	269	110	1	1	1
1	350	332	1	0	1
2	2569	2569	0	0	0
2	2506	2506	0	0	0
2	2409	2409	0	0	0
2	2218	2218	0	0	0
2	1857	1857	0	0	0
2	1829	1829	0	0	0
2	1562	1562	0	0	0
2	1470	1470	0	0	0
2	1363	1363	0	0	0
2	1030	1030	0	0	0
2	860	860	0	0	0
2	1258	1258	0	0	0
2	2246	2246	0	0	0
2	1870	1870	0	0	0
2	1799	1799	0	0	0
2	1709	1709	0	0	0
2	1674	1674	0	0	0
2	1568	1568	0	0	0
2	1527	1527	0	0	0
2	1324	1324	0	0	0
2	957	957	0	0	0
2	932	932	0	0	0
2	847	847	0	0	0
2	848	848	0	0	0
2	1850	1850	0	0	0
2	1843	1843	0	0	0
2	1535	1535	0	0	0
2	1447	1447	0	0	0
2	1384	1384	0	0	0
2	414	414	1	0	1
2	2204	2204	1	0	1
2	1063	1063	1	0	1
2	481	481	1	0	1
2	105	105	1	0	1
2	641	641	1	0	1
2	390	390	1	0	1
2	288	288	1	0	1
2	522	421	1	1	1
2	79	79	1	0	1

2	1156	748	1	1	1
2	583	486	1	1	1
2	48	48	1	0	1
2	431	272	1	1	1
2	1074	1074	1	0	1
2	393	381	1	1	1
2	10	10	1	0	1
2	53	53	1	0	1
2	80	80	1	0	1
2	35	35	1	0	1
2	1499	248	0	1	1
2	704	704	1	0	1
2	653	211	1	1	1
2	222	219	1	1	1
2	1356	606	0	1	1
3	2640	2640	0	0	0
3	2430	2430	0	0	0
3	2252	2252	0	0	0
3	2140	2140	0	0	0
3	2133	2133	0	0	0
3	1238	1238	0	0	0
3	1631	1631	0	0	0
3	2024	2024	0	0	0
3	1345	1345	0	0	0
3	1136	1136	0	0	0
3	845	845	0	0	0
3	491	422	1	1	1
3	162	162	1	0	1
3	1298	84	1	1	1
3	121	100	1	1	1
3	2	2	1	0	1
3	62	47	1	1	1
3	265	242	1	1	1
3	547	456	1	1	1
3	341	268	1	1	1
3	318	318	1	0	1
3	195	32	1	1	1
3	469	467	1	1	1
3	93	47	1	1	1
3	515	390	1	1	1
3	183	183	1	0	1

3	105	105	1	0	1
3	128	115	1	1	1
3	164	164	1	0	1
3	129	93	1	1	1
3	122	120	1	1	1
3	80	80	1	0	1
3	677	677	1	0	1
3	73	64	1	1	1
3	168	168	1	0	1
3	74	74	1	0	1
3	16	16	1	0	1
3	248	157	1	1	1
3	732	625	1	1	1
3	105	48	1	1	1
3	392	273	1	1	1
3	63	63	1	0	1
3	97	76	1	1	1
3	153	113	1	1	1
3	363	363	1	0	1

Legenda:

grupa-grupa bolesti

1-ALL

2-AML nizak rizik

3-AML visok rizik

t1-vrijeme u danima do smrti ili kraja istraživanja

t2-vrijeme do povratka bolesti, smrti ili kraja istraživanja

d1-indikator smrti

1-smrt, 0-živ

d2-indikator povratka bolesti

1-bolest se vratila, 0-zdrav

d3-indikator

1-povratak bolesti ili smrt

0-živ i zdrav

Literatura

[1] M.D. Diez, Survival Analysis in R, dostupno na

https://www.openintro.org/download.php?file=survival_analysis_in_R&referrer=/stat/surv.php (kolovoz 2015.)

[2] J. P. Klein, M. L. Moeschberger, Survival Analysis, Techniques for Censored and Truncated dana, Springer, New York, 2003.

[3] Kaplan-Meier and Nelson-Aalen with right-censored and left truncated data, dostupno na

http://www.ms.uky.edu/~mai/sta635/rightcensorlefttruncate.pdf (kolovoz 2015.)

[4] M.Huzak, Vjerojatnost i matematička statistika, dostupno na http://aktuari.math.pmf.unizg.hr/docs/vms.pdf, (kolovoz 2015.)

[5] N. Sarapa, Teorija vjerojatnosti, Školska knjiga, Zagreb, 2002.

[6] O. Aalen, O. Borgan, H. K. Gjessing, Survival and Event History Analysis, Springer, New York, 2008.

[7] Pointwise Confidence Limits in the OUTSURV=Data set, http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_lifetest_a0000000262.htm, (kolovoz 2015.)

[8] Statistical Survival Analysis, Kaplan-Meier and Nelson-Aalen with right-censored and left truncated data, dostupno na

http://www.ms.uky.edu/~mai/sta635/rightcensorlefttruncate.pdf (kolovoz 2015.)

[9] Survival analysis, Introduction, dostupno na

http://www.amstat.org/chapters/northeasternillinois/pastevents/presentations/summer05_Ibrahim_J.pdf (kolovoz 2015.)

Sažetak

U ovom radu obrađene su tehnike neparametarske procjene osnovnih veličina analize preživljavanja za lijevo cenzurirane i desno odrezane podatke. Analiza preživljavanja je skup tehnika kojima se procjenjuje i opisuje vrijeme potrebno do pojave jednog ili više određenih događaja. Osnovne veličine analize preživljavanja su funkcija preživljavanja, funkcija rizika, kumulativna funkcija rizika i očekivano trajanje života. One su detaljno definirane u prvom poglavlju. Prije njih navedene su osnovne definicije i pojmovi vjerojatnosti i statistike. Također, u istom poglavlju objašnjeni su tipovi podataka koji se obrađuju u analizi preživljavanja. Podaci mogu biti cenzurirani i odrezani. Postoji više vrsta svakog tipa, tako imamo lijevo / desno cenzurirane podatke, lijevo / desno odrezane podatke, intervalno cenzurirane podatke, te kombinacije. Veoma je bitno da prepoznamo s kakvim oblikom podataka radimo, jer se ovisno o njima mijenja i tehnika kojom procjenjujemo.

U drugom poglavlju obrađeni su procjenitelji funkcije preživljavanja i kumulativne funkcije rizika za desno cenzurirane podatke, Produkt-Limit i Nelson-Aalen procjenitelji. Kako postoji veza između osnovnih funkcija analize preživljavanja, pomoću ova dva procjenitelja možemo izraziti i ostale veličine. Nadalje, pokazano je kako se mogu odrediti pouzdani intervali i područja za ove procjenitelje. U zadnjem poglavlju se tehnike procjenjivanja primjenjuju na konkretnim podacima vezanim za oporavak nakon presađivanja koštane srži. Za obradu podataka je korišten programski jezik R.

Summary

This paper deals with the nonparametric estimation of basic quantities for right censored and left truncated survival data. Survival analysis is a set of methods for analysing time duration until one or more events happen. The basic quantities of survival analysis are the survival function, the hazard function, the cumulative hazard function, and the mean residual life function. They are consider in the first chapter. Before that, we present the basic definitions and terms of probability and statistics. At the end of the same chapter, we deal with the types of survival data. There are left / right censored data, left / right truncated data, interval censored data, and combination of them. It is important to know which type of dana we have, because depending on that, we choose the techniques for estimation.

In the second chapter we present the estimators of survival and cumulative hazard function for right censored data, Product-Limit and Nelson-Aalen estimators. Because of the ties between the basic quantities, with this two estimators we can express the other quantities, too. Afterwards, we show how to determine the confidence intervals and bands for this estimators. In the last chapter this tehniques are applied on the data about recovery after bone marrow transplantation, for what we used a programming language R.

Životopis

Rođena sam u Zagrebu 16.5.1991. godine. Pohađala sam OŠ Dragutina Domjanića (1998.-2006.g) i SŠ Dragutina Stražimira (2006.-2010.g), smjer opća gimnazija, u Svetom Ivanu Zelini. Prirodoslovno-matematički fakultet u Zagrebu, preddiplomski studij Matematika, upisujem 2010. godine. Titulu prvostupnika dobivam 2013.godine, nakon čega upisujem diplomski studij Financijska i poslovna matematika, također na Prirodoslovno-matematičkom fakultetu u Zagrebu.