

Peatükk 7

Elulemusanalüüs

Kui palju aastaid kellelgi veel elada on jäänud? Kas teistmoodi toitudes, teistmoodi patsiente ravides või vähem maailma asjade pärast muretse-des on võimalik oma eluiga pikendada? Sellistele küsimustele vastust otsib elulemus- või elukestvusanalüüs.

7.1 Elulemusfunktsioon

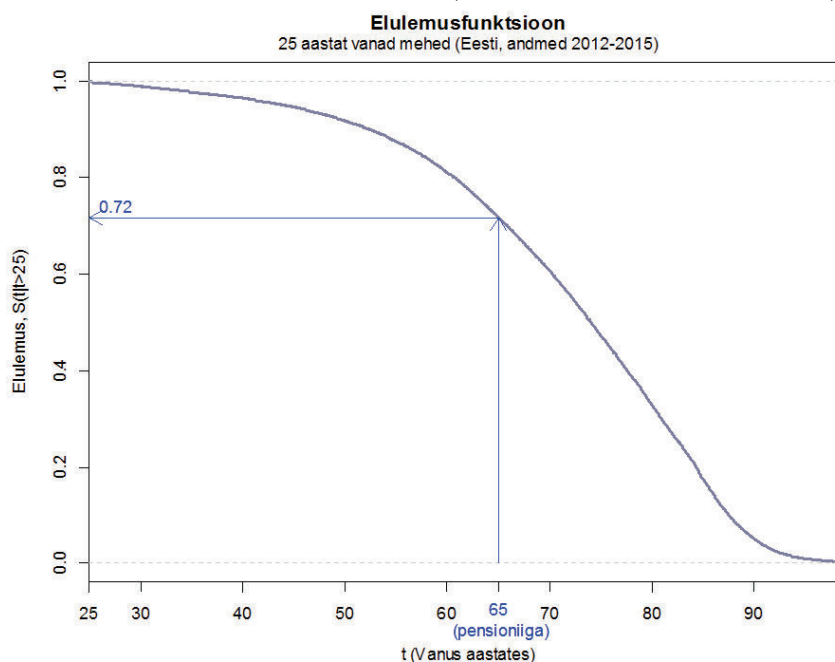
Kui kaua keegi elab? Inimese eluiga T on juhuslik suurus — mõned inimesed elavad kauem, mõned vähem. On kahtlane, et me kunagi suudaksime inimese tulevast eluiga päris täpselt prognoosida — mingisugune määramatus või teadmatus jääb ikka. Aga ka juhuslikke suuruste käitumist saab kirjeldada ja juhuslikke suuruseid saab võrrelda — mõnel viisil elavad inimesed püsivad keskmiselt kauem elus, mõnel viisil ravitud patsientidel on suurem tõenäosus veel kümme aastat elus püsida.

Eluea T jaotust kirjeldavad arstid elulemusfunktsiooni $S(t) := P(T > t)$ abil. Statistikud kasutavad sagedamini jaotusfunktsiooni $F(t) := P(T \leq t) = 1 - S(t)$, kuid alati asjade positiivsele küljele keskenduvad arstid peavad eetilisemaks rääkida elulemusfunktsioonist (elukestvuse jaotusfunktsioon näitab tõenäosust, et keegi on mingiks ajaks ära surnud, elulemusfunktsioon aga keskendub tõenäosusele veel elus olla selles vanuses).

Joonisel 7.1 on toodud 25 a vanuste meeste (\approx tudengi vanus ülikooli lõpetamise ajal) elulemusfunktsioon (Eesti andmed). Sellelt jooniselt näed ka seda, kuidas elulemusfunktsiooni graafikult leida tõenäosust elada pensionieani (72% 25 aasta vanustest Eesti meestest võiks elada kuni vanaduspensionieani — muidugi kui elukeskkonnas mingit suurt muutust ei tule, ei tule sõda ega leiutata mingit uut imerohtu).

Märkus: sageli ei vaadelda eluiga elulemusanalüüsis mitte sünnist alates. Paljude raskete haiguste puhul huvitatakse pigem sellest, kui kaua patsient peale diagnoosi elus püsis — kas ta on kolm või neli kuud peale diagnoosi ikka veel elus või mitte? Samuti võidakse mõnede ohtlike raviprotseduuride puhul huvi tunda selle vastu, kui paljud patsiendid on veel elus 3 päeva, 3 kuud või 3 aastat pärast raviprotseduuri (näiteks peale südame siirdamist).

Joonis 7.1: Elulemusfunktsioon (25 aasta vanused Eesti mehed)



Kui teame elukestvusfunktsiooni, võime leida palju teisigi kasulikke näitajaid. Näiteks seda, kui kaua keskmiselt eesti mehed saavad pensionipõlve nautida:

$$\int_{65}^{\infty} S(t) dt \approx 10,$$

ehk keskmiselt veedab eesti mees 10 aastat oma elust pensionärina.

Miks võisime oma keskmist pensioniloleku aja arvutust teha ülaltoodud viisil? Proovime järgnevalt tuletada ülaltoodud arvutuse tegemiseks vajalikud tulemused.

Esmalt vaatame, kuidas saame arvutada oodatavat tulevast eluiga, $E(T)$, elulemusfunktsiooni kasutades:

$$\begin{aligned} E(T) &= \int_0^{\infty} x f_T(x) dx \\ &= \int_0^{\infty} \int_0^{\infty} I(y < x) dy f_T(x) dx, \end{aligned}$$

sest $\int_0^{\infty} I(y < x) dy = \int_0^x 1 dy = x$.

Nüüd aga võime muuta integreerimise järjekorda:

$$\begin{aligned} E(T) &= \int_0^{\infty} \int_0^{\infty} I(y < x) f_T(x) dx, dy \\ &= \int_0^{\infty} \left(\int_0^y 0 \cdot f_T(x) dx + \int_y^{\infty} 1 \cdot f_T(x) dx \right) dy \\ &= \int_0^{\infty} (0 + F_T(\infty) - F_T(y)) dy \\ &= \int_0^{\infty} (1 - F_T(y)) dy \\ &= \int_0^{\infty} S(y) dy \end{aligned}$$

Oletame, et teame kellegi elulemusfunktsiooni. Milliseks muutub tema elulemusfunktsioon kui teame, et ta püsis elus t_0 aastat ehk milline on tinglik elulemusfunktsioon $S_{T|T>t_0}(t) := P(T > t | T > t_0)$? Selle leidmiseks võime kasutada tavalist tingliku tõenäosuse leidmise reeglit:

$$\begin{aligned} S_{T|T>t_0}(t) &= \frac{P(T > t \cap T > t_0)}{P(T > t_0)} \\ &= \begin{cases} \frac{S(t)}{S(t_0)}, & \text{kui } t > t_0 \\ 1, & \text{kui } t \leq t_0, \end{cases} \end{aligned}$$

sest kui $t < t_0$ siis $P(T > t \cap T > t_0) = P(T > t_0)$.

Kui soovime leida, kui kaua keegi vanuseni t_0 (näiteks pensioniiga) elanud inimesed veel keskmiselt elavad, $E(T - t_0 | T > t_0)$ (mitu aastat pensionär

keskmiselt enne surma pensioni saab) võime seda arvutada nii:

$$\begin{aligned}
 E(T - t_0 | T > t_0) &= E(T | T > t_0) - t_0 \\
 &= \int_0^\infty S_{T|T>t_0}(y) dy - t_0 \\
 &= \left(\int_0^{t_0} S_{T|T>t_0}(y) dy + \int_{t_0}^\infty S_{T|T>t_0}(y) dy \right) - t_0 \\
 &= \left(\int_0^{t_0} 1 dy + \frac{1}{S(t_0)} \int_{t_0}^\infty S(y) dy \right) - t_0 \\
 &= \frac{1}{S(t_0)} \int_{t_0}^\infty S(y) dy
 \end{aligned}$$

Kui t_0 tähistaks pensionile jäämise aega, siis saaksime ülaltoodud valemi abil leida, kui kaua saavad pensioni keskmiselt pensionärid. Aga mitte kõik inimesed ei ela pensionini. Mitu aastat saab aga keskmine 25.a vanune eesti mees pensionipõlve nautida? Siin saame kasutada valemit, mis seob keskväärtuse ja tinglikud keskväärtused:

$$\begin{aligned}
 E(\text{aeg penionil}) &= E(\text{aeg penionil} | \text{elab pensionini})P(\text{elab pensionini}) + \\
 &\quad + E(\text{aeg penionil} | \text{ei ela pensionini})P(\text{ei ela pensionini}) \\
 &= \frac{1}{S(t_0)} \int_{t_0}^\infty S(y) dy \cdot S(t_0) + 0 \cdot (1 - S(t_0)) \\
 &= \int_{t_0}^\infty S(y) dy.
 \end{aligned}$$

Miks me üritame keskmist eluiga (elada jäänud aastaid) leida elulemusfunktsiooni abil? Miks me tavalist lähenemist ei kasuta — valimi keskmist? Sellel on üks lihtne põhjus.

Nimelt on elulemusanalüüsi tegemiseks kasutatavate andmetel tavaliselt üks viga. Mõnedel patsientidel võib vahel vedada ja nad võivad üsna pikaks ajaks ellu jääda. Patsiendi õnn võib aga statistikule peavalu tekitada — me ei tea nende inimeste surmaaega (sest nad on analüüsi tegemise hetkel ikka veel elus). Seega on meie valimis osad kirjed/inimesed/patsiendid, kelle kohta uuritava tunnuse (eluiga) väärtus pole teada. Teame küll seda, kui kaua nad on elus püsinud — näiteks mõnel patsiendil diagnoositi vähk kümme aastat tagasi ja ta on ikka elus ja tegus. Kui kaua aga ta võiks veel tulevikus elada, seda me ei tea.

Vahel võime puuduvaid andmeid lihtsalt ignoreerida — võime teha näo, et oleme võtnud lihtsalt väiksema valimi. Paraku oleks selline käitumine kirjeldatud situatsioonis väga vale. Nimelt on tõenäolisemalt puudu nende inimeste/patsientide eluead, kes kauem elus püsivad. Mis juhtub aga valimi keskmisega, kui me suurimad väärtused valimist minema viskame? Selliselt kahjustatud valimi keskmine ei kirjeldaks ju enam adekvaatselt populatsiooni keskväärtust!

Kui aga puuduvate eluigade tõttu on valimiga kehvasti, siis miks me keskväärtuse hindamise ülesannet üldse lootusetuks ülesandeks ei lahterda? Sellel on lihtne põhjus – vahel, teatud eeldustel, saab elulemusfunktsiooni siiski korrektselt hinnata. Isegi siis, kui meil osad inimesed ikka veel elus püsivad. Seejärel on aga võimalik hinnatud elulemusfunktsiooni kasutada teiste asjade — näiteks keskmise eluea — leidmiseks.

Selleks, et mõista, kuna võiksime puuduvate eluigadega probleemiga hakkama saada (ja millal need puuduvad eluead ikkagi tõsiseks probleemiks võivad kujuneda) selleks peame veidi rääkima andmete tsenseerimisest.

7.2 Mitteinformatiivne tsenseerimine

Vaatlus on tsenseeritud (*censored*), kui teame, et uuritava tunnuse väärtus on suurem (või väiksem) mingist arvust, aga me ei tea kui palju suurem või väiksem. Näiteks kui mõõdame inimeste kaale kaaluga, mis suudab mõõta kaalu kuni 120kg. Selgub aga, et paaril inimese kaalud meie valimis on suuremad — nende kohta teame vaid, et nende kaal on rohkem kui 120kg — aga kui palju rohkem, seda me ei tea. Sellisel juhul ütleme, et nende inimeste kaalud on tsenseeritud.

Vahel on ka tsenseeritud andmete põhjal võimalik korrektselt hinnata meid huvitavaid suuruseid (nagu elulemusfunktsiooni), aga mitte alati. Kui tsenseerimine on mitteinformatiivne (*non-informative*), siis eksisteerib mitmeid erinevaid meetodeid elulemusfunktsiooni korrektseks hindamiseks. Kui aga on tegemist informatiivse tsenseerimisega siis oleme suurtes raskustes.

Millal on tsenseerimine mitteinformatiivne? Proovime seda alljärgnevalt selgitada.

Kasutame järgmiseid tähistusi:

T - eluiga (tegelik, mittetsenseeritud. juhuslik suurus — me ei tea ette, kuna keegi sureb)

C - tsenseerimisaeg (kui keegi püsiks igavesti elus, siis millal ta tsenseeritakse — kas tänu sellele, et uurijatel lõppeb kannatus ehk uuringuks etenähtud aeg saaks läbi või selle pärast, et inimene kolib välismaale ja Eesti

arstid ei tea peale patsiendi emigreerumist mis temast edasi sai vms).

Kui $T \leq C$ siis teame täpselt, kuna inimene suri — me näeme inimese surmaaega enne kui uuring läbi saab või enne kui ta jõuab välismaale kolida ja meie jälgimise alt kaduda.

Kui $T > C$ siis me inimese surmaaega ei tea — inimene kaob uurijate vaateväljast enne oma surma (näiteks sellepärast, et uuring sai läbi). Teisisõne, inimese eluiga tsenseeritakse. Teame vaid, et ta on elus püsinud vähemalt C aastat.

Kui nüüd $\forall t_0, t$ korral kehtib tingimus

$$P(T > t | T > t_0, C = t_0) = P(T > t | T > t_0, C > t_0) = P(T > t | T > t_0)$$

siis ütleme, et tsenseerimine on mitteinformatiivne. Mitteinformatiivne tsenseerimine tähendab seega, et kui mõned inimesed on püsinud elus t_0 aastat, ja kui osad neist tsenseeritakse siis kui nad t_0 aasta vanuseks saavad ja teisi ei tsenseerita, siis mõlemas grupis on tulevaste eluigade jaotus sama. Ehk sisuliselt: inimese tsenseerimine/mittetsenseerimine ei tohi sõltuda inimese tervislikust seisundist, küll aga võib sõltuda patsiendi tervisest sõltumatust asjaoludest (näiteks sellest, et uuringuks ettenähtud aeg sai läbi või sellest, et patsiendi perearst läks pensionile ja tema uus perearst ei olnud huvitatud koostööst uuringu korraldajatega).

Keerukamate elulemusanalüüsi mudelite puhul võime tingimusele lisada veel teisi X -tunnuseid (näiteks patsiendi sugu) ja siis kasutatakse mitteinformatiivse tsenseerimise defineerimisel tinglikku sõltumatust. Kui

$$P(T > t | T > t_0, X, C = t_0) = P(T > t | T > t_0, X, C > t_0)$$

siis on tegemist mitteinformatiivse tsenseerimisega, kui aga antud tingimus pole täidetud, siis räägime informatiivsest tsenseerimisest.

7.3 Kaplan-Meieri hinnang elulemusfunktsioonile

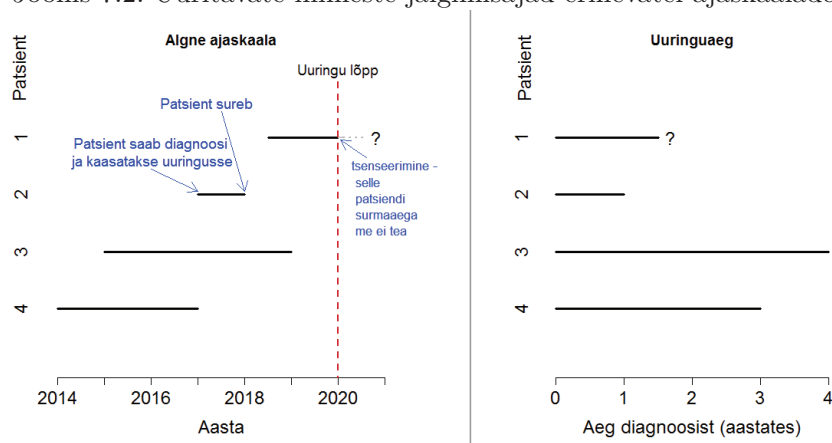
Kaplan Meieri hinnang kasutab ära seda, et ka tsenseeritud eluigade korral saame me ikkagi hinnata tinglikke üleelamistõenäosuseid — kui keegi ona elus püsinud kuni ajahetkeni t_i , siis milline on tema tõenäosus elada üle ka ajahetk t_i seda saame hinnata tavalise suhtelise sageduse abil (kui suur osa nendest inimestest kes on elus ja jälgimise all vahetult enne ajahetke t_i jäävad ellu ka pärast ajahetke t_i). Leitud tinglikke tõenäosuste hinnanguid kasutades on aga võimalik konstrueerida elulemusfunktsiooni hinnang:

$$\hat{P}(T > t_i) = \hat{P}(T \geq t_i) \cdot \hat{P}(T > t_i | T \geq t_i).$$

Vaatame Kaplan-Meieri hinnangu leidmist alljärgnealt veidi detailsemalt.

Kõigepealt tuleks valida analüüsiks sobiv ajaskaala (aeg sünnist ehk vanus või aeg diagnoosist või aeg operatsioonist vms). Vaata ka joonist 7.2.

Joonis 7.2: Uuritavate inimeste jälgimisajad erinevatel ajaskaaladel



Seejärel järjestame jälgimisajad (aeg meid huvitavast sündmusest patsiendi surma või tsenseerimiseni):

$$t_{(1)} \leq t_{(2)} \leq \dots \leq t_{(n)}.$$

Olgu $t_{(j)}$ väikseim jälgimisaeg, mis lõpeb tsenseerimisega. Kuni ajani $t_{(j)}$ võime elulemusfunktsiooni hinnata tavapärasel viisil — kasutades suhtelist sagedust. Sest me teame iga ajahetke $t < t_{(j)}$ jaoks kui suur osa uuringusse kaasatututest elasid selle ajahetke üle ja kui paljud ei elanud sellest ajahetkest kauem. Kui vaadata joonisel ?? kujutatud nädisandmestikku, siis saame näiteks kenasti hinnata elulemusfunktsiooni väärtust kohal $\hat{S}(2) = 3/4$, sest me teame, et 3 patsienti neljast elas kauem kui 2 aastat peale diagnoosi. Kuidas leida aga elulemusfunktsiooni väärtust kohal 3, $S(3)$? Selleks kasutame valemit

$$S(3) = P(T > 3) = P(T > 3 | T > 2) \cdot P(T > 2) = P(T > 3 | T > 2) \cdot S(2).$$

Elulemusfunktsiooni hinnangut kohal 2 me juba teame, aga tinglikku tõenäosust $P(T > 3 | T > 2)$ võime hinnata mitteinformatiivse tsenseerimise

korral järgmisel moel:

$$\begin{aligned} P(T > 3 | T > 2) &= P(T > 3 | T > 2, C > 2) \\ &\approx 1/2 \end{aligned}$$

Sest neid, kes elasid vähemalt kaks aastat peale diagnoosi saamist ja keda ei tsenseeritud aastal 2 või varem on meie uuringus 2 tükki ja neist ainult üks oli elus peale kolme aasta möödumist. Seega elulemusfunktsiooni hinnang kohal 3 on $\hat{S}(3) = 3/4 \cdot 1/2 = 3/8$.

Jätkates samal moel võime leida elulemusfunktsiooni hinnangud kuni viimase teadaoleva surmahetkeni. Kui kõige pikem jälgimisaeg lõppeb surmaga, siis jõuab elulemusfunktsiooni hinanng nulli ja me oleme leidnud elulemusfunktsiooni hinnangu kõigi ajahetkede t jaoks (sest elulemusfunktsioon saab vaid kahaneda, aga tõenäosus ei saa negatiivseks muutuda). Kui aga kõige pikem jälgimisaeg on tsenseeritud (me ei tea täpselt, kuna kõige kauem jälgimise all olnud inimene elas), siis suudame rekonstrueerida elulemusfunktsiooni hinnangu vaid kõigi nende ajahetkede jaoks, kus vähemalt üks patsient veel jälgimise all oli.

Üldine valem elulemusfunktsiooni hinnangu leidmiseks Kaplan-Meieri meetodil on kirja pandav järgmise valemi abil:

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

kus \hat{p}_j on hinnanguline tõenäosus elada üle ajahetk $t_{(j)}$.

Näide

Olgu patsientide surmaaajad järgmised (+ tähistab tsenseeritud vaatlust): 1; 2; 4+; 6; 7+; 7+; 9; 9; 11; 11+; 15. Kaplan-Meieri hinnang elulemusfunktsioonile ja vahetulemused on ära toodud tabelis 7.1 (d_i on ajahetkel $t_{(i)}$ aset leidnud surmade arv ja n_i tähistab samal ajal jälgimise all olevate inimeste arvu).

7.4 Usaldusintervall elulemusfunktsioonile

Viisakas on hinnangute täpsust kirjeldada — näiteks usaldusintervalli abil. Kuidas leida usaldusintervalli elulemusfunktsioonile? Selleks peame mõistma, milline on meid huvitava hinnangu jaotus.

Tabel 7.1: Kaplan-Meieri hinnang elulemusfunktsioonile, näide.

$t_{(i)}$	surm või tsenseerimine	d_i	n_i	\hat{p}_i	$\hat{S}(t_{(i)})$
1	surm	1	11	10/11	10/11
2	surm	1	10	9/10	$9/10 \cdot 10/11 = 9/11$
4	tsens.	0	9	9/9	$1 \cdot 9/11 = 9/11$
6	surm	1	8	7/8	$7/8 \cdot 9/11 = 63/88$
7	tsens.; tsens.	0	7	7/7	$1 \cdot 63/88 = 63/88$
9	surm; surm	2	5	3/5	$3/5 \cdot 63/88 = 189/440$
11	surm; tsens.	1	3	2/3	$2/3 \cdot 189/440 = 126/440$
15	surm	1	1	0	0

Leiame esmalt elulemusfunktsiooni hinanngu logaritmi jaotuse.

$$\begin{aligned} \log(\hat{S}(t)) &= \log\left(\prod_{i:t_{(i)} \leq t} \hat{p}_i\right) \\ &= \sum_{i:t_{(i)} \leq t} \log(\hat{p}_i). \end{aligned}$$

Aga

$$\log(\hat{p}_i) \underset{approx.}{\sim} N\left(\log(p_i), \frac{1-p_i}{n_i p_i}\right)$$

ja seega (ligikaudu)

$$\log(\hat{S}(t)) \sim N\left(\sum_{i:t_{(i)} \leq t} \log(p_i); \sum_{i:t_{(i)} \leq t} \frac{1-p_i}{n_i p_i}\right)$$

Aga

$$\begin{aligned} \sum_{i:t_{(i)} \leq t} \frac{1-p_i}{n_i p_i} &\approx \sum_{i:t_{(i)} \leq t} \frac{d_i/n_i}{n_i(1-d_i/n_i)} \\ &= \sum_{i:t_{(i)} \leq t} \frac{d_i}{n_i(n_i-d_i)} \end{aligned}$$

ja seega, ligikaudu,

$$\log(\hat{S}(t)) \sim N \left(\log(S(t)); \sum_{i:t_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)} \right).$$

Saadud tulemusest saame delta meetodi abil, et (ligikaudu):

$$\hat{S}(t) \sim N \left(S(t); \sum_{i:t_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)} (S(t))^2 \right)$$

Kust saame ligikaudse arvutusvalemi 95%-usaldusintervalli leidmiseks $S(t)$ -le:

$$\hat{S}(t) \pm 1,96 \cdot \hat{S}(t) \sqrt{\sum_{i:t_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}}.$$

Saadud valemit tuntakse ka Greenwoodi valemi nime all.

Eksisteerib ka teisi võimalusi, kuidas konstrueerida ligikaudseid usalduspiire elulemusfunktsioonile. Näiteks võib Greenwoodi valem produtseerida usalduspiire mis on kas 0-st väiksemad või 1-st suuremad. Sestap konstrueeritakse vahel usalduspiirid suurusele $L(t) = \ln(-\ln(S(t)))$. Kui me transformeerime saadud usalduspiire tagasi elulemusfunktsiooni usalduspiirideks kasutades valemit $S(t) = \exp(-\exp(L(t)))$ siis võime olla kindlad, et saadud usalduspiirid jäävad alati vahemikku $0 \dots 1$.

7.5 Kaplan-Meieri hinnangu leidmine R-i abil

Kaplan-Meieri hinnangut saame R-is leida lisamooduli survival abil. Alljärgnevalt kordame tabelis 7.1 tehtud arvutusi R-i abil:

```
library(survival)
```

```
eluiga=c(1, 2, 4, 6, 7, 7, 9, 9, 11, 11, 15)
surm = c(1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1)
```

```
model=survfit(Surv(eluiga, surm)~1)
```

```
summary(mudel)
plot(mudel)
```

Ülesanded

1. Intensiivravi osakonna arst märgib üles oma patsientide käekäiku peale operatsiooni. Mõned patsiendid surevad internsiivraviosakonnas, mõned aga liiguvad edasi tavapalatisse. Mis neist tavapalatis saab — kuni nad surevad ja kas surevad — seda intensiivravi osakonna arst ei tea, nende eluead on tsenseeritud. Kas antud juhul on tegemist mitteinformatiivse või informatiivse tsenseerimisega? Põhjenda oma otsust!
2. Peale kroonilise haiguse diagnoosi saamist elasid patsiendid veel niimitu aastat: 5; 9+; 12; 12; 14+; 18+; 18; 19; 22; 25.

Kõigi patsientide surmaajad pole teada. Eluead, kus numbrile järgneb „+“-sümbol tähistavad eluigasid (peale diagnoosi saamist eluspüstitud aega), mil patsient veel kindlasti elus oli. Kui kaua „+“-märgiga tähistatud inimesed veel võisid pärast märgitud aega elada — seda me ei tea.

- a. Leia Kaplan-Meieri meetodil elukestvusfunktsiooni $S(t)$ hinnang!
- b. Leia ise rehkendades 95%-usaldusintervall Kaplan-Meieri hinnangule punktis $t = 5$ (näita arvutusi)!
- c. Leia keskmine oodatav eluiga peale diagnoosi $E(T)$!
- d. Leia hinnang suurusele $E(T|T > 20)$ — mitu aastat keskmiselt võiks veel elada krooniline haige, kes on elus püsinud 20 aastat peale diagnoosi saamist!