# 3. Procjena distribucije vremena života

## 3.1 Uvod

Želimo procijeniti $F(t)$ od $T$
$\rightarrow$ *neparametarska procjena*!

"*Najjednostavnija metoda*": opaža se vrlo velik broj novorođenih osoba do njihove smrti.

*Problemi s "najjednostavnijom" metodom*:

1. ako i nađemo dovoljno velik uzorak, trebali bi ga opažati barem 100 godina
2. osobe mogu napustiti opažanje i prije eventualne smrti (zbog raznih razloga)

$\rightarrow$ *cenzurirani podaci*

## 3.2 Mehanizmi cenzuriranja

Vrste cenzuriranja:

1. desno cenzuriranje
2. lijevo cenzuriranje
3. intervalno cenzuriranje
4. slučajno cenzuriranje
5. neinformativno cenzuriranje
6. cenzuriranje tipa I
7. cenzuriranje tipa II

## 3.3 Kaplan - Meierov procjenitelj

## **3.3.2** Kaplan-Meierova procjena

$$\widehat{F}(t) = 1 - \prod_{t_j \le t} (1 - \widehat{\lambda}_j)$$

gdje je $\widehat{\lambda}_j = d_j/n_j$, $j = 1, \ldots, k$

Ovdje su;

$t_j$... vrijeme smrti

$d_j$... broj umrlih u vremenu $t_j$

$n_j$... broj živih i pod rizikom od smrti neposredno prije $t_j$ (tj. do $t_j-$)

## 3.4 Usporedba razdioba vremena života

Greenwoodova formula:

$$\mathrm{Var}[\widehat{F}(t)] \approx (1 - \widehat{F}(t))^2 \sum_{t_j \le t} \frac{d_j}{n_j(n_j - d_j)}$$

## 3.3 Nelson - Aalenov procjenitelj

Procjena integriranog hazarda:

$$\Lambda_t = \int\limits_0^t \mu_s \, ds + \sum_{t_j \leq t} \lambda_j$$

### **3.3.2** Nelson - Aalenova procjena

$$\widehat{\Lambda}_t = \sum_{t_j \leq t} \frac{d_j}{n_j}$$

$$\mathsf{Var}[\widehat{\Lambda}_t] \approx \sum_{t_j \leq t} \frac{n_j - d_j}{n_j^3}$$

**Primjer.** U tablici se nalaze podaci o 12-torici pacijenata koji su bili podvrgnuti nekoj operaciji.

| pacijent | vrijeme op. (tjedni) | opažanje prestalo (u tjednu) | raz. za prestanak |
|---|---|---|---|
| 1 | 0 | 120 | cenzurirano |
| 2 | 0 | 68 | smrt |
| 3 | 0 | 40 | smrt |
| 4 | 4 | 120 | cenzurirano |
| 5 | 5 | 35 | cenzurirano |
| 6 | 10 | 40 | smrt |
| 7 | 20 | 120 | cenzurirano |
| 8 | 44 | 115 | smrt |
| 9 | 50 | 90 | smrt |
| 10 | 63 | 98 | smrt |
| 11 | 70 | 120 | smrt |
| 12 | 80 | 110 | smrt |

Pretpostavljamo: cenzuriranje je neinformativno.

$(i)$ Procijenite funkciju doživljenja pomoću Kaplan - Meierove metode i odredite 95% p.i. za njene vrijednosti.

$(ii)$ Procijenite integrirani hazard Nelson - Aalenovom procjenom i pomoću njega procijenite funkciju doživljenja.

$(iii)$ Procijenite vjerojatnost da će pacijent doživjeti 70-ti tjedan nakon operacije.

*Rješenje*:

$(i)$: Razdoblje opažanja počinje u trenutku $t = 0$ odmah nakon operacije. Opažena relativna vremena su redom (u tjednima, '+' označava cenzuriranje):

$$120+, 68, 40, 116+, 30+, 30, 100+, 71, 40, 35, 50, 30$$

Statistike:

| $t_j$ | $n_j$ | $d_j$ | $\widehat{\lambda}_j = \frac{d_j}{n_j}$ | $1 - \widehat{\lambda}_j$ | $\prod(1 - \widehat{\lambda}_j)$ |
|---|---|---|---|---|---|
| 30 | 12 | 2 | 1/6 | 5/6 | 0.8333 |
| 35 | 9 | 1 | 1/9 | 8/9 | 0.7407 |
| 40 | 8 | 2 | 1/4 | 3/4 | 0.5556 |
| 50 | 6 | 1 | 1/6 | 5/6 | 0.4630 |
| 68 | 5 | 1 | 1/5 | 4/5 | 0.3703 |
| 71 | 4 | 1 | 1/4 | 3/4 | 0.2778 |

Kaplan-Meierova procjena $\widehat{S}(t)$ za $S(t)$:

| $t_{j-1} \leq t < t_j$ | $\widehat{S}(t) = \prod_{t_j \leq t}(1 - \hat{\lambda}_j)$ | $\pm 2 \cdot \text{s.e.}(\widehat{S}(t))$ |
|---|---|---|
| $0 \leq t < 30$ | 1 | $\pm 0$ |
| $30 \leq t < 35$ | 0.8333 | $\pm 0.2154$ |
| $35 \leq t < 40$ | 0.7407 | $\pm 0.2592$ |
| $40 \leq t < 50$ | 0.5556 | $\pm 0.2987$ |
| $50 \leq t < 68$ | 0.4630 | $\pm 0.3010$ |
| $68 \leq t < 71$ | 0.3703 | $\pm 0.2922$ |
| $71 \leq t$ | 0.2778 | $\pm 0.2717$ |

gdje je:

$$\text{s.e.}(\widehat{S}(t)) = \widehat{S}(t) \sqrt{\sum_{t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}}$$

$(ii)$: Nelson - Aalenova procjena $\tilde{S}(t)$ za $S(t)$:

| $t_{j-1} \leq t < t_j$ | $\widehat{\Lambda}(t) = \sum_{t_j \leq t} \widehat{\lambda}_j$ | $\tilde{S}(t) = e^{-\widehat{\Lambda}(t)}$ |
|---|---|---|
| $0 \leq t < 30$ | 0 | 1 |
| $30 \leq t < 35$ | 0.1667 | 0.8465 |
| $35 \leq t < 40$ | 0.2778 | 0.7575 |
| $40 \leq t < 50$ | 0.5278 | 0.5899 |
| $50 \leq t < 68$ | 0.6944 | 0.4994 |
| $68 \leq t < 71$ | 0.8944 | 0.4088 |
| $71 \leq t$ | 1.1444 | 0.3184 |

$(iii)$: $\widehat{S}(70) = 0.3703$

# 4. Coxov regresijski model

## **4.1** Ovisnost o varijablama (kovarijatama)

## **4.2** Potpuno parametarsko modeliranje

Pretpostavka: distribucija vremena života pripada nekoj parametarskoj familiji (npr. eksponencijalna, Weibullova, Gompertz-Makehamova,...)

$\rightarrow$ populacije se uspoređuju usporedbom parametara

$\rightarrow$ može se primijeniti ako postoji apriorno znanje o pripadnosti određenoj param. familiji

**4.3** Coxov poluparametarski model:

$$\lambda(t; z) = \lambda_0(t) \exp(\beta z^T)$$

gdje su:

– $\lambda_0(t)$... osnovni hazard

– $\beta$... $1 \times p$-vektor regresijskih parametara

– $z$... $1 \times p$-vektor $p$-dim, kovarijate

*Osnovna pretpostavka Coxovog modela*:

Omjeri hazarda različitih osoba s vrijednostima kovarijata $z_1$, $z_2$ ne ovise o $t$, tj.

$$\frac{\lambda(t; z_1)}{\lambda(t; z_1)} = \exp(\beta(z_1^\tau - z_2^\tau))$$

$\rightarrow$ model *proporcionalnih* hazarda

*Upotreba Coxovog modela*:

– ako nas primarno zanima istraživanje *efekta* od $z$ na $\lambda(t; z)$, a ne sama funkcija doživljenja (tj. $\lambda$)

$\rightarrow$ poluparametarski pristup

## 4.4 Parcijalna vjerodostojnost

Pretpostavka: $d_j = 1$ za svaki $j$

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp(\beta z_j^{\mathcal{T}})}{\sum_{i \notin R(t_j)} \exp(\beta z_i^{\mathcal{T}})}$$

gdje je $R(t_j)$ skup svih osoba $i$ živih i pod rizikom do $t_j-$

$\rightarrow$ procjena od $\beta$ (numeričkom) maksimizacijom $L(\beta)$

U praksi se može dogoditi:

$(a)$ $d_j > 0$
$(b)$ postoje cenzurirana opažanja u nekom od $t_j$

**Ad** $(a)$: Breslowova aproksimacija:

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp(\beta s_j^\tau)}{\left(\sum_{i \notin R(t_j)} \exp(\beta z_i^\tau)\right)^{d_j}}$$

gdje je $s_j$ zbroj svih $z$ vrijednosti osoba umlih u $t_j$

**Ad** $(b)$: Osobe cenzurirane u $t_j$ su uključene u skup $R(t_j)$.

**Primjer.** U tablici se nalaze podaci malog uzorka zaposlenih u jednoj tvornici koji predstavljaju vremena u mjesecima do prvog izostajanja s posla. Sa "+" su označena vremena napuštanja posla koja se ne smatraju izostajanjem.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| M: | 6+ | 11 | 13+ | 15 | 16+ | 19+ | 20 | |
| F: | 2+ | 4 | 7 | 8+ | 10+ | 12+ | 17 | 21+ |

Stopa hazarda se modelira po Coxovom modelu:

$$\lambda(t) = \lambda_0(t)e^{\beta x}$$

gdje su:
$x = 0$ za muškarce ($M$) i
$x = 1$ za žene ($F$).

($i$) Odredite parcijalnu vjerodostojnost od $\beta$ i MLE.
($ii$) Procijenite asimptotski 95% p.i. za $\beta$.
($iii$) Testirajte hipotezu da žene imaju veću stopu izostanka s posla od muškaraca.

$(i)$: $t_j$: 4$F$, 7$F$, 11$M$, 15$M$, 17$F$, 20$M$

Parcijalna vjerodostojnst:

$$
\begin{aligned}
& L(\beta) = \\
= \ & \prod_j \frac{e^{\beta x_j}}{\sum_{i \in R(t_j)} e^{\beta x_i}} = \\
= \ & \frac{e^{\beta(1+1+0+0+1+0)}}{(7e^{\beta \cdot 1}+7)(6e^{\beta \cdot 1}+6)(3e^{\beta \cdot 1}+6)(2e^{\beta \cdot 1}+4)(2e^{\beta \cdot 1}+2)(e^{\beta \cdot 1}+1)} \\
= \ & \frac{e^{3\beta}}{7 \cdot 6 \cdot 3 \cdot 2^2 (e^{\beta}+1)^4 (e^{\beta}+2)^2}
\end{aligned}
$$

Parcijalna log-vjerodostojnost:

$$\ell(\beta) = \log L(\beta) = 3\beta - 4\log(e^{\beta}+1) - 2\log(e^{\beta}+2) + \text{const.}$$

$$\begin{aligned} & \ell'(\beta) = 3 - \tfrac{4e^{\beta}}{1+e^{\beta}} - \tfrac{2e^{\beta}}{2+e^{\beta}} = 0 \\ \Leftrightarrow \quad & 3e^{2\beta} + e^{\beta} - 6 = 0 \\ \Leftrightarrow \quad & e^{\beta} = \tfrac{\sqrt{73}-1}{6} = 1.2573 \\ \Rightarrow \quad & \widehat{\beta} = 0.2290 \ \ \text{(MLE)} \end{aligned}$$

$(ii)$:

$$\ell''(\beta) = -\frac{4e^{\beta}}{(1+e^{\beta})^2} - \frac{4e^{\beta}}{(2+e^{\beta})^2} \;\Rightarrow\; \ell''(\widehat{\beta}) = -1.4610$$

Asimptotska razdioba od MLE:

$$\widehat{\beta} - \beta \,\dot{\sim}\, N\left(0, -\frac{1}{\ell''(\widehat{\beta})}\right) \equiv N(0, 0.8273^2)$$

95% p.i. za $\beta$:

$$\widehat{\beta} \pm 2 \cdot 0.8273 = 0.2290 \pm 1.6546$$

$(iii)$: Treba testirati hipoteze:

$$H_0: \ \ \beta = 0 \text{ vs. } H_1: \ \ \beta > 0$$

Testna statistika:

$$Z = \frac{\widehat{\beta} - 0}{\sqrt{-\ell''(0)^{-1}}} \overset{H_0}{\dot{\sim}} N(0,1)$$

Kritično područje uz $\alpha = 5\%$: $[1.64, +\infty\rangle$

$$z = \frac{0.2290}{0.8321} = 0.2752 \notin [1.64, +\infty\rangle$$

Žene nemaju značajno veću stopu izostanka s posla od muškaraca.

# 5. Markovljev proces s dva stanja

## 5.1 Pretpostavke

$$\boxed{a} \xrightarrow{\quad\mu_x\quad} \boxed{d}$$

${}_tp_x :=$ (uvjetna) vjerojatnost da osoba (proces) ostane u stanju $a$ u dobi $x+t$ ako je bila u tom stanju u dobi $x$

${}_tq_x :=$ (uvjetna) vjerojatnost da osoba (proces) bude u stanje $d$ u dobi $x+t$ ako je bila u stanju $a$ u dobi $x$

**Pretpostavka 1.**
Vjerojatnost da će se osoba dobi $x$ naći u određenom stanju u dobi $x+t$ ovisi samo o dobi $x$ i stanju u kojem se u toj dobi nalazi.

**Pretpostavka 2.**
${}_hq_{x+t} = \mu_{x+t} \cdot h + o(h)$, $h \to$ 0+, za $t \geq$ 0.

$\rightarrow$ u cilju procjene modela, restrikcija na dobni interval $[x, x+1\rangle$

**Pretpostavka 3.**
$t \mapsto \mu_{x+t} = \mu =$ konst. za $0 \leq t < 1$.

Sve ostalo (npr. stanje zdravlja, zanimanje,...) nije uključeno u model.

Znajući te faktore, možemo:

$(a)$ *stratificirati* problem, tj. za svaku kombinaciju razina faktora postaviti model
$(b)$ postaviti adekvatan *regresijski model* s tim faktorima kao kovarijatama

## **5.2** Vjerojatnosti prijelaza

$$ {}_{t+h}p_x = \cdots = {}_tp_x(1 - \mu_{x+t}h) + o(h),\ h \to 0+ $$

$\Rightarrow$

$$ {}_tp_x = \exp(-\int_0^t \mu_{x+s}\, ds) $$

## **5.3** Statistike

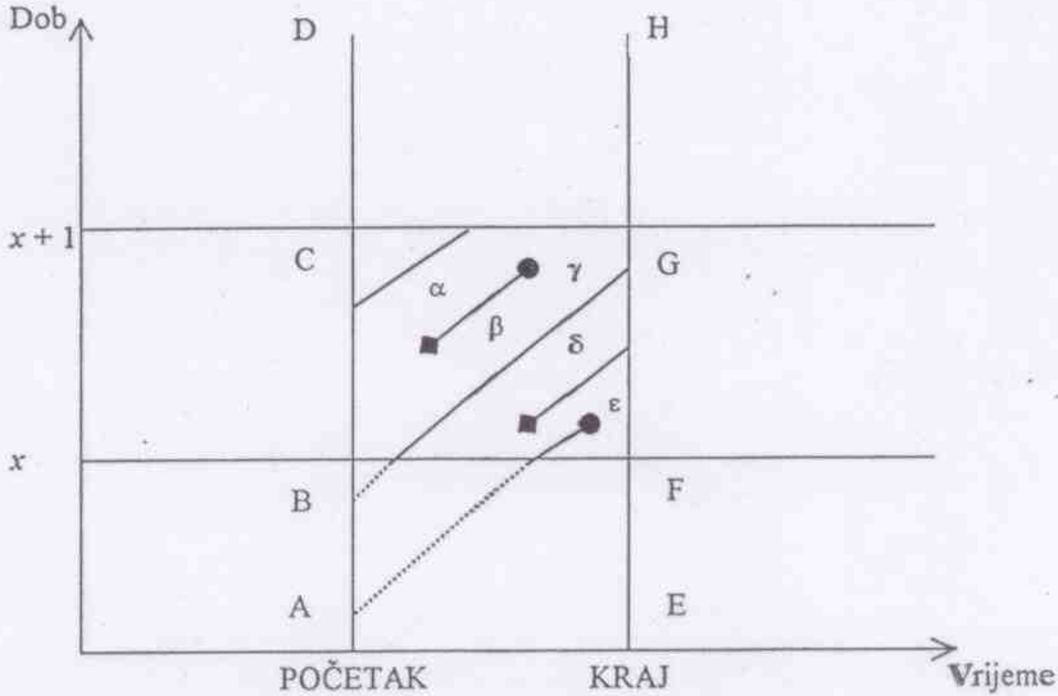
Opažamo $N$ osoba u dobi od $x$ do $x+1$
(pretp. da je $\mu_x \equiv \mu$ konst.)

$\rightarrow$ niz nezavisnih i jednako distribuiranih jedinki

$\rightarrow$ cenzuriranje tipa I

$x + a_i$ ... dob osobe $i$ na početku razdoblja opažanja ili ulaska u grupu

$x + b_i$ ... dob osobe $i$ na kraju razdoblja opažanja ako doživi tu dob




SLIKA 2

Slučajne varijable (statistike):

$D_i$ ... indikator smrti ($D_i = 1$ ako je opažena smrt osobe $i$, inače $D_i = 0$)

$x + T_i$ ... dob osobe $i$ u trenutku prestanka opažanja te osobe

$V_i := T_i - a_i$ ... ukupno vrijeme opažanja osobe $i$

$\rightarrow$ dovoljna statistika za osobu $i$: $(D_i, V_i)$

Opažena vrijednost od $(D_i, V_i)$: $(d_i, v_i)$

Vjerodostojnost od $\mu$:

$$L(\mu) = \prod_{i=1}^{N} e^{-v_i \mu} \mu^{d_i} = e^{-\mu \sum_{i=1}^{N} v_i} \cdot \mu^{\sum_{i=1}^{N} d_i} = e^{-\mu v} \mu^{d}$$

gdje su $v = \sum_{i=1}^{N} v_i$ i $d = \sum_{i=1}^{N} d_i$
opažene vrijednosti statistika:

$D = \sum_{i=1}^{N} D_i$... ukupan broj umrlih u uzorku u dobi od $x$ do $x + 1$

$V = \sum_{i=1}^{N} V_i$... ukupno vrijeme boravka opažanih osoba iz uzorka u dobnom intervalu $[x, x + 1]$

## **5.4** Procjena maksimalne vjerodostojnosti

Log-vjerodostojnost:

$$\ell(\mu) = \log L(\mu) = -\mu v + d \log \mu$$

$$\ell'(\mu) = -v + \frac{d}{\mu} = 0 \quad \Leftrightarrow \quad \widehat{\mu} = \frac{d}{v}$$

Procjenitelj MLE:

$$\tilde{\mu} = \frac{D}{V}$$

Asimptotska razdioba MLE:

$$\ell''(\mu) = -\frac{d}{\mu^2} \quad \Rightarrow \quad -\frac{1}{\ell''(\widehat{\mu})} = \frac{\widehat{\mu}}{v}$$

$$\tilde{\mu} - \mu \mathrel{\dot{\sim}} N\left(0, -\frac{1}{\ell''(\tilde{\mu})}\right) \equiv N\left(0, \frac{\tilde{\mu}}{V}\right)$$

standardna pogreška: s.e.$(\tilde{\mu}) = \sqrt{\frac{\tilde{\mu}}{V}}$

95% p.i. za $\mu$: $\tilde{\mu} \pm 2 \cdot \sqrt{\frac{\tilde{\mu}}{V}}$