11 Life data analysis

11.1 Introduction

Det som är speciellt med livslängdsdata är att man alltsomoftast inte har möjlighet att vänta tills alla enheter man testar har havererat, så man har behov av att kunna göra analyser av ofullständiga eller, som man säger, censurerade datamängder.

Vi antar att n komponenter testas och skriver T_i för komponent i:s livstid (funktionstid) och antar att $T_1, T_2, \ldots, T_n$ är oberoende och fördelade som T. Vi låter $F(t)$ vara T:s fördelningsfunktion, $R(t) = 1 - F(t)$ vara motsv överlevnadsfunktion och $z(t) = f(t)/R(t)$, där $f(t) = F'(t)$, vara felbenägenheten.

Ofta är det första man är intresserad av huruvida den aktuella fördelningen är IFR, DFR eller ev har något annan typ av felbenägenhet. Detta görs ofta med en s.k TTT-plott. TTT-plotten är en parameterfri metod. Man gör alltså inga fördelningsantaganden. Andra parameterfria metoder vi ska lära oss i denna kursen är Kaplan-Meiers estimator av överlevnadsfunktionen $R(t)$ och Nelsons estimator av $Z(t) = \int_0^t z(s)\,ds$. De två sistnämnda tittar vi bara i samband med övningsräkning.

11.2 Complete and censored data sets

11.2.1 Complete data set

Vi låter $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$ beteckna det ordnade stickprovet.

Det är praktiskt att införa $T_{(0)} = 0$ och $T_{(n+1)} = \infty$.

11.2.2 Type I censoring

Man avbryter vid en fix tidpunkt t_0. Antag att s enheter har fallerat, d.v.s

$$0 = T_{(0)} \leq T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(s)} \leq t_0 < T_{(s+1)}$$

Om de $n - s$ återstående vet man att de har fungerat i t_0 tidsenheter.

11.2.3 Type II censoring

Man avbryter då r komponenter har fallerat. Då känner man tidpunkterna

$$0 = T_{(0)} \leq T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(r)}$$

och vet dessutom att de $n - r$ återstående komponenterna har fungerat i $T_{(r)}$ tidsenheter.

11.2.4 Type III censoring

Man avbryter vid tidpunkten $\min(T_{(r)}, t_0)$.

11.2.5 Type IV censoring

Typ 4-censurering får man typiskt då de n enheterna ej sätts igång samtidigt. Här har man stokastiska censureringstider $S_1, \ldots, S_n$. Denna typ av censurering har man framförallt i medicinska sammanhang.

11.3 Nonparametric methods

11.3.2 Sample measures

11.3.3 The empirical distribution and survivor functions

11.3.4 Probability plotting paper

11.3.5 The Kaplan-Meyer estimator

Gås igenom i samband med övningsräkning.

11.3.6 Nelson's estimator of the cumulative failure rate

Gås igenom i samband med övningsräkning.

11.3.7 Total time on test plot for complete data sets

Vi ska anta att T är kontinuerligt fördelad, att $F(t)$ är strikt växande, att $\mu = \int_0^\infty R(t)\,dt < \infty$, samt att

$$F^{-1}(0) = 0 < t < F^{-1}(1)$$

I det sista villkoret innebär den vänstra likheten att $F(0) = 0$ och den högra olikheten att $F(t) < 1$. (F^{-1} existerar ju och är även den strängt växande. Tänk igenom vad det skulle innebära om $F(t)$ ej vore strikt växande.)

Definition 11.1 Den totala testtiden vid tidpunkten $t > 0$ är

$$\mathcal{T}(t) = \sum_{j=1}^{i} T_{(j)} + (n-i)t$$

där i ges av att

$$T_{(i)} \le t < T_{(i+1)}$$

(i är alltså antalet komponenter som fallerat vid tidpunkten t). □

Vi förutsätter att alla n komponenterna som testas sätts igång då $t = 0$.

Total testtid då den i:te komponenten går ner är

$$\mathcal{T}(T_{(i)}) = \sum_{j=1}^{i} T_{(j)} + (n-i)T_{(i)}$$

Speciellt är

$$\mathcal{T}(T_{(n)}) = \sum_{j=1}^{n} T_{(j)} = \sum_{j=1}^{n} T_j$$

I en TTT-plott plottar man

$$y_i = \frac{\mathcal{T}(T_{(i)})}{\mathcal{T}(T_{(n)})} \quad \text{mot} \quad x_i = \frac{i}{n}$$

för $i = 0, 1, \ldots, n$.

Ett enkelt exempel med $n = 4$ livstider:

$$6,\ 8,\ 12,\ 13$$

Här är

$$\begin{aligned}\mathcal{T}(T_{(1)}) &= 6 + 3 \cdot 6 = 24\\ \mathcal{T}(T_{(2)}) &= 6 + 8 + 2 \cdot 8 = 30\\ \mathcal{T}(T_{(3)}) &= 6 + 8 + 12 + 1 \cdot 12 = 38\\ \mathcal{T}(T_{(4)}) &= 6 + 8 + 12 + 13 = 39\end{aligned}$$

och punkterna som plottas är

$$(0,0),\ \left(\frac{1}{4}, \frac{24}{39}\right),\ \left(\frac{2}{4}, \frac{30}{39}\right),\ \left(\frac{3}{4}, \frac{38}{39}\right),\ (1,1)$$

Sats 11.1 Om $T \sim \text{Exp}(\lambda)$, så har de $n-1$ stokastiska variablerna

$$\frac{\mathcal{T}(T_{(1)})}{\mathcal{T}(T_{(n)})}, \ldots, \frac{\mathcal{T}(T_{(n-1)})}{\mathcal{T}(T_{(n)})}$$

samma fördelning som lika många ordnade oberoende $\text{U}(0,1)$-variabler.

Följdsats 11.2 Om $T \sim \text{Exp}(\lambda)$, så gäller, för $i = 1, \ldots, n-1$, att

1. $\text{Var}\left[\mathcal{T}(T_{(i)})/\mathcal{T}(T_{(n)})\right] < \infty$
2. $E\left[\mathcal{T}(T_{(i)})/\mathcal{T}(T_{(n)})\right] = i/n$

Om felbenägenheten är konstant, kan vi alltså förvänta oss att punkterna i TTT-plotten hamnar på den räta linjen $y = x$.

Det finns en motsvarande teoretisk operation på fördelningar som kallas TTT-transform och genom att jämföra den uppmätta TTT-plotten med teoretiska TTT-transformer så kan man kanske resonera sig fram till vilken typ av fördelning T har. Man kan visa (sats 11.3) att

1. $F(t)$ är IFR omm TTT-transformen är konvex,
2. $F(t)$ är DFR omm TTT-transformen är konkav.

När man har en idé om vilken typ av fördelning T har kan man skatta parametrar med gängse statistiska metoder, framförallt trolighetsmetoden (”maximum likelihood”).

11.3.8 Total time on test plot for censored data sets

Går att göra. Tekniken är baserad på Kaplan-Meierskattningen $\widehat{R}(t)$ av $R(t)$.

11.3.9 A brief comparison

R & H konstatrerar bl.a att Kaplan-Meierskattningen och Nelsons skattning är rätt så lika i princip och därför ger ungefär samma typ av information, medan TTT-plotten är annorlunda och därför kan tillföra något som de två nämnda skattningarna inte kan.

F.ö, läs själv.

11.4 Parametric methods

11.4.1 Binomial data

Känns inte så relevant.

11.4.2 Data from a homogeneous Poisson process

Antag att du observerat en HPP med intensitet λ i t tidsenheter och erhållit $N(t) = n$. Då är troligheten

$$l(\lambda) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}$$

och det är inte svårt att se att trolighetsskattningen av λ är

$$\widehat{\lambda} = \frac{N(t)}{t}$$

Låt tiderna mellen händelserna som räknas vara $T_1, T_2, \ldots$. En metod för beräkning av konfidensintervall presenterades i kapitel 7.

Man kan då n är stort också använda sig av att

$$N(t) \overset{ap}{\sim} \mathrm{N}(\lambda t, \sqrt{\lambda t})$$

d.v.s att

$$\frac{N(t)/t - \lambda}{\sqrt{\lambda/t}} = \frac{N(t) - \lambda t}{\sqrt{\lambda t}} \overset{ap}{\sim} N(0,1)$$

Obs att det R & H gör är mer sofistikerat än det simpla

$$\lambda = \widehat{\lambda} \pm z_{\alpha/2}\sqrt{\widehat{\lambda}/t}$$

Se sida 503.

11.4.3 Exponentially distributed lifetimes: complete sample

Vi antar nu att $T \sim \text{Exp}(\lambda)$ för något $\lambda > 0$.

Trolighetsfunktionen är

$$l(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda t_i} = \lambda^n e^{-\lambda \sum_i t_i}$$

Trolighetsskattningen av λ, som fås medelst maximering av trolighetsfunktionen, är

$$\widehat{\lambda}_{\text{ML}} = \frac{n}{\sum_i t_i} = \frac{n}{\mathcal{T}(T_{(n)})}$$

Man kan visa (se R & H sida 505) att

$$E\left[\widehat{\lambda}_{\text{ML}}\right] = \frac{n}{n-1}\lambda$$

Därför väljer man ibland att skatta λ väntevärdesriktigt med

$$\widehat{\lambda} = \frac{n}{n-1}\widehat{\lambda}_{\text{ML}} = \frac{n-1}{\mathcal{T}(T_{(n)})}$$

Man kan visa (se R & H sida 505) att

$$\text{Var}\left[\widehat{\lambda}\right] = \frac{\lambda^2}{n-2}$$

Man kan också visa att

$$2\lambda\mathcal{T}(T_{(n)}) = 2\lambda\sum_{i=1}^{n} T_i \sim \chi^2(2n)$$

(detta faktum är något annorlunda uttryckt i läroboken). Detta fördelningsresultat används till att konstruera konfidensintervall för λ alt testa nollhypoteser av typen $\lambda \leq \lambda_0$ eller $\lambda \geq \lambda_0$. Låt $\chi^2_{2n,1-\alpha/2}$ och $\chi^2_{2n,\alpha/2}$ vara $1-\alpha/2$- resp $\alpha/2$-kvantilen i $\chi^2(2n)$-fördelningen. Då

$$P\left(Y^2 \leq \chi^2_{2n,1-\alpha/2}\right) = P\left(Y^2 \geq \chi^2_{2n,\alpha/2}\right) = \alpha/2$$

för vilken $\chi^2(2n)$-fördelad stokastisk variabel Y^2 som helst. Härur fås

$$P\left(\chi^2_{2n,1-\alpha/2} \leq 2\lambda \sum_{i=1}^{n} T_i \leq \chi^2_{2n,\alpha/2}\right) = 1-\alpha$$

och vi inser att

$$\frac{\chi^2_{2n,1-\alpha/2}}{2\sum_{i=1}^n t_i} \leq \lambda \leq \frac{\chi^2_{2n,\alpha/2}}{2\sum_{i=1}^n t_i}$$

är det observerade konfidensintervallet med konfidensgrad $1-\alpha$. För mer detaljer, se sida 506-507 i R & H.

11.4.4 Exponentially distributed lifetimes: censored data

Typ 2-censurerade data Här kan man skriva upp och maximera troligheten, som är

$$l(\lambda) = \binom{n}{r} r! \left(\prod_{i=1}^{r} \lambda e^{-\lambda t_{(i)}}\right) \left(e^{-\lambda t_{(r)}}\right)^{n-r}$$

Faktorer som ej innehåller λ påverkar ej maximats läge och kan därför bortses ifrån i sökandet av $\operatorname{argmax}_\lambda l(\lambda)$, som f.ö lättast sker medelst logaritmering av $l(\lambda)$. Så

$$l(\lambda) \propto \lambda^r \, e^{-\lambda\left(\sum_{i=1}^r t_{(i)} + (n-r)t_{(r)}\right)} = \lambda^r \, e^{-\lambda\left(\mathcal{T}(t_{(r)})\right)}$$

Man får då att trolighetsskattningen är

$$\widehat{\lambda}_{\mathrm{ML}} = \frac{r}{\mathcal{T}(T_{(r)})} = \frac{r}{\sum_{i=1}^r T_{(i)} + (n-r)T_{(r)}}$$

Man kan visa att

$$\widehat{\lambda} = \frac{r-1}{\mathcal{T}(T_{(r)})} = \frac{r-1}{\sum_{i=1}^r T_{(i)} + (n-r)T_{(r)}}$$

skattar λ väntevärdesriktigt. Man kan också visa att

$$2\lambda\mathcal{T}(T_{(r)}) = 2\lambda\left(\sum_{i=1}^{r} T_{(i)} + (n-r)T_{(r)}\right) \sim \chi^2(2r)$$

Detta resultat används till konstruktion av konfidensintervall och genomförande av test av hypoteser om λ.