

9 Analys av livslängdsdata

9.1 Introduktion

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

Ofta är det första man är intresserad av huruvida den aktuella fördelningen är IFR, DFR eller har något annan typ av felbenägenhet. Detta görs ofta med en s.k TTT-plott.

Vi antar att n komponenter testas och skriver T_j för komponent j :s livstid (funktionstid) och antar att $T_1, T_2, \dots, T_n$ 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.

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

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

9.2 Fullständiga data

Total Time on Test Plot

Vi ska anta att T är kontinuerligt fördelad, att $F(t)$ är strikt växande, att

$$\mu = \int_0^\infty (1 - F(t)) dt < \infty$$

samt att

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

Definition 9.1 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

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

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

Vi förutsätter att alla n komponenterna sätts igång vid tidpunkten $t = 0$.

Total testtid då den i :te komponenten går ner är

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

Speciellt är

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

(obs att beteckningen $\mathcal{T}_i$ ej används i läroboken).

I en TTT-plott plottar man $y_i = \mathcal{T}_i/\mathcal{T}_n$ mot $x_i = i/n$ för $i = 0, 1, \dots, n$

Enkelt exempel med $n = 4$ livstider: 6 8 12 13. Här är

$$\mathcal{T}_1 = 6 + 3 \cdot 6 = 18$$

$$\mathcal{T}_2 = 6 + 8 + 2 \cdot 8 = 30$$

$$\mathcal{T}_3 = 6 + 8 + 12 + 1 \cdot 12 = 38$$

$$\mathcal{T}_4 = 6 + 8 + 12 + 13 = 39$$

och punkterna som plottas är

$$(0, 0), (1/4, 18/39), (2/4, 30/39), (3/4, 38/39), (1, 1)$$

Idéen bakom TTT-plottarna är följande resultat.

Sats 9.1 Om $T \sim \text{Exp}(\lambda)$, så har de $n - 1$ stokastiska variablerna

$$T_1/T_n, \dots, T_{n-1}/T_n$$

samma fördelningen som lika många ordnade oberoende $U(0, 1)$ -variabler.

Följdsats 9.1 Om $T \sim \text{Exp}(\lambda)$, så gäller, för $i = 1, \dots, n - 1$, att

1. $\text{Var}[T_i/T_n] < \infty$
2. $E[T_i/T_n] = i/n$

Om felbenägenheten är konstant, kan vi alltså förvänta oss att punkterna i TTT-plotten hamnar på den rätta 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 som T har. Man kan visa (sats 9.3) 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, t.ex trolighetsmetoden ("maximum likelihood").

Analys av exponentialfördelade livstider

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

$$\hat{\lambda}_{ML} = \frac{n}{\sum_i t_i} = \frac{n}{\bar{T}_n}$$

Man kan visa (se läroboken) att

$$E[\hat{\lambda}_{ML}] = \frac{n}{n-1} \lambda$$

Därför väljer man ibland att skatta λ väntevärdesriktigt med

$$\hat{\lambda} = \frac{n}{n-1} \hat{\lambda}_{ML} = \frac{n-1}{\mathcal{T}_n}$$

Man kan också visa att

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

(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 $\lambda \geq \lambda_0$.

9.3 Censurerade data

Censurerade data har man då inte alla livstiderna är fullständigt kända. Man skiljer ut några olika typer av censurering.

Typ 1: Man avbryter vid en fix tidpunkt t_1 . Antag att s enheter har fallerat, d.v.s

$$0 = T_{(0)} \leq T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(s)} \leq t_1 < T_{(s+1)}$$

Om de $n - s$ återstående vet man att de har fungerat i t_1 tidsenheter.

Typ 2: 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.

Typ 3: Man avbryter vid tidpunkten $\min(T_{(r)}, t_1)$.

Typ 4 är till för att hantera situationen då de n enheterna ej sätts igång samtidigt. Användbart framförallt i medicinska sammanhang.

Typ 1-censurerade $\text{Exp}(\lambda)$ -data

Vi noterar bara att den intuitiva skattningen

$$\hat{\lambda} = \frac{s-1}{\mathcal{T}(t_1)} = \frac{s-1}{\sum_{i=1}^s T_{(i)} + (n-s)t_1}$$

är vad man brukar använda.

Typ 2-censurerade $\text{Exp}(\lambda)$ -data

Här kan man skriva upp och maximera troligheten, som är

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

Man får då att trolighetsskattningen är

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

Man kan visa att

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

skattar λ väntevärdesriktigt. Man kan också visa att

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

Skattning av λ i en HPP

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

$$\hat{\lambda}_{ML} = \frac{n}{t}$$

Låt tiderna mellan händelserna som räknas vara $T_1, T_2, \dots$. Obs att $N(t) = n$ är ekvivalent med

$$T_1 + \dots + T_n \leq t < T_1 + \dots + T_n + T_{n+1}$$

Om n inte är alltför litet kan man nog tänka sig att differensen $t - \sum_{i=1}^n T_i < T_{n+1}$ är försumbar (T_{n+1} bör ju då vara tämligen liten jämför med summan $T_1 + \dots + T_n$) och göra approximationen

$$\sum_{i=1}^n T_i \approx t$$

och använda sig av faktumet

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

till att konstruera konfidensintervall för λ . Vi har isåfall att

$$2\lambda t \stackrel{ap}{\sim} \chi^2(2n)$$

Man kan då n är stort också använda sig av att

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

Kaplan-Meyerskattningen av $R(t)$

Nelsons skattning av $Z(t) = \int_0^t z(s) ds = -\ln R(t)$