Notat 5 - ST1301

15. februar 2005

1 Inversjonsmetoden

La oss fortsette eksempelet i notat 4. Vi hadde at

$$F_T(t) = 1 - \frac{a}{t+a}. \tag{1}$$

Deriverer vi får vi at tettheten til T blir

$$f_T(t) = \frac{a}{(t+a)^2}. \tag{2}$$

Observerer vi data $T_1, T_2, \ldots, T_n$ fra denne fordelingen fant vi at vi kunne beregne SME for den ukjente parameteren a ved å bruke Newton's metode til å løse estimeringsligningen vi får ved å sette den deriverte til log likelihood-funksjonen med hensyn til a lik null. Dette implementere vi som en funksjon `ahat` i R som beregner SME av a i fra observerte data.

For å undersøke om vår estimeringsmetode fungerer trenger vi å beregne et estimat $\hat{a}$ på grunnlag av simulerte data fra den modellen estimatoren faktisk er basert på. Det minste vi bør kreve er at estimatoren $\hat{a}$ ligge nærme den sanne verdien av a dersom utvalgsstørrelsen n er stor. Vi har tidligere laget oss data fra ulike fordelinger ved å bruke innebygde funksjoner i R for å simulere fra kjente fordelinger. Fordelingen gitt ved (2) svarer imidlertid ikke til noen fordeling som R kjenner. Vi trenger derfor en metode for å simulere fra (2).

Mer generelt skal vi se på en metode for å simulere fra kontinuerlige fordelinger ved bruk av den såkalte inversjonsmetoden. Metoden bygger på at vi er i stand til å simulere $U \sim \text{Unif}(0,1)$ (dette kan vi gjøre med funksjonen `runif` i R). Hvis vi kan finne en passende transformasjon $Y = g(U)$ slik at Y får fordelingen vi søker har vi løst problemet. Vi kan da simulere fra fordelingen til Y ved å simulere fra uniform fordeling og så beregne $Y = g(U)$.

La $F(y)$ være den kumulative tettheten til fordelingen vi ønsker å simulere fra.

Transformasjonen vi søker, g, skal ha intervallet fra 0 til 1 som definisjonsområdet og samme verdiområde som variabelen Y. Dette vil være tilfelle for den inverse av den kumulative fordelingsfunksjonen til fordelingen vi søker, F^{-1}, denne transformerer tall på intervallet fra 0 til 1 til tall som ligger i verdiområdet til Y. Får vi rett fordeling om vi lar $Y = F^{-1}(U)$? Kumulativ tetthet til transformasjon blir

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(F^{-1}(U) \leq y) \\ &= P(U \leq F(y)) \\ &= F_U(F(y)). \end{aligned} \tag{3}$$

Fordi U har kumulativ tetthet $F_U(u) = u$ for $0 \leq u \leq 1$ blir kumulativ tetthet til Y

$$F_Y(y) = F(y), \tag{4}$$

altså har Y fordelingen vi søker.

1.1 Generell algoritme

Hvis vi kjenner kumulativ fordeling $F(y)$ til en variabel Y og den inverse av F kan vi altså simulere en realisasjon av Y på følgende måte:

1. Simuler $U \sim \text{Unif}(0, 1)$.
2. Beregn $Y = F^{-1}(U)$.

1.2 Eksempel

I levetidsmodelleksempelet i del notat 4 hadde vi at

$$F_T(t) = 1 - \frac{a}{t + a}. \tag{5}$$

I følge inversjonsmetoden skal

$$T = F_T^{-1}(U) \tag{6}$$

har rett fordeling dersom $U \sim \text{Unif}(0, 1)$. Ligning (6) er ekvivalent med at

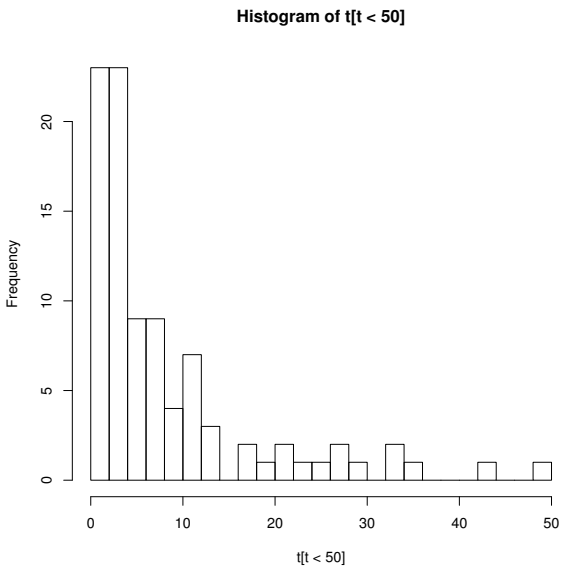
$$\begin{aligned} U &= F_T(T), \\ U &= 1 - \frac{a}{T + a}, \end{aligned} \tag{7}$$

altså transformasjonen

$$T = \frac{aU}{1-U}. \tag{8}$$

Bruker vi denne metoden kan vi simulere 100 observasjoner fra (2) på følgende måte. La oss anta at $a = 5$.

```
> a<-5
> u <- runif(100)
> t <- a*u/(1-u)
> range(t)
[1] 6.380138e-02 1.294394e+03
> hist(t[t<50],breaks=20)
```



Det kan være hensiktsmessig å lage en egen funksjon som simulerer fra levetidsmodellen med samme virkemåte som `rnorm`, `rexp` o.v.s. En slik funksjon vil kunne se slik ut

```
## Hensikt:
##   Simulere n tilfeldige variable fra fordelingentettheten
##   f(x)=a/(t+a)^2
##
rlevetid <- function(n,a) {
  u <- runif(n)
  a*u/(1-u)
}
```

Fordi dødsraten $\lambda(t) = 1/(t+a)$ går mot null når t blir stor får fordelingen en lang tynn øvre hale. I eksempelet over blir den lengste simulerte levetiden hele 1294 — langt større enn der alle fleste observasjonene.[1] Det er hensiktsmessig å ta bort disse observasjonene fra histogrammet.

Vi kan nå undersøke om estimatoren vår fungerer.

```
> ahat(t)
[1] 1.659555
[1] 2.562604
[1] 3.596082
[1] 4.461676
[1] 4.863254
[1] 4.923986
[1] 4.925144
[1] 4.925144
[1] 4.925144
```

Vi ser at algoritmen konvergerer rimelig raskt og at estimatet ser ut til å bli liggende i nærheten $a = 5$ slik det bør når n er stor.

[1] Det kan og nevnes at forventningen ikke eksisterer. Forsøker vi å beregne forventningen $E(T) = \int_0^\infty t f_T(t) dt$ finner vi at dette integralet ikke konvergerer. Hadde vi forsøkt å konstruere en estimator etter momentmetoden ville det derfor ikke ført fram.