

Estimering 1

-Punktestimering

Dekkes av kap. 8, 9.1-9.3 og 9.15/9.14.

Vi har til nå sett på en rekke forskjellige sannsynlighetsfordelinger og sett hvordan disse kan brukes til å modellere mange forskjellige typer fenomen. Vi har bl.a. sett på hvordan vi kan regne på sannsynligheter for ulike utfall, regne ut forventningsverdi, standardavvik/varsians, etc.

En grunnleggende forutsetning vi har gjort hele veien er at vi på ett eller annet vis har kjent den eksakte sannsynlighetsfordelingen til fenomenet vi har sett på - dvs at vi både har kjent den eksakte typen fordeling og de aktuelle parameterverdiene. I praksis er ofte ikke dette tilfellet - spesielt vil parameterverdiene typisk være ukjente! (Vi skal fortsatt anta at vi er i stand til å angi en (noenlunde) “korrekt” type fordeling.) *Estimering* handler om å anslå verdien på ukjente parameterverdier ut fra observerte data.

Eksempel: Levetiden til en bestemt type kretskort antas eksponensialfordelt. Vi har observert at levetiden for seks slike kort ble hhv 612, 1009, 95, 1303, 599, og 780. Hva forteller dette oss om verdien på parameteren i eksponensialfordelingen? Dvs om forventet levetid? $\square$

Eksempel: Antall ulykker per år på en oljeplattform antas Poissonfordelt. De siste fem årene har det blitt registrert hhv 5, 8, 12, 7 og 13 ulykker. Hva forteller dette oss om parameteren λ i Poisson-fordelingen? Dvs om forventet antall ulykker per år/intensiteten? $\square$

Noen grunnbegrep

Anta at vi har observasjoner/data $X_1, \dots, X_n$ fra et fenomen som er beskrevet ved en sannsynlighetsfordeling $f(x; \theta)$. Verdien på parameteren er ukjent. Dersom observasjonene $X_1, \dots, X_n$ er *uavhengige og identisk fordelte* (u.i.f.) kan de brukes til å anslå den ukjente parameteren θ .

Observasjonene er identisk fordelte dersom alle har samme sannsynlighetsfordeling $f(x; \theta)$ (med samme parameter). Observasjonene er uavhengige dersom simultanfordelingen kan skrives $f(x_1, \dots, x_n) = f(x_1) \cdots f(x_n)$ - vil i praksis si at informasjon om verdien på en av observasjonene ikke påvirker sannsynlighetsfordelingen til de øvrige.

At observasjoner er u.i.f. vil si det samme som at de er et *tilfeldig utvalg* (en tilfeldig valgt delmengde) av alle tenkelige mulige observasjoner av fenomenet som kan gjøres.

Alle tenkelige mulige observasjoner kalles *populasjonen*, og hvordan de fordeler seg er

beskrevet ved sannsynlighetsfordelingen $f(x; \theta)$. Den typiske problemstillingen er at vi basert på informasjonen i et tilfeldig utvalg, $X_1, \dots, X_n$, ønsker å si noe om hele populasjonen, f.eks. i form av å anslå parameteren θ .

Eksempel: Levetid kretskort. Populasjonen er levetiden til alle tenkelige kretskort av den aktuelle typen. Vi antar (basert på tidligere erfaring med denne typen fenomen) at denne populasjonen kan beskrives ved en eksponensialfordeling. Parameteren i eksponensialfordelingen er ukjent. Dersom observasjonene våre, $X_1, \dots, X_n$, er u.i.f., dvs et tilfeldig utvalg fra populasjonen av levetider, kan vi fra disse anslå, eller *estimere* parameteren.

Dersom $X_1, \dots, X_n$ ikke er u.i.f./tilfeldig utvalg, f.eks. dersom vi kun bruker observerte levetider som er større enn 1000, kan ikke observasjonene uten videre brukes til å estimere den ukjente parameterverdien. □

Estimatorer

Definisjon: En *estimator* er en funksjon av stokastiske variable som brukes til å anslå en ukjent parameterverdi.

Eksempel: Tre vanlig brukte estimatorer er:

$$\begin{aligned}\hat{\mu} &= \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \\ \hat{\sigma}^2 &= S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \\ \hat{p} &= \frac{X}{n}\end{aligned}$$

for hhv forventningsverdi μ og varians σ^2 i normalfordeling, og p i binomisk fordeling. $\square$

Det er vanlig å bruke gjennomsnittet til å estimere forventningsverdier også i andre fordelinger. F.eks. dersom $T_1, \dots, T_n$ er u.i.f. eksponensial med forventning β er det rimelig å estimere β med

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^n T_i.$$

Merk: Estimatorer betegner vi med store bokstaver (=før eksperimentet er utført). Innsatt observerte tallverdier, f.eks. $x_1, \dots, x_n$, (=etter eksperimentet er utført) kan vi regne ut et **estimat**, f.eks. $\hat{\mu} = \bar{x}$ eller $\hat{p} = x/n$. Estimatorer er altså stok. var. mens estimat er tallsvar.

Eksempel: Levetid kretskort. Estimatoren vi bruker til å anslå den ukjente parameteren β i eksponensialfordelingen er $\hat{\beta} = \frac{1}{n} \sum_{i=1}^n T_i$. Etter at vi har observert dataene 612, 1009, 95, 1303, 599, og 780 kan vi regne ut estimatet $\hat{\beta} = \frac{1}{6} \sum_{i=1}^6 t_i = 733$. $\square$

Egenskaper til estimatorer

Definisjon: En estimator $\hat{\theta}$ er en *forventningsrett* estimator for en parameter θ dersom:

$$E(\hat{\theta}) = \theta \quad \square$$

Eksempel: Når vi bruker estimatoren $\hat{\beta} = \frac{1}{n} \sum_{i=1}^n T_i$ for β i eksponensialfordelingen får vi:

$$E(\hat{\beta}) = E\left(\frac{1}{n} \sum_{i=1}^n T_i\right) = \frac{1}{n} \sum_{i=1}^n E(T_i) = \frac{1}{n} \sum_{i=1}^n \beta = \beta$$

Dvs $\hat{\beta} = \frac{1}{n} \sum_{i=1}^n T_i$ er forventningsrett.

Generelt ønsker vi gjerne at en estimator $\hat{\theta}$ oppfyller

1. $E(\hat{\theta}) = \theta$
2. $\text{Var}(\hat{\theta})$ minst mulig

Dersom vi har flere forventningsrette estimatører å velge mellom bruker vi den med minst varians (se øving 4).

Eksempel: Basert på u.i.f. observasjoner $X_1, \dots, X_n$ ønsker vi å estimere forventningsverdien, μ . To forslag til estimatorer: $\hat{\mu} = \bar{X}$ og $\tilde{\mu} = \tilde{X}$. Hvilken er best?

$$E(\hat{\mu}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \mu \quad (\text{alltid})$$

$$E(\tilde{\mu}) = \mu \quad (\text{dersom symmetrisk fordeling})$$

Dvs har vi en symmetrisk fordeling, f.eks. normal, uniform e.l., kan begge brukes. Har vi en usymmetrisk fordeling, f.eks. eksponential, gamma, Weibull, etc., bør vi bruke gjennomsnittet.

$$\text{Var}(\hat{\mu}) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}$$

Det kan vises at for normalfordeling er $\text{Var}(\hat{\mu}) < \text{Var}(\tilde{\mu})$. Dvs for normalfordelte data er det best å bruke gjennomsnittet (men for andre symmetrisk fordelinger kan medianen være bedre). $\square$

Sannsynlighetssmaksimerings-estimator (SME), 9.15

(Maximum likelihood estimator, MLE).

$X_1, \dots, X_n$ u.i.f. $f(x; \theta)$.

Hvordan generelt finne en estimator for θ ?

Noen ganger kan vi finne gode estimatorer fra “sunn fornuft”/intuisjon, andre ganger er dette verre. Hvordan skal vi f.eks. estimere parametrene α og β i gamma- eller Weibullfordelingen?

Eksempel: $X_1, \dots, X_n$ u.i.f. Poisson med parameter λ , og $t = 1$, dvs

$$f(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, \dots$$

Ut fra observerte data $X_1 = x_1, \dots, X_n = x_n$ ønsker vi å estimere λ .

Ide: Velg den verdien på λ som gjør det mest sannsynlig å observere det vi faktisk har observert!

$$\begin{aligned}
& P(\text{observere det vi har observert}) \\
&= P(X_1 = x_1 \cap \dots \cap X_n = x_n) \\
&= f(x_1, \dots, x_n; \lambda) \\
&\stackrel{uavh.}{=} f(x_1; \lambda) \cdots f(x_n; \lambda) \\
&= \frac{\lambda^{x_1}}{x_1!} e^{-\lambda} \cdots \frac{\lambda^{x_n}}{x_n!} e^{-\lambda} \\
&= \frac{\lambda^{x_1 + \dots + x_n}}{x_1! \cdots x_n!} e^{-n\lambda}
\end{aligned}$$

Vi ønsker å finne den λ som maksimerer $f(x_1, \dots, x_n; \lambda)$.

Triks: Det blir nesten alltid enklere regning ved å først ta logaritmen:

$$\begin{aligned}
& \ln(f(x_1, \dots, x_n; \lambda)) \\
&= \ln \left(\frac{\lambda^{\sum_{i=1}^n x_i}}{x_1! \cdots x_n!} e^{-n\lambda} \right) \\
&= \ln(\lambda^{\sum_{i=1}^n x_i}) - \ln(x_1! \cdots x_n!) + \ln(e^{-n\lambda}) \\
&= \left(\sum_{i=1}^n x_i \right) \ln(\lambda) - \ln(x_1! \cdots x_n!) - n\lambda
\end{aligned}$$

$$\frac{\partial \ln(f(x_1, \dots, x_n; \lambda))}{\partial \lambda} = \left(\sum_{i=1}^n x_i\right) \frac{1}{\lambda} - 0 - n = 0 \Rightarrow \lambda = \frac{1}{n} \left(\sum_{i=1}^n x_i\right)$$

Dvs SME for λ blir: $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$.

For eksemplet med ulykker på oljeplattform får vi estimatet $\hat{\lambda} = \frac{1}{5}(5 + 8 + 12 + 7 + 13) = \underline{\underline{9}}$. $\square$

Notasjon:

$$\begin{aligned} L(\lambda) &= f(x_1, \dots, x_n; \lambda) \\ l(\lambda) &= \ln(L(\lambda)) \end{aligned}$$

der $L(\lambda)$ kalles *likelihoodfunksjonen* og $\ln(L(\lambda))$ kalles *log-likelihoodfunksjonen*.

Merk: $L(\lambda)$ og $l(\lambda)$ har samme maksimumspunkt siden $\ln(\cdot)$ er en strengt voksende funksjon.

Vi bør også sjekke at vi har funnet et maksimumspunkt (og ikke et minimumspunkt) ved å sjekke at $\frac{\partial^2 l(\lambda)}{\partial \lambda^2} < 0$ for $\lambda = \hat{\lambda}$.

I eksemplet: $\frac{\partial^2 l(\lambda)}{\partial \lambda^2} = -\frac{\sum_{i=1}^n x_i}{\lambda^2} < 0$, dvs OK!.

SME generelt:

$X_1, \dots, X_n$ u.i.f. $f(x; \theta)$.

1. Definer likelihood-funksjonen:

$$\begin{aligned} L(\theta) &= f(x_1, \dots, x_n; \theta) \\ &\stackrel{uavh.}{=} f(x_1; \theta) \cdots f(x_n; \theta) = \prod_{i=1}^n f(x_i; \theta) \end{aligned}$$

2. Ta $\ln()$:

$$l(\theta) = \ln[L(\theta)] = \sum_{i=1}^n \ln[f(x_i; \theta)]$$

3. Deriver m.h.p. θ :

$$\frac{\partial}{\partial \theta} l(\theta) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \ln[f(x_i; \theta)]$$

4. Sett den deriverte lik null og løs mhp θ :

$$\frac{\partial}{\partial \theta} l(\theta) = 0 \quad \text{gir} \quad \hat{\theta}$$

5. Sjekk at:

$$\frac{\partial^2}{\partial \theta^2} l(\theta) < 0 \quad \text{for } \theta = \hat{\theta}$$

Eksempel: $X_1, \dots, X_n$ u.i.f. fra fordeling med tetthet

$$f(x; \theta) = \frac{\theta}{x^{\theta+1}}, x > 1$$

Finn en estimator for θ :

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta) = \prod_{i=1}^n \frac{\theta}{x_i^{\theta+1}} = \frac{\theta^n}{(\prod_{i=1}^n x_i)^{\theta+1}}$$

$$l(\theta) = \ln L(\theta) = n \ln(\theta) - (\theta + 1) \ln\left(\prod_{i=1}^n x_i\right)$$

$$= n \ln(\theta) - (\theta + 1) \sum_{i=1}^n \ln(x_i)$$

$$l'(\theta) = \frac{n}{\theta} - \sum_{i=1}^n \ln(x_i) = 0$$

$$\hat{\theta} = \frac{n}{\underline{\underline{\sum_{i=1}^n \ln(X_i)}}}$$

Sjekk: $l''(\hat{\theta}) = -\frac{n}{\hat{\theta}^2} < 0$ dvs maksimum.

Med observerte data 12.0, 11.2, 13.5, 12.3, 13.8 og 11.9 får vi estimatet:

$$\hat{\theta} = \frac{6}{\sum_{i=1}^6 \ln(x_i)} = \underline{\underline{0.397}}. \quad \square$$

Usikkerhet i estimatet:

En (punkt-)estimator $\hat{\theta}$ gir oss et anslag på en ukjent parameterverdi, men gir oss ikke noen direkte informasjon om usikkerheten i anslaget.

Kan vi anta at den estimerte verdien vi har fått er svært nær den sanne verdien, eller er det grunn til å frykte at vi er langt unna sann verdi?

Dette kan vi si noen om f.eks. ved å lage konfidensintervaller, men for å få til det må vi kjenne fordelingen til estimatoren.

Og for å kunne finne fordelingen til ulike estimatorer (som alltid er funksjoner av stokastiske variable) må vi kunne noe om fordelingen til funksjoner av stokastiske variable. Dette er omhandlet i kap. 7 i læreboka.