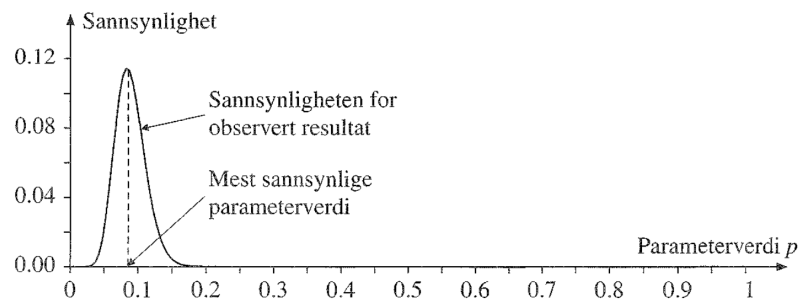
ST0103 Brukerkurs i statistikk
Høsten 2016

Momentestimatoren og sannsynlighets-maksimeringsestimatoren (SME)

Boka har bare ett eksempel med sannsynlighetsmaksimeringsestimatoren. Vi gjengir dette nedenfor, og viser så hvordan SME for p i binomisk fordeling kan finnes ved standard matematisk metode for maksimering av funksjoner.

Eksempel 6.2 I eksemplet om AluProd observerte vi 13 enheter som måtte vrakes og 147 som kunne godkjennes. Sannsynligheten for at en enkelt enhet må vrakes kalles p. Sannsynligheten for det observerte resultatet er lik

$$P(\text{observert resultat}) = \binom{160}{13} p^{13} \cdot (1-p)^{147} \qquad [6.3]$$



Figur 6.7 Logikken bak sannsynlighetsmaksimeringsestimatoren

Denne sannsynligheten er tegnet som en funksjon av parameterverdien p i figur 6.7. Hvilken verdi av p forklarer det observerte resultatet best?

Løsning: Som punktestimator skal vi velge den verdien av p som *maksimaliserer sannsynligheten* i ligning [6.3]. Figuren viser at vårt estimat skal være 0.081, som vi tidligere har funnet. (Matematisk ser vi dette ved å derivere ligning [6.3] mhp. p og finne dens nullpunkt.)

La oss finne ved regning den p som maksimerer ligning [6.3]. Det viser seg at det er lettere å maksimere den funksjonen vi får ved å ta logaritmen (naturlig logaritme ln) av [6.3]. Den må ha den samme maksimerende p siden logaritmen er en strengt voksende funksjon.

Logaritmen av høyresiden i [6.3] blir

$$\ell(p) = \ln \binom{160}{13} + 13 \ln p + 147 \ln(1-p)$$

(se grafen på neste side, laget med Wolfram Alpha).

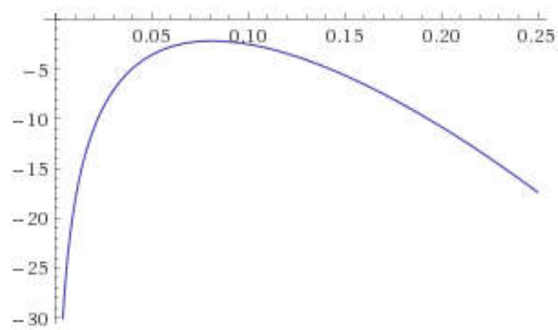
Input interpretation:

plot	$\log\left(\binom{160}{13}\right) + 13\log(p) + 147\log(1-p)$	$p = 0$ to 0.25

$\binom{n}{m}$ is the binomial coefficient

$\log(x)$ is the natural logarithm

Plot:



Ved å derivere med hensyn på p får vi

$$\ell'(p) = \frac{13}{p} - \frac{147}{1-p}$$

Settes dette lik 0 får vi

$$\begin{aligned}
\frac{13}{p} &= \frac{147}{1-p} \\
13(1-p) &= 147p \\
13 &= 160p \\
p &= \frac{13}{160} = 0.081
\end{aligned}$$

Siden dette er et *estimat* for p, vil vi isteden skrive $\hat{p} = 0.081$.

Hadde vi i regningen ovenfor erstattet 13 med x og 147 med $n - x$, ville vi endt opp med at $p = x/n$ maksimerer sannsynligheten for å observere x suksesser i n binomiske forsøk. Dette viser at den naturlige estimatoren $\hat{p} = X/n$ er en *sannsynlighetsmaksimeringsestimator*.

Fra eksamen i ST0103 desember 2014 (litt endret):

Oppgave 2

Utbruddene til en vulkan følger en poissonprosess med intensitet λ (forventet antall utbrudd pr. år).

b) Vulkanutbrudd har vært registrert siden 1. januar 1945 (tidspunkt 0). Det første registrerte utbruddet kom etter 7.9 år. Tidsintervallene mellom de følgende utbruddene var 2.5, 5.4, 10.8, 10.8, 0.3, 2.6, 8.4, 13.7 og 6.8 år. Anta at registreringen stoppet etter dette siste utbruddet. Her er et bilde av situasjonen, der kryssene på tidsaksen betyr vulkanutbrudd. Sjekk at den totale observasjonstiden blir 69.2 år.

La X være antallet utbrudd i løpet av de 69.2 år. Da er X observert til å være lik 10.

La videre tidsintervallene kalles $T_1, T_2, \ldots, T_{10}$ (der T_1 er tiden til første utbrudd, T_2 er tiden mellom første og andre, osv.) Disse er altså observert til $7.9, 2.5, \ldots, 6.8$.

Vi vil i det følgende se separat på de to måtene å observere på: antallet X i et gitt tidsrom, eller tidsintervallene $T_1, T_2, \ldots, T_{10}$.

Estimering av λ ved hjelp av antallet utbrudd X

X er Poisson-fordelt som følge av dette resultatet i læreboka:

Definisjon 5.8 (Poissonfordelingen) I løpet av de neste t tidsenhetene vil vi observere X forekomster av hendelsen A. Hvis poissonforutsetningene er oppfylt, er X poissonfordelt med parameter λt, som skrives $X \sim \text{poisson}(\lambda t)$. Sannsynlighetsfordeling, forventning og varians er lik

$$P(X = x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t}, \qquad x = 0, 1, 2, \ldots$$
$$E(X) = \lambda t, \qquad \text{Var}(X) = \lambda t$$

Momentmetoden

Siden
$$E(X) = \lambda \cdot 69.2,$$
vil vi estimere λ ved erstatte $E(X)$ med vårt "beste gjett", nemlig X (som altså er en observator). Samtidig må vi da endre λ til en estimator $\hat{\lambda}$. Vi får:
$$X = \hat{\lambda} \cdot 69.2$$

som gir
$$\hat{\lambda} = \frac{X}{69.2} = \frac{10}{69.2} = 0.145$$

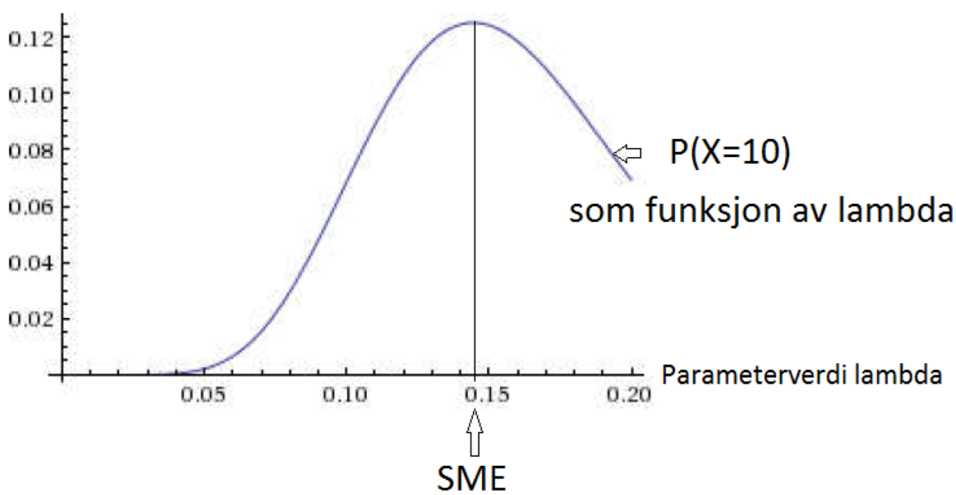
Sannsynlighetsmaksimeringsmetoden

La oss så finne SME for λ basert på X. Siden vi har observert $X = 10$, må vi sette opp sannsynligheten for dette resultatet,

$$P(X = 10) = \frac{(\lambda \cdot 69.2)^{10}}{10!} e^{-\lambda \cdot 69.2}, \quad (1)$$

som funksjon av λ. Deretter må vi finne den verdi av λ som maksimerer dette.

Grafisk har vi (Wolfram Alpha):



Vi leser av fra figuren at $\hat{\lambda} \approx 0.145$.

Matematisk skal vi maksimere uttrykket (1) for $P(X = 10)$. Tar vi logaritmen, skal vi maksimere

$$\begin{aligned} \ell(\lambda) &= 10 \ln(\lambda \cdot 69.2) - \ln(10!) - \lambda \cdot 69.2 \\ &= 10 \ln \lambda + 10 \ln 69.2 - \ln(10!) - \lambda \cdot 69.2 \end{aligned}$$

Dette gjøres ved å derivere og sette lik 0:

$$\ell'(\lambda) = \frac{10}{\lambda} - 69.2 = 0$$

$$\lambda = \frac{10}{69.2} = 0.145$$

som gir estimatet $\hat{\lambda} = 0.145$, som er det samme vi fikk med momentmetoden.

Estimering av λ ved hjelp av tidene mellom utbrudd, $T_1, T_2, \ldots, T_{10}$

Momentmetoden

For en enkel observasjon T (tid mellom to utbrudd) har vi

$$E(T) = \frac{1}{\lambda} \tag{2}$$

Basert på våre data, er vårt beste "gjett" for $E(T)$ gitt ved $\bar{T}$ (= 69.2/10). Vi erstatter derfor $E(T)$ med $\bar{T}$ i (2). Samtidig må vi da endre λ til en estimator $\hat{\lambda}$. Vi får:

$$\bar{T} = \frac{1}{\hat{\lambda}}$$

som gir

$$\hat{\lambda} = \frac{1}{\bar{T}} = \frac{1}{\frac{69.2}{10}} = \frac{10}{69.2} = 0.145$$

som er det samme vi fikk ved å bruke X.

Sannsynlighetsmaksimeringsmetoden

Ideen med SME er å finne den verdi av parameteren som maksimerer sannsynligheten for det vi har observert. Her vil det si å maksimere

$$P(T_1 = 7.9 \cap T_2 = 2.5 \cap \cdots \cap T_{10} = 6.8)$$

Men denne sannsynligheten er lik 0, siden sannsynligheten for å få en spesifikk verdi for en kontinuerlig stokastisk variabel alltid er 0.

Vi antar derfor isteden at våre data er gitt i form av små intervaller omkring de gitte dataene. For eksempel antar vi at observasjonen av $T_1 = 7.9$ svarer til at T_1 er i intervallet fra 7.9 til $7.9 + h$, der h er liten. Tilsvarende gjør vi for de andre T_i. Sannsynligheten for det vi har observert er da

$$\begin{aligned}
&P(7.9 \leq T_1 \leq 7.9 + h \ \cap \ 2.5 \leq T_2 \leq 2.5 + h \ \cap \cdots \cap 6.8 \leq T_{10} \leq 6.8 + h) \\
&= P(7.9 \leq T_1 \leq 7.9 + h) \cdot P(2.5 \leq T_2 \leq 2.5 + h) \cdots P(6.8 \leq T_{10} \leq 6.8 + h) \\
&\approx f(7.9)h \cdot f(2.5)h \cdots f(6.8)h \\
&= f(7.9)f(2.5) \cdots f(6.8)h^{10} \qquad (3)
\end{aligned}$$

der $f(t)$ er tettheten til tidene mellom utbrudd, som vi vet er eksponensialfordelte med $f(t) = \lambda e^{-\lambda t}$. Vi har her brukt at $T_1, \ldots, T_{10}$ er uavhengige, og vi har brukt det generelle resultat at for små h er

$$P(t \leq T \leq t + h) \approx f(t)h$$

Uttrykket (3) ovenfor skal altså maksimeres med hensyn på λ. Tricket vi har brukt, er å innføre en liten h, noe som gjør at vi får en positiv sannsynlighet for det vi har observert. Men siden h er vilkårlig valgt av oss, vil vi kunne maksimere (3) ved å se bort fra faktoren h^{10} og isteden maksimere den såkalte likelihood-funksjonen,

$$L(\lambda) = f(7.9)f(2.5) \cdots f(6.8) = \lambda e^{-\lambda \cdot 7.9} \lambda e^{-\lambda \cdot 2.5} \cdots \lambda e^{-\lambda \cdot 6.8}$$

$$= \lambda^{10} e^{-(7.9+2.5+\cdots 6.8)\lambda} = \lambda^{10} e^{-98.2\cdot\lambda}$$

Ved å ta logaritmen får vi

$$\ell(\lambda) = 10 \ln \lambda - 98.2 \cdot \lambda$$

som vi maksimerer ved å derivere og sette lik 0, og ender med å få det samme som før, $\hat{\lambda} = 10/69.2 = 0.145$.

Generelt finner vi likelihood-funksjonen for data $x_1, x_2, \ldots, x_n$ fra en modell med tetthet $f(x)$og ukjent parameter θ ved å la

$$L(\theta) = f(x_1) f(x_2) \cdots f(x_n)$$

og finne den verdi av θ som maksimerer dette. Vanligvis lønner det seg da å ta logaritmen først, og så derivere og sette lik 0.