SF1901: Sannolikhetslära och statistik Föreläsning 7. Statistik: statistiska inferensproblem, maximum likelihood, minsta kvadrat

Jan Grandell & Timo Koski

22.09.2008

KTH Matematik

Sammanfattning

- Vi inleder med några exempel på statistiska *inferensproblem* (inferens= slutledning):

Sammanfattning

- Vi inleder med några exempel på statistiska *inferensproblem* (inferens= slutledning):
  - opinionsundersökning

Sammanfattning

- Vi inleder med några exempel på statistiska *inferensproblem* (inferens= slutledning):
  - opinionsundersökning
  - serienummerproblemet även känt som 'the German tank problem'.

KTH Matematik

Sammanfattning

- Vi inleder med några exempel på statistiska *inferensproblem* (inferens= slutledning):
  - opinionsundersökning
  - serienummerproblemet även känt som 'the German tank problem'.
  - mätvärden med brus

KTH Matematik

Sammanfattning

- Vi inleder med några exempel på statistiska *inferensproblem* (inferens= slutledning):
  - opinionsundersökning
  - serienummerproblemet även känt som 'the German tank problem'.
  - mätvärden med brus
- Sedan definierar vi begrepp som punktskattning och stickprovsvariabel, och inför krav på deras prestanda.

KTH Matematik

Sammanfattning

- Vi inleder med några exempel på statistiska *inferensproblem* (inferens= slutledning):
  - opinionsundersökning
  - serienummerproblemet även känt som 'the German tank problem'.
  - mätvärden med brus
- Sedan definierar vi begrepp som punktskattning och stickprovsvariabel, och inför krav på deras prestanda.
- Vi inför två systematiska metoder för systematisk konstruktion av punktskattningar: *maximum-likelihood* och *minsta kvadrat* och ger exempel därpå.

KTH Matematik

Inledande om opinionsundersökning: punktskattning

Vi väljer 1000 personer ur en 'stor' population (t.ex. Sveriges befolkning). De ska svara 'JA' eller 'NEJ' till en fråga (t.ex. om medlemskap i militäralliansen NATO).

Ponera nu att

$$x = 350 \quad \text{svarat JA}$$

Vi vill veta proportionen JA-sägare i HELA populationen, dvs. punktskatta proportionen JA-sägare i HELA populationen.

En skattning av av p:

$$p^*_{obs} = \frac{x}{1000} \Rightarrow p^*_{obs} = 35\%$$

Hur stor är osäkerheten i denna skattning?

KTH Matematik

Inledande diskussion: modell för skattning

$p^*_{obs} = \frac{x}{1000} = 0.35$. Hur stor är osäkerheten i denna skattning? Vi använder oss av en sannolikhetsmodell.

URNMODELL: $p =$ andelen JA-svar i populationen. $X =$ antalet JA-svar när vi frågar 1000 individer. $N =$ populationens storlek, t.ex. $N = 7{,}3$ miljoner svenskar över 15 år (enligt SCB var Sveriges befolkning 9 182 927 invånare på nyårsafton 2007)[1].

Vi har att

$$X \in \mathrm{Hyp}(N, 1000, p)$$

[1] http://www.scb.se/templates/pressinfo___227515.asp

Inledande diskussion: modell för skattning

Vi återkallar i minnet

Sats

Om X är Hyp(N, n, p)*-fördelad med* $n/N \leq 0.1$ *så är X approximativt* Bin(n, p)*-fördelad.*

Då kan vi utan stor förlust ta

$$X \in \mathrm{Bin}(1000, p)$$

Vi uppfattar $x = 350$ som ett utfall av X. Vi kallar

$$p^* = \frac{X}{1000}$$

en *stickprovsvariabel*. $p^*_{obs} = \frac{x}{1000} = 0.35$ är ett utfall av p^*

KTH Matematik

Inledande diskussion: modell för skattning

$$X \in \text{Bin}(1000, p)$$

Vi vet att $E(X) = 1000 \cdot p$, $V(X) = 1000 \cdot p \cdot (1-p)$. Detta ger

$$E(p^*) = E\left(\frac{X}{1000}\right) = \frac{1}{1000} E(X) = \frac{1}{1000} \cdot 1000 \cdot p = p.$$

$$V(p^*) = V\left(\frac{X}{1000}\right) = \frac{1}{1000^2} V(X) = \frac{1}{1000^2} \cdot 1000 \cdot p(1-p) = \frac{p(1-p)}{1000}.$$

Svårigheten är att de ovanstående uttrycken beror på den okända (parameter) p som vi vill skatta!

KTH Matematik

Inledande diskussion: medelfelet för skattning

$$E(p^*) = p, \quad V(p^*) = \frac{p(1-p)}{1000}.$$

$$D(p^*) = \sqrt{V(p^*)} = \sqrt{\frac{p(1-p)}{1000}}$$

Vår skattning av osäkerhet i skattningen är då *medelfelet*

$$d(p^*) = \sqrt{\frac{p^*_{obs}(1-p^*_{obs})}{1000}} = \sqrt{\frac{0.35(1-0.35)}{1000}} \approx 0.0015 = 1.5\%$$

KTH Matematik

German tank problem (1)

During World War II, the Germans tried to make a complete tabulation of how much they were producing, but reports from individual factories were often late and not always reliable. British and U.S. statisticians working for military intelligence were keenly interested in estimating German war production (especially the production of Mark V German tanks (Panther)), too, but they could hardly ask the German factories to send them reports. Instead, they based their estimates on the manufacturing serial numbers of captured equipment (specifically the tank gearboxes). These numbers were consecutive and didn't vary - because that was a rational system in terms of maintenance and spare parts. These serial numbers provided a sample that was very small, but reliable.

The challenge is to choose a good estimator for the total number of tanks.

German tank problem (2)

Detta är även känt som 'the serial number problem'. Vi kan renodla det enligt följande:

Någon har levererat oss heltalen $x_1, x_2, \ldots, x_n$, som är slumpmässiga stickprov från mängden av heltal

$$\{1, 2, 3, \ldots, N\}$$

där N *är okänt*.

Uppgiften (uppgiften för statistisk inferens) är att skatta N på basis av $x_1, x_2, \ldots, x_n$. (forts.)

KTH Matematik

satta påståendet minskas.

Sannolikhetsbevis, *log.*, bevis, som ej ådagalägga slutsatsens sanning, utan blott dess sannolikhet. Hit räknas h y p o t e s-, a n a l o g i- och i n d u k t i o n s b e v i s e n. S—e.

http://runeberg.org/nfcd/0387.html

Nordisk familjebok / Uggleupplagan. 24. Ryssläder - Sekretär, sid. 725-726 (1916)

Statistisk inferens: f.d. sannolikhetsbevis

I Nordisk familjebok (1916) avses med ett induktionsbevis en utsaga om en egenskap hos en hel population som baserar sig på ett studium av en (kanske relativt liten) del av populationen.

Exempel

På en laboration vill man bestämma den fysikaliska konstanten μ. Vi gör upprepade mätningar av μ t.ex.

$$\text{mätvärde} = \mu + \text{slumpmässigt mätfel}$$

och erhåller följande mätvärden:

$$x_1, x_2, \ldots, x_n$$

Problem

Hur skall vi *skatta* μ så bra som möjligt. Här liksom i de föregående två exemplen arbetar vi med *punktskattning* (till skillnad från *intervallskattning* som kommer att presenteras längre fram i kursen).

KTH Matematik

Vi uppfattar mätvärdena som utfall av n st. oberoende och lika fördelade s.v. $X_1, X_2, \ldots, X_n$ med $E(X_i) = \mu$ och $V(X_i) = \sigma^2$.
En *punktskattning* μ^*_{obs} av μ är en funktion av mätvärdena:

$$\mu^*_{\text{obs}} = \mu^*(x_1, \ldots, x_n).$$

KTH Matematik

När vill vill analysera en skattning ersätter vi observationerna med de underliggande stokastiska variablerna. Vi säger då att

$$\mu^* = \mu^*(X_1, \ldots, X_n)$$

är en *stickprovsvariabel*.

Stickprovsvariabeln är själv en stokastisk variabel, vars fördelning beror av fördelningen för $X_1, X_2, \ldots, X_n$ och därmed av μ.

Om vi inte använder någon statistisk teori så väljer vi antagligen

$$\mu^*_{\text{obs}} = \bar{x} = \frac{1}{n}\sum x_i.$$

För motsvarande stickprovsvariabel $\mu^* = \overline{X}$ gäller (liksom tidigare konstaterats) att

$$E(\mu^*) = E(\overline{X}) = \mu$$

och

$$V(\mu^*) = V(\overline{X}) = \sigma^2/n.$$

Allmänt

Vi har en uppsättning data

$$x_1, x_2, \ldots, x_n$$

som ses som utfall av s.v.

$$X_1, X_2, \ldots, X_n.$$

Dessa variabler antages vara oberoende och likafördelade och deras gemensamma fördelning beror av en okänd parameter θ, t.ex. $N(\theta, \sigma)$, Po (θ), $N(\theta_1, \theta_2)$, osv.

KTH Matematik

Allmänt

- En punktskattning θ^*_{obs} av θ är en funktion $\theta^*(x_1, \ldots, x_n)$

Allmänt

- En punktskattning θ^*_{obs} av θ är en funktion $\theta^*(x_1, \ldots, x_n)$
- och motsvarande stickprovsvariabel θ^* är $\theta^*(X_1, \ldots, X_n)$

KTH Matematik

Allmänt

- En punktskattning θ^*_{obs} av θ är en funktion $\theta^*(x_1, \ldots, x_n)$
- och motsvarande stickprovsvariabel θ^* är $\theta^*(X_1, \ldots, X_n)$
- Exemplet ovan:

$$p^*_{obs} = \frac{350}{1000} = 0.35$$

  är en punktskattning.

$$p^* = \frac{X}{1000}, \quad X \in \text{Bin}(1000, p)$$

  är motsvarande stickprovsvariabel.

KTH Matematik

Vad menas med en bra skattning?

Vi ger tre kriterier:

Definition

1) *En punktskattning* θ^*_{obs} *av* θ *är väntevärdesriktig om* $E(\theta^*(X_1, \ldots, X_n)) = \theta$.

Vad menas med en bra skattning?

Vi ger tre kriterier:

Definition

1) *En punktskattning* θ^*_{obs} *av* θ *är väntevärdesriktig om* $E(\theta^*(X_1, \ldots, X_n)) = \theta$.
2) *En punktskattning* θ^*_{obs} *av* θ *är konsistent om* $P(|\theta^*(X_1, \ldots, X_n) - \theta| > \varepsilon) \to 0$ *då* $n \to \infty$.

KTH Matematik

Vad menas med en bra skattning?

Vi ger tre kriterier:

Definition

1) *En punktskattning* θ^*_{obs} *av* θ *är väntevärdesriktig om* $E(\theta^*(X_1, \ldots, X_n)) = \theta$.
2) *En punktskattning* θ^*_{obs} *av* θ *är konsistent om* $P(|\theta^*(X_1, \ldots, X_n) - \theta| > \varepsilon) \to 0$ *då* $n \to \infty$.
3) *Om* θ^*_{obs} *och* θ^{**}_{obs} *är väntevärdesriktiga skattningar av* θ *så säger man att* θ^*_{obs} *är effektivare än* θ^{**}_{obs} *om* $V(\theta^*(X_1, \ldots, X_n)) < V(\theta^{**}(X_1, \ldots, X_n))$.

KTH Matematik

Proportionen JA-sägare

$$X \in \text{Bin}(1000, p)$$

$$p^* = \frac{X}{1000}, \quad E(p^*) = p.$$

Således är p^*_{obs} väntevärdesriktig.

Skattning av väntevärdet μ

Sats

Stickprovsmedelvärdet $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ som skattning av väntevärdet μ är

1) *Väntevärdesriktig;*
2) *Konsistent;*
3) *Ej nödvändigtvis effektiv, dvs. den effektivaste möjliga skattningen.*

Bevis.

1) $E(\overline{X}) = \mu$.
2) $V(\overline{X}) = \sigma^2/n$ och stora talens lag gäller.
3) Det finns motexempel (den intresserade hänvisas till Blom et al.)

$\square$

KTH Matematik

Skattning av σ^2

Sats

Stickprovsvariansen $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ *som skattning av* σ^2 *är*

1) *Väntevärdesriktig;*
2) *Konsistent;*
3) *Ej nödvändigtvis effektiv.*

1) används ofta som motivering för att man dividerar med $n-1$, men det är en dålig motivering, eftersom man oftast vill skatta σ. s som skattning av σ är dock ej väntevärdesriktig.

Maximum-likelihood-metoden: Likelihoodfunktion

Antag att X_i har täthetsfunktionen $f_X(x, \theta)$, θ okänd.
Vi ska nu studera en systematisk metod att hitta skattningar. Idén är att skatta θ så att utfallet blir så "troligt" som möjligt.

Definition

$$L(\theta) = f_{X_1}(x_1, \theta) \cdot \cdots \cdot f_{X_n}(x_n, \theta)$$

kallas Likelihood-funktionen.

Observera att likelihoodfunktionen betraktas som en funktion av θ, inte av $x_1, \ldots, x_n$.

KTH Matematik

Maximum-likelihood-metoden: Diskreta fallet

Antag att X_i har sannolikhetsfunktionen $p_X(x, \theta)$, θ okänd.

$$L(\theta) = p_{X_1}(x_1, \theta) \cdot \cdots \cdot p_{X_n}(x_n, \theta)$$

kallas *Likelihood*-funktionen.

Maximum-likelihood-metoden

Det diskreta fallet:

$$L(\theta) = p_{X_1}(x_1, \theta) \cdot \dots \cdot p_{X_n}(x_n, \theta)$$

Definition

Det värde θ^*_{obs} för vilket $L(\theta)$ antar sitt största värde kallas ML-skattningen av θ.

Den grundläggande tanken: vi väljer ett θ^*_{obs} sådant att sannolikheten för de observerade data blir maximal.

KTH Matematik

Maximum-likelihood-metoden

$$L(\theta) = f_{X_1}(x_1, \theta) \cdot \dots \cdot f_{X_n}(x_n, \theta)$$

Definition

Det värde θ^*_{obs} för vilket $L(\theta)$ antar sitt största värde kallas ML-skattningen av θ.

För stora stickprov är denna skattning i allmänhet mycket bra.

Maximum-likelihood-metoden för $\mathrm{Bin}(1000, p)$

Låt oss återkalla i minnet den inledande diskussionen om skattning av proportionen JA-sägare i en stor population. Vi har observerat $x = 350$ och tar detta som ett utfall av X med

$$X \in \mathrm{Bin}(1000, p)$$

Då är likelihoodfunktionen för p

$$L(p) = p_X(x) = \binom{1000}{x} p^x (1-p)^{1000-x}$$

Vi bildar *loglikelihoodfunktionen* $\ln L(p)$, dvs.

$$\ln L(p) = \ln \binom{1000}{x} + x \ln p + (1000 - x) \ln 1 - p$$

KTH Matematik

Maximum-likelihood-metoden för $\mathrm{Bin}(1000, p)$

För att maximera $L(p)$ kan vi ekvivalent maximera

$$\ln L(p) = \ln \binom{1000}{x} + x \ln p + (1000 - x) \ln 1 - p$$

För detta deriverar vi $\ln L(p)$ m.a.p. p

$$\frac{d}{dp} \ln L(p) = x \frac{1}{p} - (1000 - x) \frac{1}{1 - p}$$

och löser $\frac{d}{dp} \ln L(p) = 0$ m.a.p. p.

$$x \frac{1}{p} - (1000 - x) \frac{1}{1 - p} = 0 \Leftrightarrow x \frac{1}{p} = (1000 - x) \frac{1}{1 - p}$$

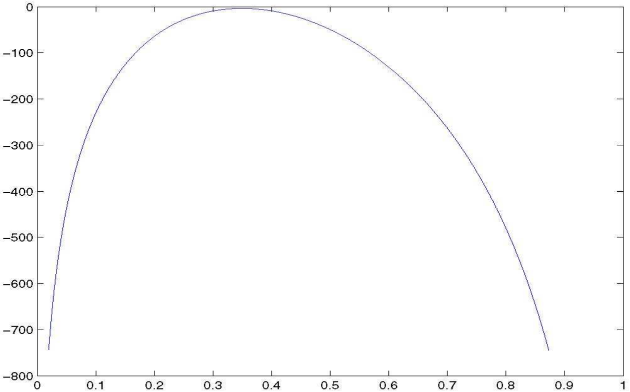
$$\Leftrightarrow (1 - p)x = p(1000 - x) \Leftrightarrow x - px = p1000 - px$$

dvs. maximum likelihood - skattningen är

$$p^*_{obs} = \frac{x}{1000} = \frac{35}{1000} = 0.35.$$

KTH Matematik

Plot av loglikelihoodfunktionen $\ln L(p)$ för $\mathrm{Bin}(1000, p)$ med $x = 350$



$$\ln L(p) = \ln \binom{1000}{350} + 350 \ln p + (1000 - 350) \ln 1 - p, \quad 0 < p < 1.$$

Maximum-likelihood-metoden

Exempel

X_i är $N(\theta, \sigma)$, dvs.

$$f(x, \theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2}.$$

Vi kan t.ex. ha

$$\text{mätvärde} = \theta + \text{ slumpmässigt normalfördelat mätfel}$$

$$X_i = \theta + \sigma Z_i, \quad Z_i \in N(0,1), \quad i = 1, 2, \ldots, n$$

Vi observerar $x_1, \ldots, x_n$. Då fås

$$L(\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_1-\theta}{\sigma}\right)^2} \cdot \ldots \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_n-\theta}{\sigma}\right)^2} = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2}\sum_1^n \left(\frac{x_i-\theta}{\sigma}\right)^2}$$

Maximum-likelihood-metoden Exempel

Vi antar att σ är känt.

$$\ln L(\theta) = -\ln(\sigma^n (2\pi)^{n/2}) - \frac{1}{2\sigma^2}\sum_1^n (x_i - \theta)^2$$

$$\frac{d \ \ln L(\theta)}{d\theta} = \frac{1}{2\sigma^2}\sum_1^n 2(x_i - \theta).$$

$$\frac{d \ \ln L(\theta)}{d\theta} = 0$$

ger

$$\sum_1^n x_i = n\theta,$$

dvs.

$$\theta^*_{\text{obs}} = \overline{x}.$$

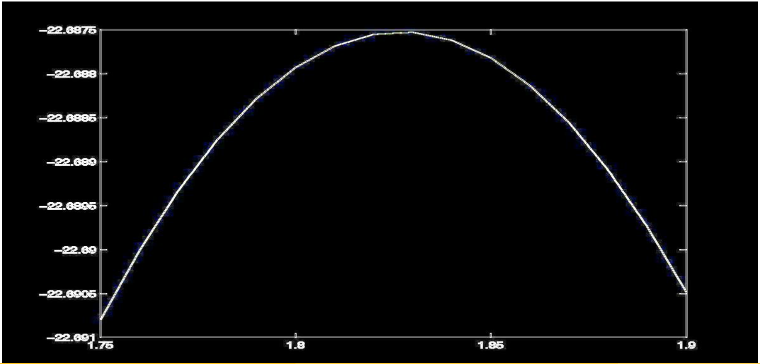
I detta fall är θ^*_{obs} *effektiv*!

$$\ln L(\theta) = -\log(3^{10}\cdot(2\pi)^{10/2}) - \frac{1}{2\cdot 3^2}\sum_{1}^{10}(x_i-\theta)^2$$

$x_1 = 3.28 \quad x_2 = -0.21 \quad x_{31} = 3.69 \quad x_4 = -2.15 \quad x_5 = 3.38$

$x_6 = 3.89 \quad x_7 = 3.14 \quad x_8 = -1.04 \quad x_9 = 0.96 \quad x_{10} = 3.33$

$\overline{x} = 1.83$



German tank problem (3)

θ = total number of tanks (true population value)

n = number of tanks captured

m = largest serial number of the captured tanks

The following *estimator* (=*stickprovsvariabel* i vår terminologi) was invented and used by statisticians in the the WW II allied military intelligence[2] :

$$\theta^* = [(n+1)/n]m \leftrightarrow \theta^* = m + (m/n)$$

which is interpreted as adding the average size of the gap to the highest serial number.

[2] Economic Warfare Division of the American Embassy in London

Betrakta ett fiktivt exempel[3]. Antag att vi har $n = 15$ och att serienumren är

5 10 15 20 25 30 35 40 45 50 55 60 65 70 75

så att $m = 75$. Då ger formeln ovan *punktskattningen*

$$\theta^*_{\text{obs}} = 75 + (75/15) = 80$$

Vi kommer nedan att tolka stickprovsvariabeln $\theta^* = m + (m/n)$ som en för väntevärdesriktighet korrigerad maximum-likelihood-skattning av θ i $U(0, \theta)$ med n observationer !

[3] http://en.wikipedia.org/wiki/German_tank_problem

Minsta-kvadrat-metoden

Om vi inte känner fördelningen helt kan inte ML-metoden användas. Ibland ger den även upphov till svåra beräkningsproblem. Man kan då gå tillväga på följande sätt:
Låt $x_1, \ldots, x_n$ vara ett stickprov från en fördelning med $E(X) = \mu(\theta)$ där $\mu(\theta)$ är en *känd* funktion av en *okänd* parameter θ.
Sätt $Q(\theta) = \sum_{i=1}^{n}(x_i - \mu(\theta))^2$ och minimera $Q(\theta)$ map. θ. Lösningen θ^*_{obs} till detta problem kallas *MK-skattningen* av θ.

KTH Matematik

Minsta-kvadrat-metoden

$$Q(\theta) = \sum_{i=1}^{n} (x_i - \mu(\theta))^2$$

$$\frac{d}{d\theta} Q(\theta) = 2 \sum_{i=1}^{n} (x_i - \mu(\theta)) \cdot \left(-\frac{d}{d\theta} \mu(\theta) \right)$$

$$\frac{d}{d\theta} Q(\theta) = 0 \Leftrightarrow -2 \frac{d}{d\theta} \mu(\theta) \sum_{i=1}^{n} (x_i - \mu(\theta)) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} (x_i - \mu(\theta)) = 0 \Leftrightarrow \sum_{i=1}^{n} x_i - n\mu(\theta) = 0$$

$$\mu(\theta) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Ifall den inversa funktionen $\mu^{-1}(\theta)$ existerar ges MK-skattningen θ^*_{obs} av

$$\theta^*_{obs} = \mu^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) = \mu^{-1} (\overline{x})$$

KTH Matematik

Minsta-kvadrat-metoden

$x_1, x_2, \ldots, x_n$ utfall av s.v. $X_1, X_2, \ldots, X_n$, $X_i \in U(0, \theta)$. Kursens formelsamling eller en enkel härledning ger

$$E\left(X_i\right) = \frac{\theta}{2}, \quad V\left(X_i\right) = \frac{\theta^2}{12}$$

dvs. vi tar $\mu(\theta) = \frac{\theta}{2}$ i $Q(\theta) = \sum_{i=1}^{n}(x_i - \mu(\theta))^2$, dvs.

$$Q(\theta) = \sum_{i=1}^{n}(x_i - \frac{\theta}{2})^2$$

MK-skattningen θ^* ges enligt ovan av

$$\theta^*_{obs} = \mu^{-1}\left(\overline{x}\right) = 2\overline{x}$$

KTH Matematik

$$\theta^*_{obs} = \mu^{-1}(\overline{x}) = 2\overline{x}$$

Denna punktskattning är väntevärdesriktig:

$$E(\theta^*) = 2E\left(\overline{X}\right) = 2\frac{\theta}{2} = \theta$$

och konsistent (Tjebysjovs olikhet)

$$V(\theta^*) = V\left(2\overline{X}\right) = 4\frac{1}{n^2} \cdot n\frac{\theta^2}{12}$$

$$= \frac{\theta^2}{3n}.$$

KTH Matematik

Maximum likelihood : samma exempel

$x_1, x_2, \ldots, x_n$ utfall av s.v. $X_1, X_2, \ldots, X_n$, $X_i \in U(0, \theta)$. Vad är maximum likelihood skattningen av θ?

$$f_X(x) = \begin{cases} \frac{1}{\theta} & \text{för } 0 \leq x \leq \theta \\ 0 & \text{annars.} \end{cases}$$

$$L(\theta) = f_{X_1}(x_1, \theta) \cdot \cdots \cdot f_{X_n}(x_n, \theta) =$$

och eftersom $x \geq \theta \Rightarrow f_X(x) = 0$

$$= \begin{cases} \frac{1}{\theta^n} & \text{för } \theta \geq \max(x_1, x_2, \ldots, x_n) \\ 0 & \text{annars.} \end{cases}$$

$L(\theta)$ är en avtagande funktion av θ och detta ger (utan derivering)

$$\theta_{obs}^{ML} = \max(x_1, x_2, \ldots, x_n).$$

Korrigerad maximum likelihood i $U(0, \theta)$

$X_1, X_2, \ldots, X_n$, $X_i \in U(0, \theta)$. Maximum likelihood skattningen av θ är

$$\theta_{obs}^{ML} = \max(x_1, x_2, \ldots, x_n).$$

Man kan checka (Blom et.al. sid. 258) att den korrigerade skattningen

$$\frac{n+1}{n}\theta_{obs}^{ML}$$

är väntevärdesriktig, konsistent samt effektivare än minst-kvadrat skattningen.

KTH Matematik

Korrigerad maximum likelihood i $U(0,\theta)$ & the German tank problem (5)

Den korrigerade skattningen

$$\frac{n+1}{n}\theta_{obs}^{ML} = \max(x_1, x_2, \ldots, x_n) + \frac{\max(x_1, x_2, \ldots, x_n)}{n}$$

är ingenting annat än den ovan framlagda skattningen i 'the German tank problem'. MEN: vi har då gett oss in på en kontinuerlig approximation av det ursprungliga problemet, dvs. vi har approximerat en likformig sannolikhetsfunktion på heltalen $\{1, 2, \ldots, \theta\}$ med en likformig täthet på $(0, \theta)$. Detta låter sig göras om θ är stort.
Exakt räknat kan man visa för det diskreta problemet att

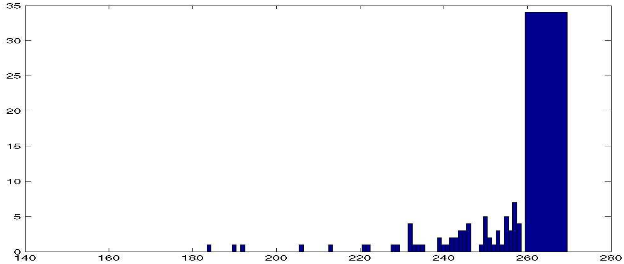
$$\theta_{obs}^{*} = \max(x_1, x_2, \ldots, x_n) + \frac{\max(x_1, x_2, \ldots, x_n)}{n} - 1$$

är en väntevärdesriktig och effektiv skattning.

KTH Matematik

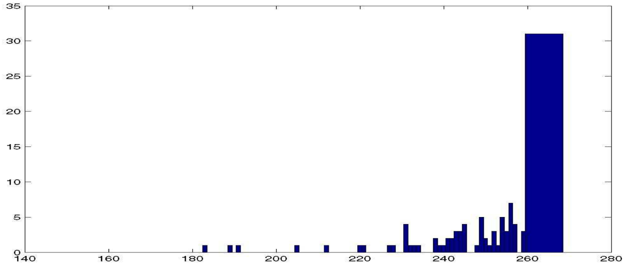
Simulerad korrigerad maximum likelihood & the German tank problem (6)

Vi drar hundra gånger tio värden $x_1, x_2, \ldots, x_{10}$ ur $1, 2, \ldots, \theta$ med $\theta = 245$ med hjälp av slumptalsgeneratorn *unidrnd* i MATLAB Statistics toolbox Sedan tar vi maximum $y_i = \max(x_1, x_2, \ldots, x_{10})$, $i = 1, \ldots, 100$ för vart och ett av dessa hundra sampel och bildar $\theta^*_{obs\ \text{nr}\ i} = y_i + y_i/10$. Histogrammet för de hundra θ^*_{obs} syns i bilden. $\overline{\theta^*_{obs}} = 250.64$.



Simulerad korrigerad maximum likelihood & the German tank problem (7)

Samma som ovan men med $\theta^*_{\underline{obs \text{ nr}} \; i} = y_i + y_i/10 - 1$. Histogrammet för de hundra θ^*_{obs} syns i bilden. $\overline{\theta^*_{obs}} = 249.65$.



German tank problem (8)

By using the formula $\theta^* = \theta^{ML} + (\theta^{ML}/n)$, statisticians reportedly estimated that the Germans produced 246 tanks per month between June 1940 and September 1942. At that time, standard intelligence estimates had believed the number was at around 1,400. After the war, the allies captured German production records of the Ministry, which was in charge of Germany's war production, showing that the true number of tanks produced in those three years was 245 per month, almost exactly what the statisticians had calculated, and less than one fifth of what standard intelligence had thought likely, and were more accurate and timely than Germany's own estimates.
Emboldened, the allies attacked the western front in 1944 and overcame the Panzers on their way to Berlin. And so it was that statisticians won the war - in their own estimation, at any rate.

KTH Matematik

German tank problem (9): lätt tillgängliga referenser

- G. Davies: How a statistical formula won the war.
  The Guardian, Thursday July 20 2006

  http://www.guardian.co.uk/world/2006/jul/20/secondworldwar.tvandradio

- Robert Matthews: Hidden truths.
  New Scientist 23 May 1998

  http://www.newscientist.com/article/mg15821355.000-hidden-truths.html

KTH Matematik

Definition

(1) $\underline{\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n.}$

Bilaga: räkneoperationer med summor

Sats

(2) $\underline{\sum_{i=1}^{n} a \cdot x_i = a \sum_{i=1}^{n} x_i.}$

Bevis. Definitionen (1) ger $\sum_{i=1}^{n} a \cdot x_i = ax_1 + ax_2 + \ldots + ax_n$
$= a\,(x_1 + x_2 + \ldots + x_n) = a \sum_{i=1}^{n} x_i.$
Exempel: $x_i = 1$, $i = 1 \ldots, n$
$\sum_{i=1}^{n} ax_i = a \sum_{i=1}^{n} 1 = a\,(1 + 1 + \ldots + 1) = a \cdot n.$

Bilaga: räkneoperationer med summor

Sats

(3) $\underline{\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.}$

Bevis. Definition (1) ger

$\sum_{i=1}^{n}(x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + \ldots + (x_n + y_n)$

$= x_1 + x_2 + \ldots + x_n + y_1 + y_2 + \ldots + y_n = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$

KTH Matematik

Bilaga: räkneoperationer med summor

Sats

(4) $\underline{\sum_{i=1}^{n}(ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i}$

Bevis. Detta fås av (3) och (2).

Sats

(5) $\underline{\sum_{i=1}^{n}(x_i + y_i)^2 = \sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2.}$

Bevis. Använd $(x_i + y_i)^2 = x_i^2 + 2x_i y_i + y_i^2$ och (4) samt (2) med $a = 2$.