Het schatten van de Duitse oorlogsproductie: maximum likelihood versus de momentenmethode

Rik Lopuhaä

TU Delft

30 januari, 2015

Inleiding

- Begin 1943: Economische Oorlogsafdeling van de Amerikaanse ambassade in Londen begint met het analyseren van *merktekens* en *serienummers* op buitgemaakt Duits oorlogsmateriaal
- **Doel:** beter inzicht te verkijgen in de Duitse oorlogsproductie (hoeveel, wanneer en waar) en oorlogsterkte
- Eerst *banden* van trucks, auto's en vliegtuigen
  Later *tanks, trucks, kanonnen, raketten*

- Maandcodes

	Jan	Feb	Mar	Apr	Mei	Jun	Jul	Aug	Sep	Okt	Nov	Dec
Dunlop	T	I	E	B	R	A	P	O	L	N	U	D
Fulda	F	U	L	D	A	M	U	N	S	T	E	R
Phoenix	F	O	N	I	X	H	A	M	B	U	R	G
Sempirit	A	B	C	D	E	F	G	H	I	J	K	L

- De serienummers werden gedecodeerd en vertaald naar een steekproef van n getallen uit $1, 2, \ldots, K$

  De *onbekende* K interpreteren we als het totaal aantal banden

- **Doel:** schat K op basis van de n buitgemaakte serienummers.

Achtergrond Literatuur

Richard Ruggles and Henry Brodie,
An Empirical Approach to Economic Intelligence in World War II,
Journal of the American Statistical Association, Vol **42**, nr 237 (March 1947), pp.72-91

Statistisch schattingsprobleem

- Gegeven een vaas met ballen genummerd $1, 2, \ldots, K$
- K is onbekend
- Trek willekeurig n ballen uit vaas
- Schat K op basis van de nummers op de getrokken ballen

Schattingsmethode 1

- Idee: het gemiddelde $\bar{x}$ van de getrokken nummers is ongeveer gelijk aan het gemiddelde van de nummers in de vaas

Schattingsmethode 1

- Idee: het gemiddelde $\bar{x}$ van de getrokken nummers is ongeveer gelijk aan het gemiddelde van de nummers in de vaas
- Gemiddelde van de nummers in de vaas is

$$\frac{1+2+\cdots+K}{K} = \frac{\frac{1}{2}K(K+1)}{K} = \frac{1}{2}(K+1)$$

Schattingsmethode 1

- Idee: het gemiddelde $\bar{x}$ van de getrokken nummers is ongeveer gelijk aan het gemiddelde van de nummers in de vaas
- Gemiddelde van de nummers in de vaas is

$$\frac{1+2+\cdots+K}{K} = \frac{\frac{1}{2}K(K+1)}{K} = \frac{1}{2}(K+1)$$

- Kortom $\bar{x} \approx \frac{K+1}{2}$ zodat $2\bar{x} - 1 \approx K$
- Conclusie: schat het aantal ballen in de vaas door

$$2\bar{x} - 1$$

Schattingsmethode 2

- Idee: de n getrokken nummers liggen min of meer gelijkmatig verspreid tussen 0 and $K+1$, zodat voor het grootste nummer m geldt:

$$m \approx \frac{n}{n+1} \times (K+1) \quad \text{zodat} \quad \frac{n+1}{n} \times m - 1 \approx N$$

- Conclusie: schat het aantal ballen in de vaas door

$$\frac{n+1}{n} \times \max - 1$$

Intermezzo

Intermezzo

Welke schatter is beter?

$$S_1 = 2\bar{x} - 1 \text{ of } S_2 = \frac{n+1}{n}\max - 1$$

A. Schatter S_1 is beter

B. Schatter S_2 is beter

C. Beide schatters zijn even goed

Momentenmethode

Stel dat stochastische variabele X een kansverdeling heeft met onbekende parameters en dat we op grond van waargenomen gegevens (steekproef) een schatting moeten geven voor de onbekende parameters.

Momentenmethode (Karl Pearson, 1857 - 1936)

- *Druk de momenten van X (zoals $\mathbb{E}[X]$, $\mathbb{E}[X^2]$, etc.) uit in de onbekende parameters*
- *Stel deze gelijk aan de corresponderende steekproefmomenten (zoals $\frac{1}{n}\sum x_i$, $\frac{1}{n}\sum x_i^2$, etc.)*
- *Los de onbekende parameter op uit de vergelijkingen*

Momentenmethode voor het schatten van het aantal ballen

- als X het nummer is op een willekeurig getrokken bal uit de volle vaas, dan is X een stochastische variabele met verwachting *(eerste moment van X)*

$$\begin{aligned}\mathbb{E}[X] &= 1 \times \mathbb{P}(X = 1) + 2 \times \mathbb{P}(X = 2) + \cdots + K \times \mathbb{P}(X = K) \\ &= 1 \times \frac{1}{K} + 2 \times \frac{1}{K} + \cdots + K \times \frac{1}{K} \\ &= \frac{1 + 2 + \cdots + K}{K} = \frac{1}{2}(K + 1)\end{aligned}$$

Momentenmethode voor het schatten van het aantal ballen

- als X het nummer is op een willekeurig getrokken bal uit de volle vaas, dan is X een stochastische variabele met verwachting *(eerste moment van X)*

$$\begin{aligned}\mathbb{E}[X] &= 1 \times \mathbb{P}(X = 1) + 2 \times \mathbb{P}(X = 2) + \cdots + K \times \mathbb{P}(X = K) \\ &= 1 \times \frac{1}{K} + 2 \times \frac{1}{K} + \cdots + K \times \frac{1}{K} \\ &= \frac{1 + 2 + \cdots + K}{K} = \frac{1}{2}(K + 1)\end{aligned}$$

- Stel $\bar{x}$ gelijk aan $\mathbb{E}[X]$ en los K op uit deze vergelijking

Maximum Likelihood
Inleiding: kiezen uit 2 dobbelstenen

- Twee dobbelstenen:
  - D_1 met 5 WIT en 1 ROOD
  - D_2 met 1 WIT en 5 ROOD
- Iemand kiest één van de twee dobbelsteen en doet drie keer hetzelfde experiment: *gooien tot ROOD boven komt*
- Informatie:
  - gekozen dobbelsteen onbekend
  - bekend zijn benodigde aantallen worpen in de drie experimenten

$$7, 4 \text{ en } 10$$

Maximum Likelihood
Inleiding: kiezen uit 2 dobbelstenen

- Twee dobbelstenen:
  - D_1 met 5 WIT en 1 ROOD
  - D_2 met 1 WIT en 5 ROOD
- Iemand kiest één van de twee dobbelsteen en doet drie keer hetzelfde experiment: *gooien tot ROOD boven komt*
- Informatie:
  - gekozen dobbelsteen onbekend
  - bekend zijn benodigde aantallen worpen in de drie experimenten

$$7, 4 \text{ en } 10$$

VRAAG: Met welke dobbelsteen is er gegooid?

Kiezen uit 2 dobbelstenen

- Kans op 7, 4 en 10 met dobbelsteen D_1:

$$\left(\frac{5}{6}\right)^6 \frac{1}{6} \times \left(\frac{5}{6}\right)^3 \frac{1}{6} \times \left(\frac{5}{6}\right)^9 \frac{1}{6} = \frac{5^{18}}{6^{21}} = 0.0001738937.$$

- Kans op 7, 4 en 10 met dobbelsteen D_2:

$$\left(\frac{1}{6}\right)^6 \frac{5}{6} \times \left(\frac{1}{6}\right)^3 \frac{5}{6} \times \left(\frac{1}{6}\right)^9 \frac{5}{6} = \frac{5^3}{6^{21}} = 5.7 \times 10^{-15}.$$

- Kans op 7, 4, en 10 is 5^{15} keer groter voor dobbelsteen D_1!
  De waargenomen data zijn het *meest waarschijnlijk* bij dobbelsteen D_1

Maximum Likelihood

Stel dat we op grond van waargenomen gegevens, kortweg data genoemd, een schatting moeten geven voor een onbekende parameter.

Het principe van Maximum Likelihood (Ronald A. Fisher, 1890-1962)

Volgens het principe van maximum likelihood nemen we als schatting die waarde van de onbekende parameter waarvoor de kans op de data het grootst is.

Maximum Likelihood schatting voor het aantal ballen

- Uit de vaas met nummers $1, 2, \ldots, K$ trekken we vijf ballen:

  $$40,\ 28,\ 7,\ 44 \text{ en } 18$$

  Wat is maximum likelihood schatting voor K?

Maximum Likelihood schatting voor het aantal ballen

- Uit de vaas met nummers $1, 2, \ldots, K$ trekken we vijf ballen:

  $$40,\ 28,\ 7,\ 44 \text{ en } 18$$

  Wat is maximum likelihood schatting voor K?

- De kans op de data *(Likelihood)*

$$L(N) = \begin{cases} 0 & \text{, voor } K = 1, 2, \ldots, 43; \\ \dfrac{1}{K(K-1)(K-2)(K-3)(K-4)} & \text{, voor } K = 44, 45, \ldots. \end{cases}$$

  De kans op de data is maximaal voor $K = 44$

Maximum Likelihood schatting voor het aantal ballen

- Uit de vaas met nummers $1, 2, \ldots, K$ trekken we vijf ballen:

  $$40,\ 28,\ 7,\ 44 \text{ en } 18$$

  Wat is maximum likelihood schatting voor K?

- De kans op de data *(Likelihood)*

$$L(N) = \begin{cases} 0 & \text{, voor } K = 1, 2, \ldots, 43; \\ \dfrac{1}{K(K-1)(K-2)(K-3)(K-4)} & \text{, voor } K = 44, 45, \ldots. \end{cases}$$

  De kans op de data is maximaal voor $K = 44$

- In het algemeen: de ML schatting is voor het aantal ballen het grootste nummer in de steekproef

Zuiverheidscorrectie van de ML schatter

- Als M het grootste nummer is van een steekproef zonder teruglegging uit de nummers $1, 2, \ldots, K$, dan is de verwachtingswaarde

$$\begin{aligned}\mathbb{E}[M] &= n \times \mathbb{P}(M = n) + \cdots + K \times \mathbb{P}(M = K) \\ &= \sum_{j=n}^{K} j \times \frac{\binom{j-1}{n-1}}{\binom{K}{n}} = \cdots = \frac{n}{n+1}(K+1)\end{aligned}$$

Zuiverheidscorrectie van de ML schatter

- Als M het grootste nummer is van een steekproef zonder teruglegging uit de nummers $1, 2, \ldots, K$, dan is de verwachtingswaarde

$$\begin{aligned}\mathbb{E}[M] &= n \times \mathbb{P}(M = n) + \cdots + K \times \mathbb{P}(M = K) \\ &= \sum_{j=n}^{K} j \times \frac{\binom{j-1}{n-1}}{\binom{K}{n}} = \cdots = \frac{n}{n+1}(K+1)\end{aligned}$$

- Kies constanten a and b (onafhankelijk van K) zodat $aM + b$ een zuivere schatter is voor K, d.w.z.

$$\mathbb{E}[aM + b] = K.$$

Dit levert

$$a = \frac{n+1}{n} \quad \text{en} \quad b = -1$$

Twee schattingsmethoden

Schatten van het aantal ballen in een vaas met nummers $1, 2, \ldots, K$:

- Momentenmethode-schatting:

$$s_1 = 2\bar{x} - 1$$

- Aangepaste maximum likelihood-schatting:

$$s_2 = \frac{n+1}{n} \max - 1$$

- Welke is nu beter?

Intermezzo

Intermezzo

Welke schatter is nu beter?

$$S_1 = 2\bar{x} - 1 \text{ of } S_2 = \frac{n+1}{n}\max -1$$

A. Schatter S_1 is beter, want de momentenmethode gebruikt alle gegevens

B. Schatter S_2 is beter, want maximum likelihood is een beter principe

Simulatie

We kiezen $K = 1000$ en $n = 10$ en voer uit op de computer

Stap 1 Trek 10 getallen zonder teruglegging uit $\{1, 2, \ldots, 1000\}$

Stap 2 Bereken

$$s_1 = 2\bar{x} - 1$$

$$s_2 = \frac{n+1}{n} \max -1$$

Stap 3 Herhaal 5000 keer stappen 1 en 2.

Intermezzo

Welke schatter is nu beter?

$$S_1 = 2\bar{x} - 1 \text{ of } S_2 = \frac{n+1}{n}\max - 1$$

A. Schatter S_1 is beter, want zijn kansverdeling is beter gespreid rond $K = 1000$

B. Schatter S_2 is beter, want zijn kansverdeling is scheef richting $K = 1000$

Intermezzo

Welke schatter is nu beter?

$$S_1 = 2\bar{x} - 1 \text{ of } S_2 = \frac{n+1}{n}\max - 1$$

A. Schatter S_1 is beter, want zijn kansverdeling is beter gespreid rond $K = 1000$

B. Schatter S_2 is beter, want zijn kansverdeling is scheef richting $K = 1000$

Nog wat theorie

- Beide schatters zijn zuiver:

$$\mathbb{E}[S_1] = \mathbb{E}[2\bar{X} - 1] = K$$
$$\mathbb{E}[S_2] = \mathbb{E}\left[\frac{n+1}{n}\max -1\right] = K$$

- Schatter S_1 heeft een grotere variantie dan schatter S_2:

$$\frac{\mathbb{V}(S_1)}{\mathbb{V}(S_2)} = \frac{n+2}{3}$$

Tabel: Gemiddelde maandelijkse productie banden in 1943.

Type band	schatting	werkelijk
Truck en auto	147 000	159 000
Vliegtuig	28 500	26 400
Totaal	175 500	186 100

Tabel: Gemiddelde maandelijkse productie banden in 1943.

Type band	schatting	werkelijk	geheime dienst
Truck en auto	147 000	159 000	
Vliegtuig	28 500	26 400	
Totaal	175 500	186 100	900 000 – 1 200 000

Tabel: Productie van trucks in 1942.

Type truck	schatting	werkelijk
Lichte truck	16 500	14 436
Medium truck	62 300	53 439
Zware truck	18 500	11 952
Totaal	97 300	79 827

Tabel: Productie van trucks in 1942.

Type truck	schatting	werkelijk	geheime dienst
Lichte truck	16 500	14 436	
Medium truck	62 300	53 439	
Zware truck	18 500	11 952	
Totaal	97 300	79 827	200 000

Tabel: Gemiddelde maandelijkse productie van tanks in 1940-1942.

Datum	schatting	werkelijk
Juni 1940	169	122
Juni 1941	244	271
Augustus 1942	327	342

Tabel: Gemiddelde maandelijkse productie van tanks in 1940-1942.

Datum	schatting	werkelijk	geheime dienst
Juni 1940	169	122	1000
Juni 1941	244	271	1550
Augustus 1942	327	342	1550

Schatten van de kans op zwangerschap

- Beschouw aantal cycli tot en met zwangerschap
- Als p is kans op zwangerschap tijdens een cyclus, dan is

$$P(\text{zwangerschap in } k\text{-de cyclus}) = (1-p)^{k-1}p, \quad \text{voor } k = 1, 2, \ldots$$

Schatten van de kans op zwangerschap

- Beschouw aantal cycli tot en met zwangerschap
- Als p is kans op zwangerschap tijdens een cyclus, dan is

$$P(\text{zwangerschap in } k\text{-de cyclus}) = (1-p)^{k-1}p, \quad \text{voor } k = 1, 2, \ldots$$

- Schat p, apart voor rokers en niet-rokers, aan de hand van de data

Aantal cycli	1	2	3	4	5	6	7	8	9	10	11	12	>12
Rokers	29	16	17	4	3	9	4	5	1	1	1	3	7
Niet-rokers	198	107	55	38	18	22	7	9	5	3	6	6	12

Maximum Likelihood schatting voor kans op zwangerschap

- Merk op

$$P(\text{zwangerschap na de 12-de cyclus}) = (1-p)^{12}.$$

Maximum Likelihood schatting voor kans op zwangerschap

- Merk op

$$P(\text{zwangerschap na de 12-de cyclus}) = (1-p)^{12}.$$

- Dan geldt

Gebeurtenis	Kans
29 keer zwangerschap in cyclus 1	p^{29}
16 keer zwangerschap in cyclus 2	$\{(1-p)p\}^{16}$
17 keer zwangerschap in cyclus 3	$\{(1-p)^2p\}^{17}$
$\vdots$	$\vdots$
7 keer zwangerschap na cyclus 12	$\{(1-p)^{12}\}^{7}$

Maximum Likelihood schatting voor kans op zwangerschap

- Merk op

$$P(\text{zwangerschap na de 12-de cyclus}) = (1-p)^{12}.$$

- Dan geldt

Gebeurtenis	Kans
29 keer zwangerschap in cyclus 1	p^{29}
16 keer zwangerschap in cyclus 2	$\{(1-p)p\}^{16}$
17 keer zwangerschap in cyclus 3	$\{(1-p)^2p\}^{17}$
$\vdots$	$\vdots$
7 keer zwangerschap na cyclus 12	$\{(1-p)^{12}\}^{7}$

- Zodat de likelihood (de kans op de data) wordt gegeven door

$$\begin{aligned} L(p) &= C \times p^{29} \times \{(1-p)p\}^{16} \times \{(1-p)^2p\}^{17} \times \cdots \times \{(1-p)^{12}\}^{7} \\ &= C \times p^{93} \times (1-p)^{322}. \end{aligned}$$

Maximum Likelihood schatting voor kans op zwangerschap

- Merk op

$$P(\text{zwangerschap na de 12-de cyclus}) = (1-p)^{12}.$$

- Dan geldt

Gebeurtenis	Kans
29 keer zwangerschap in cyclus 1	p^{29}
16 keer zwangerschap in cyclus 2	$\{(1-p)p\}^{16}$
17 keer zwangerschap in cyclus 3	$\{(1-p)^2p\}^{17}$
$\vdots$	$\vdots$
7 keer zwangerschap na cyclus 12	$\{(1-p)^{12}\}^{7}$

- Zodat de likelihood (de kans op de data) wordt gegeven door

$$\begin{aligned} L(p) &= C \times p^{29} \times \{(1-p)p\}^{16} \times \{(1-p)^2p\}^{17} \times \cdots \times \{(1-p)^{12}\}^{7} \\ &= C \times p^{93} \times (1-p)^{322}. \end{aligned}$$

- Oplossen van $L'(p) = 0$ geeft maximum likelihood schatting $p = 0.224$.

Maximum Likelihood schatter voor dalende kansdichtheid

- Men observeert $x_1, x_2, \ldots, x_n \in [0, \infty)$
- Realisaties van onafhankelijke stochasten met dalende kansdichtheid f.
- De (niet-parametrische) maximum likelihood schater voor f is de functie $\widehat{f}_n$ die de likelihood

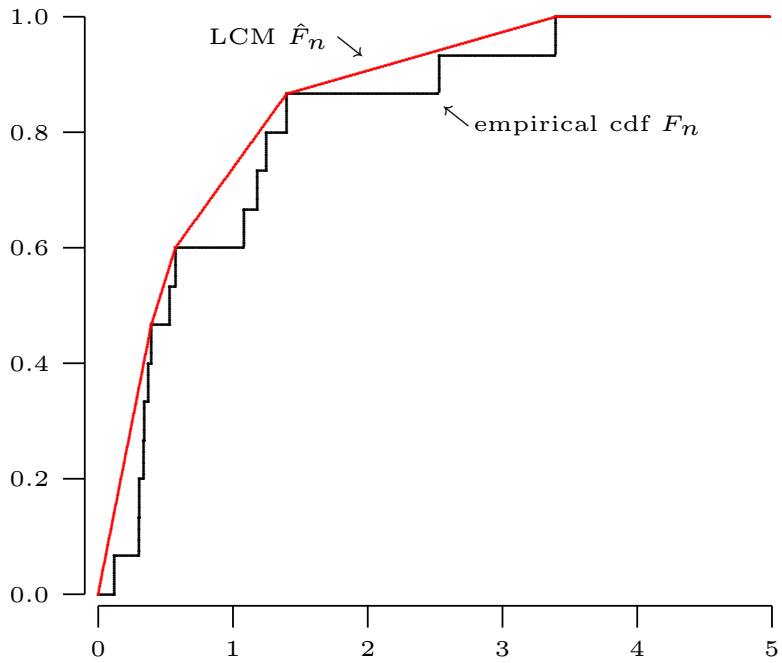
$$L(f) = \prod_{i=1}^{n} f(x_i)$$

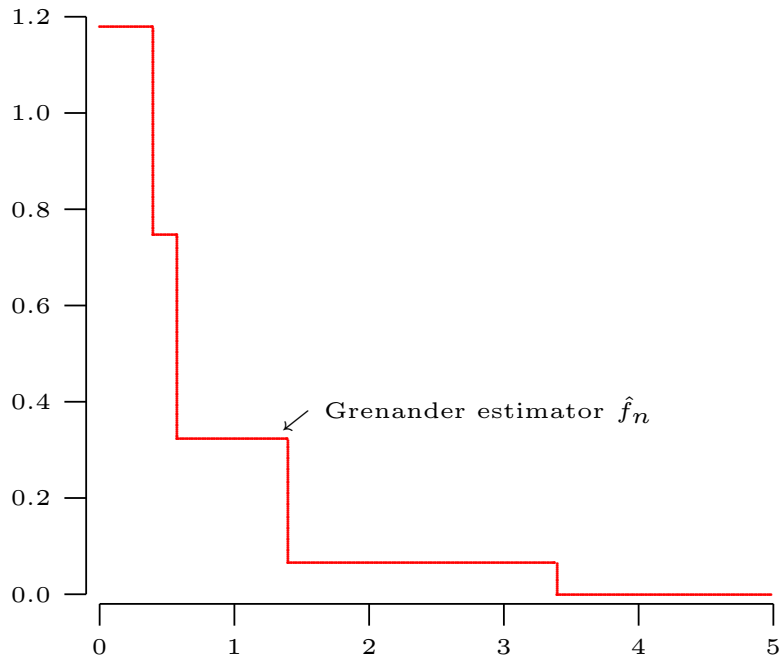
maximaliseert over alle dalende kansdichtheden f op $[0, \infty)$.

- *Grenander (1956):*

$\widehat{f}_n$ is de linker-afgeleide van de kleinste concave majorant van de empirische verdelingsfunctie

$$F_n(t) = \frac{\text{aantal } x_i \leq t}{n}$$


1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
LCM $\hat{F}_n$
empirical cdf F_n


1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
Grenander estimator $\hat{f}_n$

Hartelijke dank voor
uw aandacht