75570 TEORETISK STATISTIKK - VÅREN 1995

BOOTSTRAPPING

Bo Lindqvist

1 Bootstrap-estimator for varians

1.1 Estimering av variansen til en estimator

La $\boldsymbol{X} = (X_1, \ldots, X_n)$ være u.i.f. observasjoner av en stokastisk variabel X med sannsynlighetstetthet/punktsannsynlighet $f(x|\theta)$, der θ er den ukjente parameter (muligens en vektor). Anta at $\tau(\theta)$ (en reell funksjon av θ) skal estimeres.

La $T_n \equiv T_n(\boldsymbol{X})$ være en estimator for $\tau(\theta)$. I dette kapitlet skal vi gå ut fra at T_n er konsistent (f.eks. er T_n en MLE eller momentestimator). For å kunne si noe om usikkerheten i estimatet, er det av interesse å kjenne *variansen* $\mathrm{Var}_\theta T_n$, eller eventuelt et estimat $\widehat{(\mathrm{Var}_\theta T_n)}$ for denne. Det finnes nå flere muligheter. La i det følgende $\hat{\theta}$ være en konsistent estimator for θ, f.eks. MLE.

(i) Anta at vi kan regne ut $\mathrm{Var}_\theta T_n$ eksakt som funksjon av θ. Da kan vi sette inn $\hat{\theta}$ for θ, dvs. bruke estimatoren

$$\widehat{(\mathrm{Var}_\theta T_n)} = \mathrm{Var}_{\hat{\theta}} T_n$$

For eksempel kan $X_1, \ldots, X_n$ være u.i.f. Bernoulli(θ) og $\tau(\theta) = \theta$. Det er da naturlig å bruke $T_n = \sum X_i/n$, slik at $\mathrm{Var}_\theta T_n = \frac{\theta(1-\theta)}{n}$ kan beregnes eksakt. Vi får da estimatet

$$\widehat{(\mathrm{Var}_\theta T_n)} = \frac{\hat{\theta}(1-\hat{\theta})}{n}$$

(ii) Hvis (i) ikke er gjennomførbart, kan vi prøve å finne asymptotisk fordeling for T_n, f.eks. gitt ved

$$\sqrt{n}(T_n - \tau(\theta)) \xrightarrow{d} N(0, J^2(\theta))$$

Da vil $\mathrm{Var}_\theta T_n \approx J^2(\theta)/n$ for store n og vi kan estimere ved

$$\widehat{(\mathrm{Var}_\theta T_n)} = \frac{J^2(\hat{\theta})}{n}$$

der vi altså har satt inn $\hat{\theta}$ for θ i J^2-funksjonen.

Dersom θ er endimensjonal og $\hat{\theta}$ er MLE, vet vi fra et eget notat at

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \frac{1}{I_1(\theta)})$$

der $I_1(\theta)$ er Fisher-informasjonen i én observasjon. Som anvist i Oppgave 1(e) i Øving 7 kan $\mathrm{Var}_\theta \hat{\theta}$ estimeres ved enten

$$\frac{1}{nI_1(\hat{\theta})}$$

som finnes ved å sette inn estimatet $\hat{\theta}$ i $I_1(\theta)$, eller ved

$$\frac{1}{n\widehat{I_1(\theta)}} \tag{1}$$

der $\widehat{I_1(\theta)}$ er den såkalte *observerte Fisher-informasjonen* (se boka side 326), som består i at forventningsverdien som definerer $I_1(\theta)$, dvs.

$$I_1(\theta) = -\mathrm{E}[\frac{d^2}{d\theta^2} \ln f(X|\theta)]$$

estimeres ved et vanlig gjennomsnitt basert på $X_1, \ldots, X_n$, og der øvrige θer som måtte forekomme erstattes med $\hat{\theta}$. Estimatoren i nevneren i (1) ser dermed slik ut:

$$\widehat{I_1(\theta)} = -\frac{1}{n}\sum_{i=1}^{n} \frac{d^2}{d\theta^2} \ln f(X_i|\theta)|_{\theta=\hat{\theta}}$$

(iii) Hvis n er for liten til at asymptotiske tilnærminger er gode, kan vi bruke *Jackknife-estimatoren*, $(\widehat{\mathrm{Var}_\theta T_n})_{\mathrm{JACK}}$, som er behandlet i et eget notat.

(iv) Alternativt kan vi bruke ulike varianter av den såkalte *bootstrap* estimatoren for variansen til T_n, $(\widehat{\mathrm{Var}_\theta T_n})_{\mathrm{BOOT}}$, som er temaet for dette kapitlet.

1.2 Stokastisk simulering for beregning av varians

La oss først se på følgende ideelle situasjon. Anta at vi kjenner den sanne verdi θ_0 av θ. Da har vi fullstendig kjennskap til sannsynlighetsfordelingen til observasjonene X_i, den er nemlig gitt ved tettheten $f(x|\theta_0)$. (Det forutsettes at vi bruker den riktige parametriske modellen $f(\cdot|\cdot)$.) I prinsippet kan vi da regne ut $\mathrm{Var}_{\theta_0} T_n$, og dersom vi også kan gjennomføre det i praksis, er vi ferdige. Anta imidlertid at dette ikke lar seg gjøre analytisk. Siden fordelingen for Xene er kjent, kan variansen da approksimeres med så stor nøyaktighet vi ønsker ved hjelp av *stokastisk simulering*, også kalt *Monte Carlo* simulering.

Ideen ved en stokastisk simulering er at man ved hjelp av en *random generator* simulerer uavhengige ("tenkte") realisasjoner av X ved hjelp av en datamaskin (jfr. simulering av spesielle fordelinger i MINITAB elller S-PLUS). Gitt en slik mulighet for å generere

observasjoner av X, kan vi nå også generere realisasjoner av $T_n(\boldsymbol{X})$. Dette gjøres ved først å generere $\boldsymbol{X} = (X_1, \ldots, X_n)$ ved å trekke n uavhengige verdier for X, og så ganske enkelt regne ut $T_n(\boldsymbol{X})$. Et anslag for $\mathrm{Var}_{\theta_0} T_n$ får vi ved på denne måten å generere et (stort) antall k realisasjoner av T_n og beregne den empiriske varians for disse.

Skjematisk kan vi sette opp dette slik, idet $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(k)}$ er de simulerte realisasjoner av vektoren $\boldsymbol{X}$.

$$\begin{array}{lll} \boldsymbol{X}^{(1)} = (X_1^{(1)}, \ldots, X_n^{(1)}) & \text{gir} & T_n^{(1)} \equiv T_n(\boldsymbol{X}^{(1)}) \\ \boldsymbol{X}^{(2)} = (X_1^{(2)}, \ldots, X_n^{(2)}) & \text{gir} & T_n^{(2)} \equiv T_n(\boldsymbol{X}^{(2)}) \\ \vdots & \vdots & \vdots \\ \boldsymbol{X}^{(k)} = (X_1^{(k)}, \ldots, X_n^{(k)}) & \text{gir} & T_n^{(k)} \equiv T_n(\boldsymbol{X}^{(k)}) \end{array}$$

La nå

$$T_n^{(\cdot)} = \frac{1}{k} \sum_{i=1}^{k} T_n^{(i)}$$

Anslaget for $\mathrm{Var}_{\theta_0} T_n$ er da gitt ved den empiriske varians for de simulerte verdier $T_n^{(i)}$, nemlig:

$$\frac{1}{k-1} \sum_{i=1}^{k} (T_n^{(i)} - T_n^{(\cdot)})^2 \tag{2}$$

Når $k \to \infty$ vil dette konvergere mot den eksakte verdi for $\mathrm{Var}_{\theta_0} T_n$. Dette følger av at empirisk varians er en konsistent estimator for varians. (Vi forutsetter her at variansen virkelig eksisterer, og at vår random generator virker tilfredsstillende). Siden vi selv kan velge k, kan vi estimere $\mathrm{Var}_{\theta_0} T_n$ med så stor nøyaktighet vi ønsker. Merk at n hele tiden holdes fast, lik den aktuelle utvalgsstørrelsen.

1.3 Bootstrapping

I praksis er θ ukjent (ellers hadde vi ikke noe estimeringsproblem!) *Bootstrap-prinsippet* går ut på at vi gjennomgår det samme programmet som ovenfor, men med den ovenfor *kjente* sannsynlighetsfordelingen $f(x|\theta_0)$ erstattet med en sannsynlighetsfordeling som er *estimert ut fra observasjonene* $\boldsymbol{x} = (x_1, \ldots, x_n)$. Det er ulike måter å gjøre dette på. I hovedsak skilles det mellom *parametrisk* og *ikke-parametrisk* bootstrap.

Parametrisk bootstrap

Her estimerer man først θ ut fra de opprinnelige observasjonene $\boldsymbol{x} = (x_1, \ldots, x_n)$. La oss kalle estimatoren $\hat{\theta}$. Denne kan f.eks. være maximum likelihood estimatoren (MLE). Deretter går vi fram nøyaktig som i tilfellet med kjent θ_0, bare at θ_0 er erstattet med estimatet $\hat{\theta}$.

Siden vi nå ikke trekker fra den virkelige underliggende fordeling $f(x|\theta_0)$, men fra den estimerte fordeling $f(x|\hat{\theta})$, er det vanlig å sette en * på de genererte observasjonene av $\boldsymbol{X}$,

T_n, etc. Vi får følgende skjema, der altså $\boldsymbol{X}^{*(1)}, \ldots, \boldsymbol{X}^{*(k)}$ er de simulerte realisasjoner av vektoren $\boldsymbol{X}^* = (X_1^*, \ldots, X_n^*)$ med komponenter trukket fra fordelingen $f(x|\hat{\theta})$.

$$\begin{array}{lll} \boldsymbol{X}^{*(1)} = (X_1^{*(1)}, \ldots, X_n^{*(1)}) & \text{gir} & T_n^{*(1)} \equiv T_n(\boldsymbol{X}^{*(1)}) \\ \boldsymbol{X}^{*(2)} = (X_1^{*(2)}, \ldots, X_n^{*(2)}) & \text{gir} & T_n^{*(2)} \equiv T_n(\boldsymbol{X}^{*(2)}) \\ \vdots & \vdots & \vdots \\ \boldsymbol{X}^{*(k)} = (X_1^{*(k)}, \ldots, X_n^{*(k)}) & \text{gir} & T_n^{*(k)} \equiv T_n(\boldsymbol{X}^{*(k)}) \end{array}$$

La nå

$$T_n^{*(\cdot)} = \frac{1}{k} \sum_{i=1}^{k} T_n^{*(i)}$$

Vi definerer da

$$\widehat{(\mathrm{Var}_\theta T_n)}_{\mathrm{BOOT}} = \frac{1}{k-1} \sum_{i=1}^{k} (T_n^{*(i)} - T_n^{*(\cdot)})^2 \tag{3}$$

Praksis har vist at antallet k av Bootstrap-utvalg $\boldsymbol{X}^{*(i)}$ ikke behøver å være så stort for å få et brukbart variansestimat. Vanlig brukte verdier er fra 50 til 200, men helt ned til 20 kan fungere bra.

Merk at selv om $k \to \infty$ i (3), vil høyresiden ikke konvergere mot $\mathrm{Var}_{\theta_0} T_n$ (der θ_0 er den sanne verdi på parameteren), men mot estimatet $\mathrm{Var}_{\hat{\theta}} T_n$, som blir en funksjon av de opprinnelige observasjonene $x_1, \ldots, x_n$. Dette kommer selvsagt av at "observasjonene"i $\boldsymbol{X}^*$ er trukket fra fordelingen $f(x|\hat{\theta})$. Merk også at n hele tiden er fast, lik utvalgsstørrelsen i vårt opprinnelige utvalg.

Estimatet i (3) vil p.g.a. bootstrap-trekningen variere dersom vi går gjennom det samme programmet flere ganger, selv om k ikke endres. For å få en entydig definisjon av $\widehat{(\mathrm{Var}_\theta T_n)}_{\mathrm{BOOT}}$ må vi la $k \to \infty$ i (3). Den presise definisjonen av bootstrap-estimatoren for varians er dermed

$$\widehat{(\mathrm{Var}_\theta T_n)}_{\mathrm{BOOT}} = \mathrm{Var}_{\hat{\theta}} T_n \tag{4}$$

Merk at dette er det samme estimatet som vi foreslo i (i) på side 1, for tilfellet der $\mathrm{Var}_\theta T_n$ kunne beregnes analytisk. I praksis må vi imidlertid som regel bruke Monte Carlo estimatet (3).

Eksempel

Vi studerer det samme eksempel som i notatet om Jackknifing (hentet fra Oppgave 1 i øving 7, våren 1995).

Her er $X_1, \ldots, X_n$ u.i.f. $\mathrm{N}(\theta, \theta^2)$, med θ som ukjent parameter og $\tau(\theta) = \theta$ som skal estimeres. Vi har $n = 10$ observasjoner gitt ved:

22.2	10.0	8.3	25.3	15.1
8.8	16.6	18.0	−15.9	−4.5

MLE $\hat{\theta}$ for θ faller da sammen med estimatoren T_n for $\tau(\theta)$, og er gitt ved

$$\hat{\theta} = T_n = \frac{-\bar{X}_n \pm \sqrt{(\bar{X}_n)^2 + 4Y_n}}{2} \tag{5}$$

der

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

og $+(-)$ velges foran rottegnet hvis $\bar{X}_n > (< 0)$. Med de gitte dataene får vi

$$\hat{\theta} = T_n = 11.387$$

Ved parametrisk bootstrapping må vi trekke nye sett $\boldsymbol{X}^* = (X_1^*, \ldots, X_n^*$ fra fordelingen $N(T_n, T_n^2)$, dvs. $N(11.387, 11.387^2)$, regne ut T_n^* for hvert sett, og til slutt beregne den empiriske varians for de funne T_n^*.

Tabell 1 på neste side gir resultater for $k = 20$ bootstrap-utvalg. Den empiriske varians for tallene i høyre kolonne danner nå estimatet for variansen, nemlig

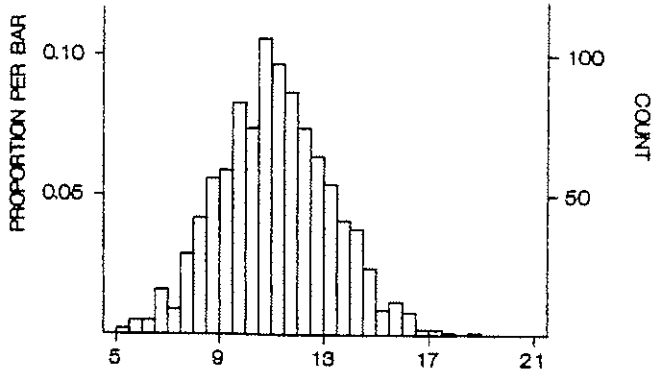
$$(\widehat{\mathrm{Var}_{\hat{\theta}} T_n})_{\mathrm{BOOT}} = 3.646$$

Ved å øke k til 50 fikk vi estimatet 3.625, og ved $k = 1000$ fikk vi estimatet 4.53, som må antas å være det beste av de tre anslagene for bootstrap-estimatoren (som altså ideelt sett skal være basert på $k = \infty$).

Histogrammet for de beregnede verdier av T_n^* i tilfellet $k = 1000$ er gjengitt i Figur 1. Siden histogrammer approksimerer de underliggende tettheter, vil dette være en approksimasjon til fordelingen for T_n for parameterverdien $\theta = \hat{\theta}$. Da $\hat{\theta}$ er vårt estimat for den underliggende virkelige parameter θ_0, vil vi samtidig oppfatte histogrammet som et (bootstrap) estimat for den virkelige fordelingen for T_n under θ_0. Som figuren viser, virker denne å være rimelig Gaussisk.

l	X_1^*	X_2^*	X_3^*	X_4^*	X_5^*	X_6^*	X_7^*	X_8^*	X_9^*	X_{10}^*	$T_n^{*(l)}$
1	17.4	24.2	18.9	-0.8	3.5	14.2	16.4	12.0	-0.5	9.1	9.39
2	-3.8	14.6	35.6	17.7	17.2	22.5	-11.2	21.7	20.3	-9.8	14.03
3	3.2	4.9	2.0	0.2	14.0	15.4	5.8	10.5	34.0	9.2	9.54
4	5.9	4.3	25.8	11.0	-4.3	4.7	15.1	1.8	23.0	28.7	10.93
5	12.4	9.3	23.7	17.3	2.2	18.8	-10.3	11.8	13.6	24.6	10.79
6	-7.2	14.7	31.5	14.7	4.6	8.4	10.3	6.4	5.8	10.7	9.53
7	24.5	21.2	17.1	29.3	5.1	15.1	0.9	4.2	13.5	-2.8	11.13
8	-0.8	-2.8	27.3	20.2	2.8	14.9	2.8	17.1	26.3	17.2	11.27
9	17.7	29.7	8.6	-7.3	13.4	13.2	30.0	15.4	31.0	21.5	13.67
10	20.9	44.9	10.8	19.0	7.2	-0.9	13.7	2.5	15.4	7.6	12.87
11	13.8	-4.5	6.4	1.8	9.9	13.7	5.2	9.8	18.3	21.3	8.19
12	17.0	-6.1	2.9	34.6	11.4	19.1	-7.1	29.9	-0.2	19.6	13.23
13	-9.2	11.4	8.3	11.8	20.1	3.1	19.2	-3.1	14.2	-3.2	8.88
14	4.8	5.8	17.4	9.1	26.6	9.5	30.9	22.8	13.1	-8.7	11.85
15	-1.6	-4.2	-12.9	16.9	1.0	30.8	3.6	1.5	20.3	16.0	11.37
16	14.8	-12.3	3.3	-1.4	16.9	17.1	-8.5	-15.6	7.4	8.5	10.43
17	16.2	-5.7	-6.2	11.1	17.8	10.6	4.5	27.5	6.2	26.8	11.05
18	10.2	8.2	12.3	8.1	-7.5	4.9	6.0	-8.8	22.4	13.1	8.28
19	17.4	20.2	1.4	22.9	8.1	22.6	13.0	-23.5	35.2	25.8	15.10
20	-1.2	-5.4	24.3	-4.8	14.4	13.7	-11.1	15.2	24.6	2.4	11.00

Tabell 1: De k=20 bootstraputvalgene fra $N(11.387, 11.387^2)$.



Figur 1: Histogram for T_n^* basert på $k = 1000$ parametriske bootstrap-utvalg.

"Vanlig" (ikke-parametrisk) bootstrap

Her genereres de nye datasettene $\boldsymbol{X}^{*(l)} = (X_1^{*(l)}, \ldots, X_n^{*(l)})$ $(l = 1, \ldots, k)$ fra den såkalte *empiriske fordeling* basert på de opprinnelige observasjonene $x_1, \ldots, x_n$.

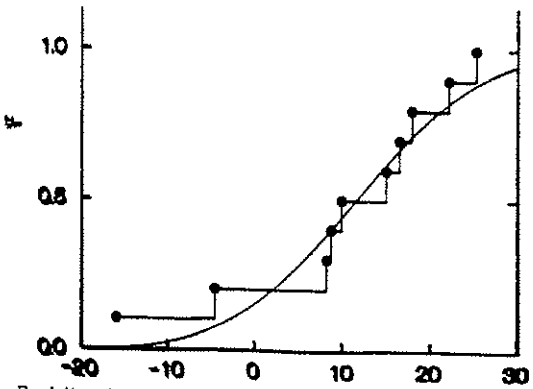
Den empiriske fordelingen er definert som en diskret fordeling med mulige verdier gitt ved de opprinnelige observasjonene $x_1, \ldots, x_n$ og med sannsynlighet $\frac{1}{n}$ for hver. Det er igjen vanlig å bruke X^* som navn på en stokastisk variabel med denne fordelingen. Skjematisk kan vi skrive denne fordelingen som

Mulige verdier for X^*	x_1	x_2	...	x_n
$P(X^* = \cdot)$	$\frac{1}{n}$	$\frac{1}{n}$	...	$\frac{1}{n}$

dvs. at vi har en *uniform* sannsynlighetsmodell (se Høyland, bind I), med utfallsrom $\{x_1, \ldots, x_n\}$.

Det er nå lett å tegne opp fordelingsfunksjonen for X^*. Dette gjøres ved først å avsette de n observasjonene $x_1, \ldots, x_n$ på x-aksen, og så tegne kurven som starter i 0 i $-\infty$ og hopper opp en høyde $\frac{1}{n}$ i hvert punkt x_i. Det kan vises at denne kurven, som vanligvis kalles $F_n(x)$, vil konvergere mot den virkelige fordelingsfunksjonen for X når $n \to \infty$. Dette viser at når n blir stor, vil trekking av observasjoner fra X^* tilnærmet svare til trekning fra den virkelige (ukjente) fordeling for X-ene (tidligere kalt $f(x|\theta_0)$.)

Figur 2 viser den empiriske fordelingsfunksjon for dataene i eksemplet ovenfor, tegnet sammen med fordelingsfunksjonen for $N(11.387, 11.387^2)$ som ble brukt i parametrisk bootstrapping. Forskjellen mellom parametrisk og ikke-parametrisk bootstrap kan nå kort beskrives ved at førstnevnte metode går ut på å simulere data X^* fra Gauss-kurven i Figur 2, mens den siste går ut på å simulere fra trappekurven.



Figur 2: Fordelingsfunksjoner for trekking av bootstraputvalg. Kontinuerlig kurve for parametrisk bootstrap ($N(11.387, 11.387^2)$) og trappefunksjon for den empiriske fordelingsfunksjon basert på de opprinnelige dataene.

Hvordan genererer man så vektorene $\boldsymbol{X}^{*(l)}$ $(l = 1, \ldots, k)$ fra den empiriske fordelingen? Jo, siden X^* har en uniform sannsynlighetsfordeling over $\{x_1, \ldots, x_n\}$, simuleres uavhengige realisasjoner av denne ganske enkelt ved tilfeldig trekking *med* tilbakelegging fra mengden $\{x_1, \ldots, x_n\}$. En vektor $\boldsymbol{X}^{*(l)}$ simuleres derfor ved å trekke n ganger med tilbakelegging fra denne mengden.

Når dette er gjort, og vi har beregnet k verdier $T_n^{*(1)}, \ldots, T_n^{*(k)}$, gir formel (3) variansestimatet $(\widehat{\mathrm{Var}T_n})_{\mathrm{BOOT}}$.

Som for den parametriske bootstrap vil den presise definisjonen av bootstrap-estimatoren for varians være grensen for (3) når $k \to \infty$. Siden X^*ene nå trekkes fra den empiriske fordelingen for $x_1, \ldots, x_n$ istedenfor fra $f(x|\hat{\theta})$, vil vi endre (4) til

$$(\widehat{\mathrm{Var}_* T_n})_{\mathrm{BOOT}} = \mathrm{Var}_*(T_n^*) \tag{6}$$

der $*$ som indeks på Var betegner at variansen er beregnet med hensyn på bootstrapfordelingen (her den empiriske fordelingen over $\{x_1, \ldots, x_n\}$) og $T_n^* = T_n(\boldsymbol{X}^*)$. Merk at dette er den vanlige måten å angi bootstrap-estimatoren på også i det parametriske tilfellet. I det følgende vil en $*$ i f.eks. $\mathrm{E}_*, P_*, \mathrm{Var}_*$ bety at forventningen, sannsynligheten eller variansen er beregnet for et bootstrap-utvalg $\boldsymbol{X}^*$.

Eksempel (forts.)

Vi har trukket ut $k = 20$ utvalg (med tilbakelegging) på $n = 10$ fra mengden av de 10 observerte verdier. Dette kan selvsagt tenkes gjort ved at vi har plassert 10 lapper med de gitte tallene i en hatt, og så trukket tilfeldig med tilbakelegging. Isteden har vi selvsagt brukt en random generator i et dataprogram. (Såvel MINITAB som S-PLUS egner seg her.)

Det første utvalget ble

5, 3, 9, 2, 10, 3, 5, 6, 2, 3,

slik at vektoren $\boldsymbol{X}^{*(1)}$ ble

(15.1, 8.3, -15.9, 10.0, -4.5, 8.3, 15.1, 8.8, 10.0, 8.3)

og den beregnede verdi $T_n^{*(1)} \equiv T_n(\boldsymbol{X}^{*(1)}) = 8.29$.

Resultatet av alle de 20 simulerte settene er gitt i tabellen på neste side.

Det ikke-parametriske bootstrapestimatet for variansen til T_n framkommer nå ved å beregne den empiriske varians for verdiene i den høyre kolonnen. Dette blir 3.712. Vi økte så antallet bootstrap-utvalg til $k = 50$ og fikk estimatet 2.990. Endelig lot vi $k = 1000$ og fikk variansestimatet 2.452 (som er det vi vil notere som bootstrap-estimatet i dette tilfellet). Histogrammet for de 1000 verdiene av T_n^* er gitt i Figur 3. Det kan ses at spredningen er noe mindre enn i Figur 1.

Vi står da med følgende estimerte varianser:

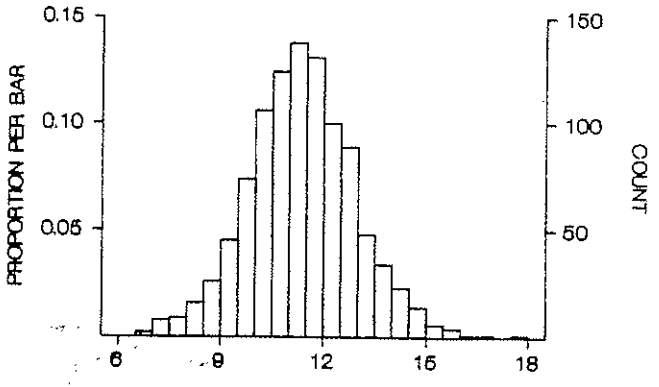
Estimeringsmetode	Estimator	Estimat
Cramer-Rao (ML-teori)	$\theta^2/30$	4.32
Jackknife	Se notat	2.32
Par. bootstrap	(3), k=1000	4.53
Ikkepar. bootstrap	(3), k=1000	2.45

Resultatene viser at variansestimatet ved Cramer-Rao's nedre grense og ved parametrisk bootstrap er forholdsvis like, mens jackknife og ikke-parametrisk bootstrap også kommer forholdsvis likt ut. Det er imidlertid litt sprik mellom de to gruppene, noe som antageligvis kan forklares som følger. Vi merker oss nemlig at de to estimatene som eksplisitt bruker normalfordeling ved beregningen (Cramer-Rao og parametrisk bootstrap) er nær hverandre, mens det samme gjelder de to som ikke bruker fordelinger i det hele tatt ved beregningen (Jackknife og bootstrap). Forskjellen i estimatene kan dermed komme av at vårt utvalg på 10 observasjoner tilfeldigvis var litt "ikke-normalt".

Det er også verdt å merke seg at med bare 10 observasjoner vil estimater for varians naturlig ha en stor grad av usikkerhet.

l	X_1^*	X_2^*	X_3^*	X_4^*	X_5^*	X_6^*	X_7^*	X_8^*	X_9^*	X_{10}^*	$T_n^{*(l)}$
1	15.1	8.3	-15.9	10.0	-4.5	8.3	15.1	8.8	10.0	8.3	8.29
2	10.0	8.8	8.3	16.6	25.3	10.0	15.1	22.2	-4.5	15.1	9.88
3	25.3	-15.9	22.2	15.1	-15.9	-15.9	15.1	25.3	10.0	-4.5	14.85
4	25.3	18.0	15.1	25.3	22.2	22.2	25.3	8.8	-4.5	22.2	13.04
5	16.6	-4.5	8.8	-15.9	8.3	22.2	22.2	-4.5	16.6	25.3	12.11
6	10.0	10.0	10.0	8.3	18.0	16.6	22.2	16.6	8.8	-4.5	8.93
7	10.0	-15.9	8.8	-15.9	25.3	8.8	-4.5	15.1	-4.5	-4.5	11.91
8	10.0	22.2	-4.5	18.0	8.8	15.1	8.3	18.0	8.3	16.6	9.27
9	10.0	-4.5	8.8	15.1	22.2	25.3	8.3	8.3	8.3	18.0	9.66
10	22.2	10.0	10.0	8.8	16.6	10.0	-15.9	10.0	-15.9	8.8	10.67
11	16.6	25.3	18.0	-15.9	18.0	25.3	25.3	-15.9	8.3	15.1	14.01
12	-15.9	25.3	10.0	25.3	18.0	22.2	15.1	-15.9	10.0	10.0	13.24
13	22.2	16.6	25.3	25.3	16.6	15.1	16.6	22.2	25.3	8.3	12.61
14	-4.5	22.2	18.0	16.6	10.0	22.2	16.6	16.6	25.3	10.0	11.28
15	-15.9	-15.9	10.0	8.8	25.3	18.0	25.3	10.0	15.1	-4.5	12.87
16	16.6	-15.9	10.0	8.8	16.6	25.3	22.2	16.6	10.0	18.0	11.55
17	8.3	18.0	15.1	10.0	10.0	8.3	22.2	8.8	15.1	8.3	8.43
18	8.3	10.0	-15.9	25.3	15.1	10.0	18.0	8.3	22.2	8.3	10.78
19	-15.9	16.6	25.3	-15.9	-4.5	22.2	16.6	22.2	-4.5	10.0	13.58
20	16.6	8.3	22.2	-4.5	25.3	15.1	10.0	-4.5	18.0	8.8	10.30

Tabell 2: De k=20 bootstraputvalgene fra den empiriske fordeling.



Figur 3: Histogram for T_n^* basert på $k = 1000$ ikke-parametriske bootstrap-utvalg.

2 Bootstrap konfidensintervaller

La situasjonen være som i Kapitel 1. For enkelthetsskyld antar vi nå at θ er endimensjonal og at det er θ selv som skal estimeres, med $\hat{\theta}$ som en gitt konsistent estimator.

Generelt kan det imidlertid godt være flere ukjente parametre i modellen, hvor vi skal estimere en funksjon $\tau(\theta)$. For enkelthets skyld vil vi bare se på $\tau(\theta) = \theta$ her.

Som i kapitel 1 lager vi bootstrap-simuleringer av $\hat{\theta}$, $\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(k)}$, der k er et valgt (stort) tall. For konfidensintervaller bør dette være større enn for variansestimering, f.eks. $k = 1000$ eller mer.

De simulerte verdier $\hat{\theta}^{*(l)}$ sorteres nå i stigende rekkefølge. Vi vil anta i det følgende at

$$\hat{\theta}^{*(1)} \leq \hat{\theta}^{*(2)} \leq \ldots \leq \hat{\theta}^{*(k)}$$

Persentilmetoden

Anta at vi ønsker et konfidensintervall for θ med konfidenskoeffisient $1 - \alpha$. *Persentil-intervallet* er nå gitt ved

$$[\hat{\theta}^{*(L)}, \hat{\theta}^{*(U)}]$$

der

$$L = [\frac{\alpha}{2}k] \text{ og } U = [(1 - \frac{\alpha}{2})k]$$

($[a]$ betyr heltallsdelen av a, dvs. f.eks. $[5.37] = 5$).

Med ord betyr dette at vi kutter ut andelen $\alpha/2$ av de laveste og $\alpha/2$ av de høyeste verdiene for $\hat{\theta}^*$, og lar konfidensintervallet være det intervallet av verdier som da gjenstår.

Hvis vi ser på et histogram for $\hat{\theta}^{*(1)}, \hat{\theta}^{*(2)}, \ldots, \hat{\theta}^{*(k)}$ (Fig. 4), vil arealene på henholdsvis venstre side av $\hat{\theta}^{*(L)}$ og høyre side av $\hat{\theta}^{*(U)}$ begge begge være (ca.) $\alpha/2$.

Eksempel (forts.)

Med $k = 1000$ simulerte verdier for $\hat{\theta}^*$, og $\alpha = 0.10$ får vi $U = 50$ og $L = 950$. Ved å sortere dataene som ligger til grunn for histogrammet i Figur 4 (dvs. den parametriske bootstrappingen fra Figur 1), får vi

$$\hat{\theta}^{*(50)} = 7.81 \text{ og } \hat{\theta}^{*(950)} = 14.71$$

Et 90% persentil-intervall for θ blir dermed $[7.81, 14.71]$.

Til sammenligning vil asymptotisk teori gi *standard-intervallet*

$$\hat{\theta} \pm 1.645 \cdot \sqrt{\widehat{Var}_\theta \hat{\theta}}$$

som med Cramer-Rao (MLE) variansestimatet 4.32 (se tidligere tabell) gir intervallet 11.39 ± 3.42, dvs. $[7.97, 14.81]$. Dette er ganske nært persentilintervallet.

BC-intervallet

Dette er en forbedret metode for bootstrap-konfidensintervaller, (BC="Bias Corrected").

Man estimerer først "bias-konstanten" z_0 fra relasjonen

$$\Phi(z_0) = P_*(\hat{\theta}^* < \hat{\theta}) \tag{7}$$

Her er Φ som vanlig fordelingsfunksjonen i $N(0,1)$, mens $*$ i P_* betegner som i formel (6) at sannsynligheten beregnes med hensyn på bootstrap-fordelingen (enten denne er parametrisk eller ikke-parametrisk). Vi ser at $z_0 = 0.50$ svarer til at $\hat{\theta}$ er medianen i bootstrap-fordelingen til $\hat{\theta}^*$. Dersom sannsynlighetsfordelingen for $\hat{\theta}^*$ er skjev i forhold til $\hat{\theta}$ ($P_*(\hat{\theta}^* < \hat{\theta}) \neq 0.50$), tyder det på at den opprinnelige estimatoren $\hat{\theta}$ har en usymmetrisk fordeling, og BC-intervallet skal justere for denne skjevheten.

Høyresiden i (7) kan estimeres ut fra histogrammet for $\hat{\theta}^*$ (arealet opp til $\hat{\theta}$), eller fra de ordnede $\hat{\theta}^{*(l)}$-verdier ved å telle opp antall $\hat{\theta}^{*(l)}$ som er $< \hat{\theta}$ og dele på k).

Altså:

$$P_*(\hat{\theta}^* < \hat{\theta}) \approx \frac{l}{k}$$

der l er størst mulig med $\hat{\theta}^{*(l)} < \hat{\theta}$. Nå vil z_0 finnes fra (7) ved å bruke tabell (eller dataprogram) for Φ^{-1}.

BC-metoden modifiserer persentilmetoden ved at vi velger

$$L_{BC} = [\Phi(2z_0 - u_{\alpha/2})k] \text{ og } U_{BC} = [\Phi(2z_0 + u_{\alpha/2})k]$$

og lar konfidensintervallet være

$$[\hat{\theta}^{*(L_{BC})}, \hat{\theta}^{*(U_{BC})}]$$

Legg merke til at $z_0 = 0$ gir persentil-intervallet.

Eksempel (forts.)

Ta utgangspunkt i den parametriske bootstrappingen med $k = 1000$ verdier for $\hat{\theta}^*$. Ved å se på de sorterte verdier for $\hat{\theta}^*$ igjen, viser det seg at

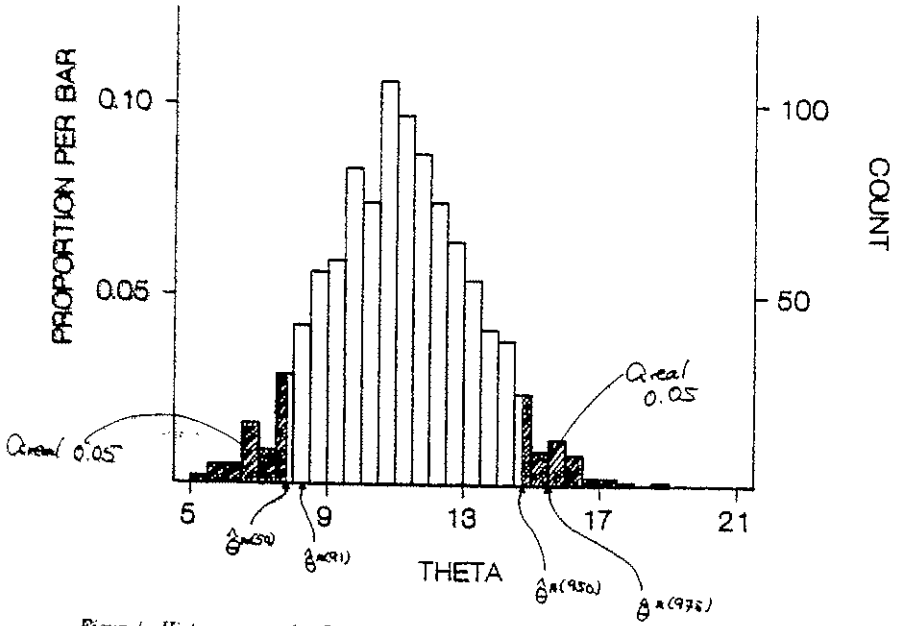
$$P_*(\hat{\theta}^* < \hat{\theta}) \approx 0.562$$

(dvs. 562 av de 1000 $\hat{\theta}^*$ var mindre enn $\hat{\theta} = 11.387$.) Dette gir $z_0 = 0.156$ og dermed

$$L_{BC} = [\Phi(-1.333) \cdot 1000] = 91 \text{ og } U_{BC} = [\Phi(1.957) \cdot 1000] = 975$$

Konfidensintervallet blir dermed $[\hat{\theta}^{*(91)}, \hat{\theta}^{*(975)}] = [8.32, 15.53]$, som er noe forskjøvet mot høyre i forhold til persentil-intervallet. Dette henger sammen med at histogrammet for $\hat{\theta}^*$ ikke er helt symmetrisk.

De to bootstrap-intervallene er tegnet inn på Figur 4.



Figur 4: Histogrammet fra figur 1, inntegnet persentil- og BC konfidensintervallene.

Begrunnelse for persentil- og BC-intervallene

Anta først at $\hat{\theta} \sim N(\theta, \tau^2)$ der τ er kjent og ikke avhenger av θ. Da vil standard konfidensintervallet for θ fra Statistikk 1 være gitt ved

$$[\hat{\theta}_L, \hat{\theta}_U] \equiv [\hat{\theta} - u_{\alpha/2}\tau, \hat{\theta} + u_{\alpha/2}\tau] \tag{8}$$

Det er lett å sjekke at dette intervallet er såkalt *korrekt*, i den forstand at

$$P_\theta(\theta < \hat{\theta}_L) = \frac{\alpha}{2} \text{ og } P_\theta(\theta > \hat{\theta}_U) = \frac{\alpha}{2}$$

Poenget ved "korrekthet" er at disse sannsynlighetene er like store, dvs. at det er like stor sannsynlighet, $\alpha/2$, for at θ ligger til venstre for $\hat{\theta}_L$ som at den ligger til høyre for $\hat{\theta}_U$. Dette er et naturlig krav til tosidige intervaller, da de to sidene dermed "behandles likt". Vi brukte dette som et overordnet prinsipp i Statistikk 1, mens f.eks. konfidensintervallet for σ^2 som vi utledet på forelesning fra sannsynlighetskvote-testen for $H_0 : \sigma^2 = sigma_0^2$, ikke har denne egenskapen.

Anta som neste mulighet at $\hat{\theta}$ *ikke* selv er normalfordelt, men at det fins en strengt voksende funksjon $g(\hat{\theta})$ som er normalfordelt. Merk at generelt vil vi ha $\mathrm{E}_\theta[g(\hat{\theta})] \neq g(\mathrm{E}_\theta(\hat{\theta}))$,

så det er ikke urimelig at $g(\hat{\theta})$ har en forventning forskjellig fra $g(\theta)$ (selv om $\hat{\theta}$ selv skulle være forventningsrett).

For å kunne møte denne muligheten antar vi derfor at

$$g(\hat{\theta}) \sim N(g(\theta) - z_0\tau, \tau^2) \tag{9}$$

der z_0, τ er konstanter som for enkelthets skyld antas å ikke avhenge av θ. Her angir altså z_0 hvor mange standardavvik forventningen ligger fra $g(\theta)$.

I praksis vil nok både z_0 og τ kunne avhenge av θ, men det viser seg at vi ofte får gode intervaller selv om de antas konstante. Bradley Efron har forøvrig også utviklet en BC_a-metode (a for *accelerated*) som tar vare på tilfellet der τ avhenger av θ. Vi skal ikke gå inn på denne metoden her.

Hvis $g(\cdot), z_0, \tau$ var kjente i (9), er det forholdsvis rett fram å konstruere et $1-\alpha$ konfidensintervall for θ. Siden $g(\hat{\theta}) + z_0\tau$ er forventningsrett for $g(\theta)$, ville vi nemlig ved samme resonnement som i begynnelsen av avsnittet få et korrekt konfidensintervall for $g(\theta)$ fra relasjonen

$$P_\theta(g(\hat{\theta}) + z_0\tau - u_{\alpha/2}\tau < g(\theta) < g(\hat{\theta}) + z_0\tau + u_{\alpha/2}\tau) = 1 - \alpha$$

Siden $g(\cdot)$ var antatt strengt voksende, ville dette være ekvivalent med

$$P_\theta(g^{-1}(g(\hat{\theta}) + z_0\tau - u_{\alpha/2}\tau) < \theta < g^{-1}(g(\hat{\theta}) + z_0\tau + u_{\alpha/2}\tau)) = 1 - \alpha \tag{10}$$

noe som gir et $1-\alpha$ konfidensintervall for θ. Det er verdt å merke seg at også dette intervallet for θ er korrekt (i den tidligere omtalte forstand). F.eks. får vi for den øvre grensen (tilsvarende for den nedre):

$$P_\theta(\theta > g^{-1}(g(\hat{\theta}) + z_0\tau + u_{\alpha/2}\tau)) = P_\theta(g(\theta) > g(\hat{\theta}) + z_0\tau + u_{\alpha/2}\tau) = \alpha/2$$

på grunn av antagelsen (9).

Vi skal nå se hvordan bootstrapping kan brukes til å finne de to grensene i konfidensintervallet gitt ved (10) også dersom $g(\cdot), z_0, \tau$ er *ukjente*.

La $\hat{\theta}^* = \hat{\theta}(\boldsymbol{X}^*)$ være gitt som tidligere, der $\boldsymbol{X}^*$ er trukket fra en estimert fordeling for $\boldsymbol{X}$ (parametrisk eller ikke-parametrisk). Hvis vi betrakter parametrisk bootstrapping, husker vi at den besto i at $\boldsymbol{X}^*$ ble trukket fra fordelingen som framkommer ved å skifte ut θ med $\hat{\theta}$. Dersom vi gjør dette i (9), ser vi at vi vil få

$$g(\hat{\theta}^*) \sim N(g(\hat{\theta}) - z_0\tau, \tau^2) \tag{11}$$

(Dersom $\boldsymbol{X}^*$ trekkes ved ikke-parametrisk bootstrapping, vil vi anta at (11) likevel er *tilnærmet* oppfylt. Det betyr at vi kan bruke de beskrevne metodene også ved ikke-parametrisk bootstrapping). Merk at i (11) er $\hat{\theta}$ å betrakte som en konstant, mens $\hat{\theta}^*$ er stokastisk (basert på bootstrap-utvalget $\boldsymbol{X}^*$).

Vi skal se at grensene i intervallet (10) kan beskrives som kvantiler i fordelingen til $\hat{\theta}^*$, slik at de kan leses direkte av fra et histogram for $\hat{\theta}^*$, eller (ekvivalent) fra de ordnede verdier $\hat{\theta}^{*(l)}$ som vi brukte i forrige avsnitt. Vi får nemlig fra (11) at

$$\begin{aligned} P_*(\hat{\theta}^* < g^{-1}(g(\hat{\theta}) + z_0\tau - u_{\alpha/2}\tau)) &= P_*(g(\hat{\theta}^*) < g(\hat{\theta}) + z_0\tau - u_{\alpha/2}\tau) \\ &= \Phi(2z_0 - u_{\alpha/2}) \end{aligned} \tag{12}$$

Tilsvarende fås

$$P_*(\hat{\theta}^* < g^{-1}(g(\hat{\theta}) + z_0\tau + u_{\alpha/2}\tau)) = \Phi(2z_0 + u_{\alpha/2}) \tag{13}$$

Dersom z_0 var kjent, ville vi nå kunne beregne konfidensintervallet (10) ved å avsette arealer henholdsvis $\Phi(2z_0 - u_{\alpha/2})$ og $\Phi(2z_0 + u_{\alpha/2})$ fra venstre i histogrammet for $\hat{\theta}^*$, (se f.eks. Figur 4). *Persentilmetoden* består i å anta at $z_0 = 0$, og vi ser da at (12) og (13) gir persentil-intervallet som ble presentert i forrige avsnitt.

BC-metoden baserer seg på at z_0 ikke nødvendigvis er 0, og for å kunne bruke (12) og (13) må vi altså estimere z_0. I avsnittet om BC-intervallet så vi at dette kan gjøres fra ligning (7), og det gjenstår derfor bare å vise denne, dvs.

$$\Phi(z_0) = P_*(\hat{\theta}^* < \hat{\theta})$$

Dette er en enkel konsekvens av (11), som siden g er strengt voksende gir

$$P_*(\hat{\theta}^* < \hat{\theta}) = P_*(g(\hat{\theta}^*) < g(\hat{\theta})) = \Phi(z_0)$$

Sluttkommentar

Vi ser at standardintervallet (8), persentilintervallet og BC- intervallet faller sammen under forutsetningen

$$\hat{\theta} \sim N(\theta, \tau^2) \tag{14}$$

og at de i den angitte rekkefølgen svarer til gradvise generaliseringer av denne antagelsen. Det bare såvidt omtalte BC_a intervallet er en ytterligere forbedring av de nevnte metodene. Merk at *asymptotisk* vil $\hat{\theta}$ som regel tilfredsstille (14). Det betyr at de omtalte intervalltypene vil være asymptotisk ekvivalente. For små eller middels n vil bootstrap-intervallene imidlertid være mer nøyaktige enn standard-intervallet, da de er konstruert for å fange opp avvik fra den eksakte normal-modell.

Du kan lese om bootstrapping i oversiktsartikkelen

> BRADLEY EFRON AND ROBERT TIBSHIRANI: "Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy", *Statistical Science, Vol. 1, p. 54-57, 1986*

som du finner på biblioteket eller i boken

> BRADLEY EFRON AND ROBERT TIBSHIRANI: "An Introduction to the Bootstrap". Chapman & Hall.

som selges i bokhandelen.