Bootstrapping – Tilleggslitteratur for STK2120

Geir Storvik

Mars 2011

1 Introduksjon

Dette notatet er et supplement til STK1120. Læreboka [Devore and Berk, 2007] diskuterer bootstrapping, men stoffet er fordelt over ulike kapitler og er noe overfladisk. Dette notatet er ment å samle hovedideene bak bootstrapping. Samtidig gir den eksempler på hvordan bootstrapping kan utføres i R. Alle koder er tilgjengelige fra kursets hjemmeside.

2 Bootstrapping

Anta vi har observert $x_1, \ldots, x_n \sim F$. Vi ønsker å estimere en ukjent parameter θ med $\hat{\theta} = \hat{\theta}(\mathbf{x})$ der $\mathbf{x} = (x_1, \ldots, x_n)$. Noen av de spørsmål vi gjerne stiller er:

(*a*) Er $\hat{\theta}$ forventningsrett, og hvis ikke hva er skjevheten til estimatet?

(*b*) Hva er usikkerheten til $\hat{\theta}$?

(*c*) Hvordan kan vi laget et konfidensintervall for θ?

Alle disse er knyttet til fordelingsegenskapene til $\hat{\theta}$ som avhenger av F. Et problem her er at F er ukjent, noe som medfører at også fordelingsegenskapene til $\hat{\theta}$ er ukjente. For å kunne gi noen svar på spørsmålene ovenfor, må vi derfor gjøre noen tilnærminger. Devore and Berk [2007, Avsnitt 7.4] omhandler asymptotiske tilnærminger for maksimum likelihood estimater, som kan være nyttige i mange tilfeller. Men hva hvis antagelsene som ligger til grunn for disse asymptotiske tilnærmingene ikke er tilstede? *Bootstrapping* er da et alternativ.

Ideen bak *bootstrapping* [Efron, 1982] er enkel: Siden vi ikke kjenner F, la oss bruke et *estimat* $\widehat{F}$ for F i stedet, og så se på egenskapene til $\hat{\theta}$ under $\widehat{F}$. Det er to hovedvalg med hensyn på hvordan vi estimerer F.

En mulighet er å anta at F tilhører en klasse av fordelinger beskrevet ved en eller flere parametre. Vi beskriver ofte da F med F_η der η er (en vektor av) parametre som beskriver fordelingen. Et estimat for F oppnås da ved å bruke estimater på de ukjente parametre, dvs $\widehat{F} = F_{\hat{\eta}}$. Et eksempel kan være å anta normalfordelingen med forventning μ og standard avvik σ. F kan da estimeres ved å innsette maksimum likelihood estimatene $\hat{\mu}$ og $\hat{\sigma}$ for μ og σ. Bootstrapping basert på slike antagelser kalles for *parametrisk bootstrapping*. Devore and Berk [2007] behandler dette (noe overfladisk) på sidene 339-340.

Et alternativ til parametrisk bootstrapping er *ikke-parametrisk bootstrapping*. Ikke-parametrisk bootstrapping omhandles delvis i Devore and Berk [2007, avsnitt 8.5]. Emnet vil imidlertid bli mer detaljert beskrevet i dette notatet. Idéen i dette tilfellet er å gjøre minimale antagelser på F i utgangspunktet. Et mulig estimat for F i det tilfellet er den *empirisk kumulative fordelingsfunksjon* som er gitt ved

$$\widehat{F}(x) = \frac{1}{n}(\#x_i \leq x) \tag{1}$$

Merk at under $\widehat{F}$ så er $\Pr(X = x_i) = \frac{1}{n}$ for $i = 1, \ldots, n.$ og at trekking fra $\widehat{F}$ svarer til å *trekke fra* $x_1, ..., x_n$ *med tilbakelegging*. Dette er en viktig egenskap ved $\widehat{F}$ som vi vil utnytte når vi skal utføre de nødvendige beregninger involvert i bootstrapping.

Vi vil starte med et eksempel for å illustrere idéen.

Eksempel Anta vi er interessert i $\theta(F) = \text{median}(F) = F^{-1}(0.5)$. Det naturlige estimat for θ er den empiriske median, $\hat{\theta} = \text{median}(x_1, \ldots, x_n)$. Vi skal se hvordan vi kan estimere standardfeilen til estimatet ved hjelp av ikke-parametrisk bootstrapping. Det er da naturlig å bruke den empiriske kumulative sannsynlighetsfordelingen, som definert i (1). La oss nå trekke $x_1^*, \ldots, x_n^*$ fra $\widehat{F}$ og estimere $\hat{\theta}$ med $\theta^* = \text{median}(x_1^*, \ldots, x_n^*)$. Denne prosedyren kan vi gjenta B ganger, som gir oss $\theta_1^*, \ldots, \theta_B^*$. Hvis $\widehat{F} \approx F$, vil variabiliteten til θ-ene gi en god beskrivelse av variabiliteten i $\hat{\theta}$. Spesielt estimerer vi standardfeilen til $\hat{\theta}$, $\sigma_{\hat{\theta}}$, ved

$$s_{\hat{\theta}} \approx \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\theta_b^* - \bar{\theta}^*)^2}$$

der $\bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\theta_b^*$. Dette estimatet kalles *bootstrap estimatet* for $\sigma_{\hat{\theta}}$. □

La oss nå diskutere bootstrap idéen i mer generelle former. Vår interesse ligger i en parameter θ. Vi vil i denne sammenheng definere en parameter til å være en hvilken

som helst funksjon $\theta(F)$ av sannsynlighetsfordelingen F. Merk at dette bl.a. inkluderer momenter som for kontinuerlige variable er definert ved

$$\theta(F) = \int x^k f(x)\, dx$$

($f(x) = F'(x)$) og kvantiler

$$\theta(F) = F^{-1}(p).$$

Parameteren θ kan estimeres ved $\hat{\theta} = \hat{\theta}(\mathbf{x})$, en funksjon av data. Egenskapene til denne estimatoren er av interesse. Vi vil først se på *forventningsskjevheten* til estimatoren som er definert ved

$$\beta_{\hat{\theta}} = \mathsf{E}^F[\hat{\theta}(\mathbf{X})] - \theta(F). \tag{2}$$

Problemet med å beregne forventningsskjevheten er at F er ukjent. *Bootstrap* estimatet for forventningsskjevheten oppnår vi ved å erstatte den ukjente F med et estimat, som gir oss

$$\mathsf{E}^{\widehat{F}}[\hat{\theta}(\mathbf{X}^*)] - \theta(\widehat{F}). \tag{3}$$

Med $\theta(\widehat{F})$ mener vi den tilsvarende egenskap i $\widehat{F}$ fordelingen, mens $\mathsf{E}^{\widehat{F}}[\hat{\theta}(\mathbf{X})]$ er forventningen til $\hat{\theta}(\mathbf{X})$ når $X_1, \ldots, X_n$ er sampler fra fordelingen $\widehat{F}$.

Det største problemet med estimatet (3) er beregning av $\mathsf{E}^{\widehat{F}}[\hat{\theta}(\mathbf{X})]$. I prinsippet kan denne beregnes siden $\widehat{F}$ er kjent, men i praksis blir det fort et stort regnestykke. Merk at $\widehat{F}$ er en diskret fordeling med n mulige utfall. For hele vektoren $\mathbf{x}^* = (x_1^*, ..., x_n^*)$ blir det dermed n^n mulige utfall, et stort tall selv for moderate n. Et alternativ da er å benytte oss av at uttrykket vi ønsker å beregne er en forventning, og forventninger kan vi estimere ved hjelp av et sample fra den aktuelle fordelingen. La $\mathbf{x}_1^*, ..., \mathbf{x}_B^*$ være B uavhengige sampler, hver av størrelse n, fra fordelingen $\widehat{F}$ og la $\theta_b^* = \hat{\theta}(\mathbf{x}_b^*), b = 1, ..., B$. Da kan $\mathsf{E}^{\widehat{F}}[\hat{\theta}(\mathbf{X}^*)]$ tilnærmes ved

$$\mathsf{E}^{\widehat{F}}[\hat{\theta}(\mathbf{X}^*)] \approx \widehat{\mathsf{E}}^{\widehat{F}}[\hat{\theta}] = \frac{1}{B}\sum_{b=1}^{B} \theta_b^*. \tag{4}$$

En slik tilnærming kalles ofte for en *Monte Carlo integrasjon* og er et svært nyttig numerisk verktøy. Bruken av *integrasjon* her kommer av at forventninger i praksis er integraler. Merk at vi her er i en situasjon hvor vi selv kan velge B, dvs vi kan bestemme nøyaktigheten på tilnærmingen vår.

Monte Carlo integrasjonen medfører at vi må simulere θ^*. En prosedyre for dette er:

(*a*) Simuler $x_1^*, \ldots, x_n^*$ u.i.f. fra $\widehat{F}$,

(*b*) Sett $\theta^* = \hat{\theta}(\mathbf{x}^*)$ der $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$.

Det endelige estimatet på forventningsskjevhetet $\beta_{\hat{\theta}}$ blir dermed

$$b_{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\theta_b^* - \theta(\widehat{F}). \tag{5}$$

Merk at *to* tilnærminger er involvert i beregning av dette estimatet. En Monte Carlo variabilitet inngår ved estimering av $\mathsf{E}^{\widehat{F}}[\theta^*]$. Den feilen vi her innfører vil ofte være neglisjerbar i forhold til den andre tilnærmingen vi gjør ved å erstatte F med $\widehat{F}$.

Selv om det ikke alltid er slik, vil i mange tilfeller $\hat{\theta}(\mathbf{x}) = \theta(\widehat{F})$. I eksemplet ovenfor var $\hat{\theta}(\mathbf{x})$ lik den empiriske medianen, som også er den teoretiske medianen til den empiriske kumulative fordeling, $\theta(\widehat{F})$.

Ovenfor har vi beskrevet hvordan forventningsskjevheten til en estimator $\hat{\theta}$ kan estimeres. Like enkelt kan andre egenskaper ved estimatoren analyseres. Anta f.eks. vi er interessert i standardfeilen til estimatoren. Vi har

$$\sigma_{\hat{\theta}} = \sqrt{\mathsf{E}^F[(\hat{\theta}(\mathbf{X}) - \mathsf{E}^F[\hat{\theta}(\mathbf{X})])^2]}$$

med tilhørende bootstrap estimat

$$\sqrt{\mathsf{E}^{\widehat{F}}[(\hat{\theta}(\mathbf{X}^*) - \mathsf{E}^{\widehat{F}}[\hat{\theta}(\mathbf{X}^*)])^2]}. \tag{6}$$

Også i dette tilfellet kan estimatet i prinsippet beregnes analytisk, men i praksis vil vi tilnærme estimatet ved Monte Carlo integrasjon:

$$s_{\hat{\theta}} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\theta_b^* - \bar{\theta}^*)^2}. \tag{7}$$

3 Bootstrapping i praksis

I dette avsnittet vil vi se på hvordan ikke-parametrisk bootstrapping kan utføres i praksis. Vi vil illustrere det med eksempler og også vise hvordan vi kan utføre bootstrapping i programpakken R.

Eksempel For å evaluere en behandlingsmetode, ble overlevelsestiden til 16 mus observert. For hver mus ble det trukket tilfeldig om den skulle behandles eller ikke. Et spørsmål av interesse er om behandlingsmetoden forlenget levetiden. Datasettet er hentet fra Efron and Tibshirani [1993] og er vist i Tabell 1.

Gruppe	Overlevelsestider								
Behandling	94	197	16	38	99	141	23		
Kontroll	52	104	146	10	51	30	40	27	46

Table 1: Overlevelsestider (i dager) til 16 mus tilfeldig allokert til behandlingsgruppen eller kontrollgruppen. Data fra Efron and Tibshirani [1993, side 11].

La x_i være i-te måling i behandlingsgruppen og y_i i-te måling i kontrollgruppen. Vi vil konsentrere oss om analyse av behandlingsgruppen. En analyse av forskjellen mellom gruppene vil bli gitt som oppgave.

Anta $x_1, \ldots, x_7$ er uavhengige identisk fordelte med sannsynlighetfordeling F. Vi er interessert i både forventning og median for fordelingen F. De naturlige estimater er empirisk gjennomsnitt og median:

$$\overline{x} = 86.86, \quad \mathsf{med}(x_1, \ldots, x_7) = 94.0.$$

Det empiriske standardavviket til $x_1, \ldots, x_7$ er 66.77 som indikerer en stor variabilitet i dataene.

Hva er så forventningsskjevhet og standardfeil til disse estimatene? Vi kan benytte bootstrap metodikken for å besvare dette. I Figur 1 er det gitt R kommandoer for beregning av ikke-parametriske bootstrap estimater for forventningsskjevheten og standardfeilen til $\overline{x}$ og $\mathsf{med}(\mathbf{x})$. Merk at en sentral funksjon her er `sample` som kan brukes for å trekke fra en vektor av verdier med (og hvis ønskelig også uten) tilbakelegging. En kjøring av kommandoene i Figur 1 ga

$$\begin{array}{llll} b_{\overline{x}} & = & 0.19 & s_{\overline{x}} = 23.94 \\ b_{\mathsf{med}(\mathbf{x})} & = -14.64 & & s_{\mathsf{med}(\mathbf{x})} = 37.31 \end{array}$$

Sammenliknet med variabiliteten, er forventningsskjevheten til gjennomsnittet svært nær null. Dette er ingen tilfeldighet, da en kan vise at bootstrap skjevheten til et gjennomsnitt

```
#Data
x <- c(94,197,16,38,99,141,23)
n <- length(x)
#Estimates
mean.hat <- mean(x);med.hat <- median(x)
B <- 1000
mean.star <- rep(NA,B);med.star <- rep(NA,B)
for(b in 1:B)
 {
   #Sampling from F.hat
   x.star <- sample(x,size=n,replace=T)
   #Calculating estimates based on bootstrap samples
   mean.star[b] <- mean(x.star);med.star[b] <- median(x.star)
 }
#Calculating bootstrap estimates
bias.mean <- mean(mean.star)-mean.hat;sd.mean <- sd(mean.star)
bias.med <- mean(med.star)-med.hat;sd.med <- sd(med.star)
```

Figure 1: R kommandoer for beregning av bootstrap estimater for forventningsskjevheten og standardfeil til $\overline{x}$ og $\mathsf{med}(\mathbf{x})$ basert på data i Tabell 1.

alltid er lik null. Vårt (lille) avvik fra null i dette tilfellet skyldes at vi bruker Monte Carlo integrasjon til å beregne bootstrap skjevheten som introduserer en (tilfeldig) feil.

Verdiene for medianen indikerer derimot noe forventningsskjevhet, dog ikke overbevisende stor i forhold til variabiliteten. Merk forøvrig at de samme simuleringer $\mathbf{x}^*$ kan brukes for både beregning av bootstrap simulerte gjennomsnitt og medianer. I mer kompliserte situasjoner, der simulering av $\mathbf{x}^*$-ene kan være beregningstunge, er dette en klar fordel. □

La oss så se på hvordan *parametrisk* bootstrapping kan utføres. Som et konkret eksempel, la oss se på tilpasningen av nedbørsmengder under 227 stormer i Illinois fra 1960 til 1964 til en gamma fordeling. Data, som er gitt i Rice [1995, Oppgave 42, kapittel 10], var samlet inn og analysert i et forsøk på å karakterisere den naturlige variabilitet i nedbør fra storm til storm. Momentestimatene for parametrene λ og α i gamma fordelingen er gitt ved

$$\hat{\lambda} = \frac{\overline{X}}{\hat{\sigma}^2} \quad \hat{\alpha} = \frac{\overline{X}^2}{\hat{\sigma}^2}$$

Vi ønsker å si noe om forventningsskjevhet og usikkerhet til disse estimatene. Hvis $F_{\lambda,\alpha}$ beskriver den sanne gamma fordelingen, vil vi i dette tilfellet lage oss et estimat for $F_{\lambda,\alpha}$

```
#Read data, treating * as missing
x <- scan("../data/ILLRAIN.DAT",na.strings="*")
#Remove missing data
x <- x[!is.na(x)];n <- length(x)
#Moment estimates
lambda.hat <- mean(x)/var(x);alpha.hat <- mean(x)^2/var(x)
B <- 1000
lambda.star <- rep(NA,B);alpha.star <- rep(NA,B)
for(b in 1:B)
{
 #Draw from Gamma distribution
 x.star <- rgamma(n,shape=alpha.hat,rate=lambda.hat)
 #Calculate estimates based on bootstrap samples
 lambda.star[b] <- mean(x.star)/var(x.star)
 alpha.star[b] <- mean(x.star)^2/var(x.star)
}
#Calculate bootstrap estimates
bias.lambda <- mean(lambda.star)-lambda.hat
sd.lambda <- sd(lambda.star)
bias.alpha <- mean(alpha.star)-alpha.hat
sd.alpha <- sd(alpha.star)
```

Figure 2: R kommandoer for beregning av bootstrap estimater for forventningsskjevheten og standardfeil til $\hat{\lambda}$ og $\hat{\alpha}$ basert på nedbørsdata.

ved $F_{\hat{\lambda},\hat{\alpha}}$. Figur 2 viser R kommandoer for å utføre parametrisk bootstrapping i dette eksemplet. Merk likheten med kommandoene i Figur 1 som illustrerte ikke-parametrisk bootstrapping. Den vesentlige forskjellen ligger i at mens for ikke-parametrisk boostrapping bruker vi funksjonen `sample` for simulering mens for parametrisk bootstrapping bruker vi en funksjon for å simulere fra en gitt parametrisk fordeling (`rgamma` i dette tilfellet).

Kjøring av kommandoene i Figur 2 ga

$$\begin{array}{llll} b_{\hat{\lambda}} & = 0.088 & s_{\hat{\lambda}} & = 0.354 \\ b_{\hat{\alpha}} & = 0.0151 & s_{\hat{\alpha}} & = 0.0659 \end{array}$$

I forhold til usikkerheten i estimatene, er forventningsskjevheten liten.

4 Konfidensintervaller

Når vi vil si noe om egenskapene til en estimator, ønsker vi ofte også å rapportere et konfidensintervall for parameteren θ. Det er flere måter å gjøre dette på. Den mest vanlige metoden er å anta at $\hat{\theta}$ er tilnærmet normalfordelt og dermed bruke

$$\hat{\theta} \pm z(\alpha/2) \cdot s_{\hat{\theta}}$$

som et $100(1-\alpha)\%$ konfidensintervall. Her er $z(\alpha/2)$ øvre $\alpha/2$-fraktilen i standard normalfordelingen (dvs $\Pr(Z > z(\alpha/2)) = \alpha/2$). I situasjoner der $\hat{\theta}$ er et gjennomsnitt av (uavhengige) tilfeldige variable, kan sentralgrenseteoremet brukes for å vise at $\hat{\theta}$ er asymptotisk normalfordelt (dvs når antall observasjoner går mot uendelig). Men også i mange andre situasjoner (f.eks. når $\hat{\theta}$ er en sannsynlighetsmaksimeringsestimator) vil normalitetstilnærmingen kunne brukes. For å finne $s_{\hat{\theta}}$, kan bootstrap teknikken, som tidligere beskrevet, benyttes.

Det er imidlertid andre teknikker for å konstruere konfidensintervaller som ikke baserer seg på normaltilnærmingen. Innenfor bootstrapverdenen eksisterer det flere alternativer. Vi vil konsentrere oss om en av de enkleste, *standard bootstrap konfidensintervall* (*basic bootstrap confidence intervals* på engelsk). Anta som før $\hat{\theta}$ er et estimat for parameteren θ. Definer $\Delta = \hat{\theta} - \theta$, og anta for øyeblikket at fordelingen til Δ er kjent. La $\underline{\delta}$ og $\overline{\delta}$ være $\alpha/2$ og $1-\alpha/2$ kvantilene for denne fordelingen, dvs

$$\begin{aligned} P(\hat{\theta} - \theta \leq \underline{\delta}) &= \frac{\alpha}{2} \\ P(\hat{\theta} - \theta \leq \overline{\delta}) &= 1 - \frac{\alpha}{2} \end{aligned}$$

Da er

$$P(\underline{\delta} \leq \hat{\theta} - \theta \leq \overline{\delta}) = 1 - \alpha$$

og ved manipulering av ulikhetene får vi at

$$P(\hat{\theta} - \overline{\delta} \leq \theta \leq \hat{\theta} - \underline{\delta}) = 1 - \alpha$$

Dette viser at $(\hat{\theta} - \overline{\delta}, \hat{\theta} - \underline{\delta})$ er et $100(1-\alpha)\%$ konfidensintervall for θ.

Problemet med bruk av dette intervallet er at fordelingen til Δ typisk er ukjent. Bootstrapidéen her er å tilnærme fordelingen til $\hat{\theta} - \theta$ med fordelingen til $\theta^* - \hat{\theta}$ der θ^* er et bootstrap sample av $\hat{\theta}$ basert på $\widehat{F}$. Spesielt, hvis $\theta_1^*, ..., \theta_B^*$ er B bootstrap sampler av $\hat{\theta}$, så er bootstrap estimatet av $\underline{\delta}$ lik $\alpha/2$-kvantilen i den empiriske fordelingen til $\theta_1^* - \hat{\theta}, ..., \theta_B^* - \hat{\theta}$.

```
alpha <- 0.05
#Konfidensintervall basert på normalapproksimasjon
z <- qnorm(1-0.025)
ki.mean.norm <- c(mean.hat-z*sd.mean,mean.hat+z*sd.mean)
ki.med.norm <- c(med.hat-z*sd.med,med.hat+z*sd.med)
#Standard bootstrap konfidensintervall
k1 <- as.integer(B*alpha/2);k2 <- as.integer(B*(1-alpha/2))
mean.sort <- sort(mean.star)
deltaL <- mean.sort[k1]-mean.hat
deltaU <- mean.sort[k2]-mean.hat
ki.mean.basic <- c(mean.hat-deltaU,mean.hat-deltaL)
med.sort <- sort(med.star)
deltaL <- med.sort[k1]-med.hat
deltaU <- med.sort[k2]-med.hat
ki.med.basic <- c(med.hat-deltaU,med.hat-deltaL)
```

Figure 3: R kommandoer for beregning av bootstrap intervaller for forventnings og median basert på data i Tabell 1.

θ	$\mathsf{E}^F[x]$	$\mathsf{med}(F)$
Normaltilnærming	$[40.7, 133.1]$	$[18.9, 169.1]$
Bootstrap intervall	$[42.7, 128.4]$	$[47.0, 165.0]$

Table 2: Konfidensintervaller for forventningsskjevheten og standardfeil til $\overline{x}$ og $\mathsf{med}(\mathbf{x})$ basert på data i Tabell 1.

Eksempel (mus, forts.) Basert på de tidligere simuleringene av θ^*, kan beregning av bootstrap konfidensintervall utføres i R ved kommandoene gitt i Figur 3. Merk at for $B = 1000$, og med $\alpha = 0.05$, så er $\alpha/2$ kvantilen gitt ved den 25. verdien av $\theta^* - \hat{\theta}$. Tilsvarende er $1 - \alpha/2$ kvantilen gitt ved den 975. minste verdien. I Tabell 4 er konfidensintervall for $\mathsf{E}^F[x]$ og $\mathsf{med}(F)$ basert på normaltilnærmingen og bootstrap angitt. Intervallene for forventningen er ikke så ulike hverandre, noe som indikerer at normaltilnærmingen kan være rimelig for $\overline{x}$. Noe større forskjeller får vi for medianen. Både gjennomsnittet og medianen er asymptotisk normale. For at tilnærmingen skal være god, krever imidlertid medianen typisk flere observasjoner enn gjennomsnittet. Avvikene i konfindensintervallene illustrerer dette.

□

5 Et eksempel med bivariate observasjoner

Vi vil i dette avsnittet se på et eksempel som illustrerer bootstrap metoden for bivariate observasjoner. Tabell 3 gir diameter ved bryst høyde (i fot) og alder (antall år) av 26 kastanjetrær, hentet fra [Rice, 1995, Oppgave 37, kapittel 14].

Obs	Age	DBH	Obs	Age	DBH
1	4	0.8	14	23	4.7
2	5	0.8	15	25	6.5
3	8	1.0	16	28	6.0
4	8	3.0	17	29	4.5
5	10	2.0	18	30	6.0
6	10	3.5	19	30	7.0
7	12	4.9	20	33	8.0
8	13	3.5	21	34	6.5
9	14	2.5	22	35	7.0
10	16	4.5	23	38	5.0
11	18	4.6	24	38	7.0
12	20	5.5	25	40	7.5
13	22	5.8	26	42	7.5

Table 3: Diameter ved bryst høyde og alder av 26 kastanjetrær, se [Rice, 1995, Oppgave 37, kapittel 14].

Vi er i dette tilfellet interessert i korrelasjonen ρ mellom de to målingene. La $\mathbf{x}_i = (x_{1,i}, x_{2,i})$ være `Age` og `DBH` målingene. Det naturlige estimatet for korrelasjonen er den empiriske korrelasjonen definert ved

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)(x_{2,i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{1,i} - \bar{x})^2 \sum_{i=1}^{n}(x_{2,i} - \bar{x}_2)^2}}.$$

Når vi skal lage bootstrap simuleringer av $\hat{\rho}$, må vi nå lage et estimat $\widehat{F}$ for den *bivariate* sannsynlighetsfordelingen F. Dette gjør vi tilsvarende som for det en-dimensjonale tilfellet ved at under $\widehat{F}$ så er $\Pr(\mathbf{X} = \mathbf{x}_i) = \frac{1}{n}$ for $i = 1, \ldots, n$. Simulering av $\hat{\rho}$ kan utføres i R ved kommandoer gitt i Figur 5. Merk her trikset med å trekke en indeks først som gjør at vi enkelt kan trekke par $(x_{1,i}, x_{2,i})$. Bootstrap estimater på forventningsskjevhet og standardfeil for $\hat{\rho}$ ble $b_{\hat{\rho}} = -0.002$, $s_{\hat{\rho}} = 0.043$, Som vi ser, er $\hat{\rho}$ nær forventningsrett.

Et 95% standard bootstrap konfidensintervall for ρ er $[0.820, 0.984]$ som gir en klar indikasjon på at det er en positiv sammenheng mellom de to variablene. Til sammenlikning ville en normaltilnærming gi et konfidensintervall på $[0.800, 0.963]$, som er et intervall (litt) forskjøvet nedover i forhold til bootstrap intervallet. Figur 4 viser et histogram av de

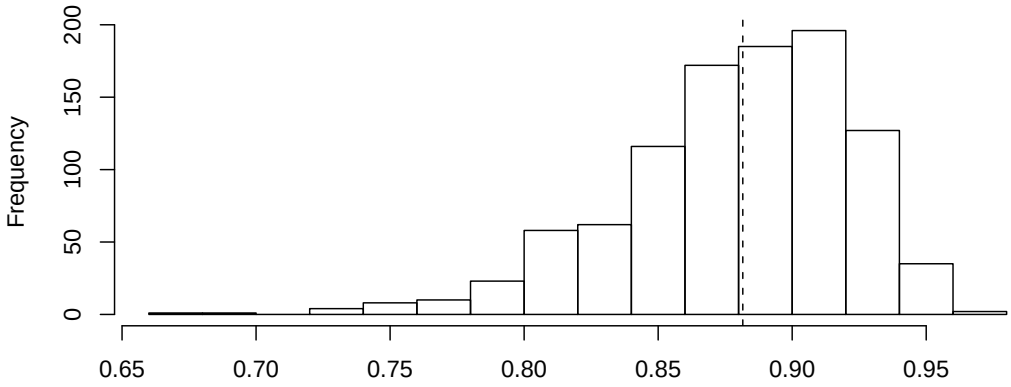

Figure 4: Histogram av ρ^*.

simulerte ρ^* verdiene. Dette plottet indikerer at $\hat{\rho}$ har en skjev fordeling, som forklarer forskjellen mellom bootstrap intervallet og normaltilnærmingsintervallet. For å bruke normaltilnærmingen i dette tilfellet, ville en ha transformert $\hat{\rho}$ først til noe som ligner mer på en normalfordelt variabel. Merk imidlertid at bootstrap metodikken automatisk tar hensyn til den asymmetrien som er tilstede, dvs ingen transformasjon er nødvendig.

I diskusjonen ovenfor har vi brukt den empiriske sannsynlighetsfordelingen som estimat for F. La oss nå se på parametrisk bootstrapping. I så tilfelle må vi velge en klasse av *bivariate* fordelinger $F_{\boldsymbol{\eta}}$. Vi vil anta at observasjonene følger en (bivariat) normalfordeling. Her består $\boldsymbol{\eta}$ av fem parametre, forventning og varians til hver av variablene og deres innbyrdes korrelasjon. Disse kan estimeres på vanlig måte. Simuleringer fra $F_{\hat{\boldsymbol{\eta}}}$ kan nå utføres ved å simulere 26 variable fra den tilpassede bivariate normalfordelingen. For å kunne simulere slike variable, vil vi benytte oss av følgende egenskaper ved den bivariate fordelingen [Devore and Berk, 2007, side 255-256]:

$$\begin{aligned} X_1 \sim & N(\mu_{X_1}, \sigma_{X_1}) \\ X_2|X_1 \sim & N(\mu_{X_2} + \rho(x_1 - \mu_{X_1})\sigma_{X_2}/\sigma_{X_1}, \sigma_{X_2}\sqrt{1-\rho^2}) \end{aligned}$$

En kan dermed først simulere X_1 fra dens marginale normalfordeling og deretter $X_2|X_1$ fra den betingede normalfordeling. Figur 6 viser R kommandoer for den parametriske bootstrappingen.

Simuleringene basert på kommandoene i Figur 6 ga en estimert forventningsskjevhet på -0.00361 og et standardfeil på 0.049. Et 95% bootstrap konfidensintervall er $[0.815, 1.004]$ mens et intervall basert på normalapproksimasjon av $\hat{\rho}$ er $[0.785, 0.978]$.

```
#Read data
dbhtrees <- read.table("DBHTREES.DAT",col.names=c("age","dbh"))
n <- dim(dbhtrees)[1]
#Calculate estimate
rho.hat <- cor(dbhtrees$age,dbhtrees$dbh)
#Initialization for Bootstrap sampling
B <- 1000
rho.star <- rep(NA,B)
#Loop for Bootstrap sampling
for(b in 1:B)
{
 #First simulate index
 ind <- sample(1:n,n,replace=T)
 #Pick out right age and dbh
 age.star <- dbhtrees$age[ind];dbh.star <- dbhtrees$dbh[ind]
 #Estimate rho based on bootstrap sample
 rho.star[b] <- cor(age.star,dbh.star)
}
#Calculate Bootstrap estimates
bias.rho <- mean(rho.star)-rho.hat
sd.rho <- sd(rho.star)
#Confidence intervals based on normal approximation
alpha <- 0.05
z <- qnorm(1-alpha/2)
ki.norm <- c(rho.hat-z*sd.rho,rho.hat+z*sd.rho)
#Standard bootstrap connfidence intervall
k1 <- as.integer(B*alpha/2)
k2 <- as.integer(B*(1-alpha/2))
rho.sort <- sort(rho.star)
deltaL <- rho.sort[k1]-rho.hat
deltaU <- rho.sort[k2]-rho.hat
ki.basic <- c(rho.hat-deltaU,rho.hat-deltaL)
```

Figure 5: Bootstrap analyse av sammenhengen mellom alder og brysthøyde av kastanjetrær.

```
#Read data
dbhtrees <- read.table("DBHTREES.DAT",col.names=c("age","dbh"))
n <- dim(dbhtrees)[1]
#Calculate estimate
mu.hat <- c(mean(dbhtrees$age),mean(dbhtrees$dbh))
sd.hat <- c(sd(dbhtrees$age),sd(dbhtrees$dbh))
rho.hat <- cor(dbhtrees$age,dbhtrees$dbh)
#Initialization for Bootstrap sampling
B <- 1000
rho.star <- rep(NA,B)
#Loop for Bootstrap sampling
for(b in 1:B)
 {
  #Simulate age.star from the marginal normal distribution
  age.star <- rnorm(n,mu.hat[1],sd.hat[1])
  #Simulate dbh from the conditional normal distribution
  mu.cond <- mu.hat[2]+
             rho.hat*sd.hat[2]*(age.star-mu.hat[1])/sd.hat[1]
  sd.cond <- sd.hat[2]*sqrt(1-rho.hat^2)
  dbh.star <- rnorm(n,mu.cond,sd.cond)
  #Estimate rho based on bootstrap sample
  rho.star[b] <- cor(age.star,dbh.star)
 }
#Calculate Bootstrap estimates
bias.rho <- mean(rho.star)-rho.hat
sd.rho <- sd(rho.star)
#Confidence intervals based on normal approximation
alpha <- 0.05
z <- qnorm(1-alpha/2)
ki.norm <- c(rho.hat-z*sd.rho,rho.hat+z*sd.rho)
#Standard bootstrap connfidence intervall
k1 <- as.integer(B*alpha/2)
k2 <- as.integer(B*(1-alpha/2))
rho.sort <- sort(rho.star)
deltaL <- rho.sort[k1]-rho.hat
deltaU <- rho.sort[k2]-rho.hat
ki.basic <- c(rho.hat-deltaU,rho.hat-deltaL)
```

Figure 6: R kommandoer for simulering av ρ^* under normalantagelse.

6 Biblioteket `boot`

Implementering av bootstrap metoden er ofte svært enkelt. I R kan dette gjøres enda enklere ved å utnytte at det finnes ferdige rutiner som gjør det meste for oss. Bootstrapping-rutiner i R ligger i biblioteket `boot` som kan gjøres tilgjengelig ved kommandoen

```
> library(boot)
```

De mest aktuelle funksjonene er `boot`, som kan brukes for å generere bootstrap sampler for oss, og `boot.ci` som kan brukes for å beregne konfidensintervall. Den vanligste formen for bruk av disse er

```
boot.res <- boot(data,statistic,R)
boot.ci(boot.res,conf=0.95,type="basic")
```

Her er `data` de orginale data, `statistic` er en funksjon som beregner estimatet $\hat{\theta}$ mens `R` er antall bootstrap replikasjoner (vi har brukt B). Merk at hvis vi skal bruke ikke-parametrisk bootstrapping, må funksjonen som regner ut estimatet ha to input-parametre, en som angir de opprinnelige data og en som er en indeks-vektor som viser hvilke data som er trukket ut i bootstrap-samplet. Se nedenfor et eksempel på dette. En full beskrivelse av funksjonen `boot` kan fås ved kommandoen

```
> help(boot)
```

Eksempel (mus, forts.) La oss se på hvordan funksjonen `boot` kan brukes på muse-dataene i Tabell 1 (fremdeles kun behandlingsgruppen). I Figur 7 følger en utskrift av et slikt kall basert på gjennomsnittet.

```
x <- c(94,197,16,38,99,141,23)
#First define function for mean
#Note the use of the index vector
mean.boot <- function(x,ind){mean(x[ind])}
boot.mean <- boot(x,mean.boot,1000)
```

Figure 7: Bruk av `boot` kommandoen på data i Tabell 1 for ikke-parametrisk bootstrapping.

Når vi kaller på funksjonen `boot`, blir resultatene lagret i et *dataobjekt*, som vi i dette tilfellet har kalt `boot.mean`. Vi kan få en utskrift av resultatene ved kommandoen

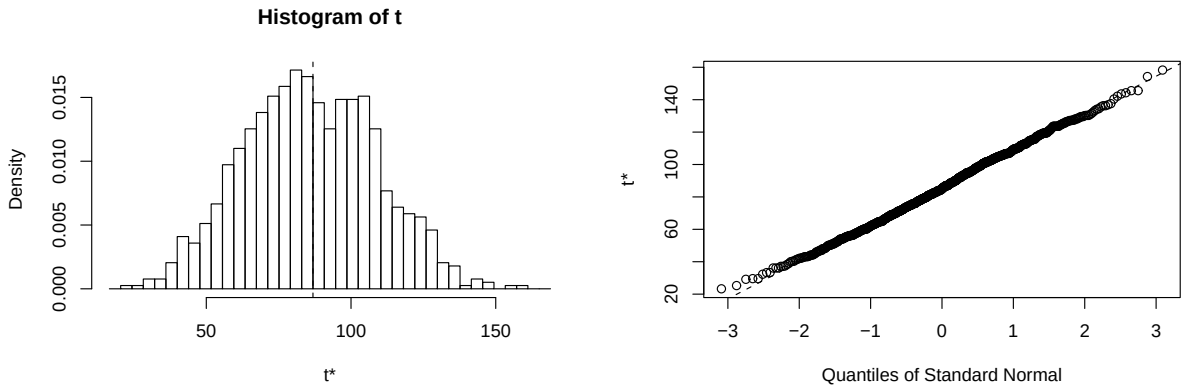

Figure 8: Eksempel på bruk av `plot` kommandoen på et dataobjekt som er generert av funksjonen `boot`.

```
print(boot.mean)
```

som gir resultatet

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = x, statistic = mean.boot, R = 1000)


Bootstrap Statistics :
    original    bias    std. error
t1* 86.85714 -1.074571    22.90002
```

Som vi ser, får vi omtrent de samme resultatene for forventningsskjevhet og standardfeil som med våre egne R kommandoer. Vi kan også gi kommandoen

```
plot(boot.mean)
```

som gir et plot som vist i Figur 8. Venstre panel viser her et histogram av de bootstrap-simulerte θ^*-ene (med den observerte $\hat{\theta}$ indikert ved en stiplet vertikal linje) mens høyre panel lager et Q-Q plot av de samme variablene i forhold til normalfordelingen. Slike plot kan brukes for å undersøke fordelingen til θ^*.

Også konfidensintervall kan beregnes ved en ferdigfunksjon, `boot.ci`. Ved kallet

```
boot.ci(boot.mean,conf=0.95,type="basic")
```

får vi følgende resultat:

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot.mean, conf = 0.95, type = "basic")

Intervals :
Level      Basic
95%   ( 44.14, 131.28 )
Calculations and Intervals on Original Scale
```

Igjen får vi nogenlunde samsvarende resultater med våre egne kommandoer. □

La oss så se på hvordan `boot` funksjonen kan brukes til parametrisk bootstrapping. I dette tilfellet vil et typisk kall se slik ut:

```
boot.mean <- boot(x,statistic,1000,sim="parametric",
                  ran.gen=gen.data,mle=par)
```

Her gir opsjonen `sim="parametric"` beskjed om at parametrisk bootstrapping skal brukes, `ran.gen=gen.data` sier at simuleringene skal utføres ved funksjonen `gen.data` mens `mle` inneholder parametre som brukes for simulering.

Eksempel (mus, forts.) Figur 9 viser R kommandoer for å utføre parametrisk bootstrapping basert på musedata. Vi har her antatt at F er inneholdt i klassen av normalfordelinger. En utskrift av resultatet gir

```
> print(boot.mean)

PARAMETRIC BOOTSTRAP


Call:
boot(data = x, statistic = mean.boot, R = 1000, sim = "parametric",
```

```
x <- c(94,197,16,38,99,141,23)
#Function for simulating from parametric distribution
gen.data <- function(x,mle)
  rnorm(length(x),mle$mu,mle$sd)
boot.mean <- boot(x,mean,1000,sim="parametric",
                  ran.gen=gen.data,mle=list(mu=mean(x),sd=sd(x)))
```

Figure 9: Bruk av `boot` kommandoen på data i Tabell 1 for parametrisk bootstrapping.

```
    ran.gen = gen.data, mle = list(mu = mean(x), sd = sd(x)))


Bootstrap Statistics :
    original    bias    std. error
t1* 86.85714 0.3202847    25.22716
```

og konfidensintervall er gitt ved

```
> boot.ci(boot.mean,conf=0.95,type="basic")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot.mean, conf = 0.95, type = "basic")

Intervals :
Level      Basic
95%   ( 37.46, 135.56 )
Calculations and Intervals on Original Scale
```

7 Sluttkommentarer

Den grunnleggende motivasjonen for at bootstrap metoden gir gode tilnærminger til de sanne egenskaper til estimatet $\hat{\theta}$ ligger i at $\widehat{F}$ er en rimelig tilnærming til F. En rigid matematisk utledning av hvordan bootstrap estimater oppfører seg vil ikke bli tatt opp her, men det viktigste resultatet er at bootstrap estimatene under svært generelle og milde antagelser er *asymptotisk* korrekte. Med dette mener vi at når antall observasjoner går mot uendelig, så vil bootstrap estimatene bli mer og mer korrekte.

Under diskusjonen om konfidensintervall ble også normaltilnærmingen nevnt som en mulig metode. Tilsvarende som for bootstrap metodene, er også normalfordelingstilnærmingen korrekt i asymptotisk forstand. Motivasjonen for å bruke den mer kompliserte bootstrap metodikken er at denne stort sett vil være *bedre* i den forstand at for et endelig antall observasjoner vil den feilen vi gjør under tilnærming typisk være mindre enn den vi får ved en normaltilnærming. For mer diskusjon om bootstrap prinsippet, referer vi til Efron and Tibshirani [1993] og Davison and Hinkley [1997].

References

A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Applications*. Cambridge Series in Statistics and Probabilistics Mathematics. Cambidge University Press, Cambridge, 1997.

J.L. Devore and K.N. Berk. *Modern mathematical statistics with applications*. Duxbury Pr, 2007. ISBN 0534404731.

B. Efron. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Number 38 in CBMS-NSF Regional Conference Series in Applied Mathematics. Siam, Philadelphia, 1982.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

J. A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, California, 1995.