Mål for asymmetri og kurtose

Masteroppgåve i statistikk

Arne Vestrheim

Matematisk institutt
Universitetet i Bergen

13. april 2010

Føreord

Gjennom heile skulegangen min har eg likt å arbeide sjølvstendig for best læring. Diverre har det ikkje alltid vore godt tilrettelagt for denne arbeidsmåten. Difor var det ein veldig fin overgang å byrje å studere ved universitetet, då ein her i mykje større grad kunne velje arbeidsmåte sjølv. Diverre har sjølvstendig læring ved Universitetet i Bergen vorte meir og meir avgrensa av ulike reformer gjennom studietida. Difor var det igjen ei fin oppleving å byrje på master, då det i ein mastergrad er nettopp denne arbeismåten som står i fokus. Likevel følte eg at eg trengte litt oppfølgjing i byrjinga av mastergradsarbeidet, då eg hadde vore borte frå studiene i lengre tid grunna sjukdom. I denne situasjonen var rettleiar Ivar Heuch til god hjelp. Eg har alltid likt den teoretiske biten av fagfeltet, og han la fram eit forslag til oppgåve som i utgangspunktet gjekk i den retninga, men som likevel gav mogelegheiter til utrekningsorienterte og eksperimentelle problemstillingar. I tillegg la han i stor grad til rette for sjølvstendige arbeidsmetodar, men gav likevel god og detaljert oppfølgjing når det var ønskeleg. Av den grunn vil eg rette han ei stor takk for god rettleiing gjennom heile mastergradsarbeidet. Eg vil også takke Kristine Selvikvåg for hjelp til korrekturlesing og organisering av oppgåva, og eg vil takke medstudentar på Kroepelien for godt fagleg og sosialt miljø. Til slutt vil eg takke alle som har vore med i Matematisk fagutval, ein organisasjon eg har vore delaktig i gjennom heile studietida. Statistikkdelen av instituttet er fråskild rein- og utrekningsorientert matematikk, og gjennom fagutvalet har eg vorte kjent med studentar og tilsette ved alle deler av instituttet.

Arne Vestrheim,
Bergen, April 2010

Innhald

Innleiing

Dei vanlegaste observatorane for å skildre datasett, tar utganspunkt i lokaliseringa og variasjonen til datamaterialet. Dette kjem til uttrykk i dei statistiske omgrepa gjennomsnitt og varians, og mål basert på moment er gode mål for dette. Det neste steget for å skildre forma til fordelinga, vil vere å sjå på skeivleik og kurtose. Historisk er dette mål som vart innført for å skildre data som ikkje var normalfordelte, og vart fyrst nytta til å skildre finansdata. Her var det ulike sannsyn for gevinst og tap, og uteliggarar forekom ofte. Mål for asymmetri og kurtose var difor naudsynt for å lage gode modellar. Seinare har mål for asymmetri og kurtose vore brukt på mange område.

Skeivleiken er eit mål på forma til fordelinga, og skal måle i kor stor grad fordelinga er symmetrisk. Det tradisjonelle målet for skeivleik tar utgangspunkt i moment omkring forventningsverdien. Dette målet har fleire veikskapar, som til dømes at målet ikkje fangar opp kvar fordelinga er asymmetrisk. Det kan også klassifisere ei fordeling som symmetrisk, sjølv om dette ikkje er tilfellet. Ein artikkel av Critchley og Jones (2008) innfører eit nytt mål for asymmetri. Istadenfor moment omkring forventningsverdi, føreslår dei heller å ta utgangspunktet i modalverdien til fordelinga. Dette er ingen ny ide, og vart fyrst føreslått av Pearson (1895). I nyare tid er dette oppsummert og drøfta av Oja (1981) og Arnold og Groeneveld (1995). Det nye i artikkelen til Critchley og Jones (2008), er innføring av asymmetrifunksjonar. Sjølv om Averous et al. (1996) og Boshnakov (2007) også er inne på dette, føreslår dei ikkje så konkrete mål og funksjonar som Critchley og Jones (2008). Fordelen med desse måla er at dei skildrar kvar fordelinga er asymmetrisk.

Kurtose er eit omgrep som ikkje er like presist som symmetri. Både kurtosen og skeivleiken måler forma på fordelinga, men der skeivleik måler asymmetri måler kurtose spissheit og haletyngda. Med høg kurtose er tettleiken spiss med tunge halar. På same måte som for asymmetri, er det tradisjonelle målet basert på moment. Diverre er ikkje målet basert på moment fullgodt til å skildre desse eigenskapane. Fleire fordelingar manglar moment, og det er heller

ikkje alltid kurtosen samsvarar med eigenskapane han er meint å måle. Fiori og Zenga (2009) gjev ei innføring i opphavet til kurtose, medan Balanda og MacGillivray (1988) ser på utviklinga av konseptet og kva det bør tyde. Artikkelen til Critchley og Jones (2008) kjem også med eit forslag til nytt mål på kurtose. Dei bruker den deriverte til tettleiken og innfører kurtosefunksjonar.

I denne oppgåva ser vi på dei nye måla for asymmetri og kurtose Critchley og Jones (2008) innfører. I kapittel 1 definerer vi og ser på døme for det nye målet for asymmetri. Vidare ser vi på kurtose i kapittel 2, der vi også definerer og ser på døme for det nye målet. I kapittel 3 utarbeidar vi teori for estimering av dei nye måla, før vi i kapittel 4 brukar teorien og måla på datasett. I det siste kapittelet oppsummerar vi oppgåva med å sjå på vidare arbeid og gjere ein konklusjon omkring dei nye måla.

Kapittel 1

Nytt mål for asymmetri

I dette kapittelet skal vi sjå på asymmetri, og det er dette som er kjernen i artikkelen til Critchley og Jones (2008). I byrjinga definerer vi matematiske omgrep som vi treng både til det nye målet for asymmetri, men også til det meir tradisjonelle målet. Det nye målet brukar funksjonar til å skildre asymmetrien, og vi prøver å forklare ideen bak dette og kva det nye målet skildrar. Deretter ser vi på analytiske døme på asymmetrifunksjonen. Her drøftar vi eigenskapane til ulike asymmetrifunksjonar og forsøker å kartlegge generelle eigenskapar til målet. Undervegs føreslår vi eit nytt mål sjølve, som hentar inspirasjon frå fleire ulike mål for asymmetri. Til slutt i kapittelet ser vi på skalare mål for asymmetri og, samanliknar målet til Critchley og Jones (2008) det tradisjonelle målet.

1.1 Definisjonar

Vi byrjar med å definere sentrale statistiske omgrep som forventningsverdi, moment og skeivleikskoeffisienten frå Casella og Berger (2002). Deretter tar vi for oss andre matematiske eigenskapar vi treng for å definere det nye målet til Critchley og Jones (2008).

Definisjon 1.1.1. *Forventningsverdien til ein funksjon $g(x)$, med hensyn på ei tettleik $f(x)$, er integralet til produktet av funksjonen og tettleiken, det vil seie $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x$.*

Definisjon 1.1.2. *Det n-te sentralmomentet er definert som*

$$\mu_n = E\left[(X - E\left[X\right])^n\right] \ . \tag{1.1}$$

Definisjon 1.1.3. *Skeivleikskoeffisienten er definert som*

$$s = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} . \tag{1.2}$$

Definisjon 1.1.4. *Modalverdien til ein funksjon, er den verdien som gir høgast funksjonsverdi. Ein eintoppa funksjon har berre ein modalverdi, og er strengt stigande fram til modalverdien. Deretter strengt synkande.*

Definisjon 1.1.5. *Ein funksjon, $f(x)$, er rota på eit intervall, (a, b), dersom $f(a) = f(b) = 0$. Intervallet (a, b) treng ikkje vere avgrensa, dvs at (a, b) kan ta verdiane $a = -\infty$ og $b = \infty$.*

Definisjon 1.1.6. *La $f(x)$ vere ein eintoppa funksjon definert på (a, b), med modalverdi m. Då er $f_L(x)$ den venstre delen av funksjonen, definert på $x \in (a, m)$, og $f_R(x)$ er den høgre del av funksjonen, definert på $x \in (m, b)$.*

1.1.1 Asymmetrifunksjonane til Critchley og Jones

Vi skal no bruke definisjon 1.1.4)-1.1.6 frå førre avsnitt til å definere asymmetrifunksjonen til Critchley og Jones (2008). La $f(x)$ vere ein rota eintoppa tettleiksfunksjon på området (a, b), med modalverdi m. La deretter

$$X_R(p) = f_R^{-1}(pf(m)) , \tag{1.3}$$
$$X_L(p) = f_L^{-1}(pf(m)) . \tag{1.4}$$

Til slutt definerer vi avstandsfunksjonane som gir avstanden frå modalverdi til X_R og X_L,

$$\tau_R(p) = X_R(p) - m , \tag{1.5}$$
$$\tau_L(p) = m - X_L(p) . \tag{1.6}$$

På denne måten blir X_R og X_L funksjonar av p multiplisert med funksjonsverdien til modalverdien. Dette er verdiane til høgreinversen og venstreinversen til f. Variabelen p er definert på $(0, 1)$, medan funksjonsverdien i m er konstant. Høgre- og venstreinversen får dermed riktig definisjonsområde. Avstandsfunksjonane blir på same måte også ein funksjon av p. Sjå figur 1.1 for forklaring. Asymmetrifunksjonane til Critchley og Jones (2008) er definert som

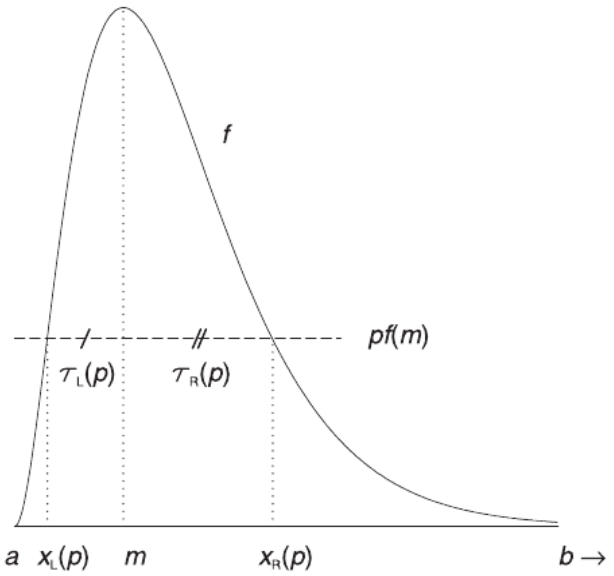
Definisjon 1.1.7.

$$\rho(p) = \frac{\tau_{R(p)}}{\tau_{L(p)}} , \qquad 0 < p < 1 . \tag{1.7}$$

Definisjon 1.1.8.

$$\gamma^*(p) = \frac{\rho(p) - 1}{\rho(p) + 1} = \frac{\tau_{R(p)} - \tau_{L(p)}}{\tau_{R(p)} + \tau_{L(p)}}\,, \qquad 0 < p < 1\,. \tag{1.8}$$

Vi har her definert asymmetrifunksjonar for nivå, til motsetnad frå eit skalart mål for ein heil tettleik. Funksjonen $\rho(p)$ gir oss tilhøvet mellom avstanden til den høgre delen og avstanden til den venstre delen av tettleiken, og kan ta verdiar på intervallet $(0, \infty)$. Funksjonen $\gamma^*(p)$ gir oss ein normalisert ulikskap i forholdet mellom avstanden til den høgre- og til den venstre delen, og tar verdiar på intervallet $(-1, 1)$. Fordelen med dette målet er at det skildrar skeivleiken for heile fordelinga, og ikkje berre ein samla skalar for heile asymmetrieb. På den måten vil vi fange opp kvar fordelinga er skeiv og kvar ho er symmetrisk. Vi vil også kunne fange opp om fordelinga ikkje er symmetrisk sjølv om skeivleikskoeffisienten skulle vere 0.



Figur 1.1: *Vi ser at asymmetrifunksjonen tar utgangspunkt i modalverdien, m, og reknar tilhøve mellom avstanden til høgre del, τ_L, og avstanden til venstre del, τ_R, av tettleiken. Det som avgjer kvar vi reknar avstandane er $pf(m)$, der $f(m)$ er funksjonsverdien i modalverdien og p er eit tal mellom null og ein. På denne måten blir avstandsfunksjonane og asymmetrifunksjonane definert på intervallet $p \in (0, 1)$. (Critchley og Jones (2008))*

1.2 Døme på asymmetrifunksjonar

I avsnitt 1.1.1 definerte vi to funksjonar for asymmetri. I dette avsnittet vil vi finne døme på desse funksjonane for ulike fordelingar, og drøfte kva eigenskapar målet har.

1.2.1 Tettleik av andregradspolynomen

Vi byrjar med å sjå på sannsynstettleiken,

$$f(x) = \begin{cases} -x^2 + 1 \,, & -1 < x < 0 \\ -\frac{1}{4}x^2 + 1 \,, & 0 < x < 2 \,. \end{cases} \tag{1.9}$$

Vi ser at f er ein tettleik sidan $\int f(x)\,\mathrm{d}x = 1$ og at $f \geq 0$ for alle x. Modalverdien er $m = 0$ og $f(m) = 1$. Vi bruker definisjon 1.1.6 og likningane (1.4)- (1.6) frå avsnitt 1.1.1 til å dele opp funksjonen i venstre- og høgredel,

$$f_L(u) = -u^2 + 1 \,, \qquad -1 < u < 0 \,, \tag{1.10}$$

$$f_R(u) = -\frac{1}{4}u^2 + 1 \,, \qquad 0 < u < 2 \,. \tag{1.11}$$

Deretter finn vi høgre- og venstreinversen, som vi skal bruke til å finne X_R og X_L,

$$f_L^{-1}(x) = -\sqrt{1-x} \,, \qquad 0 < x < 1 \,, \tag{1.12}$$

$$f_R^{-1}(x) = 2\sqrt{1-x} \,, \qquad 0 < x < 1 \,. \tag{1.13}$$

I avsnitt (1.1.1) definerte vi avstandsfunksjonane vi bruker til å rekne ut asymmetrifunksjonane ρ og γ^*, og vi skal no rekne ut desse,

$$\tau_{L(p)} = m - X_L(p) = m - f_{L^{-1}}(pf(m)) = \sqrt{1-p} \,, \tag{1.14}$$

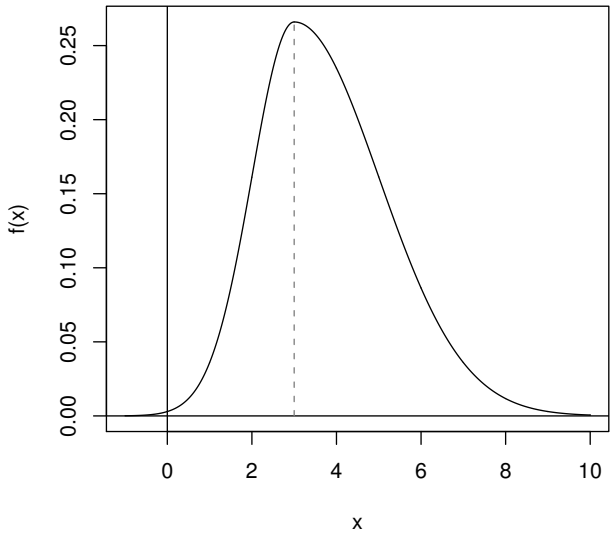
$$\tau_R = 2\sqrt{1-p} \,, \tag{1.15}$$

$$\rho(p) = \frac{\tau_R}{\tau_L} = \frac{2\sqrt{1-p}}{\sqrt{1-p}} = 2 \,. \tag{1.16}$$

Vi ser at asymmetrifunksjonen er ein konstant i dette tilfellet; det er dobbelt så stor avstand frå modalverdien 0 til X_R som til X_L.

$$\gamma^*(p) = \frac{\tau_R - \tau_L}{\tau_R + \tau_L} = \frac{2\sqrt{1-p} - \sqrt{1-p}}{2\sqrt{1-p} + \sqrt{1-p}} = \frac{1}{3} \,. \tag{1.17}$$

Denne asymmetrifunksjonen er også ein konstant, med forholdet mellom differansen i avstanden frå modalverdien og avstanden mellom X_R og X_L.



Figur 1.2: *Samansett normalfordeling med modalverdi* $m = 3$, $\sigma_1 = 1$ *og* $\sigma_2 = 2$.

1.2.2 Samansett normalfordeling

Her skal vi sjå på ei fordeling som er sett saman av to normalfordelingar med ulik varians. Sjå figur 1.2. La

$$f(x) = \begin{cases} \frac{\sqrt{2}}{\sqrt{\pi}(\sigma_1+\sigma_2)} e^{-\frac{(x-m)^2}{2\sigma_1^2}}, & -\infty < x < m \\ \frac{\sqrt{2}}{\sqrt{\pi}(\sigma_1+\sigma_2)} e^{-\frac{(x-m)^2}{2\sigma_2^2}}, & m < x < \infty \, . \end{cases} \tag{1.18}$$

der σ_1 og σ_2 er konstantar større enn null medan m er modalverdien. Vi ser at fordelinga er kontinuerleg i m, ettersom

$$f_L(m) = f_R(m) = \sqrt{2}/(\sqrt{\pi}(\sigma_1 + \sigma_2)).$$

Den deriverte er også kontinuerleg i modalverdien då

$$f_L'(m) = f_R'(m) = 0 \, .$$

Den andrederiverte til f eksisterer, men er ikkje kontinuerleg i m, ettersom

$$f_L''(m) = \sqrt{2}/(\sqrt{\pi}\sigma_1(\sigma_1 + \sigma_2)) \, ,$$

medan

$$f_R''(m) = \sqrt{2}/(\sqrt{\pi}\sigma_2(\sigma_1 + \sigma_2)) \, .$$

Avstandsfunksjonane blir

$$\tau_L(p) = \sigma_1\sqrt{\log(p^{-2})}\ , \tau_{R(p)} = \sigma_2\sqrt{\log(p^{-2})}\ ,$$

og asymmetrifunksjonane til fordelinga blir

$$\rho(p) = \frac{\sigma_1\sqrt{\log(p^{-2})}}{\sigma_2\sqrt{\log(p^{-2})}} = \frac{\sigma_2}{\sigma_1}\ , \tag{1.19}$$

og

$$\gamma^*(p) = \frac{\sigma_2 - \sigma_1}{\sigma_2 + \sigma_1}\ . \tag{1.20}$$

Teorem 1.2.1. *La $f(x)$ vere ei sannsynstettleik med modalverdi m. Dersom $f_R(m+x) = f_L(m-ax)$,der a er ein konstant, blir også asymmetrifunksjonen $\rho(p)$ ein konstant. Nærmare bestemt er $\rho(p) = \frac{1}{a}$ for alle p.*

Bevis. Vi byrjar med definisjonen av τ_R og τ_L:

$$\tau_R(p) = X_R(p) - m\ ,$$
$$\tau_L(p) = m - X_L(p)\ .$$

Vi snur om på uttrykket, slik at

$$X_{R(p)} = \tau_R(p) + m\ .$$

Vidare bruker vi definisjonane til X_R og X_L,

$$f_R(X_R(p)) = f_R(\tau_R(p) + m) = pf(m) = f_L(-a\tau_R(p) + m) = f_L(X_L(p))\ ,$$

frå føresetnadane i starten. Ut frå dette finn vi at

$$X_L(p) = -a\tau_R(p) + m = m - \tau_L(p)\ .$$

Vi snur om på den siste likninga og får dermed definisjonen av ρ:

$$\frac{\tau_R(p)}{\tau_L(p)} = \frac{1}{a} = \rho(p)\ .$$

□

Korrolar 1.2.2. *Dersom $\rho(p)$ er konstant, er også $\gamma^*(p)$ konstant.*

Bevis. Vi har $\rho(p) = a$. Då blir $\gamma^*(p) = \frac{\rho(p)-1}{\rho(p+1)} = \frac{a-1}{a+1}$, som igjen er ein konstant. □

Her har vi sett døme på asymmetrifunksjonar som er konstante, og frå teorem 1.2.1 har vi klarlagt kva som gjer dette. Dersom fordelinga vi vil måle har asymmetrifunksjon $\rho(p) = 1$, eventuelt $\gamma^*(p) = 0$, er fordelinga symmetrisk. Vi kan også seie at fordelinga har større grad av symmetri dersom asymmetrifunksjonen er ein konstant og dermed ikkje avhenger av p. Dette vil riktignok vere ein ny måte å sjå symmetri på, ettersom det vanlegaste skeivleiksmålet i størst grad vil ta hensyn til verdien til koeffisienten. På same måte kan vi seie at dersom asymmetrifunksjonen varierer mykje med p, er fordelinga lite symmetrisk, sjølv om den tradisjonelle skeivleikskoeffisienten skulle vere 0.

1.2.3 Tettleik konstruert av polynom med ulik grad

I dei fyrste døma såg vi på tettleikar med konstante asymmetrifunksjonar. No skal vi sjå på asymmetrifunksjonar som varierer med p. Vi byrjar med tettleiken som vi kan sjå i figur 1.3,

$$g(x) = \begin{cases} -ax^2 + c\,, & -\alpha < x < 0 \\ -bx^4 + c\,, & 0 < x < \beta, \end{cases} \tag{1.21}$$

der a, b, c, α, $\beta > 0$. For at $g(x)$ skal vere ei rota tettleik, må $g(x)$ vere null i endepunkta $x = -\alpha$ og $x = \beta$, slik at

$$g(-\alpha) = -a\alpha^2 + c = 0\,, \tag{1.22}$$

$$g(\beta) = -b\beta^4 + c = 0\,. \tag{1.23}$$

Dette medfører at

$$c = a\alpha^2 = b\beta^4\,. \tag{1.24}$$

Integralet over alle verdiane må også vere ein, slik at

$$\int_{-\alpha}^{\beta} g(x)\,\mathrm{d}x = -\frac{a}{3}\alpha^3 + c\alpha - \frac{b}{5}\beta^5 + c\beta = 1\,. \tag{1.25}$$

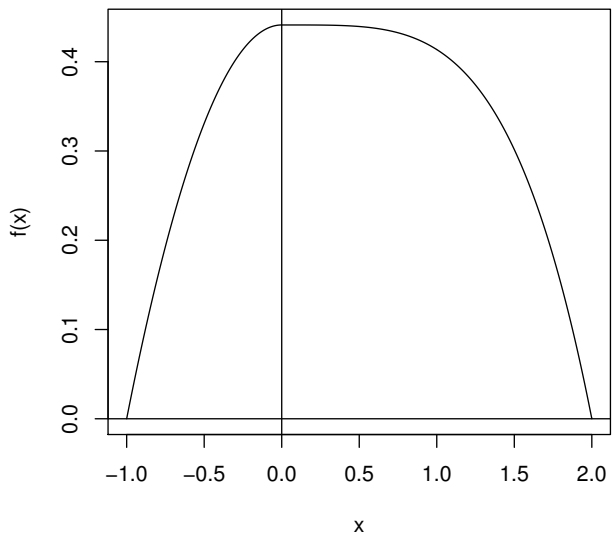
Vi set likning 1.24 inn i likning 1.25 og får

$$\frac{2}{3}c\alpha + \frac{4}{5}c\beta = 1\,. \tag{1.26}$$

Dersom vi snur om på dette, finn vi krava til α, β, a og b:

$$\frac{15}{10\alpha + 12\beta} = c\,, \tag{1.27}$$

$$a\alpha^2 = b\beta^4 = \frac{15}{10\alpha + 12\beta}\,. \tag{1.28}$$



Figur 1.3: *Fordeling sett saman av andregrads- og fjerdegradspolynomen. Her er* $\alpha = 1$, $\beta = 2$ *og* $c = \frac{15}{34}$.

Eit av krava for at asymmetrifunksjone skal vere definert, er at tettleiken må vere eintoppa. Dette tyder at han er strengt stigande fram til modalverdien og deretter synkande. Vi ser difor på den deriverte til g,

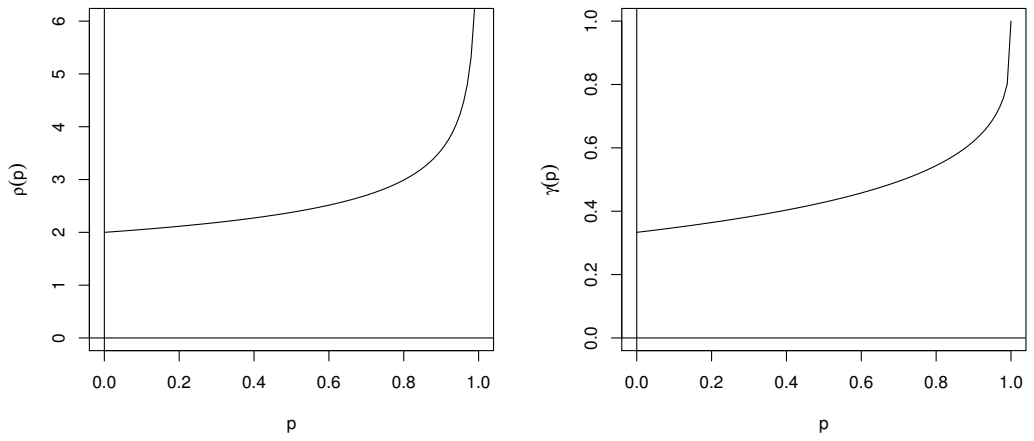
$$g'(x) = \begin{cases} -2ax\ , & -\alpha < x < 0 \\ -4bx^3\ , & 0 < x < \beta\ . \end{cases} \tag{1.29}$$

Vi ser at g' er positiv for $x < 0$ og negativ for $x > 0$. Tettleiken g er difor både rota og eintoppa, og asymmetrifunksjonen er dermed definert. Vi ser også at tettleiken er kontinuerleg, ettersom $\lim_{x\to 0^-} g(x) = \lim_{x\to 0^+} g(x) = c$. Den deriverte er også kontinuerleg då $\lim_{x\to 0^-} g'(x) = \lim_{x\to 0^+} g'(x) = 0$, og g er dermed ein glatt funksjon.

Vi har no funne forutsetningane som må ligge til grunn for at g skal vere ei sannsynstettleik og for at asymmetrifunksjonen skal vere definert. Vidare er det to måtar å la g oppfylle desse krava: Den fyrste måten er å velje a og b, og deretter bestemme α og β ut frå det. Den andre måten vil vere omvendt, og det er dette vi vil gjere i denne oppgåva. Vi byrjar med å finne høgre- og venstreinversen:

$$g_L^{-1}(x) = -(\frac{1}{a}(c-x))^{\frac{1}{2}}\ , \qquad 0 < x < c \tag{1.30}$$

$$g_R^{-1}(x) = (\frac{1}{b}(c-x))^{\frac{1}{4}}\ , \qquad 0 < x < c\ . \tag{1.31}$$



Figur 1.4: *Asymmetrifunksjonane til fordelinga sett saman av polynom av ulik grad, som kan finne i likning (1.21). Asymmetrifunksjonen $\rho(p)$ er til venstre, medan $\gamma^*(p)$ er til høgre. Konstantane er $\alpha = 1$ og $\beta = 2$.*

Vidare finn vi avstandsfunksjonane

$$\tau_L(p) = a^{-\frac{1}{2}}(c - cp)^{\frac{1}{2}} = \alpha(1-p)^{\frac{1}{2}} \,, \tag{1.32}$$

$$\tau_R(p) = b^{-\frac{1}{4}}(c - cp)^{\frac{1}{4}} = \beta(1-p)^{\frac{1}{4}} \,. \tag{1.33}$$

Under forutsetningane

$$a\alpha^2 = b\beta^4 = \frac{15}{10\alpha + 12\beta} \,, \tag{1.34}$$

blir asymmetrifunksjonane til slutt

$$\rho(p) = \frac{\beta}{\alpha}(1-p)^{-\frac{1}{4}} \,, \tag{1.35}$$

$$\gamma^*(p) = \frac{\beta - \alpha(1-p)^{\frac{1}{4}}}{\beta + \alpha(1-p)^{\frac{1}{4}}} \,, \tag{1.36}$$

Dersom vi ser på monotonieigenskapane til asymmetrifunksjonen, ser vi at funksjonen er strengt stigande ettersom

$$\rho'(p) = \frac{\beta}{4\alpha}(1-p)^{-\frac{5}{4}} > 0 \,, \tag{1.37}$$

for alle $p \in (0,1)$. Vi ser også at funksjonen er konveks, ettersom

$$\rho''(p) = \frac{5\beta}{16\alpha}(1-p)^{-\frac{9}{4}} > 0 \,, \tag{1.38}$$

for alle $p \in (0,1)$. Dette er også illustrert i figur 1.4, der vi ser at α og β opptrer som ein konstant faktor, og at monotonieigenskapane berre er avhengig av p. Asymmetrifunksjonen $\rho(p)$ er stigande og definert på $(2,\infty)$, medan $\gamma^*(p)$ også stigande, men definert på $(1/3,1)$. At asymmetrifunksjonen er stigande, er uttrykk for at asymmetrien er størst omkring modalverdien, og definisjonsområdet viser at fordelinga er skeiv mot høgre.

Dersom vi ser på grenseverdiane til asymmetrifunksjonen, $\lim_{p\to 0^+}\rho(p) = \alpha/\beta$ og $\lim_{p\to 1^-}\rho(p) = \infty$, viser dei at funksjonen er asymmetrisk med faktoren $\frac{\alpha}{\beta}$ ved røtene, men er svært asymmetrisk omkring modalverdien. Vi kan sjå av figur 1.3 at funksjonen er relativt flat på toppen, og det er dette asymmetrifunksjonen viser i grenseverdien når p går mot 1. Funksjonar med eigenskapar som dette, kan vere vanskeleg å estimere, og då særskild modalverdien. Dette kjem vi tilbake til i kapittel 3. Vi kan også undersøke monotonieigenskapane til $\gamma^*(p) = (\rho(p)-1)/(\rho(p)+1)$. Denne brøken vil som nemnd tidlegare ta verdiar mellom -1 og 1, og kan difor gi eit meir objektivt mål på kor asymmetrisk tettleiken er. Asymmetrifunksjonen $\gamma^*(p)$ er også strengt stigande sidan den deriverte,

$$\gamma^{*\prime}(p) = \frac{\alpha\beta(1-p)^{-\frac{3}{4}}}{(\beta+\alpha(1-p)^{\frac{1}{4}})^2}\,, \tag{1.39}$$

er større enn 0 for alle $p \in (0,1)$. Dette finn vi av at alle faktorane er større enn null. Vidare ser vi at også $\gamma^*(p)$ er konveks. Den andrederiverte,

$$\gamma^{*\prime\prime}(p) = \frac{3\alpha\beta^3(1-p)^{-\frac{1}{4}} + 8\alpha^2\beta^2}{4(\beta^2(1-p)^{\frac{3}{4}} + 2\alpha\beta(1-p) + \alpha^2)^2}\,, \tag{1.40}$$

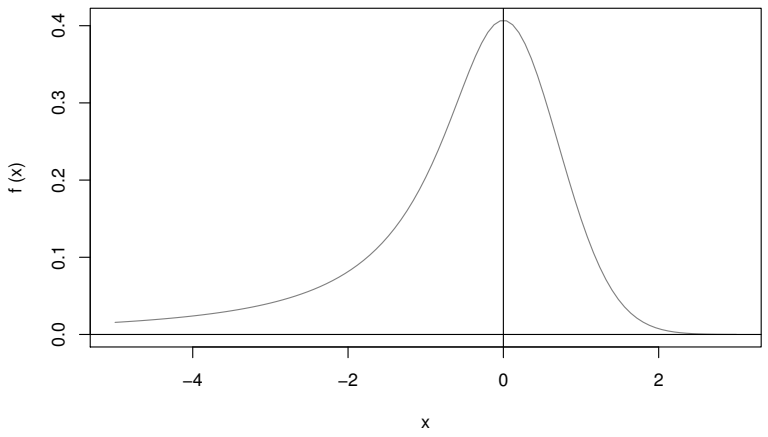
er større enn null for $p \in (0,1)$ ettersom alle faktorar og ledd er positive. Til slutt ser vi på grenseverdiane. Dei er $\lim_{p\to 0^+}\gamma^*(p) = (\beta-\alpha)/(\beta+\alpha)$ og $\lim_{p\to 1^-}\gamma^*(p) = 1$. Dette viser igjen at tettleiken er svært asymmetrisk omkring modalverdien, medan den relative skilnaden i avstand frå modalverdi avgjer asymmetri ved røtene.

Teorem 1.2.3. *Monotonieigenskapane til $\rho(p)$ og $\gamma^*(p)$ er like.*

Bevis. Vi byrjar med at $\gamma^*(p) = \frac{\rho(p)-1}{\rho(p)+1}$. Deretter finn vi den deriverte til γ^*:

$$\frac{\mathrm{d}\gamma^*}{\mathrm{d}p} = \frac{\mathrm{d}}{\mathrm{d}p}\frac{\rho(p)-1}{\rho(p)+1} = \frac{2\rho'(p)}{(\rho(p)+1)^2}\,. \tag{1.41}$$

Ettersom nemnaren alltid er større enn null, vil brøken berre endre forteikn saman med $\rho'(p)$ og dei to asymmetrifunksjonane har same monotonieigenskapar. □



Figur 1.5: *Fordeling sett saman av Cauchy- og normalfordeling som vi finn i likning (1.43).*

Sjølv om dei ulike asymmetrifunksjonane har same monotonieigenskapar, treng dei ikkje ha like krumningseigenskapar. Den andrederiverte til den normaliserte asymmetrifunksjonen,

$$\frac{\mathrm{d}^2\gamma^*}{\mathrm{d}p^2} = \frac{(\rho(p)+1)(\rho''(p)(\rho(p)+1) - 2(\rho'(p))^2)}{(\rho(p)+1)^4} \,, \tag{1.42}$$

er ikkje nødvendigvis positiv eller negativ for same verdiar av p for dei to ulike asymmetrifunksjonane.

1.2.4 Cauchy-normalfordeling

I dette avsnittet skal vi sjå på ei fordeling som er sett saman av deler av Cauchyfordelinga og normalfordelinga. Fordelinga er illustrert i figur 1.5, og gitt ved

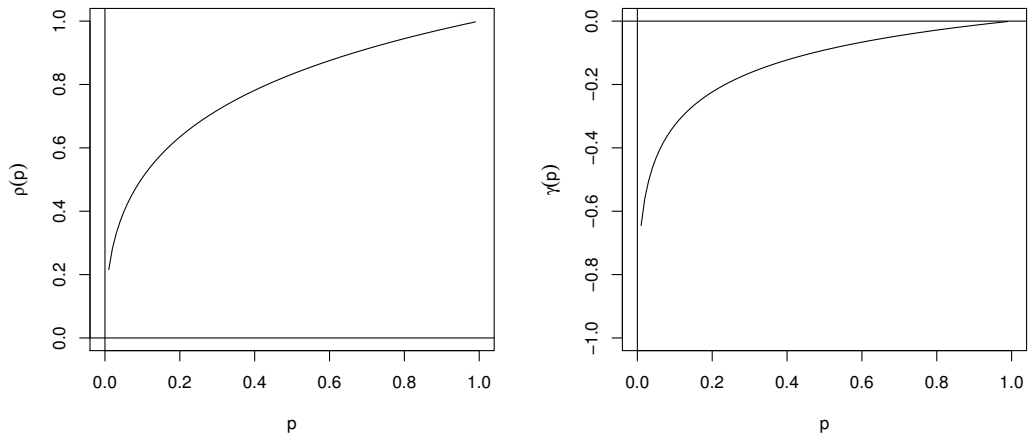
$$f(x) = \begin{cases} k\frac{1}{1+x^2} & x < 0 \\ ke^{-x^2} \,, & 0 < x \,. \end{cases} \tag{1.43}$$

der integreringskonstanten $k = 2/(\pi + \sqrt{\pi})$. Avstandsfunksjonane til f blir

$$\tau_R(p) = \sqrt{-\log(p)} \,, \tag{1.44}$$

$$\tau_L(p) = \sqrt{\frac{1}{p} - 1} \,, \tag{1.45}$$

dermed blir asymmetrifunksjonane



Figur 1.6: *Asymmetrifunksjonane til fordelinga sett saman av Cauchy- og normalfodeling som vi finn i likning (1.43), med ρ til høgre og γ^* til venstre.*

$$\rho(p) = \sqrt{\frac{-p\log(p)}{1-p}}\,, \tag{1.46}$$

$$\gamma^*(p) = \frac{\sqrt{-\log p} - \sqrt{\frac{1}{p}-1}}{\sqrt{-\log(p)} + \sqrt{\frac{1}{p}-1}}\,, \tag{1.47}$$

og den deriverte til ρ er

$$\frac{\mathrm{d}\rho}{\mathrm{d}p} = \left(\frac{-p\log(p)}{1-p}\right)^{-\frac{1}{2}} \left(\frac{1+p-\log(p)}{(1-p)^2}\right)\,. \tag{1.48}$$

Ettersom $-\log(p)$ er større enn null på heile intervallet, er også $1+p-\log(p)$ større enn null. Av same grunn er brøken

$$\left(\frac{-p\log(p)}{(1-p)}\right)^{-\frac{1}{2}}\,,$$

også større enn null, og dette er dessutan naudsynt for at den deriverte skal vere definert. Ut frå dette ser vi at den deriverte er positiv og at begge asymmetrifunksjonane er stigande. I figur 1.6, ser vi at ρ er stigande mot ein, medan γ^* er negativ og stigande mot null. Begge funksjonane uttrykker dermed at asymmetrien er størst i halane og at fordelinga er skeiv mot venstre. Når p går mot 1, ser vi kva som hender omkring modalverdien,

$$\lim_{p\to 1}\rho(p) = \lim_{p\to 1}\frac{-p\log p}{1-p} = 1\,. \tag{1.49}$$

Her kan vi sjå at tettleiken er symmetrisk omkring modalverdien. Når p går mot 0 ser vi på halane,

$$\lim_{p\to 0}\rho(p) = \lim_{p\to 1}\frac{-p\log p}{1-p} = 0\ . \tag{1.50}$$

Vi ser at venstre del av fordelinga har mykje tyngre halar enn høgre del, noko som skulle stemme bra då Cauchyfordelinga er kjend for å ha tunge halar. I neste kapittel skal vi sjå på kurtose, og det blir ofte sett på som eit mål på haletyngde. Ut frå dette dømet ser vi at haletyngde speler ei sentral rolle i symmetrimålet Critchley og Jones (2008) innfører, og er eit godt døme på at symmetri og kurtose heng saman.

1.3 Skalare mål for asymmetri

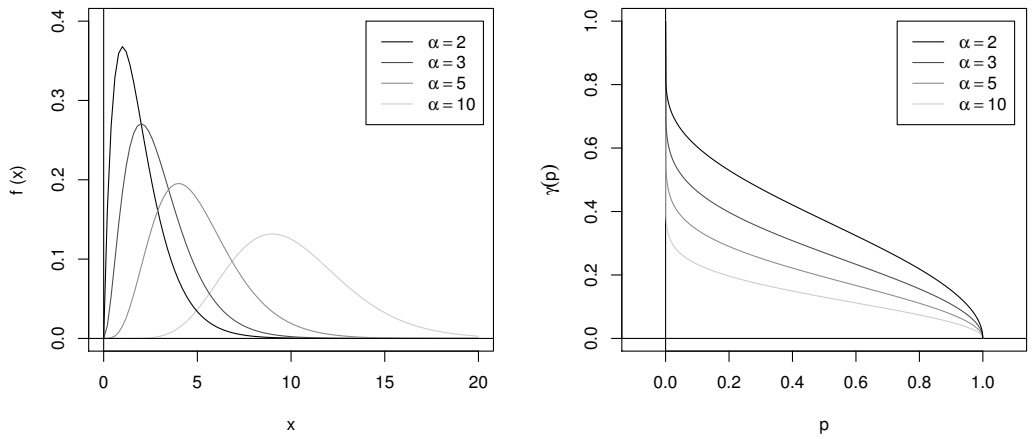
Som vi har sett, innfører Critchley og Jones (2008) nivåbaserte asymmetrifunksjonar. Likevel gir dei eit forslag for korleis ein kan konvertere desse funksjonane til skalare mål,

Definisjon 1.3.1.

$$\gamma = \int_0^1 \gamma(p)w_f(p)\,\mathrm{d}p \tag{1.51}$$

der $\gamma(p)$ er ein nivådefinert asymmetrifunksjon og $w_f(p)$ er ein tetteleik på $[0,1]$.

Ein av føresetnadane i denne definisjonen, er at integralet i likning (1.51) konvergerer. Ettersom $w(p)$ er ein tettleik på $[0,1]$ er vi sikra dette dersom $\gamma(p)$ er endeleg. Asymmetrifunksjonen $\gamma^*(p)$ oppfyller dette kravet for alle fordelingar, ettersom denne funksjonen tar verdiar på intervallet $[-1,1]$, i motsetnad til $\rho(p)$ som er definert for alle positive tal. Vi vil difor bruke $\gamma^*(p)$ når vi skal rekne ut skalare verdiar seinare. Tettleiken $w_f(p)$ styrer kva del av asymmetrifunksjonen vi skal legge vekt på. Dersom $w_f(p)$ gjer høge verdiar for p i nærleiken av ein, legg γ vekt på asymmetrien omkring modalverdien. Motsett legg γ vekt på ulikskapen i haletyngde dersom $w_f(p)$ er stor for p i nærleiken av null. Vidare i dette avsnittet vil vi bruke $w_{f(p)} = 1$, altså uniform fordeling, når vi skal samanlikne det nye målet med det gamle. På den måten legg vi like stor vekt på alle verdiar av p. Frå definisjonane 1.1.1-1.1.3 hugsar vi korleis μ_n og s var definert. Å innføre ein skeivleikskoeffisient som tar hensyn til modalverdien, men også brukar moment på same måte som den s, kan vere ein mellomting mellom γ og s. Av den grunn innfører vi to nye mål:



Figur 1.7: *Gammafordeling til venstre med tilhøyrande asymmetrifunksjonar til høgre. For alle fordelingane har vi vald* $\beta = 1$ *ettersom dette ikkje påverkar skeivleiken.*

Definisjon 1.3.2. *Det n-te modalmomentet til* X, μ_n^* *definerer vi som* $\mu_n^* = E\left[(X - m)^n\right]$ *der* m *er modalverdien til* X.

Definisjon 1.3.3. *Den modale skeivleikskoeffisienten definerer vi som*

$$s^* = \frac{\mu_3^*}{\mu_2^{*\frac{3}{2}}} \, . \tag{1.52}$$

Den modale skeivleikskoeffisienten har i stor grad dei same eigenskapane som målet basert på sentralmoment. Særskild vil halane til fordelinga bidra like mykje i begge måla. Likevel vil det bli ulikskapar dersom modalverdien ligg langt frå forventningsverdien.

Vi vil no samanlikne dei nye måla våre med tradisjonell skeivleik. På den måten vil vi kartlegge dei ulike eigenskapane ved måla. Til det brukar vi eit program som er skrive i språket R, der vi bruker funksjonen uniroot for å numerisk å rekne ut dei ulike avstandsfunksjonane med påfølgjande asymmetrifunksjon. Metodar som blir brukt er vedlagt i tillegg C.

1.3.1 Gammafordeling

Gammafordelinga er ei eintoppa rota fordeling som i tillegg er asymmetrisk. Denne fordelinga har vore referanse for asymmetri, og er gjeven ved

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} \, , \qquad 0 \leq x < \infty \, , \tag{1.53}$$

α	γ	s	s^*	μ	m
2	0.38	1.41	2.12	2	1
3	0.28	1.15	2.00	3	2
5	0.20	0.89	1.77	5	4
10	0.14	0.63	1.40	10	9

Tabell 1.1: *Skeivleikskoeffisientar for gammafordelinga. Her er α parameter i gammafordelinga, γ, s og s^* er skeivleikskoeffisientar, μ er forventningsverdi, medan m er modalverdi.*

der α og β er konstantar større enn null. Konstanten β er ein skaleringskonstant, og er difor ikkje med å avgjere skeivleiken. Dette kan visast ved hjelp av transformasjonen $y = \frac{x}{\beta}$ og teorem 2.1.4. Sentralmomenta til gammafordelinga kan reknast ut analytisk, då

$$\mu_1 = \alpha\beta \ , \ \mu_2 = \alpha\beta^2 \text{ og } \mu_3 = 2\alpha\beta^3 \ .$$

Skeivleikskoeffisienten blir såleis

$$s = \frac{\alpha\beta^3}{(\alpha\beta^2)^{\frac{3}{2}}} = \frac{2}{\sqrt{\alpha}} \ ,$$

modalverdien blir

$$m = \beta(\alpha - 1) \ ,$$

og den modale skeivleikskoeffisienten blir

$$s^* = \frac{1 + 5\alpha}{(1 + \alpha)^{\frac{3}{2}}} \ .$$

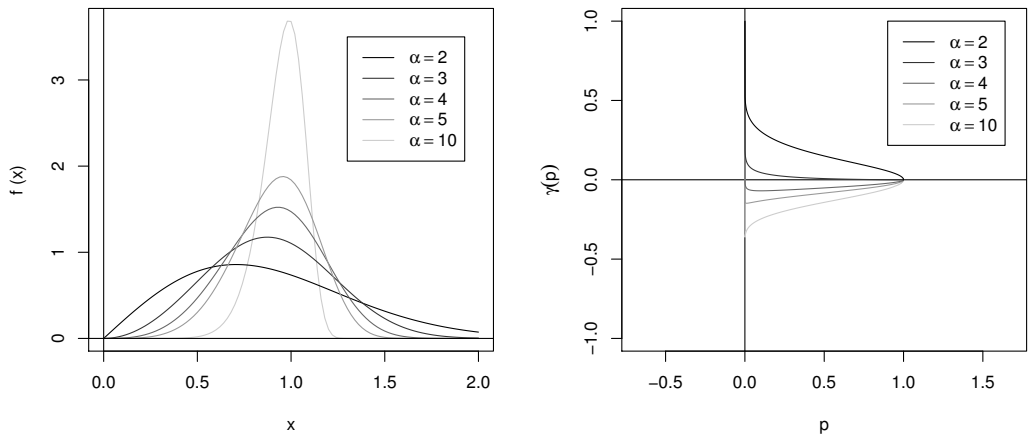
Dersom vi ser på monotonieigenskapane til skeivleikskoeffisientane, kan vi sjå at

$$\frac{\mathrm{d}s}{\mathrm{d}\alpha} = -\alpha^{-\frac{3}{2}} \ , \tag{1.54}$$

og at

$$\frac{\mathrm{d}s^*}{\mathrm{d}\alpha} = \frac{7 - 5\alpha}{2(1 + \alpha)^{\frac{5}{2}}} \ . \tag{1.55}$$

Ettersom α er ein konstant større enn null, vil s vere strengt synkande då den deriverte med hensyn på α alltid er mindre enn null. Den modale skeivleikskoeffisienten er derimot stigande for $\alpha \in (0, 7/5)$, og synkande på $\alpha \in (7/5, \infty)$. Det er diverre ikkje mogeleg å rekne ut asymmetrifunksjonen analytisk, ettersom det er umogeleg å løyse ut eksplisitt for x i likninga $y = x^{\alpha-1}e^{-\frac{x}{\beta}}$. Vi har likevel ei formeining om at γ er monoton ved å sjå på dei ulike verdiane



Figur 1.8: *Weibullfordeling til venstre med tilhøyrande asymmetrifunksjonar til høgre. Skaleringskonstanten, β, påverkar ikkje asymmetrien og er ein i alle fordelingane som er illustrert her.*

for α i figur 1.7, og frå tabell 1.1. Der kan vi også sjå at for dei same parametera, rangerer dei ulike måla asymmetrien likt, og at fordelinga er meir symmetrisk for store α. Dette ser vi gjennom at alle koeffisientane er lågare for store α. Modalverdi og forventningsverdi ligg også tett opptil kvarandre.

1.3.2 Weibullfordeling

Weibullfordelinga liknar ein del på gammafordelinga i matematisk form, men har likevel statistiske eigenskapar som kan skilje seg frå gammafordelinga ved ulike val av α,

$$f(x) = \frac{\alpha}{\beta} x^{\alpha-1} e^{-\frac{x^\alpha}{\beta}} , \qquad 0 \leq x < \infty , \tag{1.56}$$

der α og β er konstantar større enn null. På same måte som i gammafordelinga er β ein skaleringskonstant. Momenta til Weibullfordelinga er mogeleg å rekne ut, men gir til dels stygge uttrykk som er vanskeleg å tolke. Modalverdien til Weibullfordelinga er

$$m = \left(\frac{(\alpha - 1)\beta}{\alpha} \right)^{\frac{1}{\alpha}} , \tag{1.57}$$

medan forventningsverdien er

$$\mu = \beta^{\frac{1}{\alpha}} \Gamma(1 + \frac{1}{\alpha}) , \tag{1.58}$$

der Γ er gammafunksjonen. Av dette ser vi at modalverdien går mot 1 når α blir stor, uansett kva verdiar vi vel for β. Likevel styrer β ifrå kva retning

α	γ	s	s^*	μ	m
2	0.17	0.63	1.50	0.89	0.71
3	0.02	0.17	0.35	0.89	0.87
4	-0.05	-0.09	-0.36	0.91	0.93
5	-0.08	-0.25	-0.77	0.92	0.96
10	-0.15	-0.64	-1.43	0.95	0.99

Tabell 1.2: *Asymmetrikoeffisientar for Weibullfordelinga. Her er α parameteret til fordelinga, γ den ny skeivleikskoeffisienten, s^* den modale skeivleikskoeffisienten og s er den tradisjonelle skeivleikskoeffisienten. Forventningsverdien er μ og modalverdien m.*

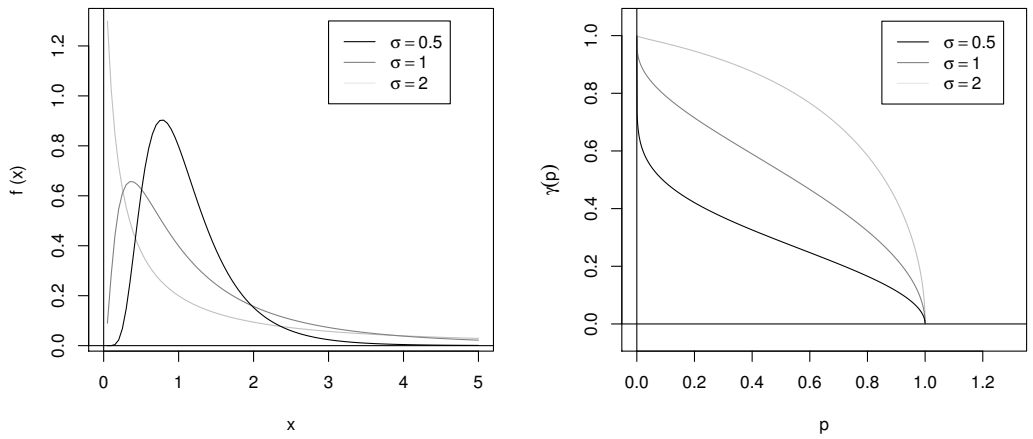
modalverdien konvergerer. Dette kan vere interessant sidan skeivleiken ikkje er avhengig av β verken for s,s^* eller γ. Av figur 1.8 og tabell 1.2 ser vi korleis forventningsverdi og modalverdi utviklar seg med α, og at skeivleiken går frå å vere positiv til negativ for kritiske verdiar av α for alle måla. Desse kritiske verdiane er alle ulike, men ligg i intervallet $\alpha \in (3, 4)$. Det kan vere verdt å merke seg at $\gamma^*(p)$ går mot 1 når p går mot 0 for alle val av α. Grunnen til dette er at fordelinga er definert på $(0, \infty)$. Likevel kan vi få negative verdiar for γ, ettersom tettleiken konvergerer raskt mot 0, særskild for store α.

1.3.3 Lognormalfordeling

Lognormalfordelinga liknar også på gammafordeling både i utsjånad, men også med tanke på definisjonsområde. Lognormalfordelinga er ei fordeling som har tunge halar, ettersom logaritmeleddet i eksponensialfunksjonen divergerer relativt sakte. Fordelinga er illustrert i figur 1.9 og er gjeven ved

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} x^{-1} e^{-\frac{(\log x)^2}{2\sigma^2}} , \qquad 0 \leq x < \infty , \tag{1.59}$$

der σ er ein konstant større enn 0. I tabell 1.3 ser vi at også for lognormalfordelinga rangerer dei ulike måla skeivleiken likt. For store σ er det stor skeivleik både for moment og γ. Vi kan likevel sjå at skeivleiken blir svært stor for moment. Det kan tyde at moment er ein dårleg metode for å klassifisere skeivleik dersom det blir svært tunge halar. Ein kan argumentere for å bruke ei anna hjelpefordeling enn uniform fordeling, dersom ein ikkje ønskjer å legge vekt på halane når ein reknar ut γ.



Figur 1.9: *Lognormalfordeling til venstre med tilhøyrande asymmetrifunksjonar til høgre.*

σ	γ	s	s^*	μ	m
0.5	0.30	1.75	1.12	1.13	0.79
1	0.52	6.18	3.94	1.65	0.37
2	0.76	414	403	7.39	0.018

Tabell 1.3: *Asymmetrikoeffisientar til lognormalfordelinga. Her er σ parameteret til fordelinga, γ den ny skeivleikskoeffisienten, s^* den modale skeivleikskoeffisienten og s er den tradisjonelle skeivleikskoeffisienten. Forventningsverdien er μ og modalverdien m.*

1.3.4 Samansett normalfordeling

I avsnitt 1.2.2 såg vi på ei fordeling som var sett saman av to normalfordelingar med ulik varians.

$$f(x) = \begin{cases} ke^{-\frac{(x-m)^2}{2\sigma_1}}, & -\infty < x < 0 \\ ke^{-\frac{(x-m)^2}{2\sigma_2}}, & 0 < x < \infty, \end{cases} \tag{1.60}$$

der $k = \frac{\sqrt{2}}{\sqrt{\pi}(\sigma_1+\sigma_2)}$. For den samansette normalfordelinga kan vi rekne ut skeivleikskoeffisientane eksakt både for moment og den integrerte asymmetrifunksjon. Dette gjer vi ikkje her, ettersom det blir eit svært stygt uttrykk for det tredje sentralmomentet. Grunnen til det er at forventningsverdien ikkje samsvarer med modalverdien for $\sigma_1 \neq \sigma_2$. Likevel er det mogeleg å finne verdiar for σ_1 og σ_2 der dei ulike måla konkluderer ulikt. Dersom vi held σ_1 konstant, vil både γ og s^* stige for alle val av σ_2, medan s vil synke i området $(0, 0.3)$. Dette ser vi i tabell 1.4 ettersom dei ulike måla rangerer

σ_2	γ	s	s^*	μ	m
0.1	-0.82	-0.93	-1.67	-0.72	0
0.3	-0.54	-0.73	-1.73	-0.54	0
0.5	-0.33	-0.50	-1.54	-0.40	0
2	0.33	0.50	1.54	0.80	0

Tabell 1.4: *Asymmetrikoeffisientar frå den samansette normalfordelinga. Her er* $\sigma_1 = 1$ *i alle radene, medan* σ_2 *er parameteret som varierer. Asymmetrikoeffisientane er* γ*,* s *og* s^**, medan* μ *er forventningsverdi og* m *er modalverdi.*

ulikt for $\sigma_2 = 0.3$. Det kan også vere verdt å merke seg at måla er uavhengig av skalering, ettersom dei viser det same for $\sigma_2 = \frac{1}{2}$ og $\sigma_2 = 2$, men med motsett forteikn.

1.3.5 Tettleik av andregradspolynomen

På same måte som i avsnitt 1.2.1 konstuerer vi ein tettleik av andregradspolynomen.

$$f(x) = \begin{cases} c(1 - \frac{x^2}{\alpha^2}) \,, & -\alpha < x < 0 \\ c(1 - \frac{x^2}{\beta^2}) \,, & 0 < x < \beta \,, \end{cases} \tag{1.61}$$

der α og β er konstantar større enn null og $c = \frac{3}{2(\alpha+\beta)}$. Vidare finn vi at $\rho(p) = \beta/\alpha$ og $\gamma^*(p) = (\beta - \alpha)/(\beta + \alpha)$. Dermed blir også

$$\gamma = \int_0^1 \gamma^*(p)\,\mathrm{d}p = \frac{\beta - \alpha}{\beta + \alpha} \,. \tag{1.62}$$

Uttrykk for s og s^* er relativt kompliserte, men er likevel mogeleg å rekne ut analytisk,

$$s = \frac{1}{1280}(-7\alpha^3 - 11\beta\alpha^2 + 11\alpha\beta^2 + 7\beta^3) \,, \tag{1.63}$$

$$s^* = \frac{1}{8}(-\alpha^3 + \beta\alpha^2 - \alpha\beta^2 + \beta^3) \,. \tag{1.64}$$

På same måte som for den samansette normalfordelinga, kan vi undersøke monotonieigenskapane til dei ulike måla for å sjå om dei rangerer skeivleiken ulikt. Dersom vi held α konstant, kan vi finne at s er synkande i området $\beta \in (0, 0.37\alpha)$, medan både s^* og γ er stigande for alle val av β.

1.3.6 Polynomen av ulik grad

Den siste fordelinga vi ser på er fordelinga sett saman av polynomen av ulik grad vi såg på i avsnitt 1.2.3,

$$g(x) = \begin{cases} c(1 - \frac{x^2}{\alpha^2}) \, , & -\alpha < x < 0 \\ c(1 - \frac{x^4}{\beta^4}) \, , & 0 < x < \beta \, , \end{cases} \tag{1.65}$$

der α, $\beta > 0$ og $c = 15/(10\alpha + 12\beta)$. Asymmetrifunskjon er i dette dømet

$$\gamma^*(p) = \frac{\beta - \alpha(1-p)^{\frac{1}{4}}}{\beta + \alpha(1-p)^{\frac{1}{4}}} \, . \tag{1.66}$$

Dette er den einaste fordelinga vi klarer å rekne ut integralet til asymmetrifunksjonen analytisk,

$$\gamma = -\frac{3\alpha^4 - 8\beta\alpha^3 + 12\beta^2\alpha^2 - 24\beta^3\alpha + 24\beta^4 \log(\beta + \alpha) - 24\beta^4 \log \beta}{3\alpha^4} \, .$$

Sjølv om uttrykket for asymmetrien er relativt stort, er det mogeleg å drøfte kva påverknad val av parameter vil påverke denne koeffisienten. Dersom vi ser på asymmetrifunksjonen er det storleiken på α i høve til β som avgjer om han er negativ eller positiv. Omkring modalverdien vil $\gamma^*(p)$ gå mot 1 for alle val av α og β. Difor må vi ha ganske stor α for at asymmetrikoeffisienten, γ, skal bli mindre enn null. For skeivleikskoeffisienten s er det annleis, då fordelinga blir rangert som relativt symmetrisk uansett kva verdiar vi vel for α og β. Dersom vi ser i tabell 1.5, kan vi sjå at for $\alpha = 1$ og $\beta = 1$ er s negativ. Dette antyder skeivleik mot venstre. Asymmetrikoeffisienten til Critchley og Jones (2008), γ, er derimot positiv og viser skeivleik mot høgre. For $\alpha = 1$ og $\beta = 2$ viser γ klar asymmetri, medan s måler fordelinga som relativt symmetrisk. Den modale skeivleikskoeffisienten blir ein mellomting, sjølv om han samsvarar best med γ for begge val av parameter. Dette viser at dei nye måla for asymmetri har eigenskapar som skil seg klart frå det tradisjonelle målet i nokre tilfelle.

1.4 Samandrag

Critchley og Jones (2008) argumenterer for å bruke asymmetrifunksjonar istadenfor skalare mål. På den måten kan vi få eit inntrykk til forma på heile fordelinga berre ved å sjå på målet for asymmetri. Likevel gjer dei også eit forslag til korleis vi kan gjere funksjonen om til ein skalar verdi. Fordelen med denne nye asymmetrikoeffisienten, er at vi kan velje kvar vi vil leggje

Par.	γ	s	s^*	μ	m
$\alpha = 1, \beta = 1$	0.12	-0.09	0.70	0.057	0
$\alpha = 1, \beta = 2$	0.43	0.04	1.02	0.48	0

Tabell 1.5: *Skeivleikskoeffisientar for fordelinga som er sett saman av polynomen av ulik grad. På same måte som i dei føregåande tabellane er γ skeivleikskoeffisenten til Critchley og Jones (2008), medan s er den tradisjonelle skeivleiken, s^* er den modale, μ er forventningsverdi og m er modalverdi.*

vekt på asymmetrien. Dette kan vere nyttig i fleire samanhengar, til dømes i finansielle data, der sannsynet for store tap ofte er større enn sannsynet for like store gevinstar. I det tilfellet er vi mest interessert i symmetrien i halane. Skeivleikskoeffisienten, s, kan vise at dataene er relativt symmetrisk fordelt, og då vil det nye målet hjelpe oss med å kartlegge kvar vi kan finne eventuell asymmetri.

Dersom vi bruker den uniforme fordelinga som hjelpefordeling, kan vi sjå at dei ulike måla konkluderer relativt likt både innafor ein klasse fordelingar, og mellom ulike klassar. Vi kan derimot få unntak dersom halane er tunge eller ikkje eksisterer. Det same gjeld for fordelingar som har relativt flat topp. Isåfall vil dei ulike måla rangere skeivleiken ulikt. Vi kan justere for dette med å velge hjelpefordeling, $w_f(p)$, som tek omsyn til det. På grunnlag av dette kan vi seie at γ er eit nyttig mål på asymmetri. Den nye asymmetrikoeffisienten γ har mange av dei same eigenskapane som den meir tradisjonelle s, men dei skil seg frå kvarandre på nokre punkt. Med skeivleikskoeffisienten γ, kan vi dessutan finne asymmetrien til fordelingar som ikkje har moment. Den modale skeivleikskoeffisenten er i så fall mindre nyttig, då den har dei same eigenskapane som den tradisjonelle skeivleiken i halane. Vi kan likevel få litt ulik rangering dersom modalverdien skil seg vesentleg frå forventningsverdien, noko som er tilfelle for svært asymmetriske fordelingar.

Kapittel 2

Kurtose

I dette kapittelet skal vi undersøke mål for kurtose. På same måte som for asymmetri, bruker Critchley og Jones (2008) funksjonar på dette målet. Vi byrjar med sjølve definisjonen Critchley og Jones (2008) innfører for kurtose. Deretter finn vi analytiske døme på kurtosefunksjonar. Til slutt føreslår vi ein eigen definisjon for kurtose, og samanliknar dei ulike måla med kvarandre.

Definisjon 2.0.1. *Kurtosekoeffisienten er definert som*

$$k = \frac{\mu_4}{\mu_2^2}\,, \tag{2.1}$$

der μ_n er det n-te sentralmomentet (Casella og Berger 2002).

Kurtosen er med andre ord fjerdemomentet dividert med kvadratet av variansen. Grunnen til at vi dividerer med variansen, er at målet ikkje skal bli påverka av skaleringa.

2.1 Gradientasymmetri

Critchley og Jones (2008) innfører eit heilt nytt konsept når det gjeld kurtose. Dei deler opp kurtosen i ein høgre- og ein venstredel, og bruker modalverdien som skiljepunkt. Deretter ser dei på den deriverte til henholdsvis f_R og f_L. På same måte som vart gjort for asymmetri i kapittel 1, framstiller dei kurtosen gjennom funksjonar.

Definisjon 2.1.1. *Vi føreset at f er ein eintoppa rota sannsynstettleik definert på (a, b) med modalverdi m. Vidare føreset vi at den deriverte til f_L er rota og eintoppa med modalverdi π_L. Då er $f_{LT}(x)$ den venstre halen til*

f, definert på (a, π_L) og $f_{LP}(x)$ den venstre delen av toppen til f definert på (π_L, m). Vidare er

$$X_{LT}(p) = {f'_{LT}}^{-1}(pf'(\pi_L)) ,$$
$$X_{LP}(p) = {f'_{LP}}^{-1}(pf'(\pi_L)) .$$

På same måte føreset vi at den deriverte til $-f_R$ er rota og eintoppa med modalverdi π_R. Då er $f_{RP}(x)$ den høgre toppen til f, definert på (m, π_R) og $f_{RT}(x)$ den venstre halen til f definert på (π_L, m).

$$X_{RP}(p) = {f'_{RP}}^{-1}(pf'(\pi_L)) ,$$
$$X_{RT}(p) = {f'_{RT}}^{-1}(pf'(\pi_L)) .$$

Vidare vil vi definere avstanden $X(p)$ har til vendepunktet til f.

$$\tau_{LT}(p) = \pi_P - X_{LT}(p) ,$$
$$\tau_{LP}(p) = X_{LP}(p) - \pi_L ,$$

og på same måte definerer vi avstandar for den høgre del av fordelinga

$$\tau_{RP}(p) = \pi_R - X_{RP}(p) ,$$
$$\tau_{RT}(p) = X_{RT}(p) - \pi_R ,$$

der R, L T og P står for høgre, venstre, hale og topp.

Frå definisjon 2.1.1 ser vi klare likskapar med framgangsmåten som vart brukt for å definere asymmetri i kapittel 1. Etterkvart skal vi definere kurtose for den høgre delen av fordelinga, f_R, og den venstre delen, f_L, basert på den deriverte eller gradienten til henholdsvis f_R og f_L.

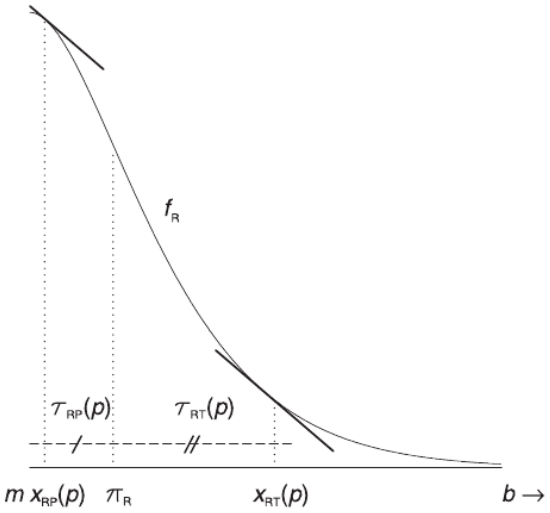
Definisjon 2.1.2. *Høgre kurtosefunksjon til ei fordeling med eintoppa, rota tettleiksfunksjon f, og eintoppa, rota høgrederivert, f'_R, er definert som*

$$\kappa_R(p) = \frac{\tau_{RT}(p)}{\tau_{RP}(p)} , \qquad 0 < p < 1 . \tag{2.2}$$

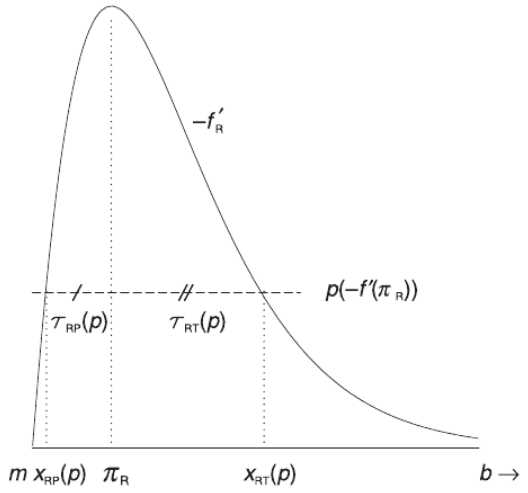
Venstrekurtosefunksjonen med dei same vilkåra, er definert som

$$\kappa_L(p) = \frac{\tau_{LT}(p)}{\tau_{LP}(p)} , \qquad 0 < p < 1 . \tag{2.3}$$

På same måte som for asymmetri, vil vi ha ein normalisert funksjon som er definert på eit endeleg intervall. Kurtosefunksjonen κ_L og κ_R vil ta verdiar på intervallet $(0, \infty)$, medan δ_L og δ_R vil ta verdiar på $(-1, 1)$



Figur 2.1: *Høgre del av ei eintoppa, rota tettleik. Her er π_R vendepunktet til tettleiken og m er modalverdien. Funksjonane $\tau_{RP}(p)$ og $\tau_{RT}(p)$ er avstandane frå vendepunktet den deriverte (Critchley og Jones (2008)).*



Figur 2.2: *Høgre del av den deriverte til ei eintoppa, rota tettleik. Den deriverte er også eintoppa og rota, men med negativt forteikn, på intervallet (m, b). Vi ser at π_R er modalverdien til den deriverte, altså vendepunktet. På same måte som for asymmetrien, er τ_{RP} og τ_{RT} avstandsfunksjonar frå eit referansepunkt, men denne gongen er det vendepunktet og inversen til den deriverte vi brukar som utgangspunkt (Critchley og Jones (2008)).*

Definisjon 2.1.3. *Normalisert høgre gradientasymmetri er definert som*

$$\delta_R^*(p) = \frac{\kappa_R(p) - 1}{\kappa_R(p) + 1} = \frac{\tau_{RT}(p) - \tau_{RP}(p)}{\tau_{RT}(p) + \tau_{RP}(p)} = \frac{X_{RT}(p) + X_{RP}(p) - 2\pi_R}{X_{RT}(p) - X_{RP}(p)} \ . \quad (2.4)$$

Normalisert venstre gradientasymmetri er definert som

$$\delta_L^*(p) = \frac{\kappa_L(p) - 1}{\kappa_L(p) + 1} = \frac{\tau_{LT}(p) - \tau_{LP}(p)}{\tau_{LT}(p) + \tau_{LP}(p)} = \frac{X_{LT}(p) + X_{LP}(p) - 2\pi_L}{X_{LP}(p) - X_{LT}(p)} \ . \quad (2.5)$$

Kurtosefunksjonane våre er altså forholdet mellom stigningstalet til tettleiken i halen og ved toppen. Dette kan vi sjå illustrert i figur 2.1 og 2.2.

Teorem 2.1.4. *La $f(x)$ vere ein sannsynstettleik slik at $f(x) = kh(x)$, der k er ein konstant. Asymmetrifunksjonen kan då reknast ut berre utfrå $h(x)$, og på same måte som for $f(x)$.*

Bevis. Vi byrjar med at

$$f(x) = kh(x) \ .$$

Vidare har vi at

$$f_R^{-1}(x) = h_R^{-1}\left(\frac{x}{k}\right) \ ,$$

$$X_R(p) = f_R^{-1}(pf(m)) = h_R^{-1}\left(\frac{p}{k}f(m)\right) = h_R^{-1}\left(\frac{p}{k}kh(m)\right) = h_R^{-1}\left(ph(m)\right) \ .$$

Det same gjeld for venstre sida av funksjonen

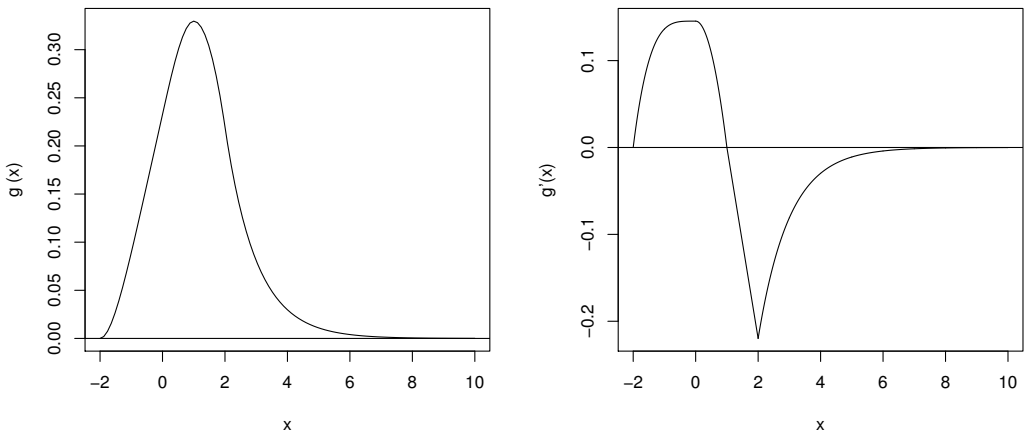
$$f_L^{-1}(x) = h_L^{-1}\left(\frac{x}{k}\right) \ ,$$

$$X_L(p) = f_L^{-1}\left(pf(m)\right) = h_L^{-1}\left(\frac{p}{k}f(m)\right) = h_L^{-1}\left(\frac{p}{k}kh(m)\right) = h_L^{-1}\left(ph(m)\right) \ .$$

Vi ser at $X_R(p)$ og $X_L(p)$ berre er avhengig av h, og det same gjeld for $\tau_R(p)$, $\tau_L(p)$ og $\rho(p)$. □

Korrolar 2.1.5. *La $f(x)$ vere ein sannsynstettleik slik at $f(x) = kh(x)$, der k er ein konstant. Gradientasymmetrien kan då reknast ut berre ved hjelp av $h(x)$.*

Bevis. Den deriverte til til tettleiken er $f'(x) = kh'(x)$. Vi kan difor rekne ut gradientasymmetrien på same måte som asymmetrien i teorem 2.1.4. □



Figur 2.3: *Sannsynstettleik sett saman av polynomen og eksponensialfordeling. Tettleiken er til venstre og den deriverte av tettleiken til høgre. Vi ser at tettleiken er rota og eintoppa på* $(-2, \infty)$ *med modalverdi* $m = 1$*. Den venstre delen av den deriverte er rota og eintoppa på* $(-2, 1)$ *med toppunkt* $\pi_L = 0$*. Høgre del av den deriverte er rota og eintoppa med negativt forteikn på* $(1, \infty)$ *og botnpunktet er* $\pi_R = 2$*. Den deriverte er kontinuerleg for alle verdiane på intervallet, men har eit knekkpunkt i* $x = 2$*.*

2.1.1 Tettleik sett saman av polynom og eksponensialfunksjon

I avsnitt 1.2 såg vi døme på asymmetrifunksjonar. Dette vidarefører vi og ser her på døme på gradientasymmetri. Fordelinga vi skal sjå på, er ei fordeling sett saman av polynomen og ein eksponensialfunksjon, sjå figur 2.3. Tettleiken er gjeven ved

$$g(x) = \begin{cases} k\left(-\frac{1}{80}x^5 + x + \frac{8}{5}\right) , & -2 < x \leq 0 \\ k\left(-\frac{1}{3}x^3 + x + \frac{8}{5}\right) , & 0 < x \leq 1 \\ k\left(-\frac{34}{45}(x-1)^2 + \frac{34}{15}\right) , & 1 < x \leq 2 \\ k\left(\frac{68}{45}e^{-(x-2)}\right) , & 2 < x < \infty \, . \end{cases} \tag{2.6}$$

der $k = \frac{540}{3713}$. Vi kan sjå at g er kontinuerleg, ettersom

$$\begin{aligned} \lim_{x \to 0^-} g(x) &= \lim_{x \to 0^+} g(x) = k\frac{8}{5} \, , \\ \lim_{x \to 1^-} g(x) &= \lim_{x \to 1^+} g(x) = k\frac{34}{15} \, , \\ \lim_{x \to 2^-} g(x) &= \lim_{x \to 2^+} g(x) = k\frac{68}{45} \, . \end{aligned}$$

Vi treng den deriverte til g for å rekne ut gradientasymmetrien,

$$g'(x) = \begin{cases} -k\left(\frac{1}{16}x^4 + 1\right) , & -2 < x \leq 0 \\ -k(x^2+1) , & 0 < x \leq 1 \\ -k\left(\frac{68}{45}(x-1)\right) , & 1 < x \leq 2 \\ -k\left(\frac{68}{45}e^{-(x-2)}\right) , & 2 < x \leq \infty . \end{cases} \tag{2.7}$$

Dersom vi ser på kontinuitetseigenskapane til g', kan vi finne at

$$\lim_{x\to 0^-} g'(x) = \lim_{x\to 0^+} g'(x) = k , \tag{2.8}$$

$$\lim_{x\to 1^-} g'(x) = \lim_{x\to 1^+} g'(x) = 0 , \tag{2.9}$$

$$\lim_{x\to 2^-} g'(x) = \lim_{x\to 2^+} g'(x) = k\frac{68}{45} . \tag{2.10}$$

Den andrederiverte er derimot ikkje kontinuerleg i verken $x = 1$ eller $x = 2$, ettersom

$$\lim_{x\to 1^-} g''(x) = -2k , \tag{2.11}$$

$$\lim_{x\to 1^+} g''(x) = -k\frac{68}{45} , \tag{2.12}$$

$$\lim_{x\to 2^-} g''(x) = -k\frac{68}{45} , \tag{2.13}$$

$$\lim_{x\to 2^-} g''(x) = k\frac{68}{45} . \tag{2.14}$$

Ifølgje teorem 2.1.4 er asymmetrifunksjonen ikkje avhengig av normaliseringskonstanten til g. Dette brukar vi til å finne venstre kurtosefunksjon,

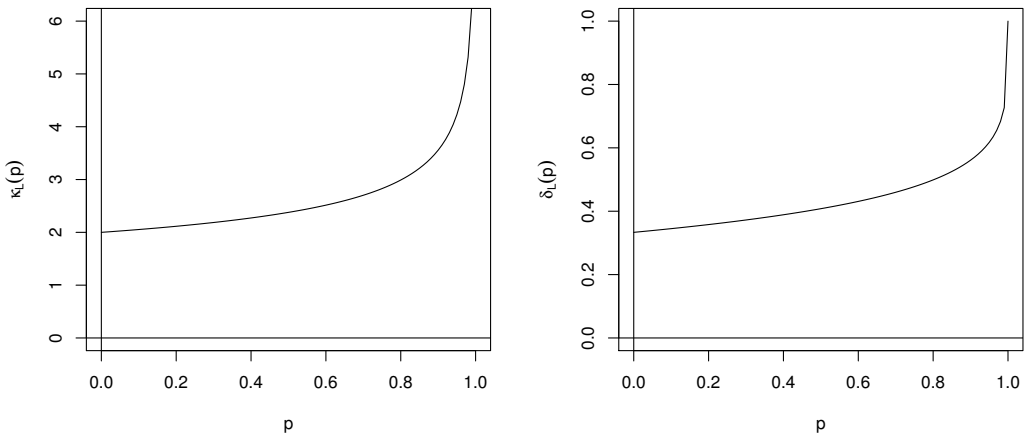
$$\tau_{LP}(p) = (1-p)^{\frac{1}{2}} , \tag{2.15}$$

$$\tau_{LT}(p) = 2(1-p)^{\frac{1}{4}} , \tag{2.16}$$

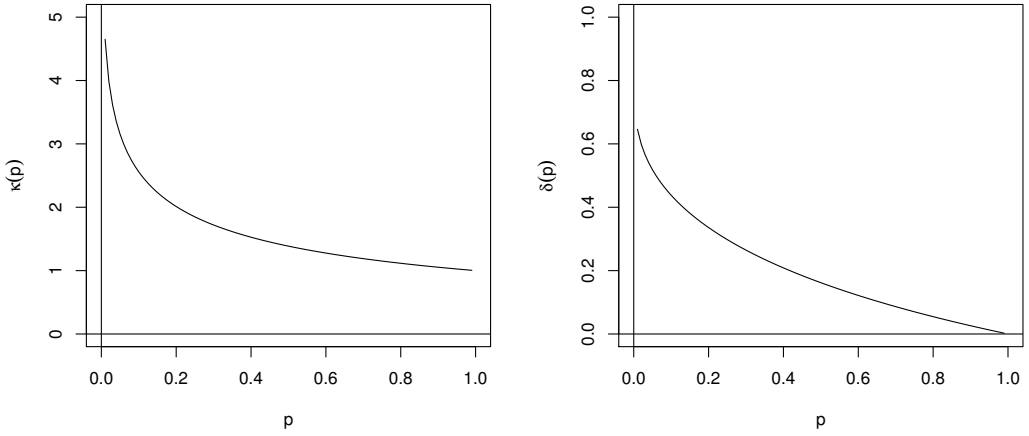
$$\kappa_L(p) = 2(1-p)^{-\frac{1}{4}} , \tag{2.17}$$

$$\delta^*_L(p) = \frac{2-(p-1)^{\frac{1}{4}}}{2+(p-1)^{\frac{1}{4}}} . \tag{2.18}$$

Vi ser her at venstre gradientasymmetri til likning (2.6) er lik asymmetrien vi fann for likning (1.21) i avsnitt 1.2.3. Dersom vi ønskjer å sjå på monotonieigenskapane til gradientasymmetrien, kan vi frå teorem 1.2.3 utleie at monotonieigenskapane til $\delta^*(p)$ berre er avhengig av $\kappa(p)$. Ettersom $\kappa'_L(p) = 2(1-p)^{-\frac{5}{4}} > 0$ for alle $0 < p < 1$, kan vi sjå at den venstre kurtosefunksjonen er stigande på heile intervallet. Dette tyder at skilnaden i monotonieigenskapane til venstre del av tettleiken er størst omkring vendepunktet,



Figur 2.4: *Venstrekurtosen til fordelinga sett saman av polynomen og eksponesialfordeling, der $\kappa_L(p)$ er til venstre og $\delta_L^*(p)$ er til høgre.*



Figur 2.5: *Høgrekurtosen til fordelinga sett saman av polynomen og eksponesialfordeling, der $\kappa_R(p)$ er til venstre og $\delta_R^*(p)$ er til høgre.*

medan den deriverte er forholdsvis lik i venstre rot og modalverdien. Venstre gradientasymmetri er illustrert i figur 2.4. Vidare finn vi høgre gradientasym-

metrifunksjon,

$$\tau_{RP}(p) = 1 - p \,, \tag{2.19}$$

$$\tau_{RT}(p) = -\log(p) \,, \tag{2.20}$$

$$\kappa_R(p) = \frac{-\log(p)}{1-p} \,, \tag{2.21}$$

$$\delta_R^*(p) = \frac{p - \log(p) - 1}{1 - \log(p) - p} \,. \tag{2.22}$$

Monotonieigenskapane til høgrekurtosen finn vi ved å derivere

$$\kappa_R'(p) = (1 - \frac{1}{p} - \log p)/(1-p)^2 \,, \tag{2.23}$$

som er negativ for alle p. Det tyder at den høgre kurtosefunksjonen er synkande på heile intervallet. Med andre ord er det stor skilnad mellom den deriverte i halen og den deriverte til høgre for modalverdien. Dette samsvarar bra med kva vi finn i figur 2.5, ettersom funksjonsverdien til den deriverte er svært låg for store x-verdiar.

2.2 Skalare mål for kurtose

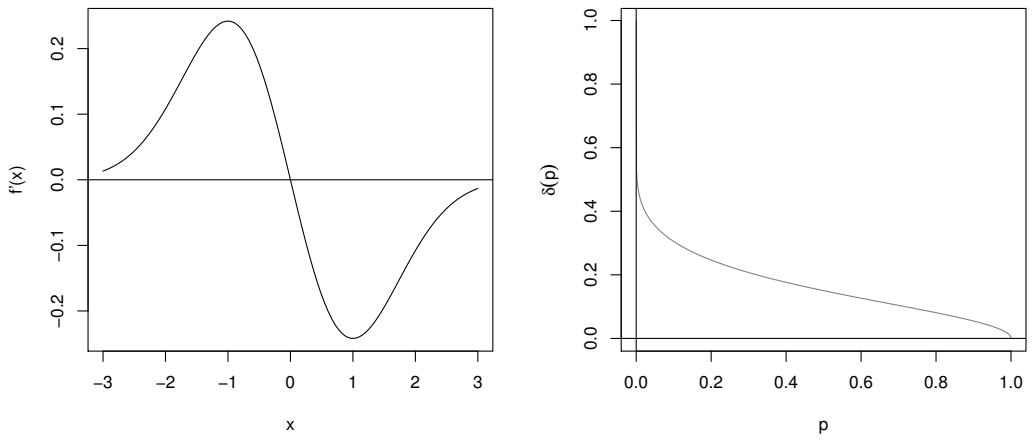
Critchley og Jones (2008) argumenterer for funksjonar for asymmetri og kurtose. Denne måten å definere asymmetri og kurtose gjer at heile forma til tettleiken blir skildra. Likevel må vi gjere funksjonane om til skalare verdiar dersom vi skal kunne gi ei fornuftig samanlikning med dei tradisjonelle måla. Ei fordeling kan med den nye definisjonen av kurtose, ha definert kurtosen for berre ei side av modalverdien. Vi har difor i stor grad vald å samanlikne fordelingar som har både venstre- og høgrekurtose. På same måte som for asymmetri, gir Critchley og Jones (2008) eit forslag til korleis vi kan rekne ut skalare verdiar for kurtosen.

Definisjon 2.2.1. *Høgre kurtosekoeffisient er definert som*

$$\delta_R = \int_0^1 \delta_R^*(p) w_f(p) \, \mathrm{d}p \,, \tag{2.24}$$

der $w_f(p)$ er ei fordeling på $(0,1)$. På same måte er venstre kurtosekoeffisient definert som

$$\delta_L = \int_0^1 \delta_L^*(p) w_f(p) \, \mathrm{d}p \,. \tag{2.25}$$



Figur 2.6: *Den deriverte til tettleiken til normalfordelinga til venstre og høgre kurtosefunksjon til høgre.*

I motsetnad til kurtosekoeffisienten som er basert på fjerdemomentet, kan kurtosekoeffisienten til Critchley og Jones (2008) bli negativ. Det tyder at avstanden frå toppen til vendepunktet, er større enn avstanden frå halen til vendepunktet. Dette tyder igjen at det er meir masse omkring modalverdien enn i halane, og kan tilsvare låg kurtose med det tradisjonelle målet. Sjølv om Critchley og Jones (2008) definerer ein måte å finne skalare verdiar for kurtosen, er dei skalare måla likevel ikkje eintydige, ettersom det både fins ein høgre- og ein venstrekurtose. Difor definerer vi to nye mål. Eit som reknar begge kurtosefunksjonane om til ein skalar verdi, og eit som tek utgangspunkt i modalmoment.

Definisjon 2.2.2. *Totalkurtose definerer vi til summen av venstre og høgre kurtosekoeffisient,*

$$\delta = \delta_R + \delta_L \ . \tag{2.26}$$

Definisjon 2.2.3. *Modalkurtose definerer vi som*

$$k^* = \frac{\mu_4^*}{\mu_2^{*2}} \ . \tag{2.27}$$

På same måte som for asymmetrien, har vi brukt den uniforme fordelinga som vektfunksjon når vi skal rekne ut skalare verdiar for kurtose.

2.2.1 Normalfordeling

Vi byrjar med å sjå på normalfordelinga. Den deriverte til tettleiken og kurtosefunksjonen er illustrert i figur 2.6, og er gjeven ved

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2^2}} , \qquad -\infty < x < \infty , \tag{2.28}$$

$$f'(x) = -\frac{1}{\sqrt{2\pi}} x e^{-\frac{x^2}{2^2}} , \qquad -\infty < x < \infty . \tag{2.29}$$

Normalfordelinga blir ofte brukt som referanse for kurtose. Den tradisjonelle kurtosekoeffisienten til normalfordelinga er $k = 3$. Som mål på kurtose, blir ofte $\gamma_2 = k - 3$ brukt. Dette målet blir kalla eksesskurtose. På den måten har vi eit kurtosemål som er basert på moment, men som også kan ta negative verdiar. Funksjonar med eksesskurtose mindre enn null vert ofte kalla leptokurtiske funksjonar. Dersom eksesskurtosen er lik null blir dei kalla mesokurtiske og med eksesskurtose større enn null platikurtiske (Balanda og MacGillivray 1988). Kurtosekoeffisienten til Critchley og Jones (2008) er $\delta_L = \delta_R = 0.34$, og variansen til normalfordelinga påverkar verken kurtosekoeffisienten k, eller kurtosefunksjoane $\kappa(p)$ og $\delta(p)$. Dette kan visast ved hjelp av teorem 2.1.4. Som vi ser er av likning 2.29, er den deriverte til normalfordelinga på lik form som Weibullfordeling med parameter $\alpha = 2$ og $\beta = 2$. Frå teorem 2.1.4 veit vi då at kurtosefunksjonen til normalfordelinga blir lik asymmetrifunksjonen til Weibullfordelinga. Ettersom normalfordeling er symmetrisk blir venstrekurtosen lik høgrekurtosen. I tillegg blir også modalverdi og forventningsverdi den same, og modalkurtosen blir lik den tradisjonelle kurtosen.

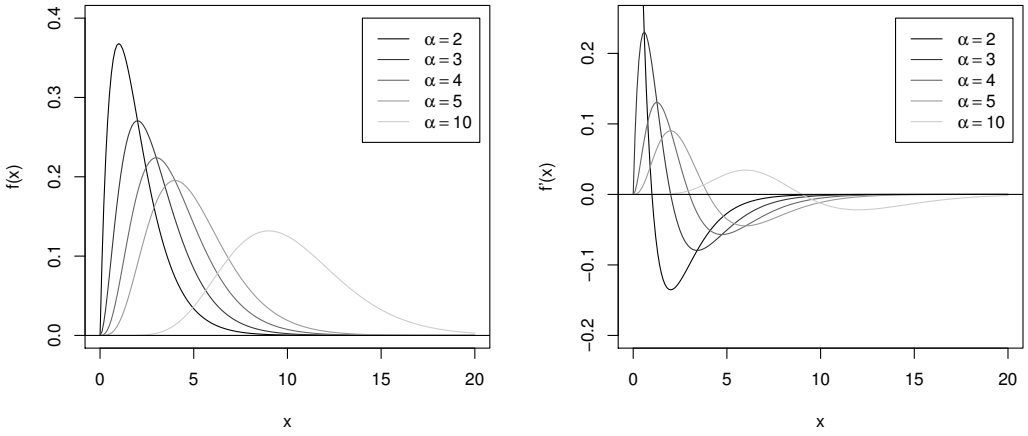
2.2.2 Gammafordeling

Gammafordelinga er ei fordeling som har mange interessante eigenskapar med tanke på asymmetri, og av den grunn ser vi også på gradientasymmetrien. Både tettleiken og den deriverte er illustrert i figur 2.7 medan gradientasymmetrien er illustrert i 2.8. Gammafordelinga er gjeven ved

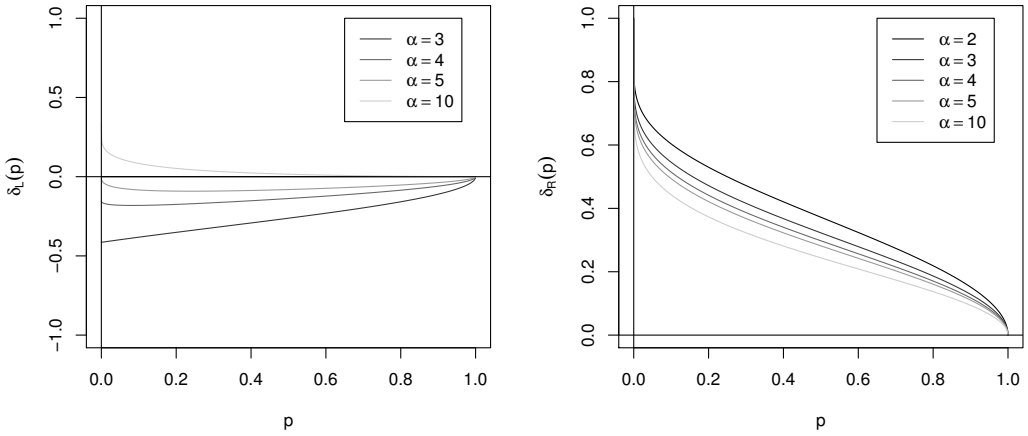
$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} , \qquad x > 0 , \tag{2.30}$$

$$f'(x) = \frac{\alpha - 1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-2} e^{-\frac{x}{\beta}} - \frac{1}{\Gamma(\alpha)\beta^{\alpha+1}} x^{\alpha-1} e^{-\frac{x}{\beta}}, \qquad x > 0 \tag{2.31}$$

der $\alpha, \beta > 0$. Vi ser at den deriverte til gammafordelinga er på forma til summen av to gammafordelingar med ulike parameter. Konstanten β er ein skaleringskonstant, og påverkar ikkje kurtosen (teorem 2.1.4). Den tradisjonelle kurtosekoeffisienten er for gammafordelinga $k = 3 + \frac{6}{\alpha}$. Det tyder at



Figur 2.7: *Tettleiken til gammafordelinga til venstre, og den deriverte til høgre.*



Figur 2.8: *Venstrekurtosen til gammafordelinga til venstre, og høgrekurtosen til høgre. Merk at venstrekurtosen ikkje eksisterer for $\alpha = 2$, då den deriverte til f ikkje er rota.*

eksesskurtosen er $\frac{6}{\alpha}$. Grunnen til kurtosen er større enn normalfordelinga, er at eksponenten i ekponentialleddet ikkje er kvadrert. Vi ser difor at den tradisjonelle kurtosen minkar saman med parameteret α. Dette er ikkje intuitivt, då vi med høgare α får tyngre halar og mindre spiss topp. Likevel er dette rimeleg, då parameteret α påverkar andremomentet like mykje som fjerdemomentet. Frå definisjon 2.0.1 hugsar vi at kurtose skal vere eit mål som er uavhengig skalereing. Modalkurtosen klarar vi diverre ikkje å rekne ut analytisk, og heller ikkje numerisk for $\alpha = 100$. Frå tabell 2.1 ser det likevel ut som om han minkar saman med α på same måte som den tradisjonelle kurtosen. Høgrekurtosen til Critchley og Jones (2008) minkar også

α	δ	δ_R	δ_L	k	k^*
2	N/A	0.38	N/A	6	5.9
3	0.08	0.33	-0.25	5	5.5
4	0.18	0.31	-0.13	4.5	5.2
5	0.22	0.29	-0.07	4.2	4.9
10	0.23	0.26	-0.03	3.6	4.1
100	0.33	0.20	0.13	3.06	N/A

Tabell 2.1: *Kurtosen til gammafordeling med ulike parameter α, totalkurtose δ, venstrekurtose δ_L, høgrekurtose δ_R, tradisjonell kurtose k og modalkurtose k^*. Venstrekurtosen og totalkurtosen eksisterer ikkje for $\alpha = 2$.*

ν	δ	δ_R	k
1	0.72	0.36	N/A
2	0.62	0.31	N/A
5	0.53	0.27	9
10	0.48	0.24	3.5
100	0.44	0.22	3.06

Tabell 2.2: *Kurtosen til t-fordeling. Her er ν talet på fridomsgrader, δ totalkurtosen, δ_R høgrekurtosekoeffisient og k den tradisjonelle kurtosekoeffisienten.*

med α, medan venstrekurtosen aukar. Totalkurtosen aukar også med α, ettersom venstrkurtosen aukar meir enn høgrekurtosen minkar. For $\alpha = 2$ er ikkje venstrekurtosen og totalkurtosen definert, ettersom den deriverte ikkje er rota.

2.2.3 t-fordeling

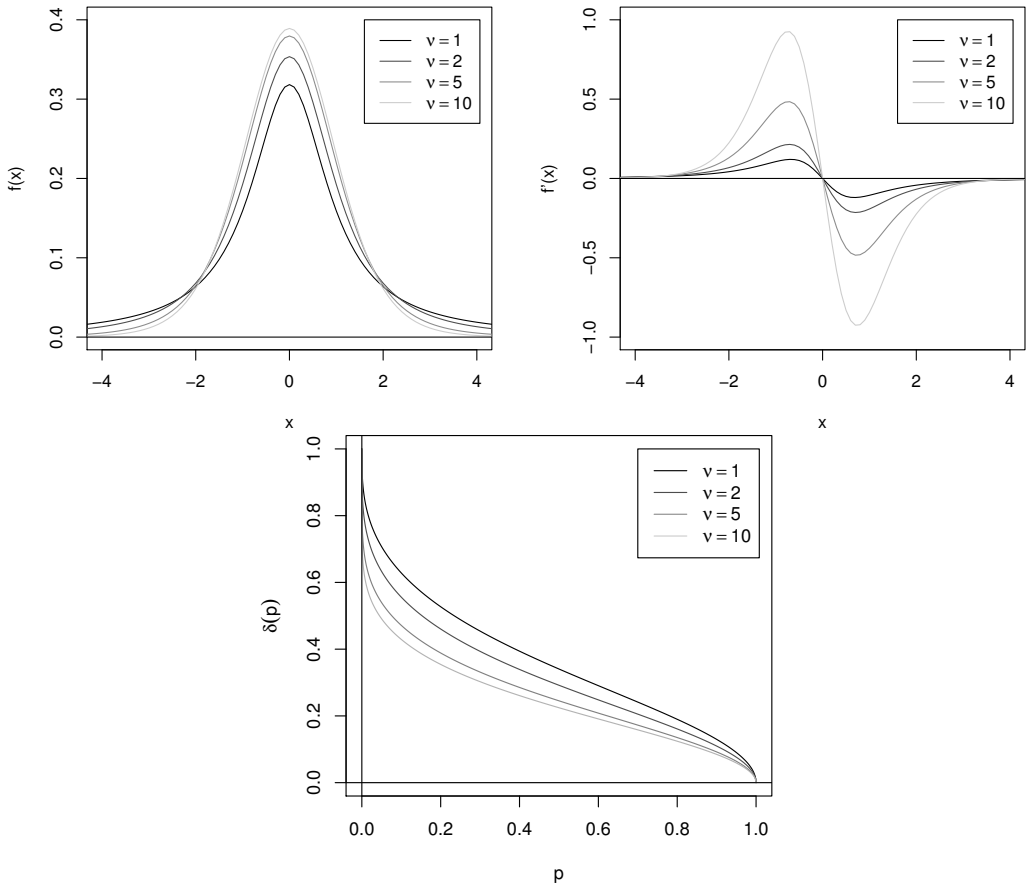
T-fordelinga er ei fordeling som blir brukt til observatorar for normalfordelte variable med ukjent varians. Fordelinga er kjent som ei fordeling med tunge halar, og observatorar basert på t-fordelinga er difor rekna som robuste for feilaktig forkasting av nullhyoptesar. Tettleiken til T-fordelinga er gjeven ved

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\sqrt{\nu\pi}} \frac{1}{(1+\frac{x^2}{\nu})^{\frac{\nu+1}{2}}}\,, \qquad -\infty < x < \infty\,, \tag{2.32}$$

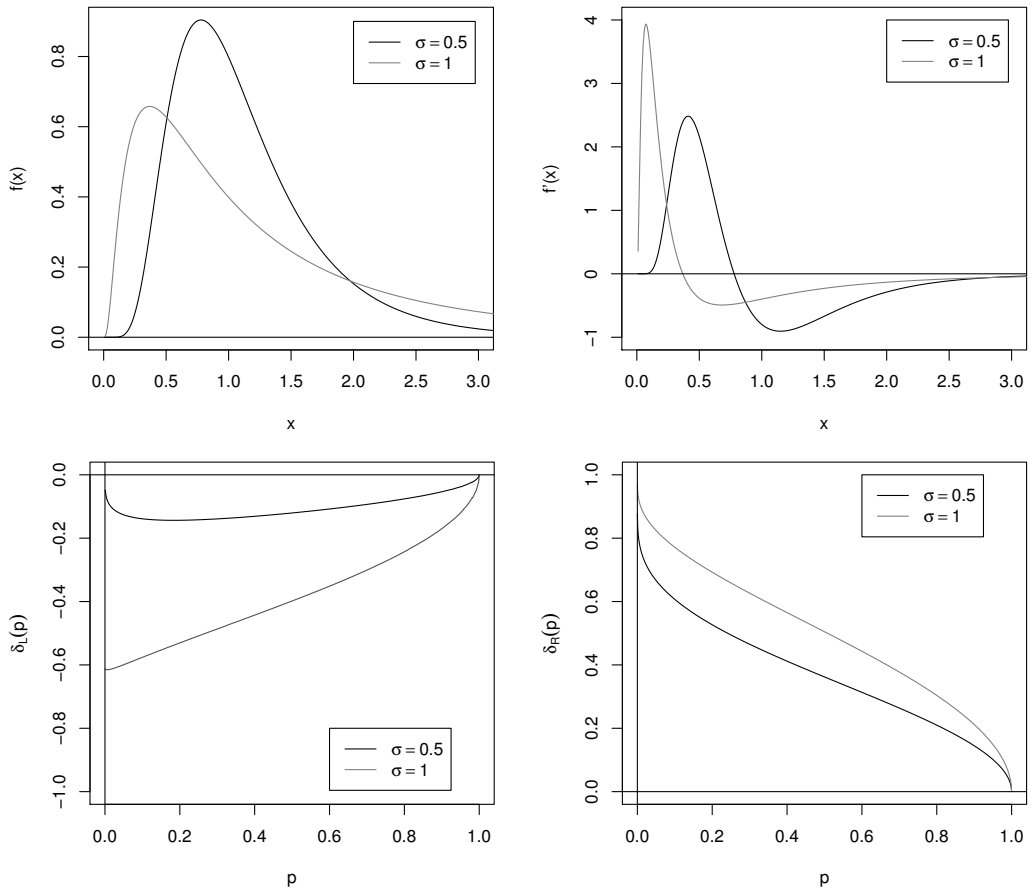
$$f'(x) = -\frac{\Gamma(\frac{\nu+3}{2})}{\Gamma(\frac{\nu}{2})} \frac{x}{\sqrt{\nu\pi}} \frac{1}{(1+\frac{x^2}{2})(1+\frac{x^2}{\nu})^{\frac{\nu+1}{2}}}\,, \qquad -\infty < x < \infty\,, \tag{2.33}$$

der ν er ein konstant større enn null. Den tradisjonelle kurtosen til fordelinga er $k = 3(\nu-2)/(\nu-4)$. Denne kurtosen er ikkje definert for $\nu \leq 4$, og

for $\nu = 4 + \epsilon$ kan kurtosen bli valfri høg for liten nok $\epsilon > 0$. I dette høvet har målet til Critchley og Jones (2008) ein fordel. Den høge kurtosen til t-fordelinga tyder at fordelinga har tunge halar. Tettleiken får meir masse på skuldrane og mindre i halane dersom vi aukar ν. Då minkar også kurtosen, og det er desse eigenskapane kurtose er meint å måle. Dette er illustrert i figur 2.9, og ettersom fordelinga er symmetrisk, er venstrekurtosen lik høgrekurtosen. Frå tabell 2.2 ser vi at for dei fridomsgradene kurtosekoeffisienten er definert, rangerer dei ulike måla kurtosen likt. Vi ser også at kurtosen synk med fridomsgradene og at k nærmar seg 3, som er den same kurtosen som normalfordelinga.



Figur 2.9: *Tettleiken til t-fordelinga oppe til venstre, den deriverte til t-fordelinga oppe til høgre og høgrekurtosen under.*



Figur 2.10: *Tettleiken til lognormalfordelinga øverst til venstre, og den deriverte øverst til høgre. Venstrekurtosen for lognormalfordelinga er nederst til venstre, og høgrekurtosen nederst til høgre.*

σ	δ	δ_R	δ_L	k	k^*	μ	m
0.5	0.33	0.30	0.03	8.90	8.41	1.13	0.78
1	0.31	0.42	-0.11	114	72	1.65	0.37

Tabell 2.3: *Kurtosen til lognormalfordeling med ulike parameter σ. Gradientasymmetrikoeffisentane δ, δ_R og δ_L er henholdsvis total- høgre- og venstrekurtose. Den tradisjonelle kurtosekoeffisienten er k medan modalkurtose er k^*. Forventningsverdi er μ og modalverdi m.*

2.2.4 Lognormalfordeling

Som vi såg i avsnitt 1.3.3 er lognormalfordelinga ei asymmetrisk fordeling med tunge halar,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{x}e^{-\frac{(\log x)^2}{2\sigma^2}}\ , \qquad x > 0\ , \tag{2.34}$$

$$f'(x) = -\frac{1}{\sqrt{2\pi}\sigma}\frac{1}{x^2}e^{-\frac{(\log x)^2}{2\sigma^2}}(1 + \frac{1}{\sigma^2}\log x), \qquad x > 0 \tag{2.35}$$

der $\sigma > 0$. I tabell 2.3 ser vi at både kurtosekoeffisienten k og modalkurtosen k^* aukar med aukande σ. Dette gjeld ikkje for totalkurtosen, då negativt bidrag frå venstrekurtosen er med på å senke den. Dette kan vere eit argument for å skilje høgre- og venstrekurtose frå kvarandre, ettersom det er tydeleg frå 2.10 at for $\sigma = 1/2$ er det vesentleg lettare halar og spissare topp. Dersom det er dette vi ynskjer kurtosen skal måle, er totalkurtosen i så måte eit dårleg mål. Vi ser vidare at modalverdi og forventning, m og μ, skil seg frå kvarandre for $\sigma = 1$, og dette fører til at det er stor skilnad i modalkurtose og kurtose. Likevel viser begge måla så høge verdiar at det ikkje er nokon tvil om at kurtosen er høg.

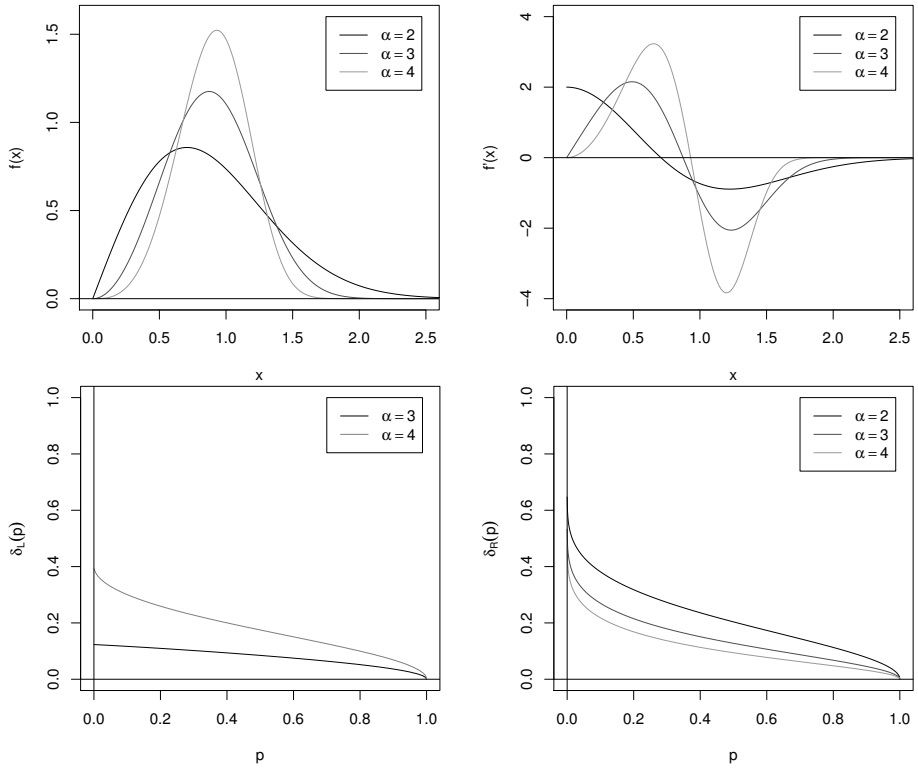
2.2.5 Weibull-fordeling

Weibullfordelinga er gjeven ved

$$f(x) = \frac{\alpha}{\beta}x^{\alpha-1}e^{-\frac{x^\alpha}{\beta}}\ , \qquad x > 0\ , \tag{2.36}$$

$$f'(x) = \frac{\alpha}{\beta}e^{-\frac{x^\alpha}{\beta}}((\alpha-1)x^{\alpha-2} - \frac{\alpha}{\beta}x^{2\alpha-2})\ , \qquad x > 0\ , \tag{2.37}$$

der $\alpha, \beta > 0$. Kurtosekoeffisienten til Weibullfordelinga er relativt komplisert, men er i motsetnad til skeivleikskoeffisienten, avhengig av β. Kurtosefunksjonane til Critchley og Jones (2008) vert derimot ikkje påverka av β (teorem 2.1.4). Vidare ser vi av tabell 2.4 at den tradisjonelle kurtosen ikkje er monoton saman med α. Dette kan vere eit argument mot den tradisjonelle kurtosen, ettersom tettleiken vi ser frå figur 2.11 får lettare halar med høg α. Likevel er tettleiken spissare med høg α, og det må vere dette som gjer at kurtosen ikkje er monoton. Dersom vi ser på kurtosekoeffisientane til Critchley og Jones (2008), kan vi sjå at δ_L er stigande med α, medan δ_R er synkande. Dette kan vere eit argument for å skilje kurtosen i to deler, ettersom dei to delene har ulike eigenskapar når det gjeld spissheit og haletyngde. I så fall har målet totalkurtose dei same veikskapane som den tradisjonelle kurtosekoeffisienten.



Figur 2.11: *Tettleiken til weibullfordelinga oppe til venstre, og den deriverte til høgre. Venstrekurtosen er nede til venstre og høgrekurtosen nede til høgre.*

α	δ	δ_R	δ_L	k	k^*	μ	m
2	N/A	0.22	N/A	3.25	3.89	0.89	0.71
3	0.23	0.15	0.08	2.73	2.77	0.89	0.87
4	0.40	0.11	0.29	2.75	2.78	0.91	0.93

Tabell 2.4: *Kurtosen til Weibullfordelinga med ulike parameter α. Gradientasymmetrikoeffisentane δ, δ_R og δ_L er henholdsvis total- høgre- og venstrekurtose. Den tradisjonelle kurtosekoeffisienten er k medan modalkurtose er k^*. Forventningsverdi er μ og modalverdi m.*

2.3 Samandrag

Høg kurtose blir ofte tolka som at fordelinga er spiss med tunge halar. Dette gjeld til ein viss grad med gradientasymmetri også. Dersom talverdien til den deriverte omkring toppen er høg, medan den er låg i halen, kan vi seie at fordelinga er spiss med tunge halar. Då vil også gradientasymmetrien vere høg. Dersom talverdien til den deriverte derimot er låg omkring toppen og høg i halane, vil gradientasymmetrien vere negativ og tilsvare låg kurtose. Likevel opplevde vi at gradientasymmetrien oppførte seg annleis enn kurtosekoeffisienten når vi skulle rangere kurtosen til ulike fordelingar med ulike parameter. Grunnen til dette var at vi innførte målet totalkurtose, der vi slo saman høgre- og venstrekurtose. Eigenskapane til høgre og venstre del av ei fordeling kan vere svært ulike, og ein bør kanskje difor halde seg til Critchley og Jones (2008) og sjå på høgre- og venstrekurtose separat dersom ein vil bruke gradientasymmetri som eit mål på kurtose.

Kapittel 3

Estimering av asymmetrifunksjonen

I dei to føregåande kapitla har vi sett på skeivleik og kurtose som teoretiske mål. I dette kapittelet skal vi sjå på teori for estimering av måla. Vi byrjar med å sjå på estimering av den tradisjonelle skeivleiken. Deretter ser vi på testing om data kjem frå eintoppa fordelingar. Dette er naudsynt, sidan definisjonane til Critchley og Jones (2008), har føresetnader om eintopping. Vidare ser vi på tettleiksestimering, før vi bruker dette til estimat av asymmetrifunksjonen. Til slutt gir vi forslag til korleis vi kan estimere gradientasymmetrien.

3.1 Empirisk skeivleik

Den mest vanlege estimatoren for skeivleik, er den empiriske skeivleikskoeffisienten som er bygd på empiriske moment (Joanes og Gill 1998).

Definisjon 3.1.1. *Det p-te empiriske sentralmomentet er definert som*

$$m_p = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^p \ . \tag{3.1}$$

Definisjon 3.1.2. *Den empiriske skeivleikskoeffisienten er*

$$g_1 = \frac{m_3}{m_2^{\frac{3}{2}}} \ . \tag{3.2}$$

Eitt av problema med denne estimatoren er at han ikkje er forventningsrett, særskild kan det vere store ulikskapar for små datasett (Joanes og Gill 1998).

Sjølv om $\lim_{n\to\infty} E(g_1) = g_1$ kan vi heller bruke

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1 \ , \tag{3.3}$$

som estimator for skeivleiken. Denne estimatoren er størren enn g_1, men går mot g_1 når n går mot uendeleg. Uheldigvis er heller ikkje denne estimatoren forventningsrett, men er under alle høve betre for små datasett.

3.2 Testing for eintopping

Ein av føresetnadane for at asymmetrifunksjonen skal vere definert, er at fordelinga er eintoppa. Det er fleire måtar vi kan gjere dette. I denne oppgåva skal vi for små datasett bruke dip-testen (Hartigan og Hartigan 1985), og for store datasett vil vi bruke ein test som basert på (Silverman 1981).

3.2.1 Dip-testen

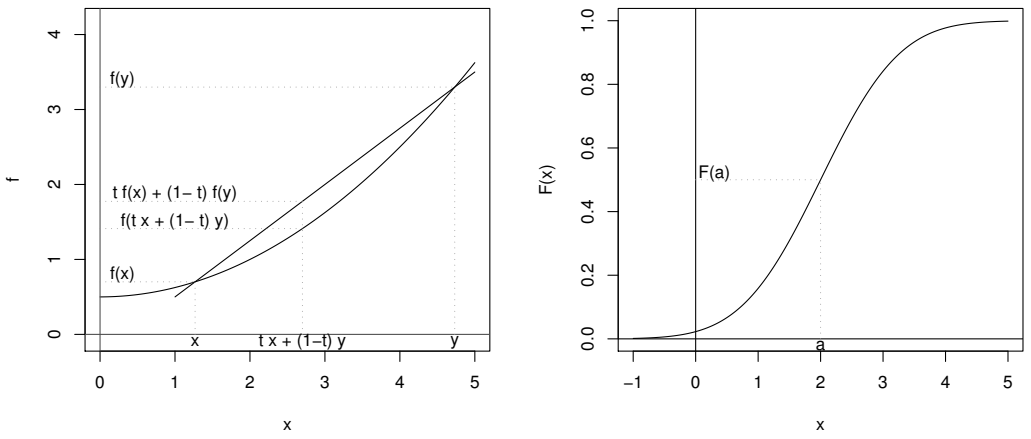
I dette delkapittelet ser vi på dip-testen for eintopping, som blir føreslått av Hartigan og Hartigan (1985). Vi byrjar med nokre definisjonar som vi treng for å definere eintoppa fordelingar.

Definisjon 3.2.1. *Ein funksjon, f er strengt konveks på (x, y) dersom $f(tx + (1-t)y) > tf(x) + (1-t)f(y)$ for alle $t \in (0, 1)$. Dersom $-f$ er konveks, er f konkav (Rudin 1976).*

Definisjon 3.2.2. *Dersom $-f$ er konveks på området (a, b), er f konkav på same område (Rudin 1976).*

Dip-testen tar utgangspunkt i fordelingsfunksjonen til ei eintoppa fordeling. Fordelingsfunksjonar kan ta verdiar mellom 0 og 1, og er i tillegg stigande, dvs at $0 \leq F(x) \leq 1$ og at $F'(x) \geq 0$ for alle x. Fordelingsfunksjonen til ei eintoppa fordeling er i tillegg strengt konveks fram til modalverdien og deretter strengt konkav. Ofte blir denne eigenskapen skildra gjennom den andrederiverte, ettersom funksjonen er konkav dersom den andrederiverte er negativ. Dette er ikkje transitivt og difor bruker vi definisjonane 3.2.1 og 3.2.2. Døme på konveks funksjon og eintoppa fordeling er illustrert i figur 3.1. Vidare fortset vi med nokre definisjonar frå (Hartigan og Hartigan 1985), som vi vil bruke til dip-testen:

Definisjon 3.2.3. *Vi definerer $\eta(F, G) = \sup_x |F(x) - G(x)|$ for alle avgrensa funksjonar F, G. Vi definerer $\eta(F, \Lambda) = \inf_{G\in\Lambda} \eta(F, G)$ for alle klassar av avgrensa funksjonar Λ. La Υ vere klassa av eintoppa funksjonar.*



Figur 3.1: *Konveks funksjons til venstre. Eintoppa fordeling som er konveks på* $(-\infty, a)$ *og konkav på* (a, ∞) *til høgre.*

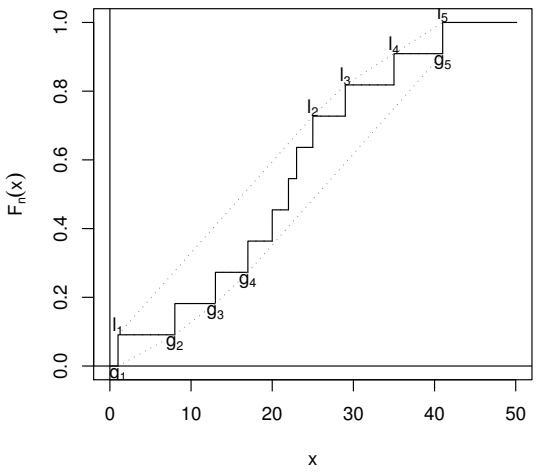
Definisjon 3.2.4. *Dipen til ei fordeling,* F*, er definert til* $D(F) = \eta(F, \Upsilon)$*.*

Definisjon 3.2.5. *Den største konvekse minoranten (g.c.m.) til* F *på området* $(-\infty, a]$ *er* $\sup G(x)$*, der* sup *vert teken over alle funksjonar* G *som er konvekse i* $(-\infty, a]$ *og ikkje større enn* F *på nokon stad. Den minste konkave majoranten (l.c.m.) til* F *på området* $[a, \infty)$ *er* $\inf G(x)$*, der* inf *vert teken over alle funksjonar* G *som er konkave i* $[a, \infty)$ *og ikkje mindre enn* F *på nokon stad.*

Definisjon 3.2.6. *Den empiriske fordelingsfunksjonen for dei identiske uavhengige observasjonane* X_1*,* X_2*,...,*X_n *er definert som*

$$F_n = \frac{1}{n} \sum_i \mathbf{1}(X_i < x) \, . \tag{3.4}$$

Vi ser frå definisjon 3.2.4 at dipen er den største skilnaden mellom fordelinga, og den eintoppa funksjonen som ligg nærmast F. Vidare ser vi at vi frå definisjon 3.2.5, at vi kan rekne ut dipen ved å ta max av den største konvekse minoranten og den minste konkave majoranten. Dersom vi har observasjonane $X_1, X_2, ..., X_n$, kan vi rekne ut ein observator for dipen. Den største konvekse minoranten til den empiriske fordelingsfunksjonen, vil vere rette liner mellom nedre knekkpunkt til funksjonen, g_i, samstundes med at linene framleis lagar ein konveks funksjon. Vidare finn vi den største avstanden mellom minoranten og den empiriske fordelingsfunksjonen. Omvendt vil den minste konkave majoranten vere rette liner mellom øvre knekkpunkt, g_j, samstundes med at linenen lagar ein konkav funksjon. Også her finn vi



Figur 3.2: *Den empiriske fordelingsfunksjonen med tilhøyrande største konvekse minorant og minste konkave majorant. Funksjonen er i heiltrekte liner, medan minoranten og majoranten er i stipla liner.*

den største avstanden mellom funksjonen og majoranten, og vi set dipen til den største av desse avstandane. Denne prosedyren gjer vi for alle intervall slik at minoranten blir rekna ut på intervallet $(-\infty, X_a]$ og majoranten blir rekna ut for $[X_a, \infty)$. Vi bruker det intervallet som gjer den minste største avstanden til å rekne ut dipen. Hartigan og Hartigan (1985) har laga ein algoritme som er effektiv for å rekne ut observator for dipen. Mange av intervalla kan forkastast enkelt, og dette vil korte ned køyretid for å rekne ut observatoren. Denne algoritmen er presentert i neste avsnitt, og er grafisk illustrert i figur 3.2.

3.2.2 Rekne ut ein observator for dipen

Når vi skal rekne ut eit estimat for asymmetrien, vil vi teste om observasjonane våre kjem får ei eintoppa fordeling. Til det testar vi hypotesen,

$$H_0: \text{ Datasettet kjem frå ei eintoppa fordeling,}$$

mot den alternative hypotesen,

$$H_1: \text{ Datasettet kjem frå ei fleirtoppa fordeling.}$$

I denne hypotesetesten vil vi bruke dipen som testobservator, og vi får forkasting for ein kritisk verdi $D > D_{crit}$.

Dersom vi har eit sett med ordna observasjonar frå ei fordeling, $X_1, X_2, ..., X_n$,

1. Sett $D = 0$, $X_L = X_1$ og $X_U = X_n$
2. Rekn ut g.c.m G og l.c.m. L for F_n på området $[X_L, X_U]$. Då blir også punkta som er i kontakt med F_n henholdsvis $g_1, g_2, ..., g_k$ og $l_1, l_2, ..., l_m$.
3. Dersom $d = \sup_i |G(g_i) - L(g_i)| > \sup_i |G(l_i) - L(l_i)|$ og sup opptrer for $l_j \leq g_i \leq l_{j+1}$, sett $X_L^0 = g_i$ og $X_U^0 = l_{j+1}$.
4. Dersom $d = \sup_i |G(l_i) - L(l_i)| \geq \sup_i |G(g_i) - L(g_i)|$ og sup opptrer for $g_i \leq l_j \leq g_{i+1}$, sett $X_L^0 = g_i$ og $X_U^0 = l_j$.
5. Dersom $d \leq D$, sett $d = D$ og hopp til punkt 8.
6. Dersom $d > D$, sett $D = \max(D, \sup_{X_L \leq x \leq X_L^0} |G(x) - F(x)|, \sup_{X_U^0 \leq x \leq X_U} |L(x) - F(x)|)$.
7. Sett $X_U = X_U^0$, $X_L = X_L^0$ og gå tilbake til 2.
8. Dersom $D > D_{crit}$ forkast H_0 : Datasettet kjem frå ei eintoppa fordeling. Dersom $D \leq D_{crit}$ hald på H_0.

Dei kritiske verdiane for dip-testen er rekna ut frå datasett som er simulert frå ei uniform fordeling. Den uniforme fordelinga er den minst eintoppa av alle eintoppa fordelingar, og minimerer dermed sannsynet for å forkaste ein nullhypotese som er riktig. I hypotesetesting vert dette kalla type 1-feil. Omvendt er type 2-feil definert som sannsynet for å forkaste ein feilaktig H_0 betinga under ein riktig H_1. Kor god dip-testen er til å unngå type 2-feil er usikkert, men sannsynet for type 1-feil er kartlagt for nokre eintoppa fordelingar i avsnitt 4.4. Det er diverre ikkje rekna ut kritiske verdiar for dip-testen for datasett større enn 5000 observasjonar. Dersom datasettet er større enn 5000 observasjonar, kan vi bruke ein metode som Silverman (1981) føreslår. Denne metoden går ut på å telje modalverdiane til den estimerte tettleiken, og er eigna til store datasett. Dette kjem vi tilbake til i avsnitt 3.3.3.

3.3 Kjerneestimering

Dersom vi fastslår at observasjonane kjem frå ei eintoppa fordeling, kan vi gå i gang med å estimere fordelinga. Ein vanleg måte å gjere dette er kjerneestimering. Ein estimator for tettleiken er

$$\hat{f}(x; h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) , \tag{3.5}$$

der K er ein funksjon som oppfyller $\int K(u)\,\mathrm{d}u = 1$ (Wand og Jones 1995). Vi kallar $K(u)$ for kjernen, og vi vil i denne oppgåva bruke ein tettleik som K. På den måten sikrar vi at $\hat{f}(x;h)$ også er ein tettleik. I tillegg må vi velje ein h, som blir kalla bandbreidde. Vidare bruker vi $\hat{f}$ som sannsynstettleik, og reknar ut asymmetrifunksjonen på same måte som i avsnitt 1.1.1. Likevel er det nokre ting vi må ta stilling til før vi kan gå i gang med dette.

3.3.1 ISE/MISE

Kvadratisk feil er eit mykje brukt mål på kor god ein estimator er. Den kvadratiske feilen til ein kontinuerleg funksjon, er definert som $ISE = \int(\hat{f}(x) - f(x))^2\,\mathrm{d}x$. Dette står for integrated square error. Vi kan også sjå på forventa kvadratisk feil, $MISE$ (mean integrated square error), eller asymptotisk forventa kvadratisk feil $AMISE$ (asymptotic mean integrated square error), som er eit godt mål dersom datasettet er stort.

3.3.2 Val av kjerne

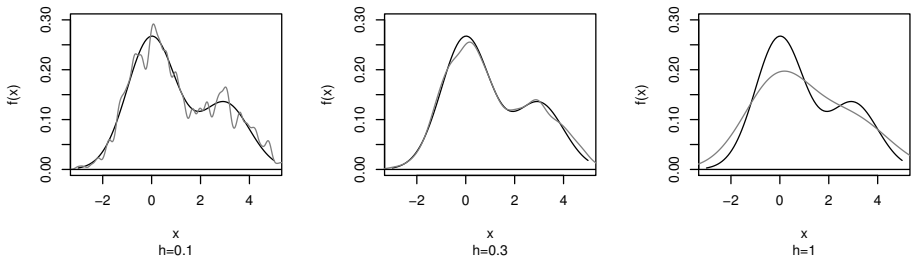
Når vi skal estimere asymmetrifunksjonen gjennom tettleiksestimering, må vi fyrst velje ein kjerne. Ein vanleg kjerne å bruke er normalfordelinga. Denne kjernen har gode rekneeigenskapar, men er ikkje nødvendigvis den kjernen som gir minst feil. Kjernar som tek utgangspunkt i betafordelinga,

$$K(x;p) = 2^{p+1}B(p+1,p+1)^{-1}(1-x^2)^p\,, \qquad -1 < x < 1\,, \tag{3.6}$$

der $B(\cdot,\cdot)$ er betafunksjonen, gir betre $MISE$-eigenskapar for $p = 1,2,3$ (Wand og Jones 1995). Når vi skal estimere asymmetrien, er riktig definisjonsområde viktig for å klare å estimere halane riktig. Dette er eitt av argumenta for å bruke asymmetrifunksjonane til Critchley og Jones (2008), og er viktig å ta hensyn til. Dersom vi veit noko om datasettet på førehand, kan vi velje ein kjerne som passar til definisjonsområdet. Isåfall kan vi bruke ein asymmetrisk kjerne dersom datasettet kjem frå ei asymmetrisk fordeling. I kapittel 4 skal vi estimere både tettleik og asymmetrifunksjon frå ulike datasett. Der kjem vi tilbake til fleire problemstillingar knytta til val av kjerne.

3.3.3 Bandbreidde

Sjølv om vi har testa om observasjonane kjem frå ei eintoppa fordeling, er det ikkje dermed gitt at den estimerte tettleiken bli eintoppa. For å få det til må vi også velje ei passande bandbreidde. Generelt sett, vil vi velje ei



Figur 3.3: *Estimerte tettleikar for 1000 observasjonar frå ei Normal-blanda fordeling. I dei ulike figurane er h bandbreidda, svart line er teoretisk tettleik medan grå line er estimert tettleik.*

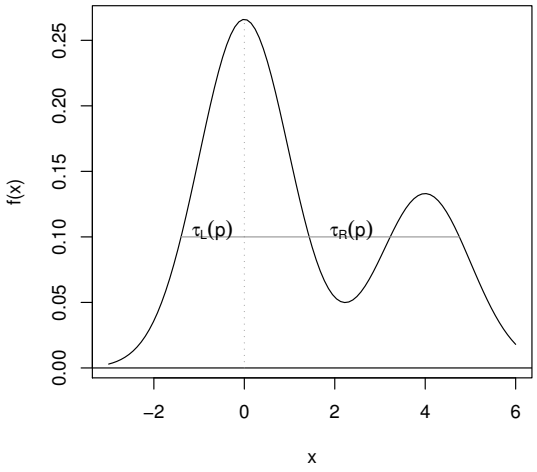
bandbreidde som gjer MISE minst mogeleg, men vi må også ta hensyn til at asymmetrifunksjonen er definert på den estimerte tettleiken. Ved val av bandbreidde på den estimerte tettleiken, kan vi bruke eigenskapen

$$h_{AMISE} \leq \left[\frac{243R(K)}{35\mu_2(K)^2 n}\right]^{\frac{1}{5}} \sigma \ , \tag{3.7}$$

der $R(K) = \int K(z)^2 \, \mathrm{d}z$, μ_2 er det andre sentralmomentet til K og σ er standardavviket til fordelinga datasettet kjem frå (Wand og Jones 1995). Standardavviket kan vi estimere på vanleg måte med

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \ . \tag{3.8}$$

Merk at denne bandbreidda er ein øvre skranke for den asymptotisk minste kvadratiske feilen, og at den beste bandbreidda som regel er mindre. Likevel er ikkje det alltid ønskjeleg å bruke den bandbreidda som gir minst kvadratisk feil. Dersom vi vel for liten bandbreidde vil tettleiken bli overestimert, og vi får fleire enn ein modalverdi. Dersom vi vel for stor bandbreidde vil tettleiken bli underestimert, og gi dårlege resultat. Vi kan i tillegg risikere å klassifisere ei fleirtoppa fordeling som eintoppa ved å velje stor bandbreidde. I figur 3.3 ser vi døme på overestimert tettleik til venstre, bra estimert tettleik i midten og underestimert tettleik til høgre. I avsnitt 3.2.2 brukte vi dip-testen for å klassifisere om ein tettleik er eintoppa. Denne testen er ikkje utvikla for datasett som er større enn 5000 observasjonar, men vi kan bruke ein test basert på Silverman (1981). Denne testen går ut på å finne ei kritisk bandbreidde som gir fleirtoppa estimert tettleik. Dersom den kritiske bandbreidda er stor, er fordelinga fleirtoppa og vi forkastar ein nullhypotese om at datasettet kjem frå ei eintoppa fordeling.



Figur 3.4: *Alternativ definisjon for skeivleik. Asymmetrifunksjonen vil i dette tilfellet ikkje vere kontinuerleg når p når den minste toppen*

Ettersom kritiske verdiar for dip-observatoren er rekna ut for ei uniform fordeling, er diptesten robust mot type 1-feil. Dette kan gje seg utslag i at faren for type 2-feil er stor. Vi kan også risikere at den estimerte tettleiken blir fleirtoppa med $\hat{h}_{AMISE}$ som bandbreidde, samstundes med at dip-testen klassifiserer datasettet som eintoppa. Då kan vi enten forkaste datasettet som eintoppa, eller vi kan utvide definisjonen for asymmetri til også å gjelde for fleirtoppa datasett. Ein måte å gjere det på, vil vere å ta utgangspunkt i den største modalverdien, og rekne asymmetrifunksjonen ut frå det. Når vi kjem på høgde med den neste modalverdien, vil ikkje inversen lenger vere eintydig. Vi vel då å bruke den største verdien til inversen av den estimerte tettleiken, og asymmetrifunksjonen gjer dermed eit hopp og er ikkje lenger kontinuerleg. Denne metoden er illustrert i figur 3.4.

3.3.4 Mål på kor vanskeleg tettleiken er å estimere

Eit vanleg mål på kor vanskeleg ein tettleik er å estimere er,

$$D(f) = \left[\sigma(f)^5 R(f'')\right]^{\frac{1}{4}} \ , \tag{3.9}$$

der $R(f'') = \int f''(x)^2 \, \mathrm{d}x$ og $\sigma(f)^2$ er variansen til f (Wand og Jones 1995). Ei fordeling med mykje krumming vil gi høg $R(f'')$, og saman med varians styrer altså dette kor vanskeleg det er å estimere fordelinga. Den fordelinga som er lettast å estimere er $Beta(4,4)$-fordelinga (Wand og Jones 1995), og blir ofte brukt som referanse for kor vanskeleg tettleiken er å estimere.

3.4 Den estimerte asymmetrifunksjonen

I dei føregåande avsnitta har vi gått gjennom teori for å kunne estimere asymmetrifunksjonen. Vi fortset i dette avsnittet med å definere den estimerte asymmetrifunksjonen, og føreslår ein metode for å bruke denne definisjonen på datasett.

Definisjon 3.4.1. *La $\hat{f}(x)$ vere eit estimat av ein rota eintoppa tettleiksfunksjon på området (a, b), med estimert modalverdi $\hat{m}$. Den estimerte modalverdien er $\hat{m} = \sup(x : \hat{f}(x; h))$. Ein føresetnad er at også $\hat{f}(x)$ er eintoppa. Då definerer vi*

$$\hat{X}_R(p) = \hat{f}_R^{-1}(p\hat{f}(\hat{m})) \ ,$$
$$\hat{X}_L(p) = \hat{f}_L^{-1}(p\hat{f}(\hat{m})) \ .$$

Til slutt lar vi

$$\hat{\tau}_R(p) = \hat{X}_R(p) - \hat{m} \ , \hat{\tau}_L(p) = \hat{m} - \hat{X}_L(p) \ .$$

På denne måten kan vi definere dei estimerte asymmetrifunksjonane,

Definisjon 3.4.2.

$$\hat{\rho}(p) = \frac{\hat{\tau}_R(p)}{\hat{\tau}_L(p)} \ , \qquad 0 < p < 1 \ , \tag{3.10}$$

$$\hat{\gamma^*}(p) = \frac{\hat{\rho}(p) - 1}{\hat{\rho}(p) + 1} = \frac{\hat{\tau}_R(p) - \hat{\tau}_L(p)}{\hat{\tau}_R(p) + \hat{\tau}_L(p)} \ , \qquad 0 < p < 1 \ . \tag{3.11}$$

3.4.1 Algoritme for å estimere asymmetrifunksjonen

1. Test om dataene kjem frå ei eintoppa fordeling med Dip-testen.
2. Velg kva kjerne du vil bruke for å estimere fordelinga.
3. Velg kva bandbreidde du vil bruke (estimat for h_{AMISE} blir foreslått).
4. Estimer fordelinga og rekn ut asymmetrifunksjonen.
5. Dersom det er ønskeleg kan du også rekne ut skeivleikskoeffisienten som er basert på asymmetrifunksjonen.

3.5 Estimering av kurtose

På same måte som for skeivleiken er det å bruke empiriske sentralmoment den mest vanlege måten å estimere kurtosen.

Definisjon 3.5.1. *Den empiriske kurtosen er*

$$g_2 = \frac{m_4}{m_2^2} \, . \tag{3.12}$$

Denne estimatoren er diverre heller ikkje forventningsrett, og difor blir ofte

$$G_2 = \frac{(n+1)(n-1)}{(n-2)(n-3)} \, g_2 \, , \tag{3.13}$$

brukt som estimator for kurtosen (An og Ahmed 2008). Denne estimatoren tar utgangspunkt i dei forventningsrette estimatorane for det fjerde og andre sentralmomentet, og er forventningsrett for normalfordelinga. Dersom vi er ute etter ein estimator for eksesskurtosen må vi trekke frå tre.

Det er mogeleg å estimere asymmetrifunksjonen ved hjelp av kjerneestimering. På same måte kan vi estimere gradientasymmetrien, men då med å estimere den deriverte til tettleiken. Ein estimator for den deriverte er

$$\hat{f}'(x;h) = \frac{1}{nh^2} \sum_{i=1}^{n} K'(\frac{x - X_i}{h}) \, . \tag{3.14}$$

Denne estimatoren er rimeleg, ettersom det er den deriverte til estimatoren for tettleiken. Optimal bandbreidde er av orden $n^{-\frac{1}{7}}$ samanlikna med $n^{-\frac{1}{5}}$ for estimat av tettleiken (Wand og Jones 1995). Sjølv om tettleiken er eintoppa, er ikkje dette tilstrekkeleg føresetnad for definisjonen til Critchley og Jones (2008), ettersom også den deriverte også må vere eintoppa. Dip-testen er diverre ikkje utvikla for dette føremålet, men vi kan bruke ei modifisering av testen til Silverman (1981). Dette gjer vi ved å estimere den deriverte, og deretter undersøke monotonieigenskapane til estimatet.

3.6 Samandrag

I dette kapittelet har vi sett på og utarbeida teori for å kunne estimere asymmetrifunksjonen til Critchley og Jones (2008), og i avsnitt 3.4.1 lagde vi ein algoritme for å gjennomføre denne estimeringa. Denne algoritmen skal vi bruke i neste kapittel for å undersøke kor godt vi klarar å estimere asymmetrifunksjonane. Algoritmen kan også brukast til å estimere gradientasymmetrifunksjonar, men vi må ta hensyn til at dip-testen ikkje fungerer på den

deriverte. Vi må også tenke på at same bandbreidde ikkje nødvendigvis er optimal for begge estimatorane, og at kjernen bør ha monotonieigenskapar som er eigna til estimering av den deriverte.

Kapittel 4

Asymmetrifunksjonen brukt på datasett

I dette kapittelet skal vi estimere asymmetrifunksjonen for datasett frå ulike fordelingar. Ettersom det er vanskeleg å finne eigenskapar til estimatoren analytisk, må vi bruke data til eksperimentelt å undersøke målet. Vi byrjar med teoretisk å klassifisere kor vanskeleg ulike fordelingar er å estimere, før vi går gjennom litt generell teori for val av kjerne. Deretter ser vi på asymmetrifunksjonen brukt på ulike datasett der vi kjenner fordelingsfunksjonen. Når vi har sett på dei estimerte asymmetrifunksjonane, vil vi sjå kor god dip-testen er til å klassifisere eintopping. Til slutt ser vi på estimat av gradientasymmetrien.

4.1 Mål på vanskegraden til fordelingane

I avsnitt 3.3.4 definerte vi eit mål på kor vanskeleg ein tettleik er å estimere. I tabell 4.1 har vi rekna ut vanskegraden til ulike fordelingar i høve til Beta(4,4)-fordelinga, som er den enklaste fordelinga å estimere. Det er data frå fordelingane i denne tabellen vi skal sjå på i denne oppgåva. Vi ser at Gamma(2,1) er den fordelinga som er vanskelegast å estimere, medan normalfordelinga er den fordelinga som er enklast. Dette gjeld for tettleiken, men det treng ikkje nødvendigvis vere slik for asymmetrien.

4.2 Optimal kjerne for estimering

I dei ulike simuleringane har vi brukt fire ulike kjernar: Normalfordelinga, epanechnikov-fordelinga, biweight, triweigth og gammafordelinga. Grunnen til at normalfordelinga ofte blir brukt, er at det blir enkelt å simulere data

Tettleik	$D(f^*)/D(f)$
Beta(4,4)	1
Normal	0.908
Polynomen(1,2)	0.859
Gamma(10,1)	0.804
Polynomen og ekponensial	0.725
Gamma(2,1)	0.378

Tabell 4.1: *Mål på kor vanskelege tettleikane er å estimere. Her er* $D(f) = (\sigma(f)^5 R(f'')^{\frac{1}{4}}$ *og* f^* *er* $Beta(4,4)$*-fordelinga.*

frå den estimerte fordelinga. Dette er ikkje nødvendigvis noko vi treng prioritere, då vi i utgangspunktet berre er interessert i å kartlegge asymmetrien til datasettet. Normalfordelinga er ei fordeling som er definert på heile tallina, og dersom vi brukar den som kjerne, vil den estimerte tettleiken også vere definert på heile tallina. Halane til den estimerte tettleiken vil vere like, og den estimerte asymmetrien vil difor vere null i halane. Difor vil vi også prøve med kjernar som er definerte på eit avgrensa område.

Tre kjernar som gir gode resultat med tanke på $MISE$, høyrer til familien

$$K_B(x;p) = 2^{2p+1}B(p+1,p+1)(1-x^2)^p 1_{|x|<1} \ . \tag{4.1}$$

Denne familien er eit spesialtilfelle av betafordelinga med parameter (p,p) og transformasjonen $x = (y+1)/2$. Dersom $p = 1$, vert kjernen kalla Epanechnikov-kjernen og er den kjernen som teoretisk gir lågast $MISE$. Ei ulempe med denne kjernen er at den deriverte ikkje er kontinuerleg i $x = \pm 1$. Då vil den deriverte til den estimerte tettleiken heller ikkje vil vere kontinuerleg, og dette vil igjen gi hopp i asymmetrifunksjonen. Gradientasymmetrien vil ikkje vere definert med denne kjernen. Kjernen vert kalla biweight dersom $p = 2$. Dette er ein effektiv kjerne med tanke på $MISE$, og den deriverte er kontinuerleg for alle x. Den andrederiverte er derimot ikkje kontinuerleg i $x = \pm 1$, og difor har vi også prøvd med triweight, $p = 4$. Kjernane som tar utgangspunkt i K_B er til motsetnad frå normalfordelinga definert på eit avgrensa område. Av den grunn vil også den estimerte tettleiken vere definert på eit avgrensa område. Dersom vi har data frå ei asymmetrisk fordeling, vil også desse kjernane estimere asymmetrien feil i halane. Likevel kan vi få betre resultat enn med normalfordelinga, då den estimerte tettleiken under alle høve ikkje vil vere symmetrisk i halane.

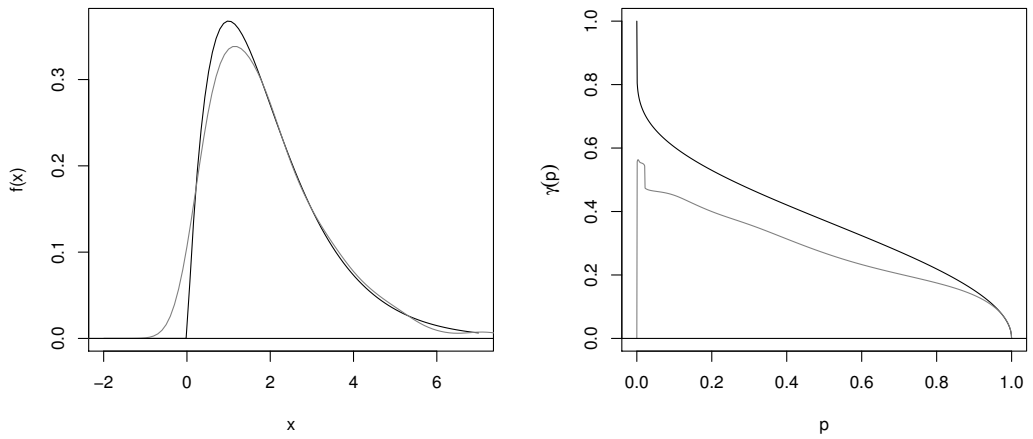
Til slutt har vi prøvd med gammafordelinga som kjerne. Generelt er asymmetriske kjernar lite brukt i kjerneestimering. Vi ønskjer likevel å sjå korleis

ein slik kjerne estimerer ein tettleik frå eit datasett vi veit er trekt frå ei asymmetrisk fordeling på $(0, \infty)$. For å sjå kor gode dei ulike kjernane er, vil vi sjå på ISE til estimert tettleik og asymmetrifunksjon. Vi vil også sjå på feilen til estimerte asymmetrikoeffisientar, men vi må hugse på at både ISE og varians representerer kvadratisk og ikkje absolutt feil.

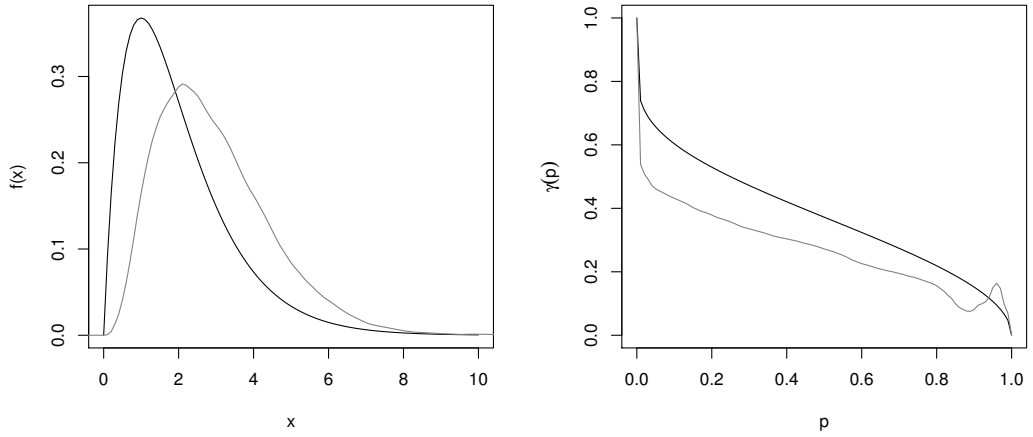
4.3 Estimerte asymmetrifunksjonar

I dette avsnittet skal vi sjå på kor bra det estimerte asymmetrimålet fungerer i praksis. Vi skal byrje med å sjå på data frå Gamma(2,1)-fordelinga, der vi skal sjå på kor bra asymmetrien er estimert, og korleis ulike kjernar påverkar dette. Grunnen til at vi har vald Gamma(2,1), er at dette er ei mykje brukt fordeling som også er relativt asymmetrisk. Vidare skal vi sjå på fordelinga sett saman av polynomen av ulik grad frå likning (1.21), då denne har ein analytisk asymmetrifunksjon. Vi skal også sjå på data frå Gamma(10,1)-fordelinga. Denne er klart mindre asymmetrisk enn Gamma(2,1), og vi ønskjer å sjå korleis estimatet fungerer på ei fordeling som i mindre grad er asymmetrisk. Til slutt skal vi sjå på data frå normalfordelinga, for å sjå om estimatoren klassifiserer fordelinga som symmetrisk.

I alle eksperimenta har vi simulert datasett på storleikane 10, 100, 1000 og 5000. Ein av grunnane til at vi har vald desse storleikane, er at kvantilar for dip-observatoren er rekna ut for desse datasetta. Vi har vald å simulere 100 datasett av kvar storleik. Dette er nok data til å få eit godt inntrykk av korleis målet fungerer, og vi legg ved varians til dei ulike estimata våre i tillegg B. Eit anna argument for å velje 100 datasett, er køyretida det tar å handsame datamaterialet. Ettersom tettleiksestimat basert på kjerneestimering er store objekt, vil dei krevje mykje kraft å handtere. Dersom vi ville sett på fleire datasett, ville vi brukt lang tid på simuleringa. Det kan argumenterast for at vi kunne sett på fleire av dei minste setta, då dei ikkje krev like mykje. Vi vil likevel påstå at vi får eit godt inntrykk av kor gode estimata er med dei dataene vi har simulert. Dei ulike verdiane vi har rekna ut er: Gjennomsnitt og varians til asymmetrikoeffisienten γ, som vi har representert med $\mathrm{E}(\hat{\gamma})$ og $\mathrm{Var}(\hat{\gamma})$. Gjennomsnitt til ISE for asymmetrifunksjonen og tettleiken, som vi har representert med $MISE_{\gamma^*}$ og $MISE_{\hat{f}}$. Gjennomsnittet til den tradisjonelle estimatoren, $\mathrm{E}(G_1)$, og til slutt gjennomsnitt for estimat av modalverdien, $\mathrm{E}(\hat{m})$. Desse variablene finn vi for dei ulike fordelingane i tabellane 4.2-4.12.



Figur 4.1: *Tettleiken til Gamma(2,1)-fordeling til venstre, og asymmetrifunksjon til høgre. Utgangspunktet er 1000 observasjonar og normalfordelinga som kjerne i tettleiksestimatet. Teoretiske funksjonar er i svart, medan estimerte er i grått.*



Figur 4.2: *Tettleiken til Gamma(2,1)-fordeling til venstre, og asymmetrifunksjon til høgre. Utgangspunktet er 1000 observasjonar og Gamma(2,1) som kjerne i tettleiksestimatet. Teoretiske funksjonar er i svart, medan estimerte er i grått.*

4.3.1 Asymmetrifunksjonen brukt på datasett frå Gamma(2,1)

Gamma(2,1)-fordelinga er ei av dei mest asymmetriske fordelingane vi har sett på i denne oppgåva. I dette forsøket skal vi estimere tettleik med kjernane normalfordeling, epanechnikov, biweight, triweight og Gamma(2,1)-fordeling. Skeivleikskoeffisienten til Critchley og Jones (2008) er for denne fordelinga $\gamma = 0.38$, medan det tradisjonelle målet er $s = \sqrt{2} \approx 1.41$.

Dersom vi ser i tabellane 4.2 - 4.5, kan vi sjå at gjennomsnittet til $\hat{\gamma}$, og dermed også asymmetrien, blir underestimert for alle datasetta. For dei store datasetta er det ikkje så stort avvik, men dei små datasetta er asymmetrien konsekvent underestimert. Dette gjeld uansett kva kjerne vi brukar, og kan ha samanheng med at kjernen er symmetrisk. Med normalfordelinga som kjerne, er definisjonsområdet til den estimerte tettleiken heile tallina. Halane blir like tunge, og når p går mot 0, vil også $\gamma^*(p)$ også gå mot 0. Dette skil seg frå den faktiske fordelinga, der det er halane som har mest asymmetri, ettersom $\gamma^*(p)$ går mot 1 når p går mot 0. Dette vil vere ein grunn til at asymmetrien er underestimert, men burde ikkje gi så stort bidrag som det vi faktisk observerer. Kjernane epanechnikov, biweigth og triweight er definert på eit avgrensa område, og det vil føre til at også den estimerte tettleiken også er definert på eit avgrensa område. Dette gjer at vi ikkje får same problem med asymmetrien i halane, sjølv om han likevel også her vil underestimere, ettersom gammafordelinga er definert for alle positive tal. I figur 4.1 ser vi at sjølv om tettleiken er bra estimert, er asymmetrien underestimert.

Dersom vi ser på $MISE$ for asymmetri og tettleik til dei ulike kjernane i tabell 4.2-4.5, er det relativt små ulikskapar mellom normalkjerne og Beta(k,k)-kjernane. Likevel er det skilnad som viser seg i at biweight- og triweightkjernen er litt betre både på å estimere tettleik og den tilhøyrande asymmetrifunksjonen. Dette gir seg utslag i betre estimat av asymmetrikoeffisientane γ og $s_{\hat{f}}$.
I tabell 4.6 har vi prøvd med Gamma(2,1) som kjerne for å sjå korleis ei asymmetrisk fordeling estimerer ei anna asymmetrisk fordeling. Dersom vi ser på definisjonen av tettleiksestimatet, $\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K((x-X_i)/h)$, vil ein asymmetrisk kjerne fordele masse asymmetrisk rundt kvart punkt. Det er ikkje nødvendigvis dette vi ønsker, men heller at definisjonsmengda til dataene skal vere meir riktig. Ved å flytte masse i ein retning, vil det

n	$\mathrm{E}(\hat{\gamma})$	$\mathrm{Var}(\hat{\gamma})$	$MISE_{\gamma^*}$	$MISE_{\hat{f}}$	$\mathrm{E}(s_{\hat{f}})$	$\mathrm{E}(G_1)$	$\mathrm{E}(\hat{m})$
10	0.096	0.016	0.11	0.039	0.38	0.89	1.59
100	0.20	$6.6 \cdot 10^{-3}$	0.045	0.014	0.94	1.27	1.32
1000	0.27	$3.1 \cdot 10^{-3}$	0.019	$4.4 \cdot 10^{-3}$	1.23	1.39	1.17
5000	0.33	$1.9 \cdot 10^{-3}$	$8.9 \cdot 10^{-3}$	$1.9 \cdot 10^{-3}$	1.32	1.40	1.09

Tabell 4.2: *Estimerte skeivleikskoeffisientar og tilhøyrande $MISE$. Utgangspunktet er gammafordeling med parameter α=2 og $\beta = 1$. Vi har brukt normalfordelinga som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Dei teoretiske asymmetrikoeffisientane er $\gamma = 0.38$ og $s = \sqrt{2}$. Modalverdien er $m = 1$.*

n	$\mathrm{E}(\hat{\gamma})$	$\mathrm{Var}(\hat{\gamma})$	$MISE_{\gamma^*}$	$MISE_{\hat{f}}$	$\mathrm{E}(s_{\hat{f}})$	$\mathrm{E}(G_1)$	$\mathrm{E}(\hat{m})$
10	0.090	0.020	0.12	0.037	0.51	0.89	1.57
100	0.24	0.013	0.043	$9.7 \cdot 10^{-3}$	1.07	1.27	1.23
1000	0.29	0.011	0.025	$2.7 \cdot 10^{-3}$	1.30	1.39	1.13
5000	0.34	$5.2 \cdot 10^{-3}$	0.012	$1.0 \cdot 10^{-3}$	1.36	1.40	1.05

Tabell 4.3: *Estimerte skeivleikskoeffisientar og tilhøyrande $MISE$. Utgangspunktet er gammafordeling med parameter α=2 og $\beta = 1$. Vi har brukt Epanechnikov-fordelinga som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Dei teoretiske asymmetrikoeffisientane er $\gamma = 0.38$ og $s = \sqrt{2}$. Modalverdien er $m = 1$.*

n	$\mathrm{E}(\hat{\gamma})$	$\mathrm{Var}(\hat{\gamma})$	$MISE_{\gamma^*}$	$MISE_{\hat{f}}$	$\mathrm{E}(s_{\hat{f}})$	$\mathrm{E}(G_1)$	$\mathrm{E}(\hat{m})$
10	0.13	0.022	0.11	0.036	0.53	0.89	1.49
100	0.26	0.014	0.037	$9.0 \cdot 10^{-3}$	1.09	1.27	1.20
1000	0.30	0.012	0.023	$2.4 \cdot 10^{-3}$	1.31	1.39	1.11
5000	0.35	$6.2 \cdot 10^{-3}$	0.012	$9.0 \cdot 10^{-4}$	1.36	1.40	1.04

Tabell 4.4: *Estimerte skeivleikskoeffisientar og tilhøyrande $MISE$. Utgangspunktet er gammafordeling med parameter α=2 og $\beta = 1$. Vi har brukt biweight-fordelinga som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Dei teoretiske asymmetrikoeffisientane er $\gamma = 0.38$ og $s = \sqrt{2}$. Modalverdien er $m = 1$.*

n	$\mathrm{E}(\hat{\gamma})$	$\mathrm{Var}(\hat{\gamma})$	$MISE_{\gamma^*}$	$MISE_{\hat{f}}$	$\mathrm{E}(s_{\hat{f}})$	$\mathrm{E}(G_1)$	$\mathrm{E}(\hat{m})$
10	0.14	0.026	0.10	0.036	0.53	0.89	1.46
100	0.26	0.017	0.040	$8.7 \cdot 10^{-3}$	1.09	1.27	1.19
1000	0.30	0.015	0.028	$2.3 \cdot 10^{-3}$	1.31	1.39	1.11
5000	0.35	$7.7 \cdot 10^{-3}$	0.014	$8.2 \cdot 10^{-4}$	1.36	1.40	1.04

Tabell 4.5: *Estimerte skeivleikskoeffisientar og tilhøyrande $MISE$. Utgangspunktet er gammafordeling med parameter α=2 og $\beta = 1$. Vi har brukt triweight-fordelinga som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Dei teoretiske asymmetrikoeffisientane er $\gamma = 0.38$ og $s = \sqrt{2}$. Modalverdien er $m = 1$.*

n	$\mathrm{E}(\hat{\gamma})$	$\mathrm{Var}(\hat{\gamma})$	$MISE_{\gamma^*}$	$MISE_{\hat{f}}$	$\mathrm{E}(s_{\hat{f}})$	$\mathrm{E}(G_1)$	$\mathrm{E}(\hat{m})$
10	0.16	0.048	0.12	0.15	0.75	0.89	3.15
100	0.21	0.036	0.078	0.094	1.25	1.27	2.34
1000	0.24	0.011	0.037	0.053	1.39	1.39	1.87
5000	0.30	$6.7 \cdot 10^{-3}$	0.018	0.034	1.40	1.40	1.61

Tabell 4.6: *Estimerte skeivleikskoeffisientar og tilhøyrande $MISE$. Utgangspunktet er gammafordeling med parameter α=2 og $\beta = 1$. Vi har brukt Gamma(2,1) som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Dei teoretiske asymmetrikoeffisientane er $\gamma = 0.38$ og $s = \sqrt{2}$. Modalverdien er $m = 1$.*

vere stor risiko for at modalverdien til $\hat{f}$ blir estimert feil i høve til den reelle modalverdien, og dette vil påverke målet vårt i stor grad. Det er dette vi opplever når vi bruker gammafordeling som kjerne i dette dømet, og av den grunn er gammafordelinga den kjernen som er dårlegast med tanke på $MISE_f$. Riktignok er ikkje $MISE_{\gamma^*}$ og γ så mykje dårlegare estimert enn ved bruk av dei andre kjernane. Heller ikkje $s_{\hat{f}}$, men dette skulle nesten berre mangle ettersom kjernen er identisk med fordelinga datasettet kjem frå. Den kvadratiske feilen for tettleiken er derimot mykje høgare enn for dei andre kjernane, og det er dette vi må legge til grunn når konkluderer med at gammafordelinga er ei dårleg fordeling til å estimere asymmetrien til gammafordelte variable. Døme på bruk av denne kjernen kan vi sjå i figur 4.2.

Asymmetrien blir i snitt underestimert for alle kjernane og alle storleikane på datasetta. Definisjonsmengda til kjernen er påpeikt som eine forklaringa, men forklarer ikkje alt. Hovudgrunnen til underestimeringa må vere ei overestimering av modalverdien. Dette går på at teori som er utvikla for kjerneestimering er laga for å minimere $MISE$ eller $AMISE$. Dette gjer ikkje nødvendigvis $\hat{f}(x;h)$ til ein forventningsrett estimator for tettleiken. Val av bandbreidde er også noko som kan gjere at at asymmetrien vert underestimert. Dersom bandbreidda er stor, vil forma på kjernen påvirke estimatet i stor grad. Bandbreidda vi har føreslått i avsnitt 3.4.1, er ein øvre skranke for $AMISE$-bandbreidda, og det er mogeleg å velje ei bandbreidde som kan gi lågare $MISE$. Likevel er vi avhengig av at $\hat{f}$ er eintoppa, og det er vanskeleg å forsvare ei bandbreidde som ofte ikkje oppfyller dette kravet. Det fins betre estimatorar for modalverdi (Bickel og Frühwirth 2006), men dette er vanskeleg å inkorporere i kjerneestimatet. For gammafordelinga får vi relativt sjeldan store observasjonar, og dersom vi får ein stor observasjon i eit lite datasett, vil dip-testen klassifisere datasettet som fleirtoppa. Summen av dette gjer at asymmetrifunksjonen blir underestimert. Vi legg merke til at estimatoren for skeivleikskoeffisienten s, også underestimerer skeivleiken.

Dersom vi ser på korleis storleiken på datasettet påverkar det nye målet for asymmetri, kan vi sjå frå tabellane 4.2 - 4.5, at skeivleikskoeffisienten kryp nærmare og nærmare teoretisk verdi jo større datasettet blir. Dette kan tyde at vi treng store datasett for å kunne fastslå om data kjem frå ei asymmetrisk fordeling. Dersom vi ser på datasetta på 10 observasjonar, er $MISE_{\gamma^*}$ på ca 0.1. Dette er kvadratisk feil og absoluttfeil vil difor ligge på 0.3. Funksjonsverdien er i snitt ca 0.38, og feilen er difor av same storleik som funksjonen. På datasetta med 5000 observasjonar tar $MISE_{\gamma^*}$ verdiar på rundt 0.01. Dette er rundt 0.1 målt i absolutt feil, og er omkring ein fjerdedel av funksjonsverdien. Her ser vi at sjølv for store datasett, er det eit relativt stort kvadratisk avvik for asymmetrifunksjonen. Denne feilen avtar heller ikkje fort saman med storleiken på datasettet. Vi estimerte i tillegg asymmetrien for datasett på 100000 observasjonar med normalfordelinga som kjerne. Her var E($\hat{\gamma}$)=0.36 med varians Var($\hat{\gamma}$) $= 8.2 \cdot 10^{-4}$, $MISE_{\gamma^*} = 4.8 \cdot 10^{-3}$ og $MISE_{\hat{f}} = 3.8 \cdot 10^{-4}$. Med andre ord var det frameleis avvik for svært store datasett, sjølv om asymmetrikoeffisienten byrja å konvergere mot riktig verdi.

For Gamma(2,1)-fordelinga vart asymmetrien underestimert, særskild for små datasett. I tillegg til dette var variansen til $\hat{\gamma}$ relativt stor. For dei minste datasetta låg standardavviket på ca 0.15, og dette er for høgt dersom forventningen til $\hat{\gamma}$ ikkje er høgare enn 0.1. For datasetta på 100 og 1000, var standardavviket 0.1, medan forventningen til $\hat{\gamma}$ var mellom 0.2 og 0.3. I desse tilfella kan vi konkludere med at datasettet kjem frå ei asymmetrisk fordeling. Frå tabell 4.1 kan vi sjå at Gamma(2,1) er ei vanskeleg fordeling å estimere og dette kan gi håp om at vi får betre resultat med andre fordelingar, sjøv om dei i mindre grad er asymmetriske.

4.3.2 Polynomen av ulik grad

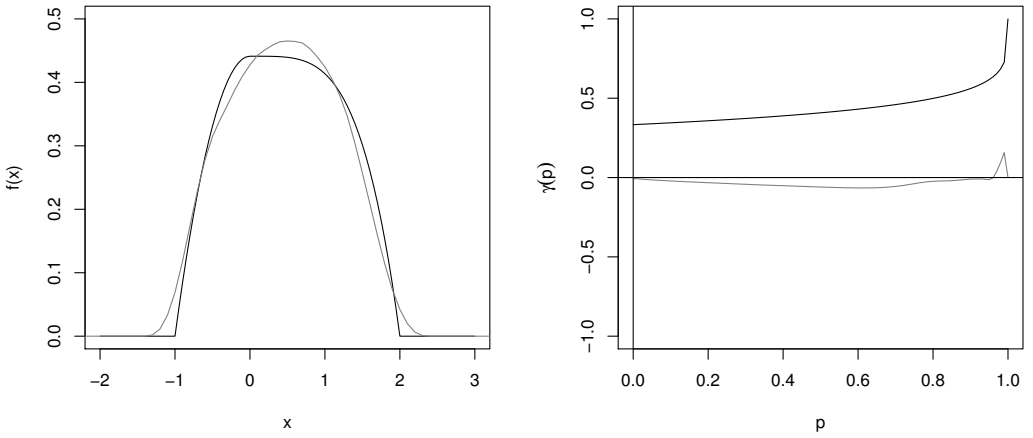
Ei av dei fyrste fordelinga vi såg på i kapittel 1 var ei fordeling sett saman av polynomen av ulik grad,

$$f(x) = \begin{cases} -ax^2 + 1\,, & -\alpha < x \leq 0 \\ -bx^4 + 1\,, & 0 < x < \beta\,. \end{cases} \tag{4.2}$$

Vi skal no analysere estimat for denne tettleiken, der vi bruker $\alpha = 1$ og $\beta = 2$. Med desse vala for parametra, blir $\gamma = 0.43$ og $s = 0.044$. Dei to måla klassifiserer dermed asymmetrien ulikt. Vidare skal vi bruke normal-, epanechnikov-, biweight- og triweightfordelinga som kjerne på same måte som i avsnitt 4.3.1. I utgangspunktet kan vi sjå for oss at kjernane som er definert

på eit avgrensa område er best, ettersom fordelinga også er definert på eit avgrensa område. Vi vil også tru at vi får betre resultat enn for Gamma(2,1) ettersom fordelinga er enklare å estimere, tabell 4.1. Dette vil i fyrste omgang gje seg utslag i ein låg $MISE_{\hat{f}}$.

Vi byrjar med å sjå på $MISE_{\hat{f}}$ i tabellane 4.7-4.10. Alle kjernane er omkring like bra, og $MISE_{\hat{f}}$ er omkring den same for både for alle kjernane, men også samanlikna med $MISE_{\hat{f}}$ for Gamma(2,1)-fordelinga. Vi ser at alle veridane for $MISE_{\hat{f}}$ er relativt gode, med veridar som ligg ca 20 prosent feil i høve til funksjonsverdien i m for dei minste datasetta og under 3 prosent for dei største setta. Likevel bommar estimatoren $\hat{\gamma}^*(p)$ grovt på asymmetrifunksjonen. Denne tettleiken er av det nye målet klassifisert som svært



Figur 4.3: *Tettleiken til fordeling sett saman av polynomen til venstre, og asymmetrifunksjon til høgre. Utgangspunktet er 1000 observasjonar og triweightfordelinga som kjerne i tettleiksestimatet. Teoretiske funksjonar er i svart, medan estimerte er i grått.*

n	$\mathrm{E}(\hat{\gamma})$	$\mathrm{Var}(\hat{\gamma})$	$MISE_{\gamma^*}$	$MISE_{\hat{f}}$	$\mathrm{E}(s_{\hat{f}})$	$\mathrm{E}(G_1)$	$\mathrm{E}(\hat{m})$
10	$1.4 \cdot 10^{-3}$	0.036	0.24	0.035	0.018	0.042	0.47
100	$-1.1 \cdot 10^{-3}$	0.071	0.28	0.010	0.043	0.058	0.47
1000	0.018	0.045	0.24	$2.2 \cdot 10^{-3}$	0.041	0.046	0.44
5000	$9.2 \cdot 10^{-3}$	0.047	0.25	$8.0 \cdot 10^{-4}$	0.041	0.044	0.45

Tabell 4.7: *Estimerte skeivleikskoeffisientar og tilhøyrande $MISE$. Utgangspunktet er fordeling sett saman av andregrads- og fjerdegradspolynomen med parameter $\alpha = 1$ og $\beta = 2$. Vi har brukt normalfordelinga som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Dei teoretiske asymmetrikoeffisientane er $\gamma = 0.43$ og $s = 0.044$. Modalverdien er $m = 0$.*

n	$\mathrm{E}(\hat{\gamma})$	$\mathrm{Var}(\hat{\gamma})$	$MISE_{\gamma^*}$	$MISE_{\hat{f}}$	$\mathrm{E}(s_{\hat{f}})$	$\mathrm{E}(G_1)$	$\mathrm{E}(\hat{m})$
10	-0.012	0.046	0.27	0.042	0.024	0.042	0.50
100	-0.084	0.081	0.38	0.011	0.049	0.058	0.58
1000	-0.023	0.051	0.29	$1.9 \cdot 10^{-3}$	0.043	0.046	0.48
5000	-0.023	0.035	0.27	$6.4 \cdot 10^{-4}$	0.043	0.044	0.48

Tabell 4.8: *Estimerte skeivleikskoeffisientar og tilhøyrande $MISE$. Utgangspunktet er fordeling sett saman av andregrads- og fjerdegradspolynomen med parameter $\alpha = 1$ og $\beta = 2$. Vi har brukt Epanechnikov-fordelinga som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Dei teoretiske asymmetrikoeffisientane er $\gamma = 0.43$ og $s = 0.044$. Modalverdien er $m = 0$.*

n	$\mathrm{E}(\hat{\gamma})$	$\mathrm{Var}(\hat{\gamma})$	$MISE_{\gamma^*}$	$MISE_{\hat{f}}$	$\mathrm{E}(s_{\hat{f}})$	$\mathrm{E}(G_1)$	$\mathrm{E}(\hat{m})$
10	-0.084	0.061	0.36	0.045	0.025	0.042	0.60
100	-0.10	0.083	0.40	0.011	0.050	0.058	0.60
1000	-0.013	0.051	0.28	$1.9 \cdot 10^{-3}$	0.043	0.046	0.47
5000	-0.022	0.037	0.27	$6.4 \cdot 10^{-4}$	0.043	0.044	0.48

Tabell 4.9: *Estimerte skeivleikskoeffisientar og tilhøyrande $MISE$. Utgangspunktet er fordeling sett saman av andregrads- og fjerdegradspolynomen med parameter $\alpha = 1$ og $\beta = 2$. Vi har brukt biweight-fordelinga som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Dei teoretiske asymmetrikoeffisientane er $\gamma = 0.43$ og $s = 0.044$. Modalverdien er $m = 0$.*

n	$\mathrm{E}(\hat{\gamma})$	$\mathrm{Var}(\hat{\gamma})$	$MISE_{\gamma^*}$	$MISE_{\hat{f}}$	$\mathrm{E}(s_{\hat{f}})$	$\mathrm{E}(G_1)$	$\mathrm{E}(\hat{m})$
10	-0.091	0.071	0.37	0.048	0.025	0.042	0.62
100	-0.13	0.085	0.44	0.012	0.051	0.058	0.64
1000	$-7.7 \cdot 10^{-3}$	0.049	0.27	$1.9 \cdot 10^{-3}$	0.044	0.046	0.47
5000	-0.023	0.034	0.27	$6.5 \cdot 10^{-4}$	0.043	0.044	0.48

Tabell 4.10: *Estimerte skeivleikskoeffisientar og tilhøyrande $MISE$. Utgangspunktet er fordeling sett saman av andregrads- og fjerdegradspolynomen med parameter $\alpha = 1$ og $\beta = 2$. Vi har brukt triweight-fordelinga som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Dei teoretiske asymmetrikoeffisientane er $\gamma = 0.43$ og $s = 0.044$. Modalverdien er $m = 0$.*

asymmetrisk, og likevel klarar ikkje estimatoren å fange opp asymmetri. I enkelte av tilfella er $MISE_{\gamma^*}$ oppe i 0.4, og er av same storleik som snittet til asymmetrifunksjonen. Det tyder at estimatoren i praksis ikkje klarar å fange opp noko av asymmetrien til den faktiske fordelinga. Det er heller ikkje noko som tyder på konvergens, og det er mest sannsynleg naudsynt med eit svært stort datasett for å få godkjente resultat.

Grunnen til dette må vere den flate toppen fordelinga har, og døme på estimert tettleik kan vi sjå i figur 4.3. Med ein svært låg derivert omkring modalverdien, blir denne vanskeleg å estimere. Dette ser vi også ettersom estimatet for modalverdi i tabellane 4.7-4.10 er overestimert konsekvent. Med mange observasjonar i området (0,1) vil eit kjerneestimat med høg bandbreidde flytte masse til høgre, ettersom kjernen er symmetrisk. Det vil ikkje nødvendigvis nødvendigvis hjelpe å senke bandbreidda, ettersom den estimerte tettleiken då gjerne får fleire toppar. På grunn av dette, er ikkje målet som blir brukt for å sjå kor vanskeleg det er å estimere ei tettleik, eit godt mål for å måle kor vanskeleg det er å estimere asymmetrifunksjonen. Eit mål på dette bør kanskje heller ta utgangspunkt i den deriverte til asymmetrifunksjonen, og då særskild for p nær ein, dvs omkring modalverdien.

Estimatet for den tradisjonelle skeivleikskoeffisienten fungerer i dette tilfellet mykje betre. Både $s_{\hat{f}}$ og særskild G_1 er gode estimatorar. Grunnen til dette er at $\bar{X}$ er ein god estimator for gjennomsnittet, og at sjølv om kjerneestimatet bommar på modalverdien, gjer ikkje det seg så stort utslag for $\mathrm{E}_{\hat{f}}(X)$. Dette er eit døme på at maksimum av tettleiksestimatet ikkje alltid er ein veldig god estimator for modalverdien.

4.3.3 Gamma(10,1)

Den tredje fordelinga vi skal sjå på er Gamma(10,1)-fordelinga. Denne fordelinga er meir symmetrisk enn Gamma(2,1), og har ein meir markert topp enn fordelinga sett saman av polynomen. Vi har vald å berre bruke biweight-kjernen i dette forsøket, då det er den kjernen som er best hittil, sjølv om alle dei symmetriske kjernane har gitt omkring like gode resultat.

Vi byrjar med å sjå på den integrerte verdien til asymmetrifunksjonen. Til motsetnad frå Gamma(2,1)-fordelinga, ser det ikkje ut som om målet konsekvent er underestimert. Dersom vi ser i tabell 4.11, ser det tilsynelatande ut som om γ konvergerer forholdsvis raskt mot den riktige verdien. Dersom vi ser på $MISE_{\gamma^*}$, kan vi sjå at det i høve til Gamma(2,1) er mykje betre veridar for dei små datasetta. For dei store setta er estimata omkring like bra,

n	$\mathrm{E}(\hat{\gamma})$	$\mathrm{Var}(\hat{\gamma})$	$MISE_{\gamma^*}$	$MISE_{\hat{f}}$	$\mathrm{E}(s_{\hat{f}})$	$\mathrm{E}(G_1)$	$\mathrm{E}(\hat{m})$
10	0.21	0.032	0.062	0.013	0.27	0.47	8.49
100	0.13	0.018	0.033	$2.1 \cdot 10^{-3}$	0.54	0.63	9.12
1000	0.14	0.018	0.027	$3.9 \cdot 10^{-4}$	0.60	0.63	9.01
5000	0.14	0.011	0.018	$1.2 \cdot 10^{-4}$	0.61	0.63	9.01

Tabell 4.11: *Estimerte skeivleikskoeffisientar og tilhøyrande $MISE$. Utgangspunktet er gammafordeling med parameter $\alpha = 10$ og $\beta = 1$. Vi har brukt biweight-fordelinga som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Dei teoretiske asymmetrikoeffisientane er $\gamma = 0.14$ og $s = \sqrt{(2/5)} \approx 0.63$. Modalverdien er $m = 9$.*

n	$\mathrm{E}(\hat{\gamma})$	$\mathrm{Var}(\hat{\gamma})$	$MISE_{\gamma^*}$	$MISE_{\hat{f}}$	$\mathrm{E}(s_{\hat{f}})$	$\mathrm{E}(G_1)$	$\mathrm{E}(\hat{m})$
10	0.0063	0.040	0.054	0.045	-0.0030	-0.0053	0.022
100	-0.021	0.034	0.043	$6.1 \cdot 10^{-3}$	-0.035	-0.041	0.037
1000	0.011	0.019	0.025	$1.1 \cdot 10^{-3}$	0.0012	-0.0013	-0.011
5000	-0.011	0.013	0.019	$3.4 \cdot 10^{-4}$	0.0012	0.0012	0.010

Tabell 4.12: *Estimerte skeivleikskoeffisientar og tilhøyrande $MISE$. Utgangspunktet er standard normalfordeling. Vi har brukt biweight-fordelinga som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Dei teoretiske asymmetrikoeffisientane er $\gamma = 0$ og $s = 0$. Modalverdien er $m = 0$.*

men her var estimatoren relativt god for Gamma(2,1) også. Kvadratisk feil for tettleiken er også liten, sjølv om skeivleikskoeffisienten til den estimerte tettleiken ikkje treff heilt for dei minste datasetta.

Grunnen til at estimatoran for asymmetri er såpass mykje betre for denne fordelinga, kan skuldast ulike årsaker. For det fyrste er fordelinga, med ei vanskegrad på $D(f) = 0.804$, ei relativ enkel å fordeling å estimere. For det andre er det ein relativt markert topp for modalverdien $m = 9$, og denne blir også godt estimert. Den siste avgjerande grunnen til at fordelinga er bra estimert sett i høve til ein asymmetrifunksjon, må vere at observasjonane som gjer fordelinga asymmetrisk kjem relativt ofte, og at dip-testen samstundes klassifiserer settet som eintoppa. Dette var ein av grunnane til at asymmetrien vart underestimert for Gamma(2,1), men det gjer seg ikkje gjeldande i like stor grad for dette datasettet som er meir symmetrisk. Konklusjonen blir difor at asymmetrifunksjonen vart godt estimert for denne fordelinga.

4.3.4 Normalfordelinga

Den siste fordelinga vi skal sjå på er normalfordelinga. Vi har også her berre sett på estimat laga av biweight-kjernen. Grunnen til at vi har vald denne fordelinga, er at ho er symmetrisk, og vi vil samanlikne korleis estimatet blir i høve til asymmetriske fordelingar.

I tabell 4.12 ser vi at estimatet for γ blir omkring null slik det skal vere. Som vi har sett tidlegare konvergerer ikkje denne estimatoren svært fort, men vi kan likevel få ein indikasjon på at fordelinga er symmetrisk. Den kvadratiske feilen for asymmetrifunksjonen $\gamma^*(p)$, er omkring som for dei andre fordelingane. Det kan difor vere nærliggande å tru at denne feilen i stor grad er absolutt, og ikkje relativ til asymmetrifunksjonen. Den kvadratiske feilen til tettleiken er også omkring den same som for dei andre fordelingane, og vi må nok bruke ein meir tilpassa bandbreidde enn den øvre skranken for h_{AMISE} for å få eit betre resultat.

4.4 Dip-testen

Dei kritiske verdiane for dip-testen er rekna ut for uniform fordeling. Dette er den minst eintoppa fordelinga som framleis er eintoppa. Grunnen til at vi vel ei slik fordeling, er at vi då minimerer sannsynet for type 1-feil, å klassifisere eit datasett som fleirtoppa sjølv om det eigentleg er eintoppa. I alle simuleringane våre har vi brukt eit signifikansnivå på fem prosent for dip-testen, men sannsynet for type 1-feil er mindre enn det, ettersom datsetta ikkje kjem frå uniform fordeling. I tabell 4.13 har vi simulert 100000 datasett av kvar størrelse, og deretter sett kor mange av dei som blir klassifisert som fleirtoppa. Ikkje overraskande er det fordelinga sett saman av polynomen av ulik grad som har flest type 1-feil. Dette kan forklarast med at denne fordelinga er den som er mest lik den uniforme fordelinga.

Fordeling	10	100	1000	5000
Gamma(2,1)	0.01851	0.00151	0.00001	0.00000
Polynomen av ulik grad	0.03126	0.01097	0.00165	0.00037
Gamma(10,1)	0.01555	0.00128	0.00002	0.00000
Normalfordelinga	0.01448	0.00113	0.00001	0.00000

Tabell 4.13: *Sannsynet for type 1-feil, det vil seie å forkaste at datasettet er eintoppa. Signifikansnivået er fem prosent, og vi har simulert 100000 datasett.*

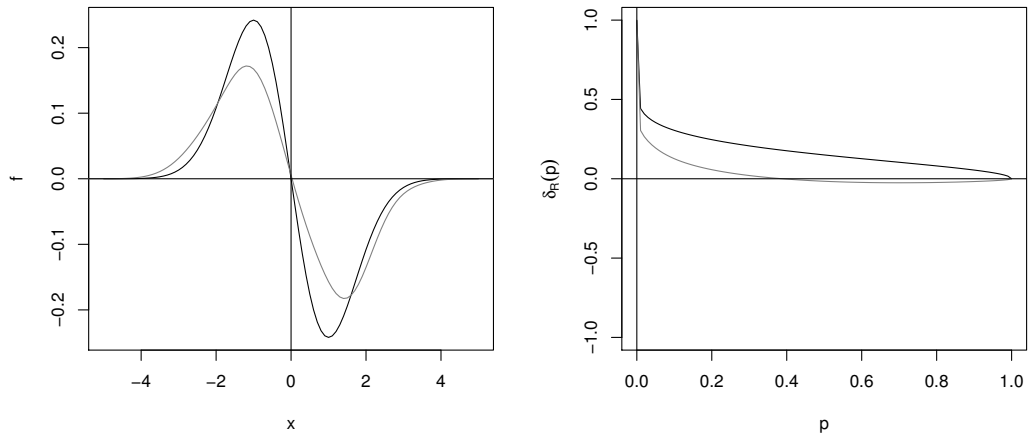
4.5 Estimat av gradientasymmetrien

I dette avsittet skal vi sjå på estimat av gradientasymmetrien til normalfordelinga og fordelinga sett saman av polynomen og eksponensialfunksjon i døme 2.6. Vi skal bruke estimat av den deriverte til å rekne ut gradientasymmetrien. På same måte som tidlegare simulerer vi datasett på 10, 100, 1000 og 5000 observasjonar, og vi har 100 datasett av kvar storleik.

For asymmetrien var biweight- og triweight-kjernen dei beste kjernane, og dette har samanheng med $MISE$-eigenskapane deira (Wand og Jones 1995). Som nemnd tidlegare er $MISE$-kriteriet eit bra mål for estimering av tettleiken, men det er dårlegare når vi skal overføre det til asymmetrifunksjonen. Dette vart tydeleg når vi estimerte den deriverte, ettersom både monotoni- og krummingseigenskapane til tettleiken er viktige når vi skal rekne ut gradientasymmetrien. På grunn av dette fungerte biweight- og triweight-kjernen dårleg, og i begge forsøka har vi difor brukt normalfordelinga som kjerne. På same måte som i avsnitt 4.3 har vi brukt $\mathrm{E}(\cdot)$ og $\mathrm{Var}(\cdot)$ til å uttrykke empirisk gjennomsnitt og varians.

4.5.1 Normalfordelinga

For normalfordelinga er høgre- og venstrekurtosen lik, og vi ser difor berre på estimat for høgrekurtosen. Av tabell 4.14 kan vi lese at forventningen til $\hat{\delta}_R$ konvergerer mot den riktige verdien når datasettet blir stort. Samstundes er variansen til $\hat{\delta}_R$ så stor, at sjølv for dei store datasetta er konfidensintervallet



Figur 4.4: *Den deriverte til normalfordelinga med til venstre, og gradientasymmetrien til høgre. Dei teoretiske funksjonane er i svart, medan estimerte i grått.*

n	$\mathrm{E}(\hat{\delta}_R)$	$\mathrm{Var}(\hat{\delta}_R)$	$MISE_{\delta_R^*}$	$MISE_{\hat{f}'}$	$\mathrm{E}(G_2)$	$\mathrm{E}(\hat{m})$	$\mathrm{E}(\hat{\pi}_R)$
10	0.12	0.027	0.035	0.081	4.35	0.015	1.30
100	0.16	0.026	0.032	0.015	3.16	-0.003	1.12
1000	0.17	0.020	0.023	0.0045	3.02	-0.0077	1.06
5000	0.18	0.015	0.018	0.0019	3.01	0.0016	1.03

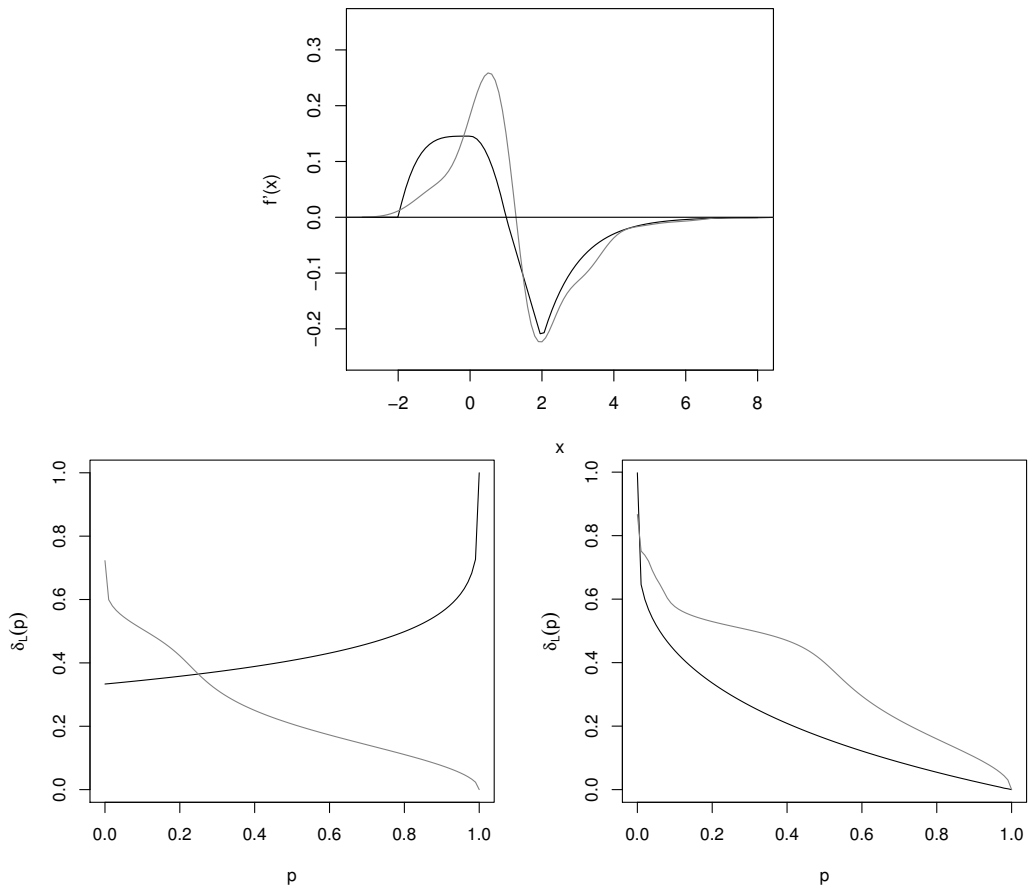
Tabell 4.14: *Estimerte kurtosekoeffisientar og tilhøyrande $MISE$. Utgangspunktet er standard normalfordeling. Vi har brukt normalfordelinga som kjerne. Den teoretiske høgre gradientasymmetrikoeffisienten er $\delta_R \approx 0.17$ og $k = 3$. Modalverdien er $m = 0$ og vendepunktet er $\pi_R = 1$.*

for den estimerte verdien relativt stort. Dette kjem også til uttrykk i $MISE_{\delta_R^*}$ som for dei store datasetta er omkring 0.02. Dette blir omkring 0.15 i absolutt feil, og er med det same storleik som gjennomsnittet til funksjonen. Den deriverte til tettleiken er betre estimert, og med $MISE$-verdiar på 10^{-3} for dei største datasetta, er den absolutte feilen omkring 0.05. Då vi estimerte asymmetrien, var eit godt estimat av modalverdien viktig for å få eit godt resultat. I denne samanhengen blir det derimot viktig å få eit godt estimat av vendepunktet, og det er her det ser ut til at feilen ligg. Vi kan sjå at estimatet for vendepunktet er for høgt for alle storleikane på datasetta, og dette gjer at gradientasymmetrien blir estimert for høgt. Dette ser vi også i figur 4.4.

4.5.2 Fordeling av polynomen og eksponensialfunksjone

For fordelinga vi såg i døme 2.6 er både høgre- og venstrkurtosen definert, men ulik. For venstre del blir gradientasymmetrien den same som asymmetrien i avsnitt 4.3.2. Ein kunne tenke seg at dei same problema som ved estimeringa av asymmetrien også skulle oppstå her. Dette er ikkje tilfellet. I asymmetrien vart modalverien estimert feil, men her er det vendepunktet som må estimerast riktig. Desse variablane har ulike statistiske eigenskapar. Modalverdien vil vere der observasjonane kjem tettast, medan vendepunktet vil vere der endringa i kor tett observasjonane kjem er størst. Det er truleg at monotonieigenskapane til kjernen vil påvirke estimat for vendepunkta.

Dersom vi ser i tabell 4.16, kan vi sjå at vendepunktet for venstre del er estimert høgare enn den reelle verdien, då det for dei største datasetta er estimert til ca 0.5, medan det eigentleg skal vere 0. Dette er motsett av kva vi opplevde med asymmetrien. Vidare ser vi at også modalverdien er estimert for høg, og dette kan vege opp for det feilestimerte vendepunktet. Likevel ser vi i tabell 4.15 at estimatet for δ_L bommar grovt på den reelle verdien.



Figur 4.5: *Den deriverte til fordelinga av polynomen og eksponensialfunksjon med tilhøyrande estimat øverst. Venstre gradientasymmetrien til same fordeling nede til venstre, og høgre gradientasymmetri nede til høgre. Dei teoretiske funksjonane er i svart, medan estimerte i grått.*

Gjennomsnittet til δ_L skil seg 50 prosent frå den riktige verdien, sjølv for dei store datasetta. For den høgre del av fordelinga, kan det sjå ut som om gradientasymmetrien blir estimert betre. Likevel konvergerer ikkje $\hat{\delta}_R$ mot riktig verdi når datasettet blir stort. Faktisk ser det ut som om feilen aukar med store datasett. Dette kan ha med at den andrederiverte til tettleiken ikkje er kontinuerleg i $x = 2$. I figur 4.5 ser vi at den høgre del av den deriverte er bra estimert, men estimatet for den venstre delen er mindre bra. Dersom vi ser på G_2, bommar også denne estimatoren ganske mykje på den teoretiske verdien. Det kan tyde på at kurtose er vanskeleg å estimere uansett kva mål eller estimator vi brukar.

n	$\mathrm{E}(\hat{\delta}_R)$	$\mathrm{Var}(\hat{\delta}_R)$	$MISE_{\delta_R^*}$	$\mathrm{E}(\hat{\delta}_L)$	$\mathrm{Var}(\hat{\delta}_L)$	$MISE_{\delta_L^*}$
10	0.16	0.027	0.043	0.13	$2.7 \cdot 10^{-2}$	0.15
100	0.22	0.011	0.015	0.20	$2.8 \cdot 10^{-3}$	0.099
1000	0.29	0.0052	0.015	0.22	$8.3 \cdot 10^{-4}$	0.095
5000	0.31	0.0029	0.016	0.24	$5.6 \cdot 10^{-4}$	0.092

Tabell 4.15: *Estimerte kurtosekoeffisientar og tilhøyrande $MISE$. Utgangspunktet er fordeling sett saman av polynomen og eksponensialfunksjon. Vi har brukt normalfordelinga som kjerne. Den teoretiske høgre gradientasymmetrikoeffisienten er $\delta_R \approx 0.20$ og den teoretiske venstre gradientasymmetrikoeffisienten er $\delta_L \approx 0.43$.*

n	$MISE_{f'}$	$\mathrm{E}(G_2)$	$\mathrm{E}(\hat{m})$	$\mathrm{E}(\hat{\pi}_R)$	$\mathrm{E}(\hat{\pi}_L)$
10	0.072	5.24	1.37	2.52	0.076
100	0.021	6.29	1.27	2.27	0.27
1000	0.029	6.62	1.22	2.04	0.42
5000	0.042	6.87	1.19	1.94	0.49

Tabell 4.16: *Estimerte verdiar for den deriverte til tettleiken sett saman av polynomen og eksponensialfunksjon. Kurtosekoeffisienten er $k = 1.8$, modalverdien er $m = 1$ og vendepunkta er $\pi_R = 2$ og $\pi_L = 0$.*

4.6 Samandrag

I dette kapittelet brukte vi teorien vi har sett på i dei føregåande kapitla til å estimere asymmetrifunksjonane og gradientasymmetrifunksjonane. Vi såg at asymmetrifunksjonen vart bra estimert for data frå Gamma(10,1)-fordelinga og normalfordelinga. For Gamma(2,1)-fordelinga, var estimatoren middels god, men for fordelinga sett saman av polynomen fungerte estimatoren for asymmetri mindre bra. Hovudgrunnen til dette var at modalverdien vart feilestimert.

Gradientasymmetrien var vanskelegare å estimere. Estimatet for gradientasymmetri var for normalfordelinga middels godt. For fordelinga sett saman av polynomen og eksponensialfunksjon var derimot estimatet mindre godt, sjølv om estimat for den deriverte ikkje hadde svært høge $MISE$-verdiar.

Kapittel 5

Vidare arbeid og konklusjon

I denne oppgåva har vi analysert og brukt dei nye måla for asymmetri og kurtose. I dette kapittelet presenterer vi ei oppsummering av måla, og føreslå vidare arbeid. Vi byrjar med å sjå på eit alternativt forslag til tettleiksestimering, før vi ser på testing av asymmetrien. Asymmetrien til Critchley og Jones (2008) er definert for ein avgrensa klasse av funksjonar, og vi føreslår difor å utvide definisjonen til å gjelde for fleire typar fordelingar. Vi presenterer også mål for asymmetri og kurtose vi ikkje har sett på i denne oppgåva. Til slutt kjem vi med ein konklusjon på arbeidet.

5.1 Differansen mellom ordna observasjonar

I kapittel 3 såg vi på teori for tettleiksestimering, og brukte dette på asymmetrifunksjonen. Her var konklusjonen at vi måtte ha relativt store datasett for å få gode estimat. Eit problem som oppstår ved store datasett, er at estimatet krev stor reknekraft. Den estimerte tettleiken er ein sum av funksjonar som er like stort som datasettet, og er dermed eit stort objekt som er vanskeleg å handtere. Når vi skal rekne ut asymmetrifunksjonen, må vi i finne inversen til denne funksjonen. Dette er ein prosess som er krevjande.

I avsnitt 4.3.1 vart det henvist til ein estimert modalverdi som er betre enn maksimum av den estimerte tettleiken. Denne estimatoren bruker avstanden mellom observasjonane til å estimere modalverdien. Dersom vi har n observasjonar, bruker vi den verdien av i som gjer intervallet $(X^{(i)}, X^{(i+\frac{n}{2})})$ minst mogeleg. Denne prosedyren gjentar vi til vi står att med to observasjonar, og den estimerte modalverdien blir snittet av desse to observasjonane (Bickel og Frühwirth 2006). Ut frå denne ideen formulerer vi teoremet:

Teorem 5.1.1. *La $X_1, X_2, ..., X_n$ vere n i.i.d. variable med kontinuerleg sannsynstettleik f. La deretter $D_i = X^{(i+k)} - X^{(i-k)}$, der k er eit heiltal mindre enn $\frac{n}{2}$, og i kan ta verdiar mellom k og $n-k$. Til slutt lar vi $\hat{f}(X^{(i)}) = (\sum_{j=1}^{n-k} D_j)^{-1} \frac{1}{D_i}$. Då vil $\lim_{n\to\infty} \hat{f}(X^{(i)}) = f(X^{(i)})$.*

Vi har ikkje prov for dette teoremet, men simuleringar frå svært store datasett antyder likevel at det er riktig. Den estimerte tettleiken konvergerer sakte mot den riktige tettleiken, under alle høve med tanke på asymmetri. Dette gjer at ein metode som bygger på denne teorien ikkje er veldig føremålsteneleg med tanke på å estimere tettleikar. For å estimere asymmetrifunksjonen kan han likevel vere god.

Teorem 5.1.2. *La $X_1, X_2, ..., X_n$ vere n i.i.d. variable med kontinuerleg sannsynstettleik f. La deretter $D_i = X^{(i+k)} - X^{(i-k)}$, der k er eit heiltal mindre enn $\frac{n}{2}$, og i kan ta verdiar mellom k og $n-k$. Vidare definerer vi*

$$i_L = \sup(i : \frac{1}{D_i} < p\frac{1}{D_j}\ , i < j)\ , \tag{5.1}$$

$$i_R = \inf(i : \frac{1}{D_i} < p\frac{1}{D_j}\ , i > j)\ . \tag{5.2}$$

Og til slutt har vi observatorane

$$\hat{\tau}_R(p) = X^{(i_L)} - m\ , \tag{5.3}$$

$$\hat{\tau}_L(p) = m - X^{(i_R)}\ , \tag{5.4}$$

der m er modalverdien til datasettet. Då vil $\lim_{n\to\infty} \hat{\tau}_R(p) = \tau_R(p)$ og $\lim_{n\to\infty} \hat{\tau}_L(p) = \tau_L(p)$.

Dette teoremet har vi heller ikkje prova, men også her har vi prøvd teorien med store datasett. Ut frå teorem 5.1.2 kan vi lage ein metode for å estimere asymmetrifunksjonen. Eit estimat som bygger på denne metoden treng berre å ordne datasettet, finne differansane mellom dei ordna observasjonane og indeksere desse differansane. Difor vil metoden krevje langt mindre reknekraft enn metoden som bygger på tettleiksestimat. Metoden konvergerer til gjengjeld sakte og krev svært mange observasjonar for å fungere bra. I denne oppgåva har vi ikkje prov for dei to teorema, og vi overlet forsøk på det til vidare arbeid.

5.2 Testing om fordeling er asymmetrisk

I kapittel 3 såg vi korleis den empiriske skeivleikskoeffisienten, g_1, er definert. Observatoren $ng_1/6$, der n er talet på observasjonar, er asymptotisk

χ^2-fordelt (Holgersson 2007). Ut frå dette er det mogeleg an å lage testar for skeivleiken, men denne testen er diverre ikkje god på alle datasett. Til dømes vil testen klassifisere symmetriske data med ulik varians som asymmetrisk (Holgersson 2007). Han vil også klassifisere asymmetriske data med skeivleik null som symmetriske. Av denne grunn vil det vere nyttig å konstruere andre testar for testing av skeivleik. I kapittel 4 såg vi at skeivleikskoeffisienten til Critchley og Jones (2008) ofte vart underestimert. Likevel var datasettet tydeleg asymmetrisk dei gongene observatoren gav utslag for skeivleik. På grunnlag av dette er det mogeleg å lage ein test som er robust mot type 1-feil. Det vil seie at sannsynet er lite for å forkaste eit datasett som symmetrisk, sjølv om det i røynda skulle vere det. Diverre er ikkje testen god med tanke på type 2-feil, å halde på ein feilaktig nullhypotese om at datasettet er symmetrisk. Ein av grunnane til dette kan vere at sjølv om datasettet er asymmetrisk, kan skeivleikskoeffisienten til Critchley og Jones (2008) vere lik null. Ein måte ein kan løyse det på vil vere å lage ein ny skeivleikskoeffisient basert på målet til Critchley og Jones (2008).

Definisjon 5.2.1. *Vi definerer den kvadratiske skeivleiken som*

$$\gamma_2 = \int_0^1 \gamma^*(p)^2 \, \mathrm{d}p \, . \tag{5.5}$$

Eit av argumenta mot tradisjonell skeivleik, er at han ikkje fangar opp asymmetri dersom skeivleikskoeffisienten er lik null. Dette er ein veikskap til skeivleikskoeffisienten til Critchley og Jones (2008) også, men den kvadratiske skeivleiken, γ_2, vil vere null viss og berre viss fordelinga er symmetrisk. Å undersøke mål og lage testar på dette overlet vi til vidare arbeid.

5.3 Utviding av definisjonen av asymmetri

Critchley og Jones (2008) definerer asymmetrien for eintoppa, rota fordelingar. Dei føreslår likevel ein definisjon for eintoppa fordelingar som ikkje er rota. I det tilfellet er definisjonen for asymmetri den same, men definisjonsmengda for p blir annleis. Dersom tettleiken er eintoppa og definert på (a, b), er asymmetrifunksjonen definert for

$$p \in \left[\max\left(\frac{f(a)}{f(m)}, \frac{f(b)}{f(m)} \right), 1 \right] \, .$$

På same måte kan vi utvide definisjonen til å gjelde for tettleikar som er definert på eit avgrensa intervall, er strengt synkande fram til eit botnpunkt

og som deretter strengt veksande. Desse fordelingane har eit eintydig botnpunkt, og det vil vere mogeleg å definere asymmetrien ut frå botnpunktet istadenfor modalverdien. Vidare utvida vi definisjonen i avsnitt 3.3.3, og denne utvidinga er det mogeleg å arbeide vidare med.

5.4 Fleire mål for asymmetri og kurtose

I denne oppgåva har vi sett på mål Critchley og Jones (2008) innfører for å skildre utsjånaden til ei fordeling. Vi har også sett på dei tradisjonelle måla for asymmetri og kurtose, og vi har føreslått mål sjølve. Dette er ikkje dei einaste måla som blir brukt til å skildre forma til ei fordeling, og

$$S_k = \frac{\mu - M}{\sigma} , \tag{5.6}$$

$$b = \frac{Q_3 + Q_1 - 2m}{Q_3 - Q_1} , \tag{5.7}$$

$$\gamma'_m = \frac{\mu - m}{\sigma} , \tag{5.8}$$

er alle vanlege mål (Arnold og Groeneveld 1995). Her står μ for forventningsverdi, σ standardavvik, m modalverdi, M median og Q for kvartilar. Det kan vere interessant å også undersøke desse måla med tanke på eigenskapar som teoretiske mål, men også med tanke på kor godt estimatorar for desse måla fungerer.

5.5 Konklusjon

Critchley og Jones (2008) argumenterer for at asymmetrien til ei fordeling ikkje berre bør målast med ein skalar koeffisient. Difor innfører dei asymmetrifunksjonar som skildrar asymmetrien i alle delane av fordeling. Desse funksjonane skildrar asymmetrien til tettleiken godt. Dersom vi kjenner asymmetrifunksjonen, modalverdi og enten høgre- eller venstre avstandsfunksjon, kan ein rekne seg tilbake til kva fordeling asymmetrifunksjonen kjem frå. Vidare innfører dei ein vektfunksjon som kan brukast til å rekne ut ein skalar verdi for asymmetrien. Denne vektfunksjonen gjer at vi kan velje sjølve kva del av asymmetrien vi vil legge vekt på, og kan vere nyttig i mange samanhengar. Likevel er det nokre ulemper med asymmetrifunksjonar. Dei kan vere vanskelege å rekne ut analytisk, og det krev difor ofte mykje for å forstå kva dei måler. I tillegg er dei berre definert for eintoppa, rota fordelingar, og sjølv om dette er ein stor klasse av fordelingar, er det naudsynt

med andre mål for asymmetri i tillegg. Asymmetrifunksjonen kan også vere vanskeleg å estimere, og krever store datasett for gode estimat. Når datasetta er små, vil estimatoren for nokre datasett underestimere asymmetrien, medan dei vil overestimere asymmetrien for andre datasett. Difor er det også vanskeleg å justere for feil i forventningsverdi til estimatoren.

Kurtose skal måle spissheit og haletyngde til ein fordeling. Om målet bygd på moment faktisk måler dette er omdiskutert. Difor innfører Critchley og Jones (2008) eit nytt mål som dei kallar gradientasymmetri. Dette målet samanliknar gradienten eller den deriverte i ulike deler av tettleiken. Ideen er at dersom talverdien til den deriverte i halen er låg, er halane tunge. På same måte vil stor talverdi til den deriverte omkring modalverdien, tyde på ein spiss tettleik. Skeivleiksmålet til Critchley og Jones (2008) kan også vere med å skildre haletyngde, ettersom stor haletyngde i eine delen av fordelinga vil påvirke asymmetrifunksjonen. Også her kan vi rekne ut skalare verdiar for gradientasymmetrien, der vi ved å bruke ein vektfunksjon kan avgjere om vi ynskjer å sjå på spissheit eller haletyngde. Det er likevel nokre ulemper for kurtosemålet til Critchley og Jones (2008). Funksjonane kan også her vere vanskeleg å rekne ut analytisk og det krev endå meir enn for skeivleik å forstå kva dei faktisk skildrar. Fordelingane som målet er definert for er ein endå mindre klasse enn målet for asymmetri, ettersom både tettleiken og den deriverte til tettleiken må vere rota og eintoppa. Estimatorar for gradientasymmetrien var sjølv for store datasett relativt svake, trass i at estimatorar for den deriverte er gode. Grunnen til dette er at gradientasymmetrien er avhengig av mange andre ting enn berre estimatet til den deriverte.

På grunnlag av dette kan vi seie at måla til Critchley og Jones (2008) kan fungere godt som teoretiske mål på skeivleik, spissheit og haletyngde. Dei er fleksible med tanke på kva eigenskapar vi er interessert i, og skildrar desse eigenskapane godt. Som empiriske mål er dei mindre gode, men kan fungere dersom datasettet som blir brukt er stort.

Tillegg A

Kvantilar for dipen

Kvantilane er rekna ut basert på utrekningen av dipen for uniformt fordelte variable og det er simulert $10^6 + 1$ datasett av kvar storleik (Maechler 2009). Storleiken på datasettet står i margen til venstre, medan storleiken på kvantilane står øverst. Kritiske verdiar for dipen kan lesast i tabellen ut frå dette. Legg merke til datasett på storleikane 7,8 og 9, der dip-observatoren ikkje er synkande saman med storleiken på settet. Utrekningar i Hartigan og Hartigan (1985) viser det same.

n	.01	.05	.10	.50	.90	0.95	.99
4	0.1250	0.1250	0.1250	0.1250	0.1874	0.2073	0.2318
5	0.1000	0.1000	0.1000	0.1216	0.1768	0.1864	0.1965
6	0.0833	0.0833	0.0833	0.1231	0.1591	0.1648	0.1919
7	0.0714	0.0726	0.0817	0.1178	0.1442	0.1599	0.1841
8	0.0625	0.0739	0.0820	0.1110	0.1418	0.1540	0.1730
9	0.0613	0.0733	0.0804	0.1042	0.1364	0.1466	0.1642
10	0.0610	0.0718	0.0780	0.0978	0.1305	0.1396	0.1597
15	0.0546	0.0610	0.0643	0.0836	0.1101	0.1188	0.1360
20	0.0474	0.0527	0.0568	0.0733	0.0972	0.1051	0.1206
30	0.0396	0.0444	0.0474	0.0615	0.0815	0.0882	0.1015
50	0.0314	0.0353	0.0377	0.0489	0.0649	0.0703	0.0812
100	0.0228	0.0257	0.0274	0.0355	0.0472	0.0511	0.0590
200	0.0165	0.0185	0.0198	0.0256	0.0340	0.0368	0.0427
500	0.0106	0.0119	0.0127	0.0165	0.0219	0.0237	0.0275
1000	0.0076	0.0085	0.0091	0.0117	0.0156	0.0169	0.0196
2000	0.0054	0.0061	0.0065	0.0084	0.0111	0.0120	0.0140
5000	0.0034	0.0039	0.0041	0.0053	0.0071	0.0077	0.0089

Tabell A.1: *Kvantilar for dipen.*

Tillegg B

Varians til estimerte variable

I kapittel 4 såg vi på ein del variable i samband med estimering av asymmetrien og kurtosen. I dette tillegget er observerte variansar til variablane vedlagt.

n	$\mathrm{Var}(ISE_{\gamma^*})$	$\mathrm{Var}(ISE_{\hat{f}})$	$\mathrm{Var}(s_{\hat{f}})$	$\mathrm{Var}(G_1)$	$\mathrm{Var}(\hat{m})$
10	8.5e-3	5.2e-4	0.087	0.34	0.25
100	1.2e-3	2.5e-5	0.10	0.18	0.030
1000	1.7e-4	9.9e-7	0.021	0.027	6.9e-3
5000	1.8e-5	6.6e-8	5.4e-3	6.1e-3	3.3e-3

Tabell B.1: *Empirisk varians for estimerte variable for asymmetri i kapittel 4. Utgangspunktet er gammafordeling med parameter α=2 og $\beta = 1$. Vi har brukt normalfordelinga som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Dei teoretiske asymmetrikoeffisientane er $\gamma = 0.38$ og $s = \sqrt{2}$. Modalverdien er $m = 1$.*

n	$\mathrm{Var}(ISE_{\gamma^*})$	$\mathrm{Var}(ISE_{\hat{f}})$	$\mathrm{Var}(s_{\hat{f}})$	$\mathrm{Var}(G_1)$	$\mathrm{Var}(\hat{m})$
10	0.011	5.5e-4	0.16	0.34	0.27
100	1.9e-3	1.9e-5	0.13	0.18	0.041
1000	1.3e-3	5.8e-7	0.024	0.027	0.016
5000	1.8e-4	3.3e-8	5.7e-3	6.1e-3	8.6e-3

Tabell B.2: *Empirisk varians for estimerte variable for asymmetri i kapittel 4. Utgangspunktet er gammafordeling med parameter α=2 og $\beta = 1$. Vi har brukt Epanechnikov-fordelinga som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Dei teoretiske asymmetrikoeffisientane er $\gamma = 0.38$ og $s = \sqrt{2}$. Modalverdien er $m = 1$.*

n	$\mathrm{Var}(ISE_{\gamma^*})$	$\mathrm{Var}(ISE_{\hat{f}})$	$\mathrm{Var}(s_{\hat{f}})$	$\mathrm{Var}(G_1)$	$\mathrm{Var}(\hat{m})$
10	8.3e-3	5.9e-4	0.17	0.34	0.24
100	2.0e-3	1.9e-5	0.14	0.18	0.044
1000	1.1e-3	5.2e-7	0.024	0.027	0.018
5000	1.3e-4	2.9e-8	5.7e-3	6.1e-3	0.010

Tabell B.3: *Empirisk varians for estimerte variable for asymmetri i kapittel 4. Utgangspunktet er gammafordeling med parameter α=2 og $\beta = 1$. Vi har brukt biweightfordelinga som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Dei teoretiske asymmetrikoeffisientane er $\gamma = 0.38$ og $s = \sqrt{2}$. Modalverdien er $m = 1$.*

n	$\mathrm{Var}(ISE_{\gamma^*})$	$\mathrm{Var}(ISE_{\hat{f}})$	$\mathrm{Var}(s_{\hat{f}})$	$\mathrm{Var}(G_1)$	$\mathrm{Var}(\hat{m})$
10	8.6e-3	6.2e-4	0.18	0.34	0.34
100	2.7e-3	1.9e-5	0.14	0.18	0.050
1000	1.6e-3	4.9e-7	0.024	0.027	0.022
5000	1.9e-5	2.7e-8	5.8e-3	6.1e-3	0.011

Tabell B.4: *Empirisk varians for estimerte variable for asymmetri i kapittel 4. Utgangspunktet er gammafordeling med parameter α=2 og $\beta = 1$. Vi har brukt triweightfordelinga som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Dei teoretiske asymmetrikoeffisientane er $\gamma = 0.38$ og $s = \sqrt{2}$. Modalverdien er $m = 1$.*

n	$\mathrm{Var}(ISE_{\gamma^*})$	$\mathrm{Var}(ISE_{\hat{f}})$	$\mathrm{Var}(s_{\hat{f}})$	$\mathrm{Var}(G_1)$	$\mathrm{Var}(\hat{m})$
10	0.020	1.8e-3	0.056	0.34	0.72
100	7.6e-3	2.3e-4	0.081	0.18	0.10
1000	1.3e-3	2.2e-5	0.019	0.027	0.027
5000	1.6e-4	3.4e-8	5.1e-3	6.1e-3	9.1e-3

Tabell B.5: *Empirisk varians for estimerte variable for asymmetri i kapittel 4. Utgangspunktet er gammafordeling med parameter α=2 og $\beta = 1$. Vi har brukt Gamma(2,1) som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Dei teoretiske asymmetrikoeffisientane er $\gamma = 0.38$ og $s = \sqrt{2}$. Modalverdien er $m = 1$.*

n	$\mathrm{Var}(ISE_{\gamma^*})$	$\mathrm{Var}(ISE_{\hat{f}})$	$\mathrm{Var}(s_{\hat{f}})$	$\mathrm{Var}(G_1)$	$\mathrm{Var}(\hat{m})$
10	0.027	5.0e-4	0.057	0.22	0.13
100	0.067	2.5e-5	0.015	0.026	0.12
1000	0.036	5.8e-7	1.3e-3	1.7e-3	0.060
5000	0.039	3.1e-8	3.9e-4	4.4e-4	0.055

Tabell B.6: *Empirisk varians for estimerte variable for asymmetri i kapittel 4. Utgangspunktet er fordeling sett saman av andregrads- og fjerdegradspolynomen med parameter $\alpha = 1$ og $\beta = 2$. Vi har brukt normalfordelinga som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Dei teoretiske asymmetrikoeffisientane er $\gamma = 0.43$ og $s = 0.044$. Modalverdien er $m = 0$.*

n	$\mathrm{Var}(ISE_{\gamma^*})$	$\mathrm{Var}(ISE_{\hat{f}})$	$\mathrm{Var}(s_{\hat{f}})$	$\mathrm{Var}(G_1)$	$\mathrm{Var}(\hat{m})$
10	0.027	8.9e-4	0.10	0.22	0.17
100	0.067	3.9e-5	0.019	0.026	0.15
1000	0.036	6.8e-7	1.5e-3	1.7e-3	0.065
5000	0.039	3.0e-8	4.2e-4	4.4e-4	0.048

Tabell B.7: *Empirisk varians for estimerte variable for asymmetri i kapittel 4. Utgangspunktet er fordeling sett saman av andregrads- og fjerdegradspolynomen med parameter* $\alpha = 1$ *og* $\beta = 2$*. Vi har brukt Epanechnikov-fordelinga som kjerne og* $\hat{h}_{AMISE}$ *som bandbreidde. Dei teoretiske asymmetrikoeffisientane er* $\gamma = 0.43$ *og* $s = 0.044$*. Modalverdien er* $m = 0$*.*

n	$\mathrm{Var}(ISE_{\gamma^*})$	$\mathrm{Var}(ISE_{\hat{f}})$	$\mathrm{Var}(s_{\hat{f}})$	$\mathrm{Var}(G_1)$	$\mathrm{Var}(\hat{m})$
10	0.069	1.0e-3	0.11	0.22	0.21
100	0.098	4.4e-5	0.020	0.026	0.17
1000	0.042	7.3e-7	1.5e-3	1.7e-3	0.060
5000	0.028	3.1e-8	4.2e-4	4.4e-4	0.054

Tabell B.8: *Empirisk varians for estimerte variable for asymmetri i kapittel 4. Utgangspunktet er fordeling sett saman av andregrads- og fjerdegradspolynomen med parameter* $\alpha = 1$ *og* $\beta = 2$*. Vi har brukt biweight-fordelinga som kjerne og* $\hat{h}_{AMISE}$ *som bandbreidde. Dei teoretiske asymmetrikoeffisientane er* $\gamma = 0.43$ *og* $s = 0.044$*. Modalverdien er* $m = 0$*.*

n	$\mathrm{Var}(ISE_{\gamma^*})$	$\mathrm{Var}(ISE_{\hat{f}})$	$\mathrm{Var}(s_{\hat{f}})$	$\mathrm{Var}(G_1)$	$\mathrm{Var}(\hat{m})$
10	0.083	1.2e-3	0.12	0.22	0.22
100	0.10	4.8e-5	0.021	0.026	0.17
1000	0.038	7.7e-7	1.5e-3	1.7e-3	0.071
5000	0.026	3.3e-8	4.2e-4	4.4e-4	0.049

Tabell B.9: *Empirisk varians for estimerte variable for asymmetri i kapittel 4. Utgangspunktet er fordeling sett saman av andregrads- og fjerdegradspolynomen med parameter* $\alpha = 1$ *og* $\beta = 2$*. Vi har brukt triweight-fordelinga som kjerne og* $\hat{h}_{AMISE}$ *som bandbreidde. Dei teoretiske asymmetrikoeffisientane er* $\gamma = 0.43$ *og* $s = 0.044$*. Modalverdien er* $m = 0$*.*

n	$\mathrm{Var}(ISE_{\gamma^*})$	$\mathrm{Var}(ISE_{\hat{f}})$	$\mathrm{Var}(s_{\hat{f}})$	$\mathrm{Var}(G_1)$	$\mathrm{Var}(\hat{m})$
10	5.1e-3	1.4e-4	0.20	0.40	0.95
100	9.1e-4	2.4e-6	0.063	0.083	0.37
1000	5.6e-4	5.2e-8	6.8e-3	7.6e-3	0.21
5000	2.6e-4	2.9e-9	1.8e-3	1.9e-4	0.079

Tabell B.10: *Empirisk varians for estimerte variable for asymmetri i kapittel 4. Utgangspunktet er gammafordeling med parameter* $\alpha = 10$ *og* $\beta = 1$*. Vi har brukt biweight-fordelinga som kjerne og* $\hat{h}_{AMISE}$ *som bandbreidde. Dei teoretiske asymmetrikoeffisientane er* $\gamma = 0.43$ *og* $s = 0.044$*. Modalverdien er* $m = 0$*.*

n	$\mathrm{Var}(ISE_{\gamma^*})$	$\mathrm{Var}(ISE_{\hat{f}})$	$\mathrm{Var}(s_{\hat{f}})$	$\mathrm{Var}(G_1)$	$\mathrm{Var}(\hat{m})$
10	2.8e-3	4.5e-2	0.17	0.34	0.25
100	3.1e-3	6.1e-3	0.031	0.041	0.080
1000	1.1e-3	1.1e-3	4.9e-3	5.5e-3	0.029
5000	4.8e-4	4.8e-4	8.4e-4	8.9e-4	0.015

Tabell B.11: *Empirisk varians for estimerte variable for asymmetri i kapittel 4. Utgangspunktet er standard normalfordeling. Vi har brukt biweight-fordelinga som kjerne og* $\hat{h}_{AMISE}$ *som bandbreidde. Dei teoretiske asymmetrikoeffisientane er* $\gamma = 0$ *og* $s = 0$*. Modalverdien er* $m = 0$*.*

n	$\mathrm{Var}(ISE_{\delta_R^*})$	$\mathrm{Var}(ISE_{\hat{f}'})$	$\mathrm{Var}(G_2)$	$\mathrm{Var}(\hat{m})$	$\mathrm{Var}(\hat{\pi}_R)$
10	4.9-3	0.081	0.55	0.16	0.33
100	1.5	0.015	0.31	0.031	0.085
1000	1.5e-3	4.5e-3	0.022	7.4e-3	0.032
5000	4.7e-4	1.9e-4	4.4e-3	2.1e-3	0.014

Tabell B.12: *Empirisk varians for estimerte variable for kurtose i kapittel 4. Utgangspunktet er standard normalfordeling. Vi har brukt normalfordelinga som kjerne og* $\hat{h}_{AMISE}$ *som bandbreidde. Den teoretiske gradientasymmetrikoeffisienten er* $\delta_R \approx 0.17$ *og kurtosekoeffisienten er* $k = 3$*. Modalverdien er* $m = 0$ *og vendepunktet er* $\pi_R = 1$*.*

n	$\text{Var}(ISE_{\delta_R^*})$	$\text{Var}(ISE_{\delta_L^*})$	$\text{Var}(ISE_{\hat{f}'})$
10	4.8-3	0.027	9.9e-3
100	4.2e-4	2.8e-3	1.1e-4
1000	2.4e-4	8.3e-4	3.4e-5
5000	1.4e-4	5.6e-4	1.2e-5

Tabell B.13: *Empirisk varians for estimerte variable for kurtose i kapittel 4. Utgangspunktet er fordeling sett saman av polynomen og eksponesialfunksjon. Vi har brukt normal-fordelinga som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Den teoretiske høgre gradientasymmetrikoeffisienten er $\delta_R \approx 0.20$ og den teoretiske venstre gradientasymmetrikoeffisienten er $\delta_L \approx 0.43$.*

n	$\text{Var}(G_2)$	$\text{Var}(\hat{m})$	$\text{Var}(\hat{\pi}_R)$	$\text{Var}(\hat{\pi}_L)$
10	1.21	0.21	0.37	0.89
100	7.30	9.8e-3	0.021	0.38
1000	2.19	1.7e-3	0.0023	0.010
10000	0.70	6.3e-4	8.9e-4	4.4e-3

Tabell B.14: *Empirisk varians for estimerte variable for kurtose i kapittel 4. Utgangspunktet er fordeling sett saman av polynomen og eksponesialfunksjon. Vi har brukt normal-fordelinga som kjerne og $\hat{h}_{AMISE}$ som bandbreidde. Kurtosekoeffisienten er $k = 1.8$, modalverdien er $m = 1$ og vendepunkta er $\pi_R = 2$ og $\pi_L = 0$.*

Tillegg C

Programkode

I dette tillegget legg vi ved ein del av programkoden vi har brukt. Metodane som er brukt i denne koden er felles for alle utrekningane vi har gjort. Datamaterialet vi har brukt, er trekt ut på førehand og vi har brukt same data for dei ulike metodane. For å finne inversen til tettleik og gradient, har vi brukt funksjonen uniroot i R, som nyttar seg av den gyldne seksjonsmetode kombinert med parabolsk interpolasjon. Vi har også brukt parabolsk interpolasjon når vi skulle integrere asymmetrifunksjonane, i form av Simpsons regel .

```
##Dokumentasjon
#Skript som reknar ut empirisk skeivleikskoeffisient,
#asymmetrifunksjon og kvadratisk feil for ulike datasett
#og med ulike kjernar.
#Reknar ut sannsyn for å forkaste at det kjem frå unimodal fordeling.

##Viktige variablar
#f: Teoretisk funksjon
#fE: Estimert funksjon
#gammas: Teoretisk asymmetrifunksjon
#gammasE: Estimert asymmetrifunksjon
#gammakoeff: Teoretisk asymmetrikoeffisient
#gammakoeffE: Estimert asymmetrikoeffisient
#ISEa: Integrert kvadratisk feil mellom teoretisk
# og estimert asymmetrifunksjon
#ISEt: Integrert kvadratisk feil mellom teoretisk
#og estimert tettleik
#s: Tradisjonell teoretisk skeivleikskoeffisient
#sE: Tradisjonell empirisk skeivleikskoeffisient
#sEt: Tradisjonell teoretisk skeivleikskoeffisient
#for estimert tettleik.
```

```
#Lastar inn pakke til Dip-testing
library(diptest)
#Lastar inn metode for empirisk skeivleik og kurtose
library(GLDEX)


##Fordeling vi har simulert frå
alfa=1
beta=2
a=15/alfa^2/(10*alfa+12*beta)
b=15/beta^4/(10*alfa+12*beta)
c=15/(10*alfa+12*beta)

f=function(x){(-a*x^2+c)*(x>-alfa & x<=0)+(-b*x^4+c)*(x>0 & x<beta)}

#Modalverdi med tilhøyrande funksjonsverdi
m=0
fM=f(m)

#Forventningsverdi, varians og tradisjonell skeivleik
EX=integrate(function(x)x*f(x),-Inf,Inf)$value
VarX=integrate(function(x)(x-EX)^2*f(x),-Inf,Inf)$value
s=integrate(function(x)(x-EX)^3*f(x),-Inf,Inf)$value/VarX^1.5

#Antal simuleringar
q=100
#Antal variable vi simulerer
n=c(10,100,1000,5000)
#Vektor for å kontrollere dipen
dipT=c(7,12,15,17)
#Estimert asymmetrikoeffisient

#Helpevektorar for integrasjon
p=seq(0,1,0.01)
l=length(p)
odd=seq(3,l-2,2)
even=seq(2,l-1,2)
xE=seq(-100,100,0.1)
lE=length(xE)
oddE=seq(3,lE-2,2)
evenE=seq(2,lE-1,2)
```

```
#Vektorar der vi lagrar numeriske utrekningar
fE1=rep(0,lE)
gammasE1=rep(0,l)
gammakoeffE=rep(0,q)
ISEa=rep(0,q)
ISEt=rep(0,q)
sE=rep(0,q)
sEt=rep(0,q)

##Teoretisk asymmetrifunksjon
gammas=function(p){(beta-alfa*(1-p)^0.25)/(beta+alfa*(1-p)^0.25)}
#Numeriske verdiar for asymmetrifunksjonen
gammas1=rep(0,l)
for(i in 1:l){gammas1[i]=gammas(p[i])}
#Asymmetrikoeffisienten rekna ut med Simpsons regel
gammakoeff=p[3]/6*(2*sum(gammas1[odd])+4*sum(gammas1[even])
+gammas1[1]+gammas1[l])

#Matrise der vi lagrar dataene våre.
POLYBETA=array(0,dim=c(4,13,3))
dimnames(POLYBETA)=list(c(10,100,1000,5000),
c("ISEa","Var(ISEa)","Dip","GammaE","Var(GammaE)","Gamma",
"ISEt","var(ISEt)","sE","var(sE)","sEt","var(sEt)","s"),
c("epanechnikov","biweight","triweight"))


##Løkke som brukar ulike kjernar til estimatet
for(w in 1:3)
{
#Teljar som vel ut kva data vi skal bruke
teller= -9
#Sender epost om at utrekninga for denne kjernen startar
system("cat polybetastart.txt | /usr/lib/sendmail -t")

##Kjerne til kjerneestimeringa
k=w
K=function(z){(1-z^2)^k/(2^(2*k+1)*beta(k+1,k+1))*(z>-1 & z<1)}
#Verdiar til bandbreidda
RK=integrate(function(x){K(x)^2},-Inf,Inf)$value
mu1K=integrate(function(x){x*K(x)},-Inf,Inf)$value
mu2K=integrate(function(x){(x-mu1K)^2*K(x)},-Inf,Inf)$value
```

```
##Løkke som brukar ulike storleikar på datasetta
for(v in 1:length(n))
{
#Teljar til testing av unimodalitet
unimodal=0
##Løkke som brukar antal datasett av kvar storleik
for(j in 1:q)
{
##Simulerte data
teller=teller+n[v]
X=Y[teller:(teller+n[v]-1)]

##Test om data kjem frå eintoppa fordeling
D=dip(X)
#Hentar frå tabell for dipfordelinga med 10 observasjonar
#og 5% signifikansnivå.
if(D>qDiptab[dipT[v],14])unimodal=unimodal+1

##Kjerneestimat
#Bandbreidde
h=(243*RK/mu2K/35/n[v])^0.2*sqrt(var(X))
fEX=function(z){K((z-X)/h)}
#Estimert funksjon
fE=function(z){(h*n[v])^(-1)*sum(fEX(z))}

##Skeivleik for den estimerte funksjonen
#Reknar ut fyrst andre og tredje moment med numerisk integrasjon
for(i in 1:lE){fE1[i]=fE(xE[i])}
muEh=fE1*xE
muE=0.2/6*(2*sum(muEh[oddE])+4*sum(muEh[evenE])+muEh[1]+muEh[lE])
mu2Eh=fE1*xE^2
mu2E=0.2/6*(2*sum(mu2Eh[oddE])+4*sum(mu2Eh[evenE])+
mu2Eh[1]+mu2Eh[lE])
mu3Eh=fE1*xE^3
mu3E=0.2/6*(2*sum(mu3Eh[oddE])+4*sum(mu3Eh[evenE])+mu3Eh[1]+
mu3Eh[lE])

#Skeivleikskoeffisienten basert på tettleiksestimatet
sEt[j]=(mu3E-3*muE*mu2E+2*muE^3)/(mu2E-muE^2)^1.5


##Modalverdi og avstandsfunksjonar for tettleiksestimatet
#Estimert modalverdi og tilhøyrande funksjonsverdi
```

```
mE=optimize(fE,lower=-5,upper=10,maximum=TRUE,tol=1e-20)$maximum
fEm=optimize(fE,lower=-5,upper=10,maximum=TRUE,tol=1e-20)$objective

#Venstre avstandsfunksjon
xLE=function(p){uniroot(function(x){fE(x)-p*fEm},
lower=-10000,upper=mE)}
tauLE=function(p){mE-xLE(p)$root}

#Høgre avstandsfunksjon
xRE=function(p){uniroot(function(x){fE(x)-p*fEm},
lower=mE,upper=10000)}
tauRE=function(p){xRE(p)$root-mE}

##Asymmetrifunksjon estimert ved hjelp av kjerneestimering
gammasE=function(p){(tauRE(p)-tauLE(p))/(tauRE(p)+tauLE(p))}

#Numeriske verdiar for asymmetrifunksjonen i gitte punkt
for(i in 1:l)gammasE1[i]=gammasE(p[i])
gammasE1[l]=0

#Asymmetrikoeffisienten rekna ut med polynomeninterpolasjon.
gammakoeffE[j]=p[3]/6*(2*sum(gammasE1[odd])
+4*sum(gammasE1[even])+gammasE1[1]+gammasE1[l])

##Integrert kvadratisk feil for asymmetrifunksjonen
#Kvadratisk feilfunksjon for asymmetrifunksjonen
SEa=function(p){(gammas-gammasE)^2}
#Verdiar til kvardratisk feilfunksjon
SEa1=(gammas1-gammasE1)^2
#Integrert kvadratisk feil
ISEa[j]=p[3]/6*(2*sum(SEa1[odd])+4*sum(SEa1[even])+SEa1[1]+SEa1[l])


##Integrert kvadratisk feil for tettleiken
#Kvadratisk feilfunksjon for tettleiken
SEt=function(x){(f(x)-fE(x))^2}
#Verdiar til kvadratisk feilfunksjon
SEt1=rep(0,lE)
for(i in 1:lE){SEt1[i]=SEt(xE[i])}
#Integrert kvadratisk feil
ISEt[j]=0.2/6*(2*sum(SEt1[oddE])+4*sum(SEt1[evenE])+SEt1[1]+SEt1[lE])

##Tradisjonelt estimat for skeivleik
```

```
sE[j]=skewness(X,method="moment")

}
#Lagring av dei ulike estimata for skeivleik og feil.
POLYBETA[v,1,w]=mean(ISEa)
POLYBETA[v,2,w]=var(ISEa)
POLYBETA[v,3,w]=unimodal
POLYBETA[v,4,w]=mean(gammakoeffE)
POLYBETA[v,5,w]=var(gammakoeffE)
POLYBETA[v,6,w]=gammakoeff
POLYBETA[v,7,w]=mean(ISEt)
POLYBETA[v,8,w]=var(ISEt)
POLYBETA[v,9,w]=mean(sE)
POLYBETA[v,10,w]=var(sE)
POLYBETA[v,11,w]=mean(sEt)
POLYBETA[v,12,w]=var(sEt)
POLYBETA[v,13,w]=s
}
}
#Lagring av resultat i R-objekt
dump("POLYBETA","polybeta.R")
#Sender epost at utrekninga er ferdig
system("cat polybetaslutt.txt | /usr/lib/sendmail -t")
```

Referansar

An, L. og S. E. Ahmed (2008). Improving the performance of kurtosis estimator. *Computational Statistics & Data Analysis 52*(5), 2669–2681.

Arnold, B. C. og R. A. Groeneveld (1995). Measuring skewness with respect to the mode. *The American Statistician 49*(1).

Averous, J., A.-L. Fougères, og M. Meste (1996). Tailweight with respect to the mode for unimodal distributions. *Statistics & Probability Letters 28*(4), 367–373.

Balanda, K. P. og H. L. MacGillivray (1988). Kurtosis: A critical review. *The American Statistician 42*(2), 111–119.

Bickel, D. R. og R. Frühwirth (2006). On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications. *Computational Statistics & Data Analysis 50*(12), 3500 – 3530.

Boshnakov, G. N. (2007). Some measures for asymmetry of distributions. *Statistics & Probability Letters 77*(11), 1111–1116.

Casella, G. og R. L. Berger (2002). *Statistical Inference; 2nd ed.* Duxbury advanced series. New Delhi: Wadsworth.

Critchley, F. og M. C. Jones (2008). Asymmetry and gradient asymmetry functions: density-based skewness and kurtosis. *Scandinavian Journal of Statistics. Theory and Applications 35*(3), 415–437.

Fiori, A. M. og M. Zenga (2009). Karl pearson and the origin of kurtosis. *International Statistical Review 77*, 40–50.

Hartigan, J. A. og P. M. Hartigan (1985). The dip test of unimodality. *The Annals of Statistics 13*(1), 70–84.

Holgersson, H. E. T. (2007). Robust testing for skewness. *Communications in Statistics - Theory and Methods 36*(1-4), 485–498.

Joanes, D. og C. Gill (1998). Comparing measures of sample skewness and kurtosis. *The Statistician 47*(1), 183–189.

Oja, H. (1981). On location, scale, skewness and kurtosis of univariate distributions. *Scandinavian Journal of Statistics*, 154–168.

Pearson, K. (1895). Contributions to the mathematical theory of evolution ii. skew variations in homogeneous material. *Transactions of the Royal Philosophical Society, Ser. A 186*, 343–414.

Rudin, W. (1976). *Principles of mathematical analysis* (Third ed.). New York: McGraw-Hill Book Co. International Series in Pure and Applied Mathematics.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B (Methodological) 43*(1), 97–99.

Wand, M. P. og M. C. Jones (1995). *Kernel smoothing*, Volume 60 of *Monographs on Statistics and Applied Probability*. London: Chapman and Hall Ltd.