Snøtetthet

Notat for TMA4240/TMA4245 Statistikk*

Institutt for matematiske fag,
NTNU

15. august 2012

I forbindelse med varsling av flom, klimaforskning og særlig kraftproduksjon er det viktig å kunne anslå hvor mye vann som er inneholdt i snøen i et gitt område, altså hvor stor mengde smeltevann en kan forvente. Dette avhenger av dybden til snøen, som er enkel å måle, og av tettheten til snøen i forhold til vann, som vanligvis er mellom 0.01 og 0.4, og er vanskeligere å måle.

Målinger av snødybde og -tetthet mellom 1976 og 2006 ved en rekke målestasjoner i Sør-Trøndelag er tilgjengelige. Vi vil her konsentrere oss om målinger fra *Rybekken* og *Sylsjødammen* i Tydal kommune. Dataene er lagret i filene `rybekken.txt` og `sylsjodammen.txt`. De fem kolonnene er, fra venstre; årstall, snødybde i cm, snøtetthet i g/cm^3, *SWE* i cm og *DOY* (henholdsvis *Snow Water Equivalent* og *Day Of Year*.) De første fem datapunktene fra Rybekken er vist i tabell 1. For de dagene hvor det er gjort flere målinger ved samme stasjon vil en, siden målingene ikke er tatt akkurat samme sted, fortsatt kunne få forskjellige måleverdier.

Tabell 1 – De første fem observasjonene i datasettet fra Rybekken.

År	Snødybde (cm)	Snøtetthet (g/cm^3)	*SWE* (cm)	*DOY*
1976	105.0	0.298	31.3	63
1976	178.0	0.342	60.9	63
1976	112.0	0.412	46.1	125
1976	180.0	0.426	76.7	125
1977	76.0	0.298	22.6	97

SWE er et mål på hvor mye vann snøen inneholder, og avhenger av tetthet, ρ, og dybde, h, via formelen

$$SWE = h \cdot \frac{\rho}{\rho_w}$$

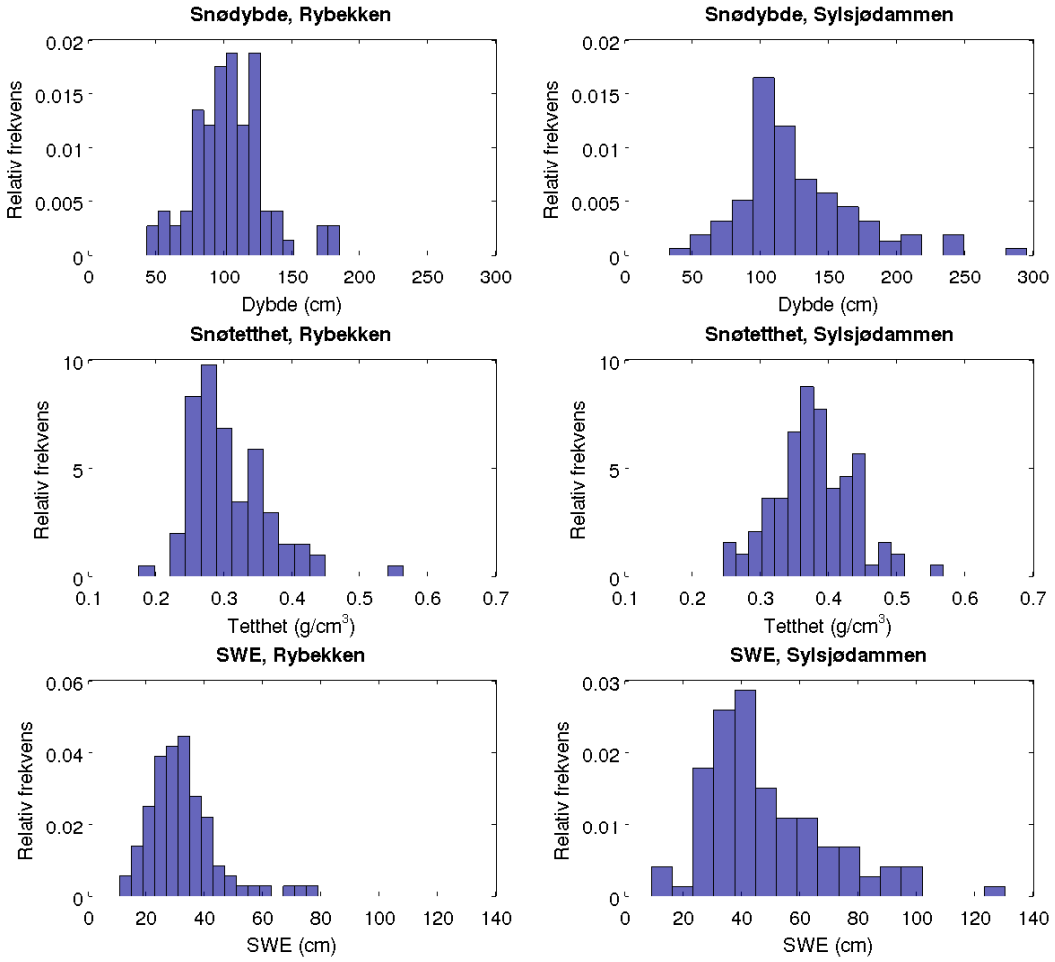
hvor ρ_w er tettheten til vann, altså 1 g/cm^3. *SWE* har samme enheter som dybde.

DOY er antall dager siden 1. januar, og antar i dette datasettet verdier mellom 38 og 132, som svarer til perioden 7. februar til 12. mai.

*Notatet er skrevet av Jacob Skauvold i samarbeid med Ingelin Steinsland, og er basert på fordypningsoppgaven *Estimating Snow Water Equivalent* av Åshild Færevåg. Oddbjørn Bruland og Knut Sand i Statkraft har gjort målinger tilgjengelige og bidratt med problemstillingen og bakgrunnsinformasjon. Dersom du finner feil eller har forslag til forbedringer, ta kontakt med Ingelin Steinsland, `ingelins@math.ntnu.no`.

Visualisering av data

For å danne oss et bilde av hvordan dataene er fordelt kan vi plotte histogrammer av snødybde og snøtetthet for de to stasjonene. figur 1 viser fire slike plott. Matlabkoden for å generere plottene er lagret i filen `histogram.m`.



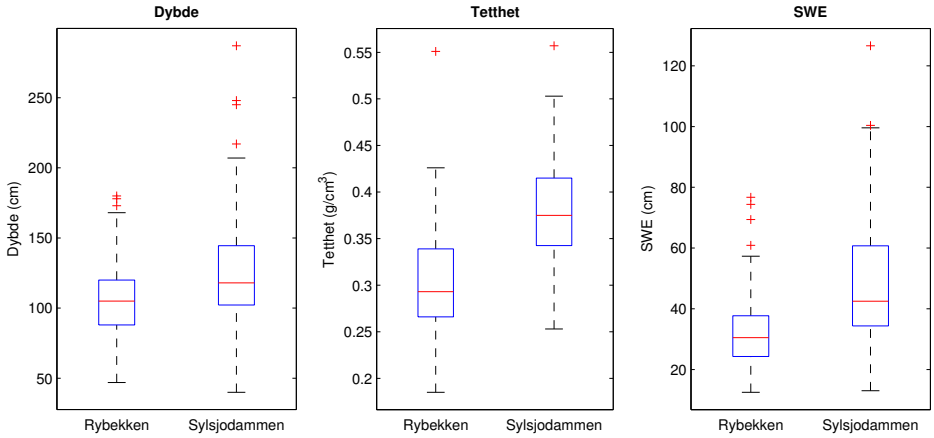
Figur 1 – Histogram av observasjonene av snødybde og -tetthet, samt *SWE*, ved Sylsjødammen og Rybekken. Figuren er laget med `histogram.m`.

Sylsjødammen ser ut til å ha noe større variasjon enn Rybekken i både dybde og tetthet. Tyngden av observasjonene ligger også tilsynelatende litt høyere for Sylsjødammen, både for dybde og for tetthet. Vi vil sammenlikne sentralitet og spredning, og forsetter derfor ved å lage box plot av dataene. Disse er vist i figur 2. Matlabkoden brukt for å lage dem er lagret i fila `boxplots.m`.

En oversikt over empirisk standardavvik og forventingsverdi for snødybde, -tetthet og *SWE* er gitt i tabell 2.

Vi kan også lage box plot med en egen boks for hvert år. Det viser på en tydelig måte at variasjonen er større enkelte år enn andre, og at antall observasjoner varierer fra år til år. Fire slike box plot er vist i figur 3.

For å synliggjøre sammenhenger mellom de ulike variablene som er målt, lager vi i figur 4



Figur 2 – Box plot av snødybde, snøtetthet og SWE ved Rybekken og Sylsjødammen. Figuren er laget med `boxplots.m`.

Tabell 2 – Empirisk forventningsverdi og standardavvik for snødybde, snøtetthet og *SWE* for Rybekken og Sylsjødammen

Stasjon	Størrelse	Forventningsverdi	Standardavvik
Rybekken	Snødybde	104.02 cm	25.77 cm
	Snøtetthet	0.3063 g/cm^3	0.0574 g/cm^3
	SWE	32.35 cm	12.12 cm
Sylsjødammen	Snødybde	126.28 cm	43.98 cm
	Snøtetthet	0.3779 g/cm^3	0.0564 g/cm^3
	SWE	48.56 cm	20.88 cm

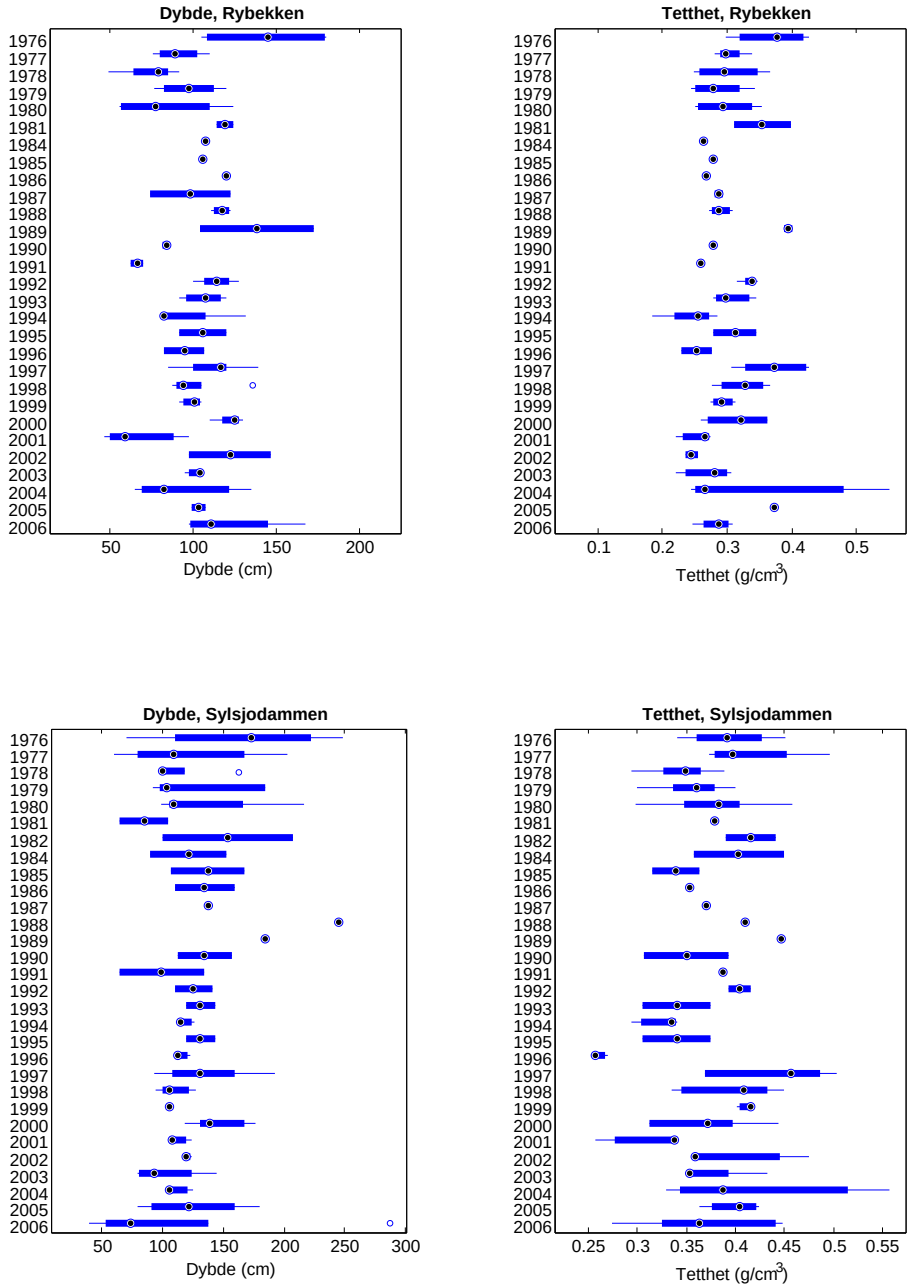
en matrise av scatter plot, hvor alle de fem variablene plottes mot hverandre. På diagonalen er vist histogrammer av observasjonene av den aktuelle variabelen. Her er dataene fra Rybekken og Sylsjødammen kombinert til ett sett med 193 observasjoner.

Sammenlikning av forventningsverdi

(W.M.M.Y. Kap. 9.8 (sammenlikning av forventningsverdier), 9.9 (observasjoner i par) og 10.5 (toutvalgstest og partest))

Med utgangspunkt i figur 1 og 2 og tabell 2 vil vi undersøke om det i gjennomsnitt var dypere snø ved Sylsjødammen enn ved Rybekken i perioden 1976-2006. Variansen er ukjent, og vi kan ikke anta at fordelingen av snødybde ved Sylsjødammen og Rybekken har samme varians (se figur 2). La $X_{11}, X_{12}, \ldots, X_{1103}$ være observasjonene av snødybde fra Sylsjødammen, og la $X_{21}, X_{22}, \ldots, X_{290}$ være observasjonene fra Rybekken. Vi antar at observasjonene er realiseringer av to normalfordelte variable, $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$ (bruk histogrammene i figur 1 og Q-Q plottene i figur 5 for å vurdere om dette er en rimelig antakelse). Da har, ifølge kap. 9.8 i W.M.M.Y., størrelsen

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$



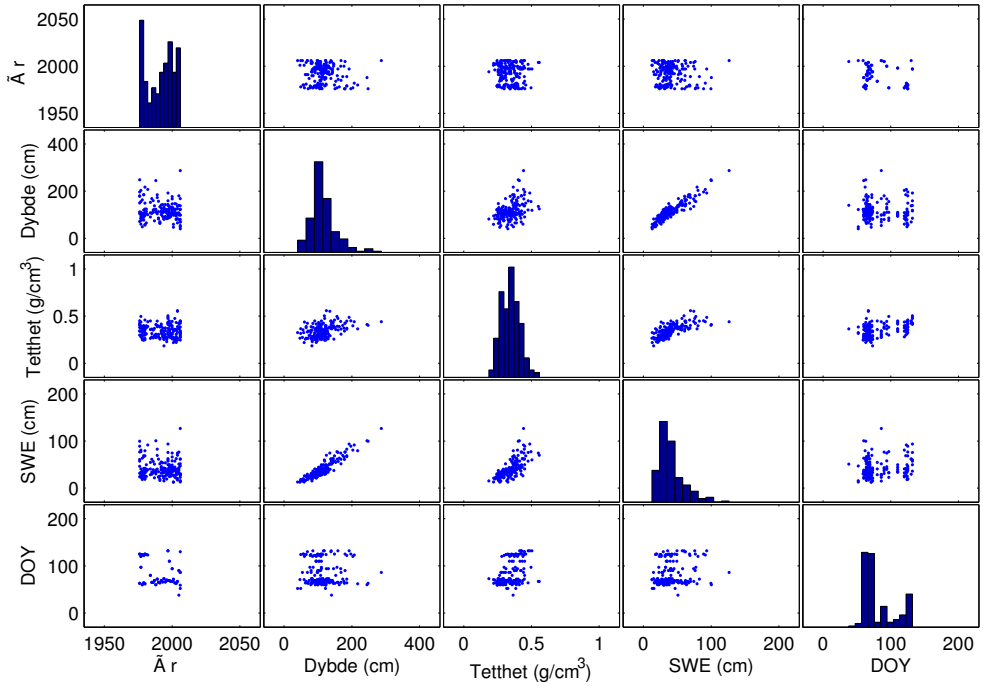
Figur 3 – Box plot av snødybde og snøtetthet ved Rybekken og Sylsjødammen, med én boks for hvert år. Plottet er tegnet horisontalt for å utnytte plassen bedre, og boksene er tegnet på en mer kompakt måte enn de i figur 2. Figuren er laget med `boxplots.m`.

hvor $\bar{X}_1$ og $\bar{X}_2$ er gjennomsnittsverdiene av X_1 og X_2, s_1^2 og s_2^2 er de forventningsrette estimatorene for σ_1^2 og σ_2^2, $n_1 = 103$ og $n_2 = 90$, tilnærmet en t-fordeling med v frihetsgrader, hvor v er gitt av Satterthwaites approksimasjon

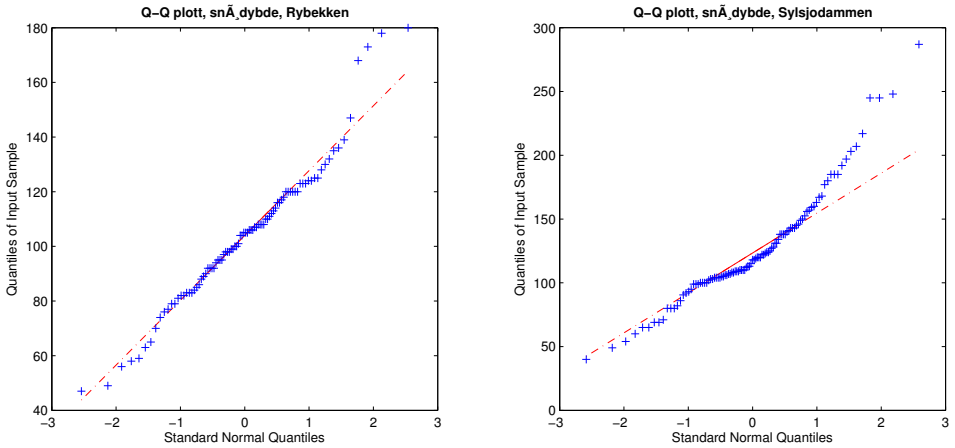
$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

som typisk rundes ned til nærmeste heltall.

Setter opp nullhypotese og alternativ hypotese:



Figur 4 – Scatter plot-matrise for kombinerte data fra Rybekken og Sylsjødammen. De tydeligste sammenhengene er mellom *SWE* og dybde, og mellom *SWE* og tetthet. Figuren er laget med `scatterplotmatrise.m`.



Figur 5 – Q-Q plott av snødybdedata fra Rybekken (venstre) og Syldsjødammen (høyre).

- H_0: $\mu_1 - \mu_2 = 0$
- H_a: $\mu_1 - \mu_2 > 0$

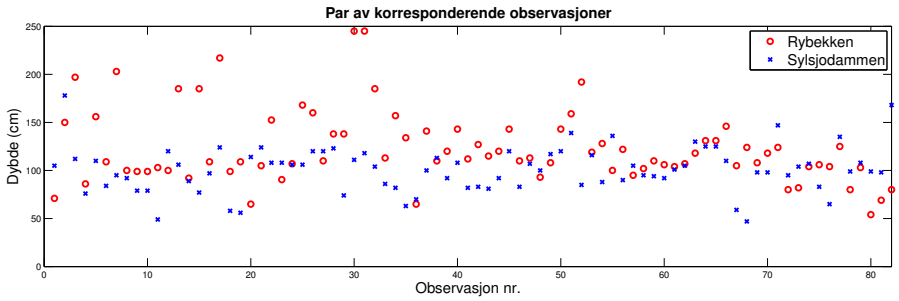
Vi utfører altså en ensidig test, hvor vi ser på den høyre halen til fordelingen til T, og bruker følgende kriterium: Forkast H_0 på signifikansnivå α dersom den observerte verdien av T er større enn den kritiske verdien $t_{1-\alpha,v}$. Et script som utfører testen er lagret i Matlabfilen `mutest.m`. Vi finner at antall frihetsgrader er $v = 168$, som på signifikansnivå $\alpha = 0.05$

gir den kritiske verdien $t_{1-\alpha,v} = 1.6540$. Den observerte verdien er imidlertid $T = 4.3522$, så vi forkaster H_0 på signifikansnivå 0.05. Forøvrig finner vi P-verdien $1.2 \cdot 10^{-5}$, og kan slå fast med ganske stor sikkerhet at det i gjennomsnitt var dypere snø ved Sylsjødammen enn ved Rybekken i perioden 1976-2006.

Her er det, på grunn av det forholdsvis store antallet observasjoner, liten forskjell på å utføre testen med en t-fordelt variabel, og å bruke en normalfordelt variabel.

Konfidensintervall for forskjellen i forventningsverdi av snødybde

Vi vil se på to ulike metoder for å finne et $100(1-\alpha)\%$ konfidensintervall for $\mu_1 - \mu_2$. De fleste av observasjonene ved Rybekken og Sylsjødammen er tatt samme dag. Hvis vi fjerner alle observasjonene fra dager hvor det ikke ble foretatt målinger ved begge stasjonene, gjenstår 82 observasjoner i hvert datasett. Vi kan se på disse som 82 par av observasjoner, hvor hvert par består av en måling fra hver stasjon, tatt på samme dag. Disse forkortede datasettene er lagret i filene `rybekken_kort.txt` og `sylsjodammen_kort.txt`. Dataene er plottet i figur 6.



Figur 6 – 82 par av korresponderende observasjoner av snødybde fra Rybekken og Sylsjødammen. Observasjonene i hvert par er målt samme dag. Figuren er laget med `parplott.m`.

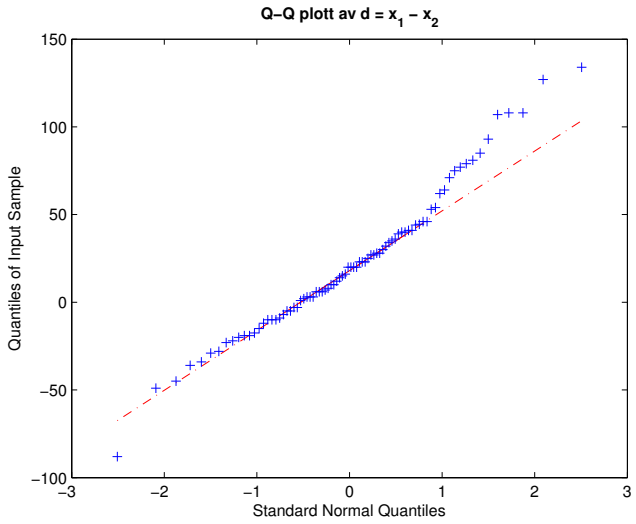
Vi kan enten gjøre som for hypotesetesten over, og anta to uavhengige normalfordelte tilfeldige variable X_1 og X_2, eller vi kan se på den tilfeldige variabelen $D = X_1 - X_2$, (se kap. 9.9 i W.M.M.Y.). Dette innebærer antakelsen at realisasjonene $d_1, d_2, \ldots, d_{82}$ av D kommer fra en normalfordeling. Vi lager et Q-Q plott (figur 7) for å kontrollere dette.

Dersom det er avhengighet mellom X_1 og X_2, vil sammenslåing av observasjoner i par gi mindre varians, men samtidig redusere antall frihetsgrader, slik at det ikke alltid er en fordel. I dette tilfellet får vi, ved å bruke Matlabkoden i fila `muci.m`, resultatene gitt i tabell 3. Her gir altså tilnærmingen med observasjoner i par et litt smalere konfidensintervall.

Tabell 3 – 95% konfidensintervall og antall frihetsgrader for forskjellen mellom μ_1 og μ_2 utledet på to forskjellige måter. For metoden med separate populasjoner er antall frihetsgrader regnet ut med Satterthwaites approksimasjon, på samme måte som for hypotesetesten over. Samme datasett er brukt for begge metodene, dvs. det forkortede datasettet er brukt, også for separate populasjoner.

Metode	95% konfidensintervall	Antall frihetsgrader
Separate populasjoner	(12.3447, 31.8992)	135
Observasjoner i par	(13.0800, 31.1639)	81

Konklusjonen blir den samme som for hypotesetesten. Siden 95%-konfidensintervallet



Figur 7 – Q-Q plott av differansene $d_i = x_{1i} - x_{2i}$, $i = 1, \ldots, 82$.

ikke inneholder null, så er forskjellen mellom forventningsverdiene signifikant forskjellig fra null på signifikansnivå 5%, dvs. det er en signifikant forskjell mellom forventningsverdiene.

Sammenlikning av varians

(W.M.M.Y. Kap. 8.7 (F-fordeling) og 9.13 (sammenlikning av varians), NB: Ikke pensum i TMA4240/4245)
Figur 2 (venstre) indikerer at snødybdedataene fra Sylsjødammen har større varians enn de fra Rybekken. Vi vil utlede et konfidensintervall for forholdet σ_1^2/σ_2^2 mellom disse. Ifølge teorem 8.8 i W.M.M.Y. følger den tilfeldige variabelen

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

en F-fordeling med $v_1 = n_1 - 1$ og $v_2 = n_2 - 1$ frihetsgrader. Finner et $100(1-\beta)\%$ konfidensintervall for forholdet mellom variansene.

$$P\left(f_{\frac{\beta}{2},v_1,v_2} \leq \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \leq f_{1-\frac{\beta}{2},v_1,v_2}\right) = 1-\beta$$

$$\Rightarrow P\left(\frac{s_1^2/s_2^2}{f_{1-\frac{\beta}{2},v_1,v_2}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2/s_2^2}{f_{\frac{\beta}{2},v_1,v_2}}\right) = 1-\beta$$

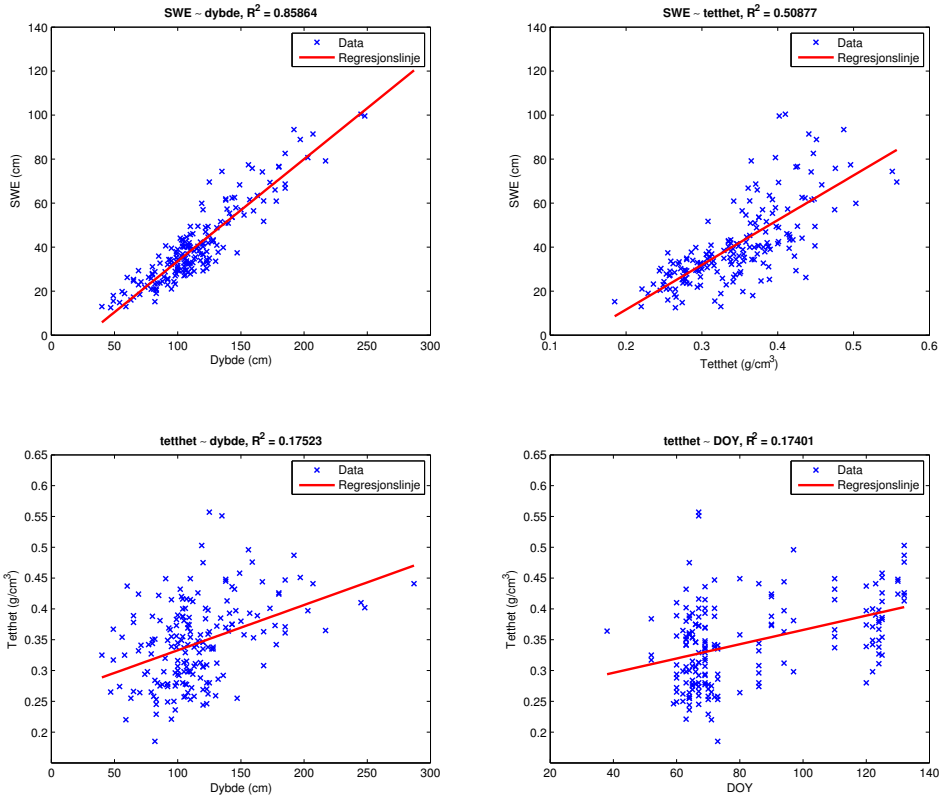
Konfidensintervallet blir

$$\left(\frac{s_1^2/s_2^2}{f_{1-\frac{\beta}{2},v_1,v_2}}, \frac{s_1^2/s_2^2}{f_{\frac{\beta}{2},v_1,v_2}}\right).$$

Ved å bruke Matlabkoden i fila `varci.m` finner vi estimatet $s_1^2/s_2^2 = 2.9123$, og 95% konfidensintervallet $(1.9380, 4.3519)$. Hvis de to populasjonene hadde samme varians $\sigma_1^2 = \sigma_2^2$, ville forholdet mellom dem vært lik 1. Siden konfidensintervallet for σ_1^2/σ_2^2 ikke inneholder 1 kan vi si, med 95% konfidens, at σ_1^2 er større enn σ_2^2, dvs. Sylsjødammen-dataene har større varians enn Rybekken-dataene.

Enkel lineær regresjon

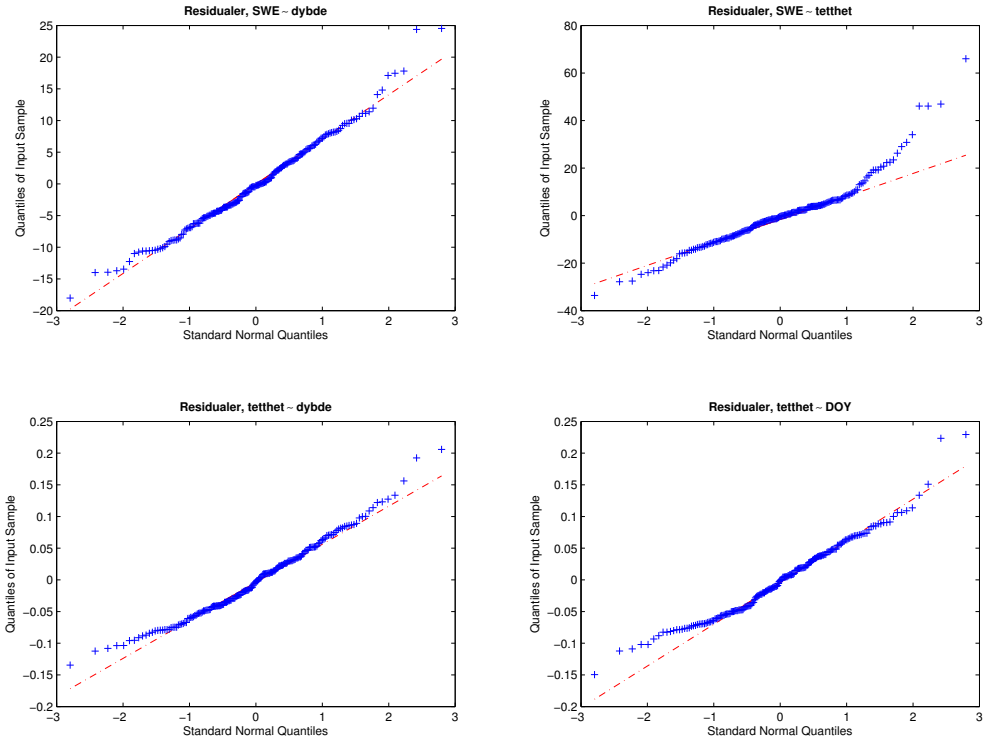
(W.M.M.Y. Kap. 11)
Utfra figur 4 ser en tydelig korrelasjon mellom SWE og dybde, og mellom SWE og tetthet. Dette er å vente, siden SWE er proporsjonal med produktet av dybde og tetthet. Vi vil finne ut hvor mye av variasjonen i SWE som kan forklares ved regresjon på kun én av disse variablene. Enkel lineær regresjon av SWE på snødybde og snøtetthet gir resultatene vist i figur 8. I tillegg ses resultatene av regresjon av snøtetthet på snødybde og DOY. Regresjon



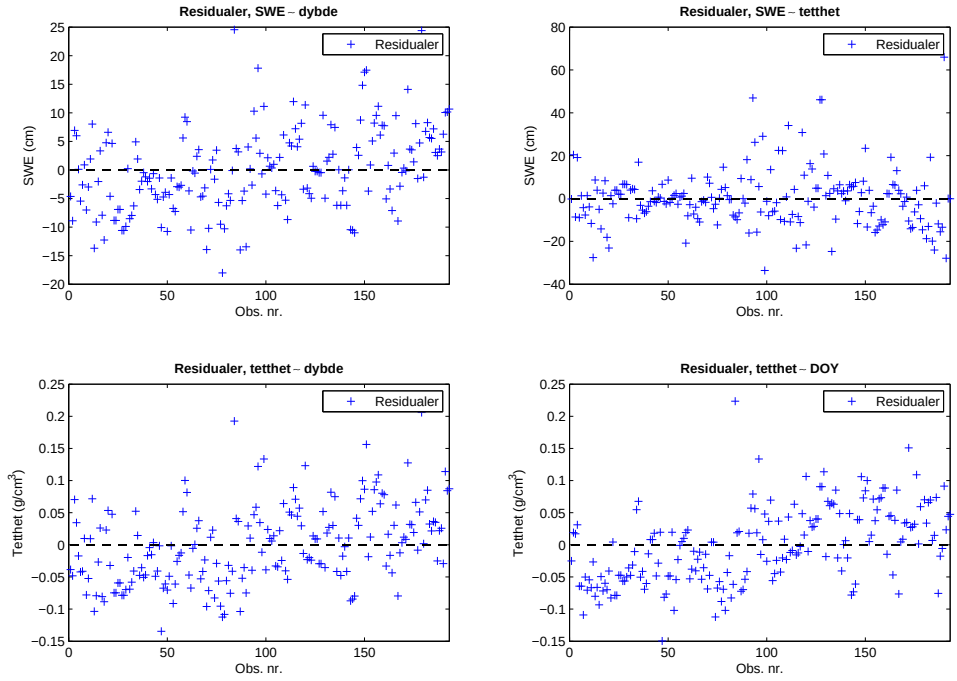
Figur 8 – Datapunkter og regresjonslinjer funnet med minste kvadraters metode for kombinerte data fra Rybekken og Sylsjødammen. Øverst, venstre: Regresjon av SWE på snødybde. Øverst, høyre: Regresjon av SWE på snøtetthet. Nederst, venstre: Regresjon av snøtetthet på snødybde. Nederst, høyre: Regresjon av snøtetthet på snøalder (DOY). Figuren er laget med `regresjon.m`

av SWE på snødybde gir $R^2 = 0.8586$, mens regresjon på snøtetthet gir $R^2 = 0.5088$. Vi kan altså forklare mer av variasjonen i SWE utfra avhengighet av dybde alene, enn utfra avhengighet av kun tetthet.

For å kontrollere antakelsen om normalfordelte residualer, er Q-Q plott av de fire settene med residualer plottet i figur 9. Antakelsene ser her ut til å holde bra, eventuelt med unntak av regresjon av SWE på tetthet, dvs. øverst til høyre på figur 9. Videre er residualplott i de fire tilfellene vist i figur 10. Disse kan brukes for å vurdere om antakelsen om konstant varians er oppfylt.



Figur 9 – Q-Q plott av residualer etter regresjon, jf. figur 8.



Figur 10 – Plott av residualer etter regresjon, jf. figur 8.