TMA4240 Statistikk
Høst 2018

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Anbefalt øving 1
Løsningsskisse

Oppgave 1

a) I) Kontinuerlige variabler: Andel stryk i %, andel jenter i %, og andel A i %.
Diskrete variabler: Årstall, kurs, og karakter

II) Vi leser av histogrammet at omtrent 125 fikk karakteren A (eksakt antall er 127).

III) Det er tydelig at karakteren C var vanligst, og dermed moden, siden den høyeste søylen i histogrammet tilhører C .

b) I) Høsten 2014 var det flest som fikk E. I 2013 var det flest som fikk C, omtrent like mange som fikk B og D, færrest fikk E og nest flest fikk A. Kanskje eksamen i 2014 var for vanskelig?

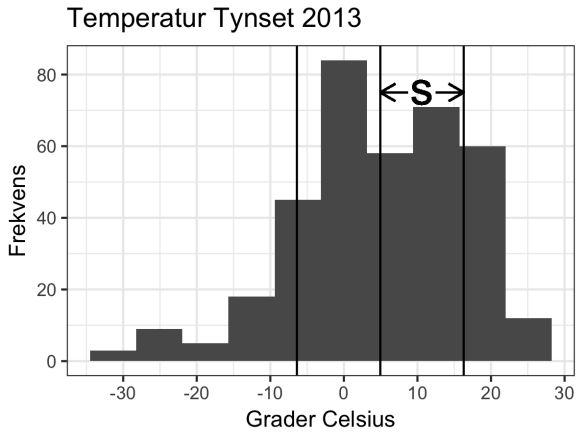
II) Gjennomsnittskarakteren er helt klart C. Karakterfordelingen over 18 år ligner noe på en normalfordeling, og er ganske symmetrisk (liten overvekt av D og E sammenlignet med B og A), men det ser ut til å være for mange som fikk C til at vi er helt fornøyde med å anta en normalfordeling.

Oppgave 2

a) I) Gjennomsnitt påvirkes av ekstremverdier, fordi verdien til alle observasjonene (temperaturene) er med i utregningen. Observasjonene er bare indirekte med i utregningen av medianen, fordi vi sorterer dataene og deretter finner den i midten. Da har det ingenting å si hvor stor den største verdien er. Medianen er altså mer robust mot ekstremverdier enn gjennomsnittet.

II) Det eksakte standarddavviket er 11.34. Et så nøyaktig tall kommer vi ikke fram til ved å se på histogrammet, men vi kan komme nærme. Først ser vi at det er større spredning i Tynsetdataene enn i Trondheimdataene ved å se på forskjell mellom minste og største temperatur i histogrammene, så vi tror standarddavviket vil være større i Tynsetdataene. Det ser ut til at standarddavviket er ca. 8 i Trondheimsdataene. Så ser vi for oss at vi legger en normalfordeling oppå histogrammet, med gjennomsnitt omtrent 5 (oppgitt i tabell). Da kan vi se at et standarddavvik mellom 10 og 15 kan passe fint (det er vanskelig å si noe mer nøyaktig enn dette kun basert på histogrammet). Vi kan også se at standarddavviket er mindre enn 20, for da vil gjennomsnittet pluss/minus ett standarddavvik dekke nesten alle

observasjonene, som vi kan se at ikke er tilfellet i f.eks. normalfordelingen. Figur 1 viser et histogram over temperatur fra Tynset der gjennomsnitt og standarddavvik er tegnet med linjer.



Figur 1: Histogram over temperaturobservasjoner fra Tynset i 2013 med gjennomsnitt og standarddavvik tegnet inn.

b) I) Eksakt stigningstall er 4.85. Ved å se på grafen ser vi at i løpet av måned stiger linja omtrent 5 grader (bruk de tynne grå linjene i plottet!), så et godt anslag på stigningstallet er 5.

II) Her betyr et stigningstall på 5 at i løpet av én måned vil det i gjennomsnitt (for dette er jo bare en tilpasning! Og ikke sannheten) bli 5 grader varmere. Så det er i gjennomsnitt 5 grader varmere 1. februar enn det var 1. januar.

III) Det ser ut som at den rette linjen passer godt til observasjonene. En rett linje vil aldri kunne tilpasse seg alle punktene, da må den være hakkete rundt februar og mars (måned 2 og 3), men ingen punkter er veldig langt unna linja. Så vi er hverken fornøyde, eller misfornøyde, og kan vi at linja passer middels godt til dataene.

c) I) I mai er spredningen størst (boksen er størst/lengst).

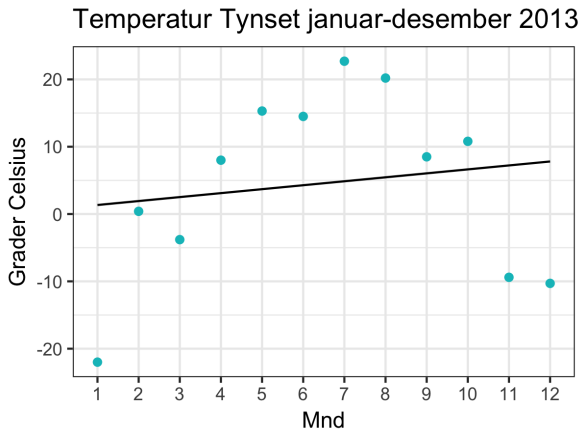
II) Basert på boksplottet er spredningen minst i oktober (boksen er minst/kortest), med mars hakk i hæl (her er det vanskelig å skille mellom dem).

Bonus: Lengden på selve boksen i et boksplott bestemmes av noe som heter interquartile range (IQR). Det er 75 %-percentilen minus 25 %-percentilen, og representerer de midterste 50 % av dataene.

d) I) Ved å bruke samme metode som i oppgave 1 c), kan vi anslå at stigningstallet er ganske midt mellom 5 og 10, men nærmere 5 enn 10, så 7 er et godt anslag. Eksakt verdi er 6.48.

II) For å finne ut hvordan en rett linje som passer dataene best mulig (en regresjonslinje) ville sett ut for alle observasjonene, kan vi tenke oss at vi har observasjonene fra januar til juli, og den rette linja fra oppgaven. Deretter legger vi til observasjonen fra 23. august også, og må tilpasse linja etter det. Den nye observasjonen ligger litt under den fra 23. juli, så vi tenker oss at linja vil få et litt lavere stigningstall,

vi må legge den litt ned. Så tar vi med 23. september, og må nå flytte mye mer på linja da det var mye kaldere da enn i august. Sånn holder vi på til til vi kommer til 23. desember. Figur 2 viser regresjonslinja for alle månedene. En rett linje passer veldig dårlig, fordi temperaturen svinger opp og ned gjennom året.
Uten å vite noe særlig om dataene, kan vi anta at en regresjonslinje vil ha stigningstall nært 0, og et gjennomsnitt nært gjennomsnittet til datapunktene. Så en linje med stigningstall 0 og skjæringspunkt 5 er en god antagelse.
Linja i Figur 2 har skjæringspunkt 0.75 og stigningstall 0.59.



Figur 2: Flat regresjonslinje (den rette linja som passer best til observasjonene).

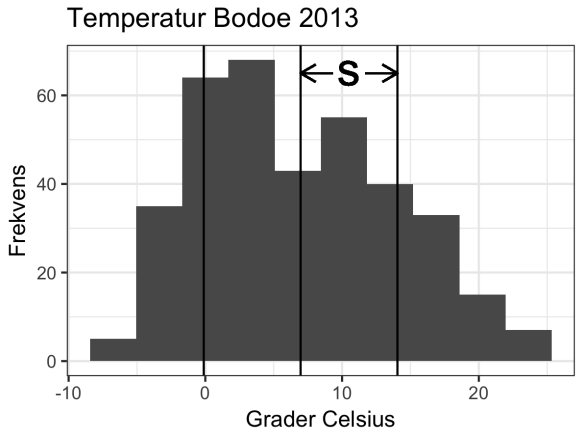
e) I) De eksakte verdiene er: Gjennomsnitt = 6.97, median = 5.6, standarddavvik = 7.08. Ved å se på grafen, kan vi se for oss at midten ligger mellom 5 og 10 grader, men nærmere 5 enn 10. Da vil vi tenke at medianen og gjennomsnittet er her et sted. Siden disse tallene var like for Trondheim og Tynset, antar vi at de er det her også (selv om dette faktisk ikke stemmer helt, men det er ikke så lett å se). Spredningen i dataene er litt mindre enn for Trondheim, så vi tror standarddavviket er litt mindre enn 10, men vi har såpass mye spredning at vi tror det må være et standarddavvik større enn 5. Da har vi altså et gjennomsnitt/median mellom 5 og 10 (nærmere 5 enn 10), og standardavvik mellom 5 og 10.

f) I) Det er størst spredning i januar.

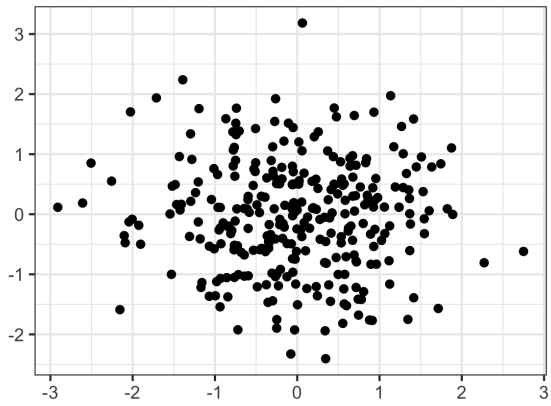
II) Det er minst spredning i juni.

g) I) Plottene viser at varmere temperatur i Trondheim betyr varmere temperatur på Tynset og i Bodø. Så det er en positiv trend her, noe som er ganske forventet (om sommeren er det varmere enn om vinteren, og det gjelder i hele Norge).

II) Det er en sterk avhengighet i begge plot. Hvis temperaturene var uavhengige ville vi ikke sett et mønster i dataene, det ville sett ut som helt tilfeldig spredte punkter, se Figur 4, som viser kaos".



Figur 3: Histogram over temperaturobservasjoner fra Bodø i 2013 med gjennomsnitt og standarddavvik tegnet inn.



Figur 4: Illustrasjon av to uavhengige variabler.