# TMA4240 Statistikk
# Høst 2018

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

**Anbefalt øving 1**

I denne øvingen skal vi analysere to ulike datasett, ett datasett med karakterstatistikk for TMA4240/TMA4245, og ett datasett med temperaturobservasjoner for Trondheim, Tynset og Bodø. Temperaturdataene er hentet fra eklima.no. Øvingen består av to oppgaver der din jobb er å tolke resultater.

## Oppgave 1

I denne oppgaven skal vi analysere et datasett med karakterstatistikk for faget TMA4240/TMA4245 Statistikk ved NTNU i perioden 2004 - 2013.

Datasettet inneholder følgende variabler:

- År: 2004 - 2013 (vår og høst alle år unntatt 2004 (kun høst) og 2013 (kun vår))
- Kurs: 1 = TMA4240 (høst), 2 = TMA4245 (vår)
- Antall av hver karakter (A,B,C,D,E)

**a**) Vi har to ulike typer variabler i dette datasettet - diskrete og kontinuerlige variabler. En diskret variabel kan bare ta bestemte verdier og vi kan telle opp hvor mange observasjoner vi har for hver mulige verdi/kategori. Eksempler på diskrete variabler er karakterene. En kontinuerlig variabel kan ta verdier i et gitt intervall, f.eks temperatur.

I) Hvilke variabler i datasettet er kontinuerlige? Hvilke er diskrete?

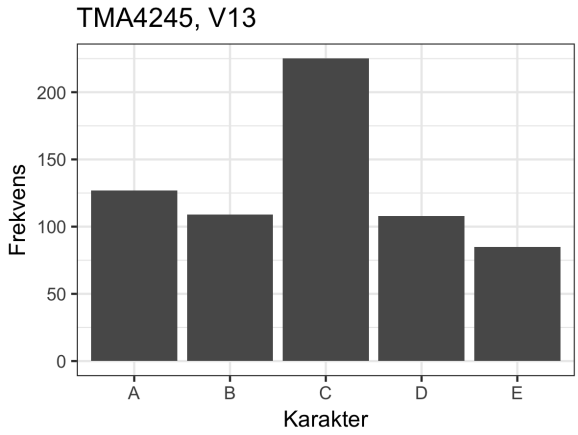
Vi har karakterdata fra 18 ulike eksamener. Merk at vi fra nå av ikke bruker alle variablene i datasettet.

II) Figur 1 viser karakterfordelingen for kurset TMA4245 våren 2013. Hvor mange fikk karakteren A?

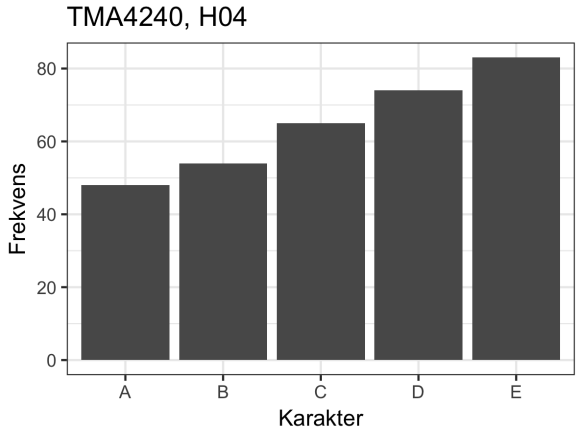
III) Hvilken karakter var vanligst, dvs., hva er moden/typetallet?

**b**) Her ser vi på karakterfordelingene i TMA4240/4245 igjen.

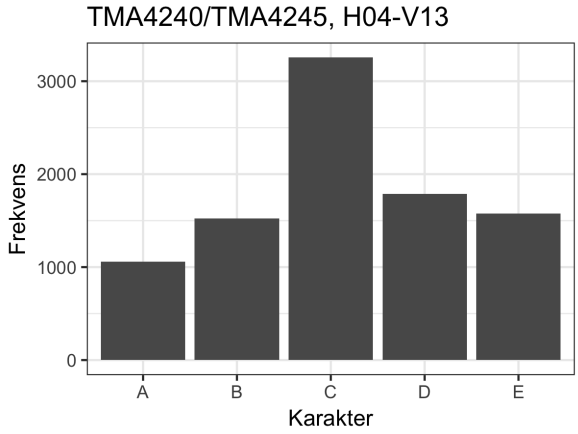
I) Se Figur 2: Hvordan var karakterfordelingen i 2004 sammenlignet med fordelingen i 2013 (oppgave 1a)?




Figur 1: Histogram med karkakterfordeling i TMA4245 vår 2013.

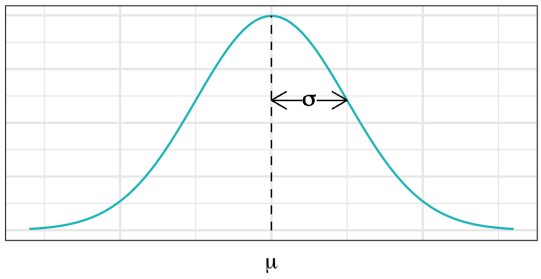



Figur 2: Histogram med karakterfordeling: TMA4240 høst 2004.




Figur 3: Histogram med karakterfordeling: TMA4240/4245 fra 2004 til 2013.

II) Se Figur 3: Hva er gjennomsnittskarakteren? Figur 4 viser en normalfordeling med




Figur 4: Normalfordeling med gjennomsnitt $\mu$ og standardavvik $\sigma$.

gjennomsnitt $\mu$ og standarddavvik $\sigma$ som ofte blir tilpasset unimodale og symmetriske fordelinger. Sammenlign karakterfordelingen for alle 18 semesterene (2004-2013) med normalfordelingen i Figur 4. Vi vil diskutere normalfordelingen mer senere i kurset. Basert på sammenligningen; er det rimelig å anta at karakterene er tilnærmet normalfordelt? Beskriv kort forskjellene.

**Oppgave 2**

I denne deloppgaven skal vi analysere data med temperaturobservasjoner for Tronheim og Tynset i perioden 01.01.2013 til 31.12.2013. Disse datasettene inneholder følgende variabler:

- Måned: 1 - 12
- Dag: 1 - 31
- Temperatur: °C

**a)** Tabell 1 viser gjennomsnittstemperaturen og mediantemperatur for året 2013 i både Trondheim og Tynset.

| | **Gjennomsnitt** | **Median** |
|---|---|---|
| **Trondheim** | 7.77 | 7.4 |
| **Tynset** | 4.95 | 5.1 |

Tabell 1: Gjennomsnittstemperatur og mediantemperatur i grader Celsius.

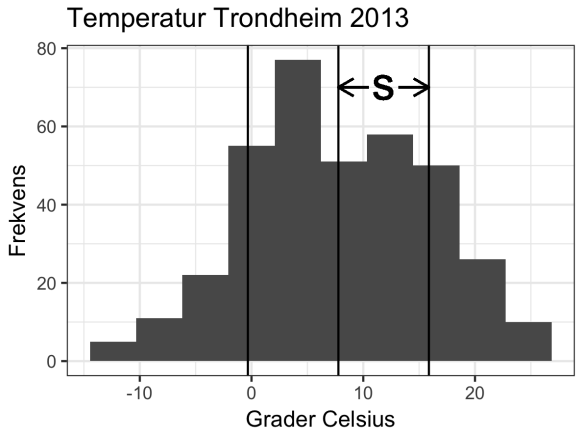
Vi ser altså at gjennomsnittstemperaturen for denne perioden er 7.8 °C i Trondheim og omtrent 5 °C på Tynset. Vi ser også at gjennomsnittstemperaturen og medianen er tilnærmet lik både i Trondheim og på Tynset.

I) Hvordan påvirkes gjennomsnitt og median av ekstreme observasjoner?

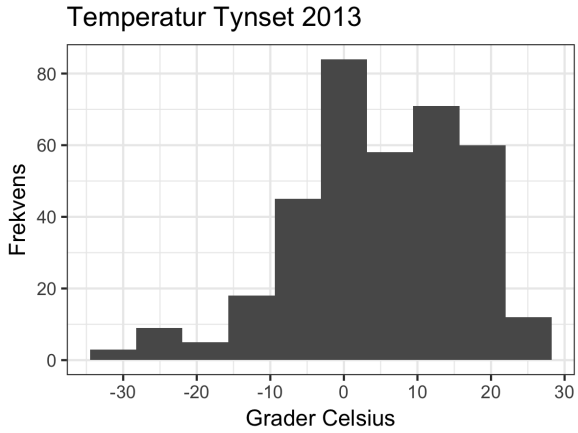
Vi ønsker å se på variasjonen i temperaturobservasjonene; hvor mye enkeltobservasjoner varierer rundt gjennomsnittsverdien, $\mu$. Standardavviket, $\sigma$, er et mål på spredningen til observasjonene i et datasett og er definert som kvadratroten til den variansen (mer

om dette senere i kurset). Figur 4 viser en normalfordeling med gjennomsnitt $\mu$ og standarddavvik $\sigma$.

II) Figur 5 viser temperaturen i Trondheim i et histogram, og gjennomsnittet og det empiriske standardavviket er indikert på figuren. Figur 6 viser tilsvarende histogram for Tynset, men uten empirisk gjennomsnitt og standarddavvik. Hva vil du anslå at det empiriske standarddavviket for temperaturen på Tynset er (obs: medianen og standarddavviket for Tynset ligger rundt 5 grader)?



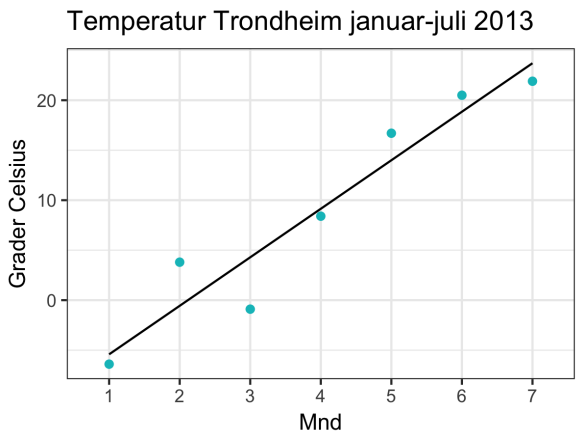

Figur 5: Temperaturobservasjoner Trondheim.




Figur 6: Temperaturobservasjoner Tynset.

**b**) Vi ønsker nå å se på temperaturutviklingen i Trondheim fra januar til juli 2013, og velger å se på temperaturobservasjonene fra 23. januar, 23. februar, 23. mars, 23. april, 23. mai, 23. juni og 23. juli. Senere i dette kurset skal vi lære hvordan vi kan finne den tilpassede linja (dette kalles lineær regresjon).

Figur 7 viser de 7 temperaturobservasjonene og den tilhørende rette linja som passer best. Skjæringspunktet på y-aksen er -10.3.



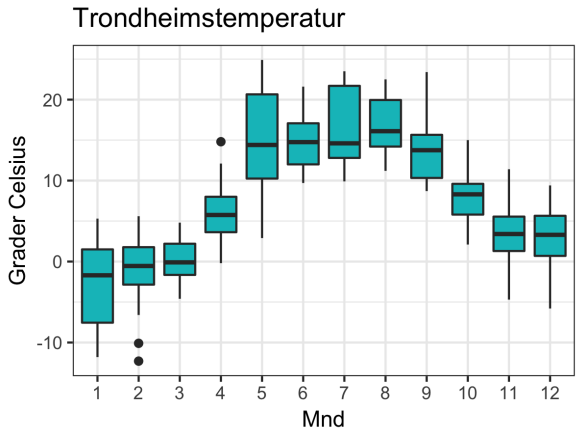

Figur 7: Temperaturobservasjoner i Trondheim, med den tilpassede linjen (den rette linjen som passer best til observasjonene).

I) Hva vil du anslå at stigningstallet er?

II) Hva betyr dette tallet i denne sammenhengen?

III) Hvordan passer den tilpassede linja til observasjonene?

**c**) For å se på spredningen i dataene for ulike kategorier av en diskret variabel kan vi lage et boksplott som viser median, kvartiler og ekstremobservasjoner fordelt på ulike kategorier av den diskrete variabelen, f.eks måned. Figur 8 viser et boksplott av temperaturen i Trondheim fordelt på måneder.



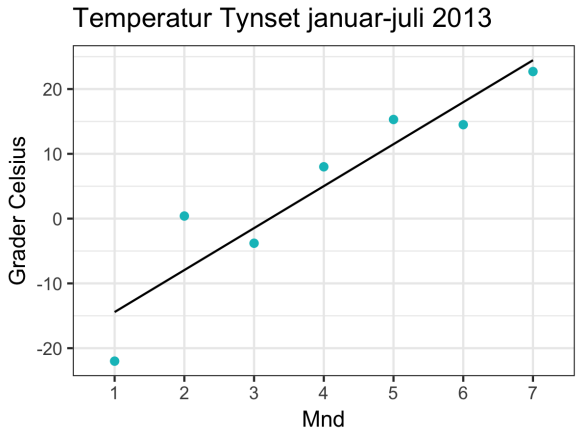

Figur 8: Boksplott av temperaturobservasjoner i Trondheim.

Standarddavviket til Trondheimstemperaturen i januar er 5.4, og i mars er standarddavviket 4.14. Vi kan se fra boksplottet at det er mindre spredning i temperaturen i mars enn i januar.

I) I hvilken måned er temperaturvariasjonen i Trondheim størst ifølge boksplottet?

II) I hvilken måned er temperaturvariasjonen i Trondheim minst ifølge boksplottet?

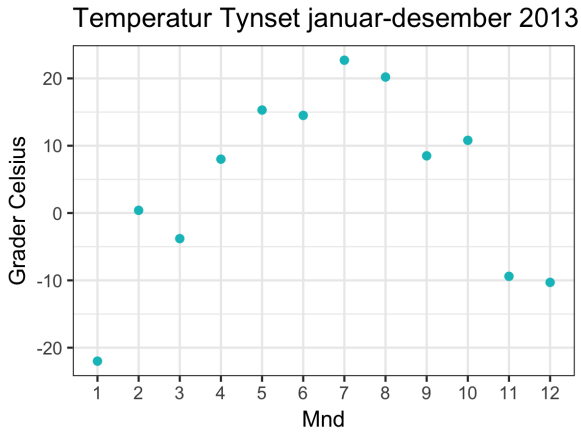
**d**) I oppgave 2c) så vi på temperaturobservasjoner for Trondheim fra januar til juli 2013, den 23. hver måned. Vi skal nå se på temperaturobservasjoner fra Tynset i den samme tidsperioden. Figur 9 viser et plot med de 7 temperaturobservasjonene sammen med den rette linja som passer dataene best. Skjæringspunktet med y-aksen er -20.9 grader.




Figur 9: Temperaturobservasjoner på Tynset, med regresjonslinje (den rette linjen som passer best til observasjonene).

I) Hva vil du anslå at stigningstallet er?

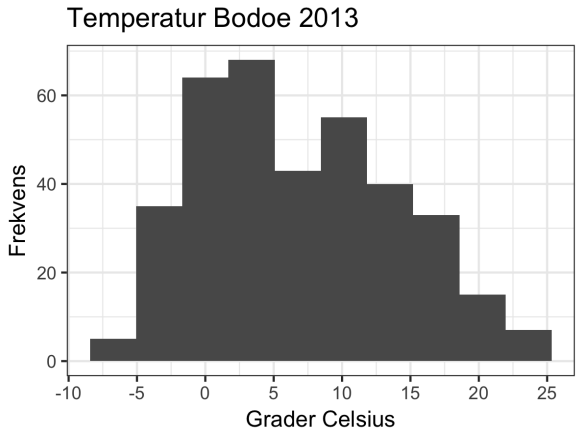
II) Hvordan ville den rette linja lagt seg om vi hadde tatt med observasjoner fra den 23. i alle årets måneder? (Hint: Figur 10 inneholder temperaturobservasjoner fra den 23. hver måned i 2013. Hvordan ville du trukket en linje i dette plottet som skal passe best mulig til observasjonene?)




Figur 10: Temperaturobservasjoner på Tynset.

**e**) I) Figur 11 viser et histogram med temperaturobservasjonene i Bodø i 2013 (tilsvarende histogrammer for Trondheim og Tynset finnes i henholdsvis Figur 5 og 6, og kan brukes til sammenligning. Datasettet for Bodø inneholder samme type informasjon

som datasettene med temperatur fra Trondheim og Tynset). Anslå gjennomsnitt, median og standarddavvik for dette histogrammet.

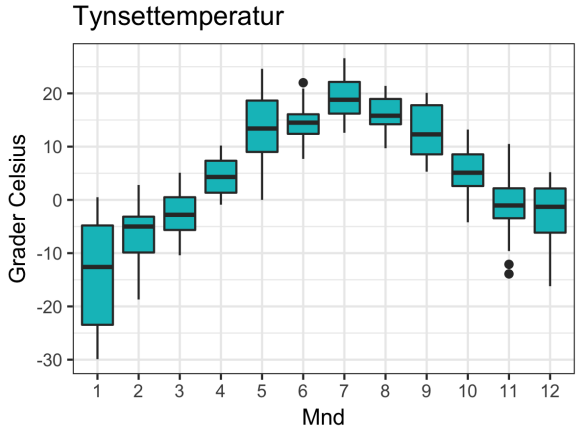



Figur 11: Histrogram over temperaturobservasjoner fra Bodø.

**f**) Figur 12 viser et boksplot med temperaturobservasjonene fra Tynset, gruppert etter måned (tilsvarende plot for Trondheim så vi på i oppgave 1d)).

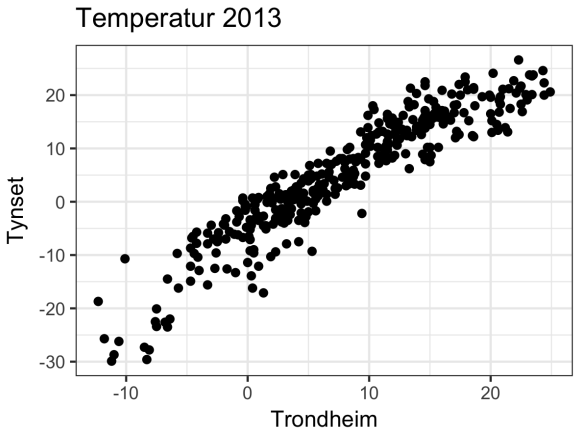
I) Bruk interkvartilbredden til å anslå i hvilken måned er temperaturvariasjonen på Tynset størst ifølge boksplottet?

II) Bruk interkvartilbredden til å anslå i hvilken måned er temperaturvariasjonen på Tynset minst ifølge boksplottet?

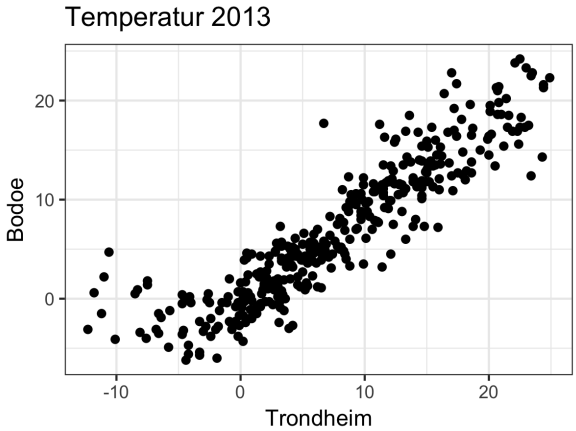



Figur 12: Boksplott av temperaturobservasjoner på Tynset.

**g**) Vi skiller mellom avhengige og uavhengige observasjoner. Vi vil nå se på avhengigheten i temperaturobservasjonene fra Trondheim, Tynset og Bodø i 2013. I Figur 13 har vi plottet temperaturen på Tynset mot temperaturen i Trondheim, og i Figur 14 har vi plottet temperaturen i Bodø mot temperaturen i Trondheim.




Figur 13: Temperaturer i 2013: Tynset mot Trondheim.




Figur 14: Temperaturer i 2013: Bodø mot Trondheim.

I) Ser du en trend i de to plottene?

II) Hva kan du si om avhengigheten i temperaturobservasjonene basert på disse plottene?

## Fasit