Øving 1 TMA4240 - Grunnleggende dataanalyse i Matlab

For grunnleggende bruk av Matlab vises til slides fra basisintroduksjon til Matlab som finnes på kursets hjemmeside.

I denne øvingen skal vi analysere to ulike datasett, ett med karakterstatistikk for TMA4240/TMA4245 og ett datasett med temperaturobservasjoner for Trondheim, Tynset og Bodø. Temperaturdataene er hentet fra `eklima.no`. Datasettene kan lastes ned fra kursets hjemmeside. Øvingen består av to oppgaver, i oppgave 1 vil alle Matlab - kommandoer bli vist samt noe tolkning av resultatene. Både oppgave 1 og 2 inneholder oppgaver som dere kan besvare i grupper på 2-3 studenter. Innlevering av øving 1 eller 7 er obligatorisk for å få gå opp til eksamen.

Datafilene leses inn til Matlab med følgende kommandoer

```
grunnkurs = load('tma42404245.txt');
trondheim = load('trondheim.txt');
bodo = load('bodo.txt');
tynset  = load('tynset.txt');
```

Ved å definere egne matriser for datafilene slipper vi å lese inn datafila til Matlab hver gang den skal brukes. Dataene fra filen `tma42404245.txt` finner vi nå i matrisa `grunnkurs`.

Oppgave 1

a)

I denne oppgaven skal vi analysere et datasett med karakterstatistikk for faget TMA4240/TMA4245 Statistikk ved NTNU i perioden 2003 - 2013.
Datasettet inneholder følgende variabler

i År - 2003 -2013 (kolonne 1)

ii kurs - 1=TMA4240, 2=TMA4245 (kolonne 2)

iii Andel stryk i % (kolonne 3)

iv Andel jenter i % (kolonne 4)

v Andel A i % (kolonne 5)

vi Karakterer (A,B,C,D,E) (kolonne 6-10) og for jenter (Aj,Bj,Cj,Dj,Ej) (kolonne 11-15)

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 `karakterer`. En kontinuerlig variabel kan ta verdier i et gitt intervall, f.eks `temperatur`.

I Hvilke variabler i datasettet `tma42404245.txt` er kontinuerlige? Hvilke er diskrete?

Dimensjonen til matrisa kan vi finne med kommandoen `size` i Matlab,

```
>> size(grunnkurs)
ans =
        19    15
```

som viser at matrisa har 19 rader og 15 kolonner. Dataene for våren 2013 finner vi i rad 19, og vi kan da opprette en ny vektor med dataene for 2013

```
>> v13 = grunnkurs(19,:);
```

Karakterfordelingen 'A-E' finner vi da som element 6-10 i denne vektoren, og vi kan skrive ut dette slik

```
>> v13(6:10)
ans =
    127   109   225   108    85
```

Vi kan plotte et histogram for karakterfordelingen våren 2013 med følgende kommandoer i Matlab

```
y = grunnkurs(19,6:10);      %henter ut alle data for TMA4245 V2013
X = {'A';'B';'C';'D';'E'};   %x-akse
bar(y);
set(gca,'XTickLabel',X);
xlabel('Karakter');
ylabel('Frekvens');
title('TMA4245 V13');
```

II Lag histogram over karakterfordelingen for kurset TMA4245 våren 2013.

b)

I denne oppgaven skal vi analysere data med temperaturobservasjoner for Trondheim og Bodø i perioden 01.01.2013 til 31.12.2013. Disse datasettene inneholder følgende variabler

i Stnr - Stasjonsnummer for Trondheim og Bodø (kolonne 1)

ii År - 2013 (kolonne 2)

iii Mnd - 1 - 12 (kolonne 3)

iv Dag - 1-31 (kolonne 4)

v Klokkeslett -14 (kolonne 5)

vi Temperatur - °C (kolonne 6)

Vi ønsker nå å se på temperaturen i Trondheim og Bodø, og oppretter to vektorer med temperaturdataene for Trondheim og Bodø

```
ttemp = trondheim(:,6);
btemp = bodo(:,6);
```

Gjennomsnitt og spredning

Vi kan se på gjennomsnittstemperaturen og median for året 2013 i både Trondheim og Bodø,

```
>> mean(ttemp)
ans = 7.7732

>> mean(btemp)
ans = 6.9674
```

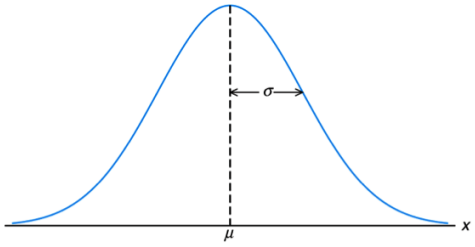
```
>> median(ttemp)
ans = 7.4000

>> median(btemp)
ans = 5.6000
```

Vi ser altså at gjennomsnittstemperaturen for denne perioden er 7.8 °C i Trondheim og $\sim$ 7 °C i Bodø Vi ser også at gjennomsnittstemperaturen og medianen i Trondheim er tilnærmet lik, mens gjennomsnittstemperaturen i Bodø er noe høyere enn medianen.

I Hvordan påvirkes gjennomsnitt og median av ekstreme observasjoner?

II Vi kan f.eks se på dette ved å skrive inn en feil i datafila, f.eks endre observasjon nr 361 fra 4.5 til 450 (btemp(361)=450;). Hvordan påvirker dette gjennomsnitt og median? Husk å endre observasjonen tilbake til opprinnelig verdi, dvs btemp(361)=4.5;

Figure 6.2 The normal curve



Copyright © 2010 Pearson Addison-Wesley. All rights reserved.

Figur 1: Normalfordeling og standardavvik σ.

Vi ønsker å se på variasjonen i temperaturobservasjonene, hvor mye varierer de enkelte observasjonene rundt gjennomsnittsverdien, μ. Standardavvik, σ, er et mål på spredningen til observasjonene i et datasett og er definert som kvadratroten til variansen (mer om dette senere i kurset). Vi kan regne ut både standardavvik og varians i Matlab,

```
%Standardavvik
```

```
>> std(ttemp)
ans =  8.1000

%Varians
>> var(ttemp)
ans = 65.6103
```

Minimums- og maksimumstemperaturene kan vi finne i Matlab ved å bruke følgende kommando

```
>> min(ttemp)
ans = -12.3000

>> max(ttemp)
ans = 24.9000
```

III Hva var maksimum og minimums-temperaturen i Bodø i 2013? Hva var standardavvik og varians til temperaturobservasjonene?

c)

Lineær regresjon

Vi ønsker nå å se på temperaturutviklingen i Trondheim fra januar til juli 2013, og vi velger da å se på temperaturobservasjonene den 23. januar, 23. februar, 23. mars, 23.april, 23.mai, 23. juni og 23. juli.

Vi ønsker å hente ut alle dataene for den 23. i hver måned fra matrisa `trondheim`. Tidligere i oppgaven brukte vi kommandoen

```
trondheim(:,6);
```

for å hente ut alle temperaturene i kolonne 6 fra matrisa `trondheim`. Nå ønsker vi i tillegg en betingelse på radene, vi vil ha alle radene der datoen er 23, dvs der kolonne 4 har verdien 23. Denne betingelsen kan vi i Matlab skrive som

```
trondheim(:,4)==23
```

Ved å bruke denne betingelsen kan vi da opprette nye vektorer som inneholder temperaturobservasjoner og månedsnummer for den 23. i hver måned med følgende Matlab-kommandoer

```
t = trondheim(trondheim(:,4)==23,6)
Mnd = trondheim(trondheim(:,4)==23,3)
```

Vi har da en vektor `t` med lengde 12 som inneholder de 12 temperaturobservasjonene og en vektor `Mnd` som inneholder månedsnummer.

Temperaturobservasjonene for perioden januar - juli finner vi som de 7 første elementene i vektoren `t`. Dette kan vi plotte med følgende kommando:

```
plot(t(1:7),'*')
xlabel('Mnd')
ylabel('Temperatur')
title('Trondheim januar – juli 2013')
```

Senere i dette kurset skal vi lære hvordan vi kan finne den rette linja som passer best til dataene, i Matlab kan vi bruke kommandoen `polyfit`, der vi tilpasser en linje av orden 1 (rett linje) til dataene.

```
>> p = polyfit(Mnd(1:7),t(1:7),1)
p =
     4.8536  −10.2714
```

Fra Matlab-kommandoen `polyfit` får vi da koeffisientene til en rett linje med stigningstall 4.9 og skjæringspunkt med y-aksen på -10.3.

For å se hvor godt dataene passer sammen med den rette linja fra `polyfit` plotter vi både den rette linja og temperaturobservasjonene i samme plott.

Vi kan plotte de syv temperaturobservasjonene med Matlab-kommandoen

```
plot(Mnd(1:7),t(1:7),'*')
```

Med kommandoen `polyval(p,x)` får vi for en vektor `x` beregnet verdier for den rette linja bestemt av p = polyfit. I vårt tilfelle vil vektoren x=Mnd. For å plotte både observasjonene og den tilpassede linja i samme plott skriver vi i Matlab

```
%plot(mnd(1:7),t(1:7),'*') plotter punkter for de fem temperaturobs.
%polyval(p,Mnd(1:7)) gir punkter pa den rette linja
plot(Mnd(1:7),t(1:7),'*',Mnd(1:7),polyval(p,Mnd(1:7)),'−')
xlabel('Mnd');
ylabel('Temperatur');
title('Trondheim');
```

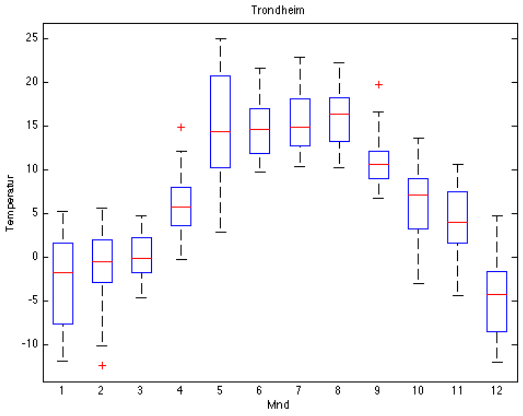
I Lag denne rette linja med kommandoene over. Hvordan passer den rette linja til de observerte dataene?

d)

Boksplott

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. Temperaturobservasjonene for Trondheim finner vi i kolonne 6 og måneden finner vi i kolonne 3. I Matlab kan vi lage et boksplott med følgende kommando

```
boxplot(trondheim(:,6),trondheim(:,3))
xlabel('Mnd');
ylabel('Temperatur');
```



Figur 2: Boksplot for temperaturobservasjoner i Trondheim i 2013

Vi kan regne ut standardavviket til temperaturobservasjonene i f.eks januar og mai.

Januar:

```
t1temp = trondheim(trondheim(:,3)==1,6);
std(t1temp)
ans =
    5.4001
```

Mai:

```
t5temp = trondheim(trondheim(:,3)==5,6);
std(t5temp)
ans =
```

```
6.3664
```

Standardavvikene for Januar og Mai viser at spredningen i dataene er større i Mai enn i Januar, noe som vi også kan se fra Figur 2.

I I hvilken måned er temperaturvariasjonen i Trondheim størst?

Oppgave 2

a)

I Lag histogram over karakterfordelingen for TMA4240 i 2003. Hvordan var karakterfordelingen i 2003 sammenlignet med i 2013 (oppgave 1a)?

II Hvordan ser karakterfordelingen for TMA4240/4245 ut for hele perioden 2003 - 2013? Lag ett histogram.

b)

I oppgave 1d så vi på temperaturobservasjoner for Trondheim i perioden januar til juli 2013. Vi skal nå se på temperaturobservasjoner for Bodø i den samme perioden.

I Plot temperaturobservasjonene for Bodø 23. januar, 23. februar, 23. mars, 23. april, 23. mai, 23. juni og 23. juli 2013.

II Tilpass en rett linje til dataene med funksjonen `polyfit`. Plot den rette linjen og observasjonene i samme plott.

c)

Vi vil nå se på temperaturdataene for Trondheim og Tynset.

I Hva var gjennomsnittstemperaturen på Tynset det siste året? Hva var standardavviket og variansen til temperaturobservasjonene?

II Sammenlign resultatene med resultatene for Trondheim i oppgave 1b.

III Hvor vil du forvente at spredningen/variansen i temperaturobservasjonene var størst? (Hint: Trondheim - kystklima, Tynset - innlandsklima)

d)

I Plot histogram for temperaturobservasjonene i hhv. Trondheim og i Tynset i 2013.

II Beskriv histogrammene

e)

I Lag boksplott for temperaturobservasjonene i Tynset, gruppert etter måned. I hvilke måneder er temperaturvariasjonen størst?

f)

Vi skiller mellom avhengige og uavhengige observasjoner. Vi vil nå se på temperaturobservasjonene i Trondheim, Bodø og på Tynset i 2013.

I Plot temperaturen i Tynset mot temperaturen i Trondheim.

II Plot temperaturen i Bodø mot temperaturen i Trondheim.

III Kan vi observere en trend i noen av disse plottene? Hvilket plott viser avhengige og hvilket viser uavhengige temperaturobservasjoner?