TMA4245 Statistikk
Vår 2013

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Øving nummer 8, blokk II — Matlabøving
Løsningsskisse

Oppgave 1

a) Ingen løsningsskisse.

b) Finn, for hvert datasett, gjennomsnittstemperatur, median, varians og standardavvik.
Løsning:

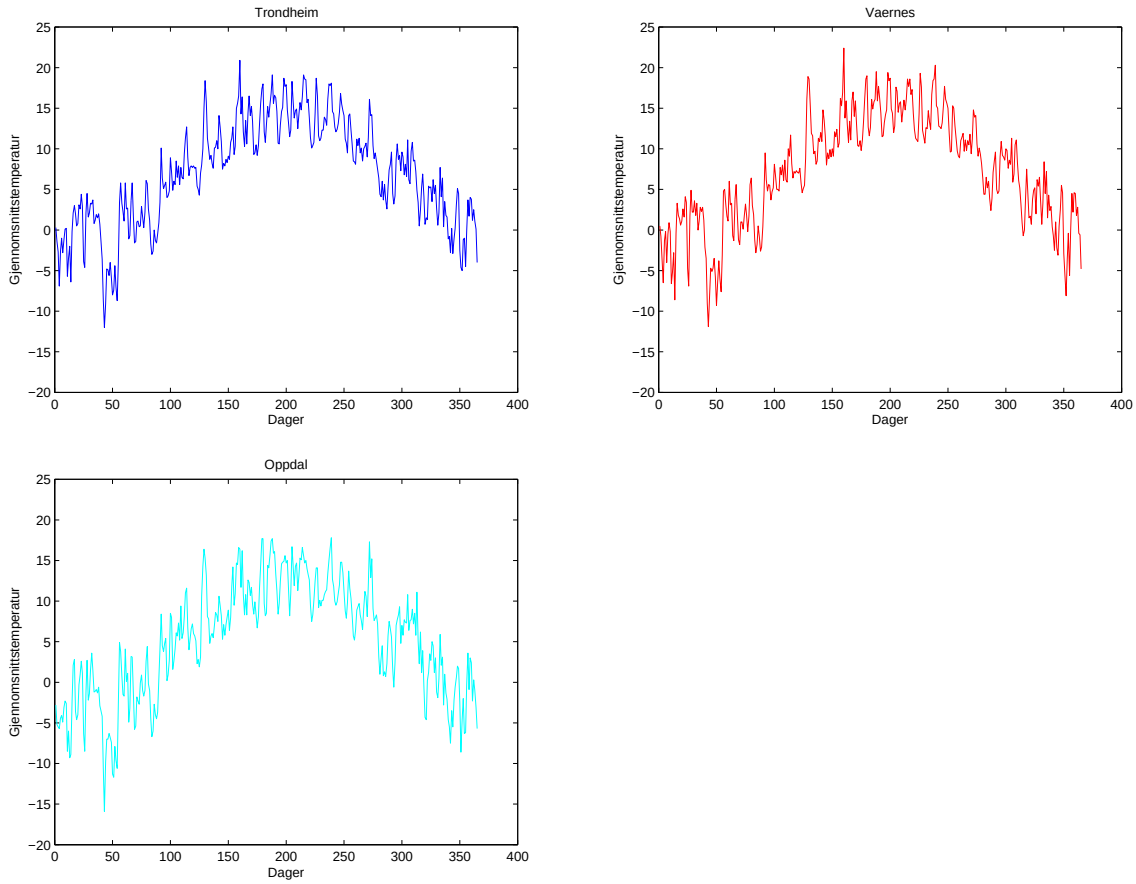
```
% Data for Trondheim:
TRD_mean=mean(TRD); TRD_median=median(TRD);
TRD_std=std(TRD); TRD_var=var(TRD);
% Data for Vaernes:
VAER_mean=mean(VAER); VAER_median=median(VAER);
VAER_std=std(VAER); VAER_var=var(VAER);
% Data for Oppdal:
OPP_mean=mean(OPP); OPP_median=median(OPP);
OPP_std=std(OPP); OPP_var=var(OPP);
```

Tabell 1: Beskrivende statistikker for dataene

Sted	*Gj.snitt*	*Median*	*St.avvik*	*Varians*
Trondheim	6.86	7.50	6.52	42.49
Værnes	7.07	7.20	6.79	46.05
Oppdal	4.98	5.80	7.00	48.96

c) Er forholdet mellom gjennomsnittsverdien og medianen i hvert datasett i overensstemmelse med forholdet mellom gjennomsnittsverdi og median for utfall av en normalfordeling? Begrunn svaret ditt.
Løsning:
Utvalg trukket fra normalfordelingen vil ha tilnærmet like gjennomsnittsverdier og medianer. Dette er ikke tilfellet for noen av datasettene, men dataene for Værnes har et mindre avvik enn Trondheim og Oppdal.

d) For hvert datasett, plott temperatur mot dagnummer (dvs. dag 1, dag 2,..., dag 365). Lag også histogrammer.
Løsning:

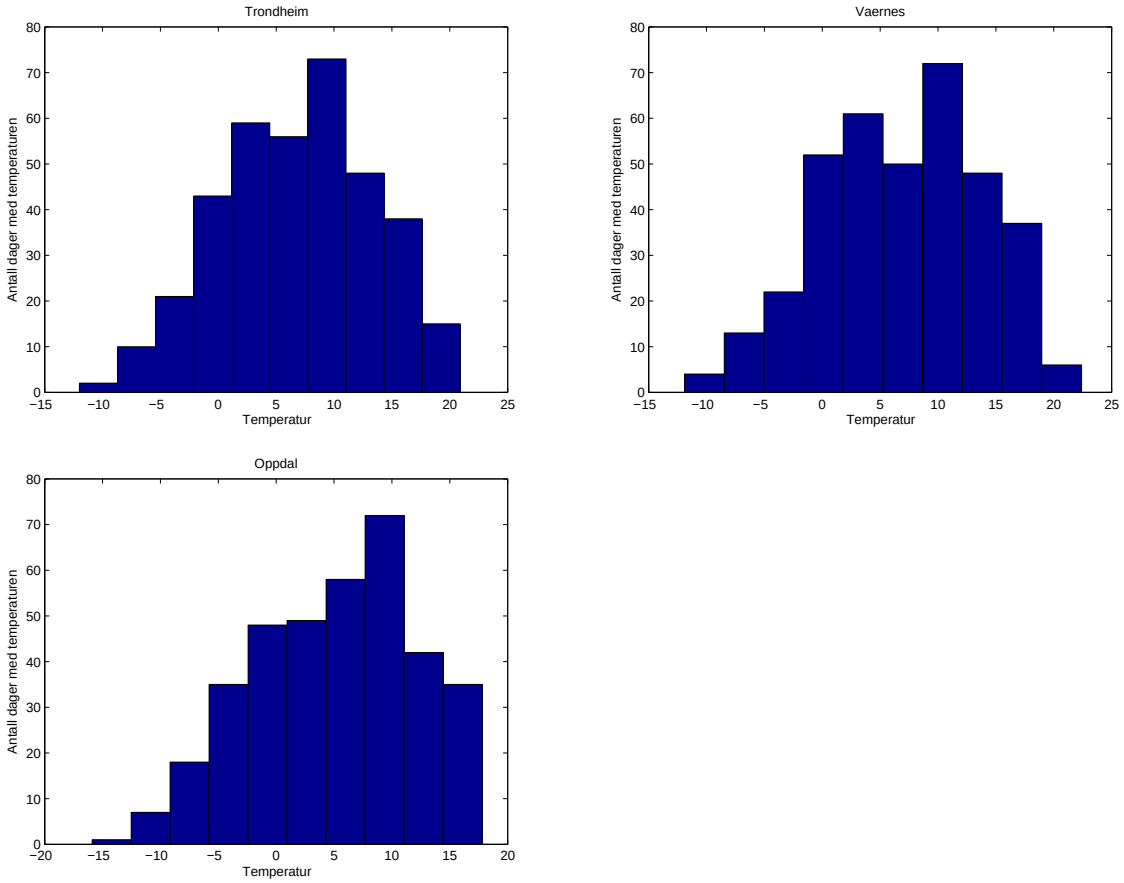


Figur 1: Plott av temperaturer for Trondheim, Værnes og Oppdal

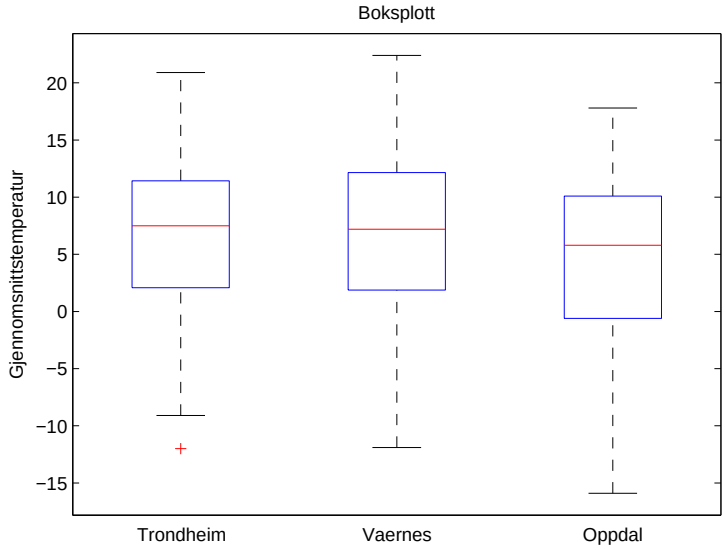
```
% Vanlige plott
plot(TRD,'b'); xlabel('Dager'); ylabel('Gjennomsnittstemperatur');
title('Trondheim'); axis([0 400 -20 25]); figure
plot(VAER,'r'); xlabel('Dager'); ylabel('Gjennomsnittstemperatur');
title('Vaernes'); axis([0 400 -20 25]); figure
plot(OPP,'c'); xlabel('Dager'); ylabel('Gjennomsnittstemperatur');
title('Oppdal'); axis([0 400 -20 25]); figure
% Histogrammer
hist(TRD); xlabel('Temperatur'); ylabel('Antall dager med temperaturen');
title('Trondheim'); figure
hist(VAER); xlabel('Temperatur'); ylabel('Antall dager med temperaturen');
title('Vaernes'); figure
hist(OPP); xlabel('Temperatur'); ylabel('Antall dager med temperaturen');
title('Oppdal');
```

Se Figur 1 og 2.

e) Presenter alle datasettene med boksplott.
Løsning:



Figur 2: Histogrammer for Trondheim, Værnes og Oppdal



Figur 3: Boksplott

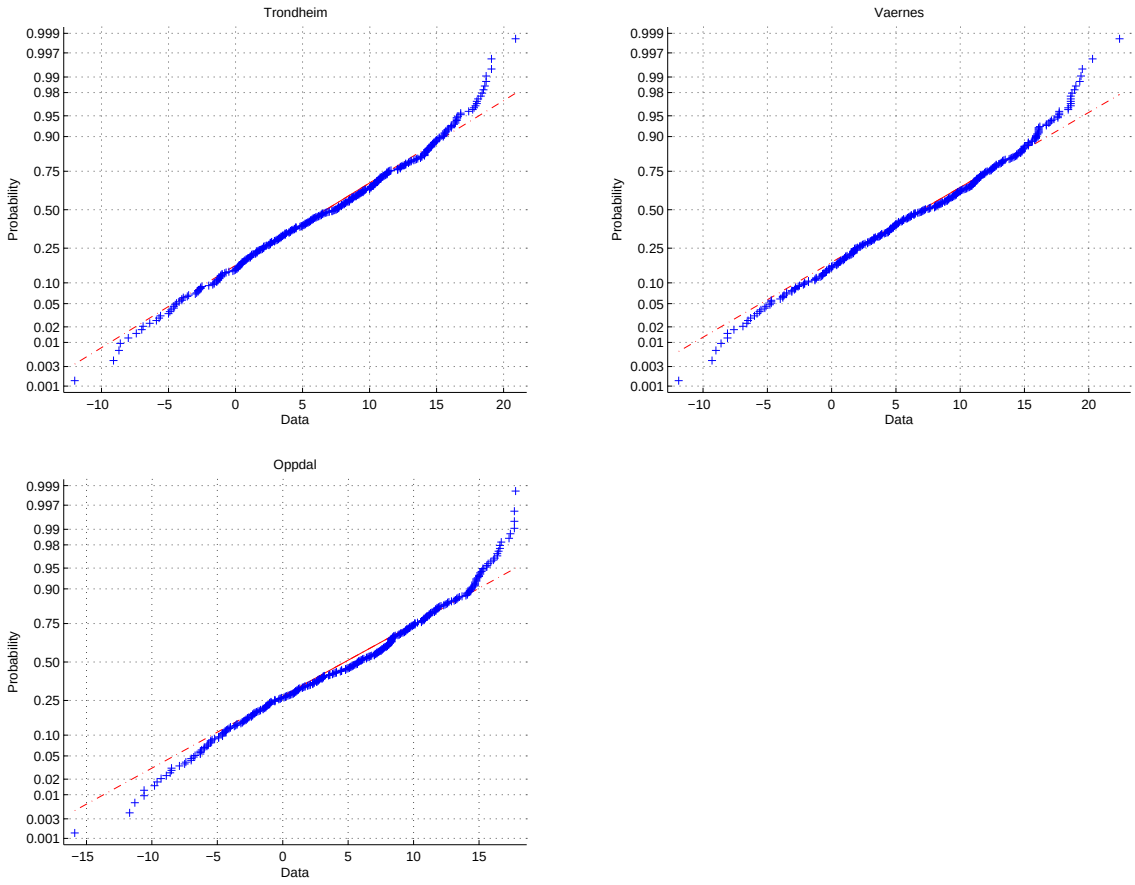
```
boxplot(MT, {'Trondheim','Vaernes','Oppdal'});
title('Boksplott');
ylabel('Gjennomsnittstemperatur');
```

Se Figur 3.

f) Hvilke konklusjoner kan du trekke basert på disse boksplottene? Kan du for eksempel se asymmetrier i dataene?
Løsning:
Vi kan se fra Figur 3 at all tre datasettene har asymmetrier, men at dette er tydeligere i dataene fra Oppdal enn de to andre stedene. Oppdal har mer variasjon sammenlighet med Trondheim og Værnes og har også lavere gjennomsnittstemperatur. Vi kan si at ca 75% av dataene fra Trondheim og Værnes er mellom $2°C$ og $12°C$, mens for Oppdal er intervallet $-2°$ og $10°C$.

g) Er det noen ekstremverdier i datasettene? Hvilken metode bruker MATLAB for å bestemme hvorvidt en observasjon er en ekstremverdi eller ikke?
Løsning:
Fra Figur 3 ser vi at det kun er ekstermverdier i dataene fra Trondheim. Fra hjelpefunksjonen i Matlab '`help boxplot`' finner vi bl.a. følgende informasjon: *I Matlab blir punkter tegnet som ekstremverdier hvis de er større enn* $Q3+W*(Q3-Q1)$ *eller mindre enn* $Q1-W*(Q3-Q1)$*, hvor* $Q1$ *og* $Q3$ *er hhv 25- og 75-prosentskvantilene. Standardverdien på 1.5 for* W *tilsvarer ca* ± 2.7 *standardavvik og* 99.3% *dekning hvis dataene er normalfordelt.*

h) Er det noen forskjeller i de tre boksplottene som tyder på at datasettene kommer fra populasjoner med forskjellige fordelinger?



Figur 4: Normal kvantil-kvantil plott for Trondheim, Værnes og Oppdal

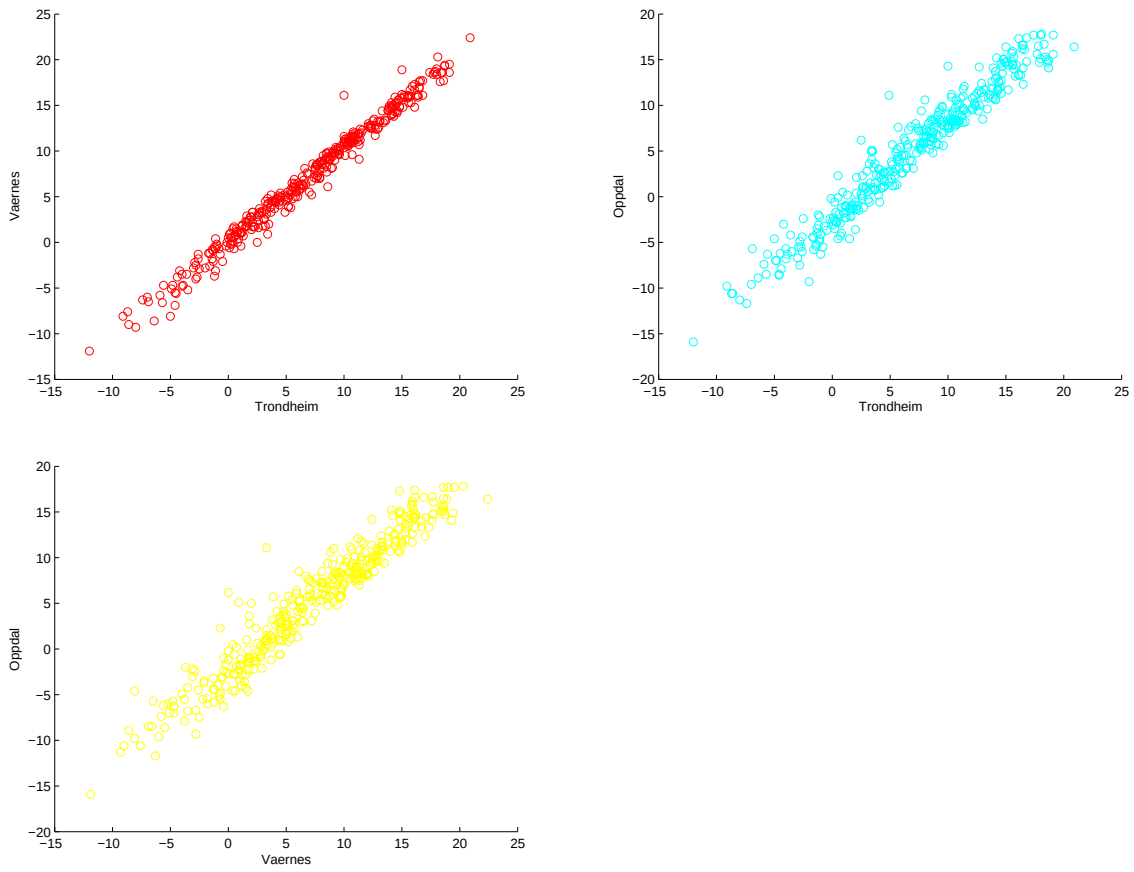
Løsning:
Figur 3 viser at dataene samlet inn i Trondheim og Værnes kommer fra samme fordeling, mens dataene fra Oppdal ser ut til å ha en annen fordeling.

i) Lag et normal kvantil-kvantil plott for å evaluere om datasettene kommer fra en normalfordeling eller ikke.
Løsning:

```
normplot(TRD); title('Trondheim'); figure;
normplot(VAER); title('Vaernes'); figure;
normplot(OPP); title('Oppdal');
```

Vi kan se i Figur 4 at alle tre datasettene avviker litt fra normalfordelingen, spesielt i halene til fordelingene. Disse avvikene er mest synlig for dataene fra Oppdal. Alle datasettene ser ut til å ha tyngre haler til høyre, siden de siste punktene i `normplot` viser en økende avastand fra linja.

j) Lag minst tre scatterplott der du plotter gjennomsnittstemperaturene fra de tre stedene mot hverandre.
Løsning:



Figur 5: Spredningsplott for Trondheim, Værnes og Oppdal

```
scatter(TRD,VAER,'r'); xlabel('Trondheim'); ylabel('Vaernes'); figure;
scatter(TRD,OPP,'c'); xlabel('Trondheim'); ylabel('Oppdal'); figure;
scatter(VAER,OPP,'y'); xlabel('Vaernes'); ylabel('Oppdal');
```

Se Figur 5.

k) Bruk MATLAB til å regne ut korrelasjonsmatrisen for de tre datasettene.
Løsning:

```
correlation_matrix=corr(MT);
```

$$\text{correlation\_matrix} = \begin{bmatrix} 1.0000 & 0.9918 & 0.9739 \\ 0.9918 & 1.0000 & 0.9650 \\ 0.9739 & 0.9650 & 1.0000 \end{bmatrix}$$

l) Er datasettene positivt korrelerte? Er dette noe du ville forventet? Forklar.
Løsning:
Ja, alle datasettene er positivt korrelerte. Vi kan se fra Figur 5 at hvert spredningsplott viser en lineær trend med positivt stigningstall og korrelasjonsmatrisen indikerer det

samme resultatet. Dette er ikke så uventet da målingene er daglige gjennomsnittstemperaturer for tre steder som ligger i samme område av landet.

Oppgave 2

a) Simuler 1000 datasett i MATLAB. Hvert datasett skal bestå av 100 utfall fra en normalfordeling med forventningsverdi 5 og standardavvik 2.
Løsning:

```
sample_size=100;
number_of_samples=1000;
mu=5; %forventning
sigma=2; %standardavvik
sample_matrix=normrnd(mu,sigma,sample_size,number_of_samples);
```

b) Regn ut gjennomsnittsverdien av alle de 1000 datasettene. Lag et histogram basert på gjennomsnittsverdiene du har regnet ut. Minner formen på histogrammet om formen til en normalfordeling? Var dette forventet? Forklar.
Løsning:

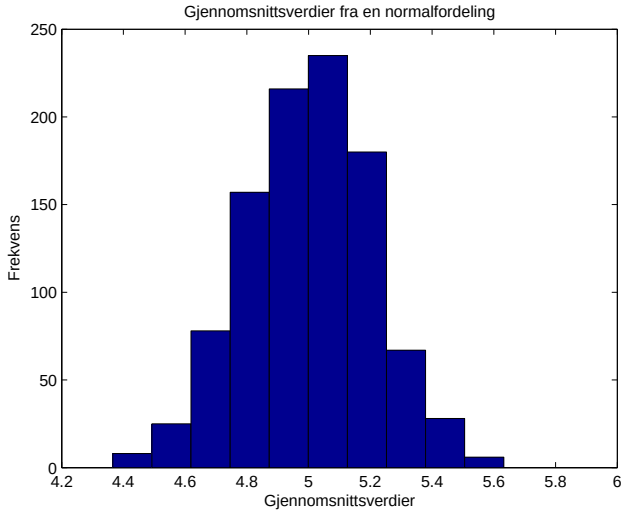
```
sample_matrix_mean=mean(sample_matrix);
hist(sample_matrix_mean);
xlabel('Gjennomsnittsverdier');
ylabel('Frekvens');
title('Gjennomsnittsverdier fra en normalfordeling');
figure
normplot(sample_matrix_mean);
title('Normal kvantil-kvantil plott for gjennomsnittsverdiene');
```

Fra Figur 6 ser vi at gjennomsnittsverdiene minner om en normalfordeling og dette støttes av kvantil-kvantil plottet i Figur 7. Dette er forventet siden vi vet fra sentralgrenseteoremet at fordelingen til $\bar{X}$ er $N(5; 4/1000)$ og at en lineær kombinasjon av normalfordelte variabler også er normalfordelt.

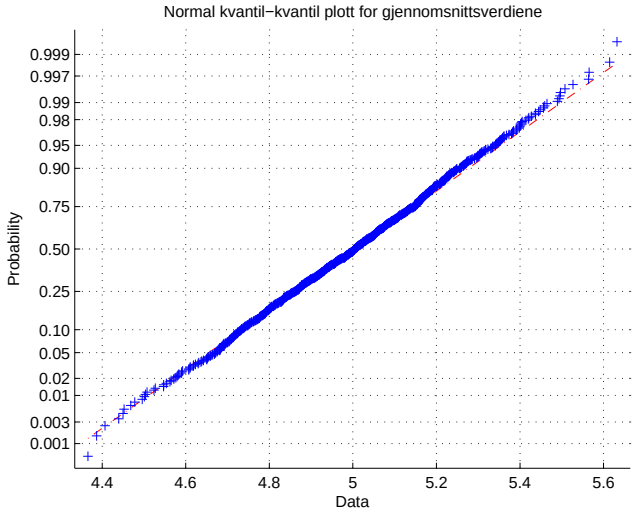
c) Gjør det samme som i **a**), men nå skal utfallene komme fra en uniformfordeling på intervallet $(0, 1)$. Prøv deg frem med 1, 2, 5, 10, 30 og 100 utfall for hvert datasett.
Løsning:

```
sample_size=[1 2 5 10 30 100];
number_of_sizes=length(sample_size);

for i=1:number_of_sizes
    unif_sample_matrix=rand(sample_size(i),1000);
    if i>1
        unif_sample_matrix_mean=mean(unif_sample_matrix);
```



Figur 6: Histogram av gjennomsnittsverdiene regnet fra 1000 utvalg av størrelse 100 fra normalfordelingen med forventning 5 og standardavvik 2

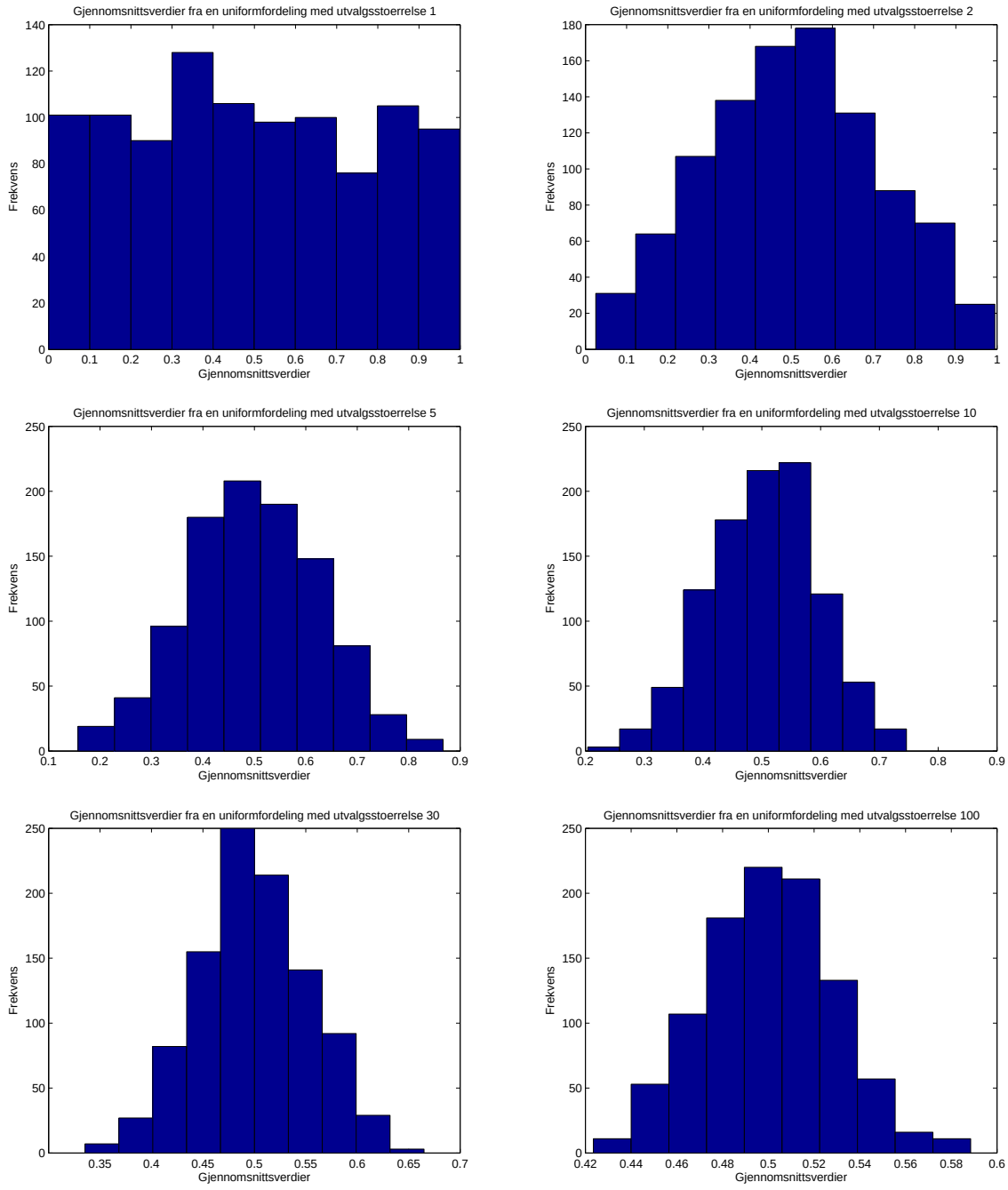


Figur 7: Normal kvantil-kvantil plott av gjennomsnittsverdiene regnet fra 1000 utvalg av størrelse 100 fra normalfordelingen med forventning 5 og standardavvik 2

```
        end
        if i==1
            unif_sample_matrix_mean=unif_sample_matrix;
        end
        samplesize_string=num2str(sample_size(i));
        figure
        hist(unif_sample_matrix_mean);
        xlabel('Gjennomsnittsverdier');
        ylabel('Frekvens');
        title(['Gjsnverdier fra en uniformford med utvalgsstr ',samplesize_string]);
    end
```

d) Hvilke av simuleringene gir et histogram som ligner en normalfordeling? Bruk sentralgrenseteoremet til å forklare resultatet du får.
Løsning:

Vi ser fra histogrammene i Figur 8 at de ligner på en normalfordeling allerede ved utvalgsstørrelse 5. Vi vet fra sentralgrenseteoremet at hvis utvalgsstørrelsen er stor nok, kan vi tilnærme fordelingen med en normalfordeling. Vårt resultat her viser at utvalgsfordelingen til gjennomsnittsverdien fra en uniformfordeling kan tilnærmes godt med en normalfordeling for utvalgsstørrelser så små som 5.



Figur 8: Histogrammer av gjennomsnittsverdiene regnet fra 1000 utvalg fra uniformfordelingen med utvalgsstørrelse 1, 2, 5, 10, 30 og 100