TMA4245 Statistikk
Vår 2015

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Øving nummer 7, blokk II (Matlab-øving 2)
Løsningsskisse

Oppgave 1

a) Intentionally left (almost) blank.

b) Utfør en beskrivende analyse av datasettet
Løsning:

```
% Data for Trondheim:
TRD_mean=mean(TRD); TRD_median=median(TRD);
TRD_std=std(TRD); TRD_var=var(TRD);
% Data for Bodo:
BO_mean=mean(BO); BO_median=median(BO);
BO_std=std(BO); BO_var=var(BO);
% Data for Tynset:
TY_mean=mean(TY); TY_median=median(TY);
TY_std=std(TY); TY_var=var(TY);
```

Tabell 1: Beskrivende statistikker for dataene

Sted	Gj.snitt	Median	St.avvik	Varians
Trondheim	7.7732	7.4000	8.1000	65.6103
Bodø	6.9674	5.6000	7.0840	50.1831
Tynset	4.9510	5.1000	11.3369	128.5260

Utvalg trukket fra normalfordelingen vil ha tilnærmet like gjennomsnittsverdier og medianer. Dette er ikke tilfellet for noen av datasettene, men dataene for Bodø har et større avvik enn Trondheim og Tynset.

```
% Vanlige plott
plot(TRD,'b'); xlabel('Dager'); ylabel('Gjennomsnittstemperatur');
title('Trondheim'); axis([0 400 -20 25]); figure
plot(BO,'r'); xlabel('Dager'); ylabel('Gjennomsnittstemperatur');
title('Bodo'); axis([0 400 -20 25]); figure
plot(TY,'c'); xlabel('Dager'); ylabel('Gjennomsnittstemperatur');
title('Tynset'); axis([0 400 -20 25]); figure
% Histogrammer
```

```
hist(TRD); xlabel('Temperatur'); ylabel('Antall dager med temperaturen');
title('Trondheim'); figure
hist(BO); xlabel('Temperatur'); ylabel('Antall dager med temperaturen');
title('Bodo'); figure
hist(TY); xlabel('Temperatur'); ylabel('Antall dager med temperaturen');
title('Tynset');
```

Se Figur 4 og 8.

c) Presenter alle datasettene med boksplott.
Løsning:

```
boxplot(MT, {'Trondheim','Bodo','Tynset'});
title('Boxplot');
ylabel('Gjennomsnittstemperatur');
```

Vi kan se fra Figur 9 at datasettene for Bodø og Tynset har noe asymmetrier. Tynset har størst varians i temperaturobservasjonene, mens det er liten forskjell i medianen av temperaturobservasjonene for de tre stedene. Tynset har de laveste temperaturobservasjonene i 2013. Vi kan si at ca 75% av dataene for Trondheim er mellom ca $2°C$ og $14°C$, for Bodø mellom $0°C$ og $11°C$ og for Tynset mellom $-2°C$ og $12°C$.

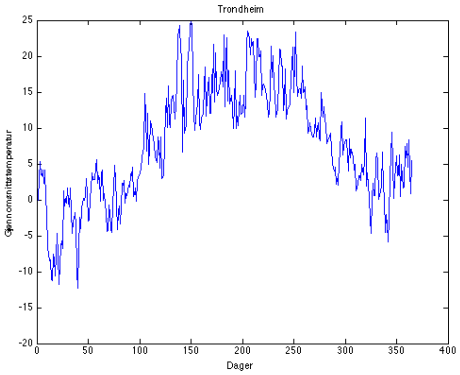
d) Er det noen ekstremverdier i datasettene? Hvilken metode bruker MATLAB for å bestemme hvorvidt en observasjon er en ekstremverdi eller ikke?
Løsning:
Fra Figur 9 ser vi at det kun er ekstermverdier i dataene fra Tynset. Fra hjelpefunksjonen i Matlab '`help boxplot`' finner vi bl.a. følgende informasjon: *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.*

e) Er det noen forskjeller i de tre boksplottene som tyder på at datasettene kommer fra populasjoner med forskjellige fordelinger?
Løsning:
Figur 9 viser at temperaturobservasjonene i de tre datasettene kommer fra forskjellige fordelinger. Fordelingen til temperaturobservasjonene for Trondheim er symmetrisk, mens fordelingen til temperaturobservasjonene for Tynset er skjev.

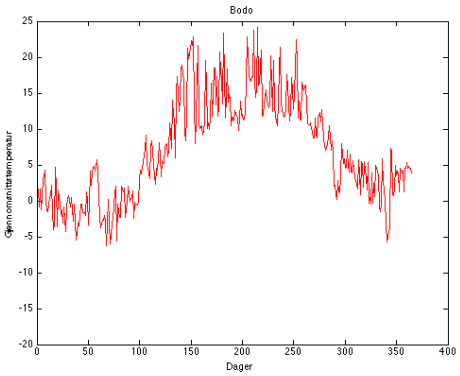
f) Lag et normal kvantil-kvantil plott for å evaluere om datasettene kommer fra en normalfordeling eller ikke.
Løsning:

```
normplot(TRD); title('Trondheim'); figure;
normplot(BO); title('Bodo'); figure;
normplot(TY); title('Tynset');
```

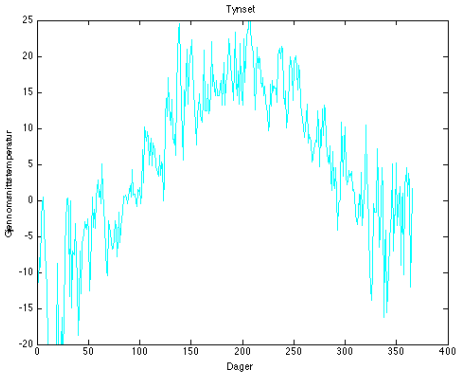
Vi kan se i Figur 13 at alle tre datasettene avviker noe fra normalfordelingen, spesielt i halene til fordelingen. Avvikene er mest synlig for dataene fra Bodø og Tynset.



Figur 1: Trondheim

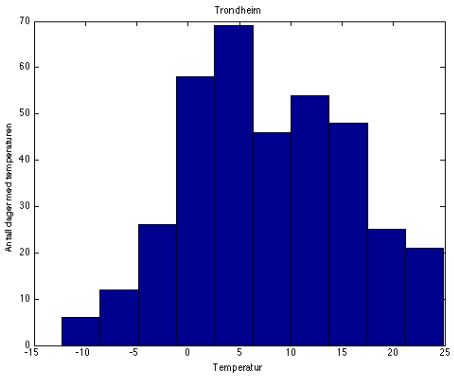


Figur 2: Bodø

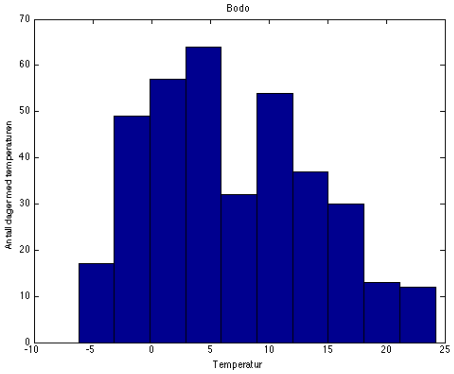


Figur 3: Tynset

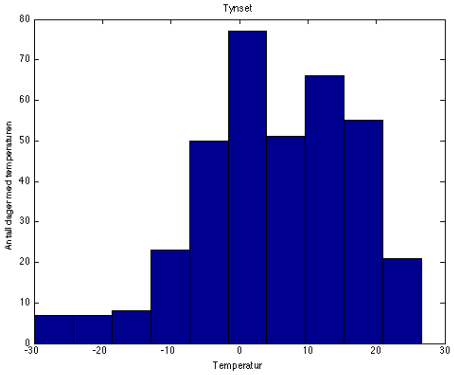
Figur 4: Temperaturobservasjoner



Figur 5: Trondheim

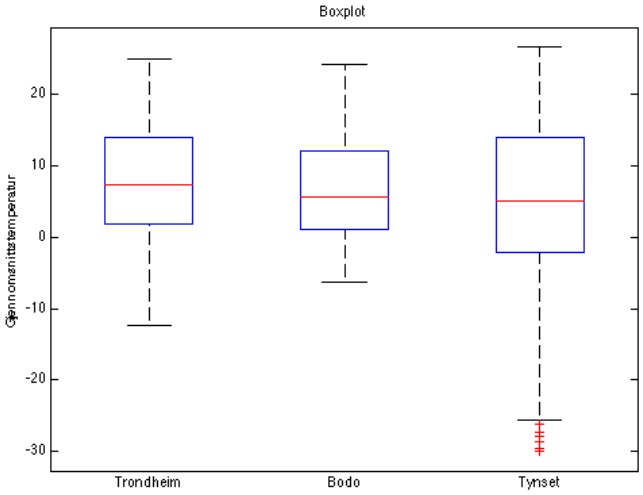


Figur 6: Bodø



Figur 7: Tynset

Figur 8: Histogrammer



Figur 9: Boksplott

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

```
scatter(TRD,BO,'r'); xlabel('Trondheim'); ylabel('Bodo'); figure;
scatter(TRD,TY,'r'); xlabel('Trondheim'); ylabel('Tynset'); figure;
scatter(BO,TY,'r'); xlabel('Bodo'); ylabel('Tynset');
```

Se Figur 17.

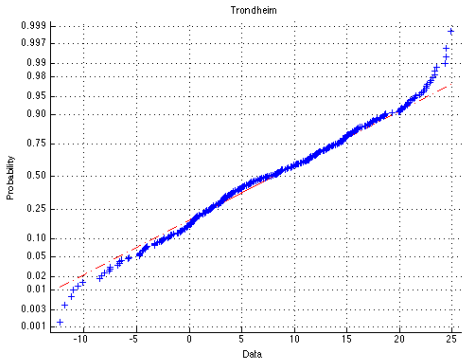
h) Bruk MATLAB til å regne ut korrelasjonsmatrisen for de tre datasettene.
Løsning:

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

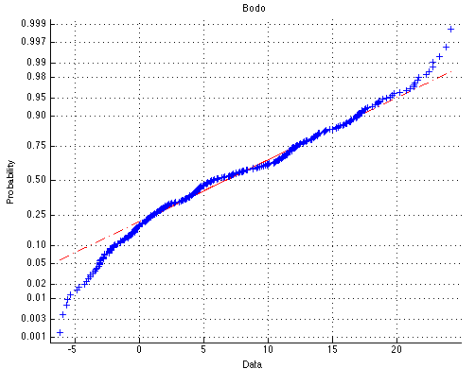
$$\text{correlation\_matrix} = \begin{bmatrix} 1.0000 & 0.9089 & 0.9304 \\ 0.9089 & 1.0000 & 0.8143 \\ 0.9304 & 0.8143 & 1.0000 \end{bmatrix}$$

i) Er datasettene positivt korrelerte? Er dette noe du ville forventet? Forklar.
Løsning:
Ja, alle datasettene er positivt korrelerte. Vi kan se fra Figur 17 at hvert spredningsplott viser en lineær trend med positivt stigningstall og korrelasjonsmatrisen indikerer det samme resultatet. Vi ser også at observasjonene for Bodø er minst korrelert med dataene for Tynset, noe vi vil forvente siden dette er målinger for steder langt fra hverandre geografisk og med ulikt klima.

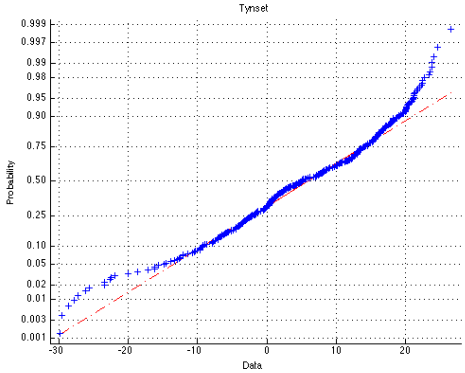
Oppgave 2



Figur 10: Trondheim

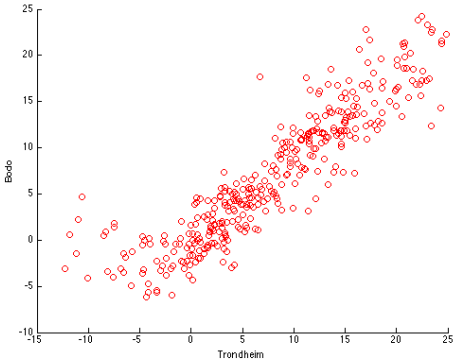


Figur 11: Bodø

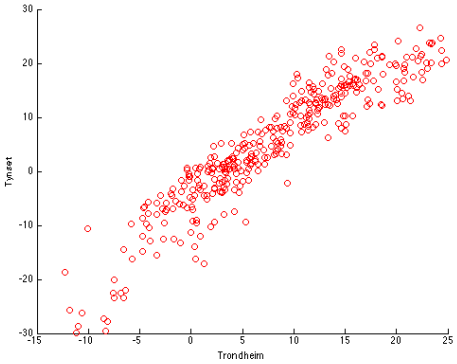


Figur 12: Tynset

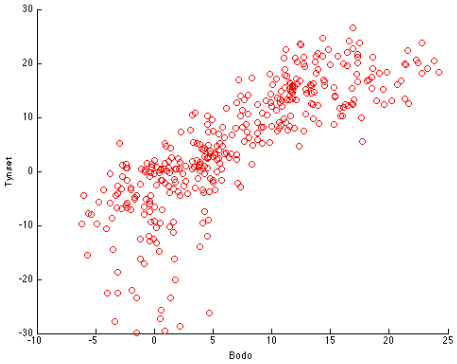
Figur 13: Normal kvantil-kvantil plott for Trondheim, Bodø og Tynset



Figur 14: Trondheim og Bodø



Figur 15: Trondheim og Tynset



Figur 16: Bodø og Tynset

Figur 17: Scatter-plot

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);
```

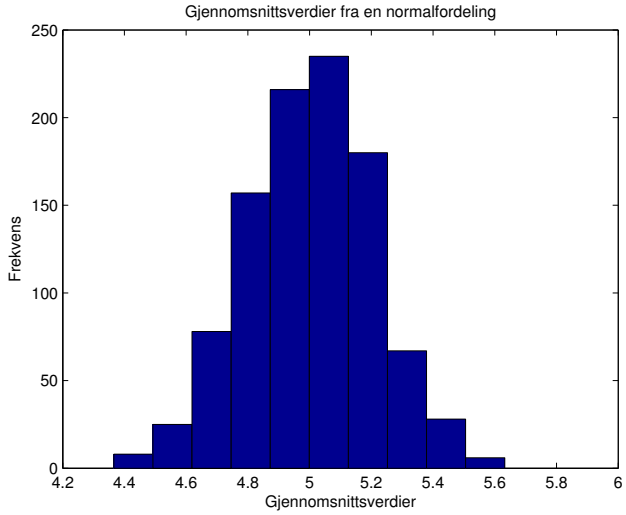
b) Regn ut gjennomsnittsverdien av alle de 1000 datasettene. Lag et histogram basert på gjennomsnittsverdiene du har regnet ut. Minner formen på histogrammet om formen til en normalfordeling? Var dette forventet? Forklar.
Løsning:

```
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');
```

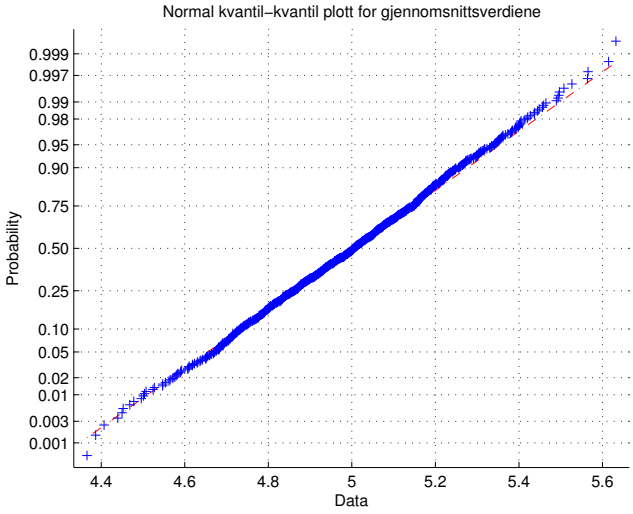
Fra Figur 20 ser vi at gjennomsnittsverdiene minner om en normalfordeling og dette støttes av kvantil-kvantil plottet i Figur 19. Dette er forventet siden vi vet fra sentralgrenseteoremet at fordelingen til $\bar{X}$ er $N(5; 4/1000)$ og at en lineær kombinasjon av normalfordelte variabler også er normalfordelt.

c) Gjør det samme som i **a**), men nå skal utfallene komme fra en binomisk fordeling med parametre $N = 5$,$p = 0.2$ og utvalgsstørrelser $n = 2, 5, 10, 20, 50, 100$.
Løsning:

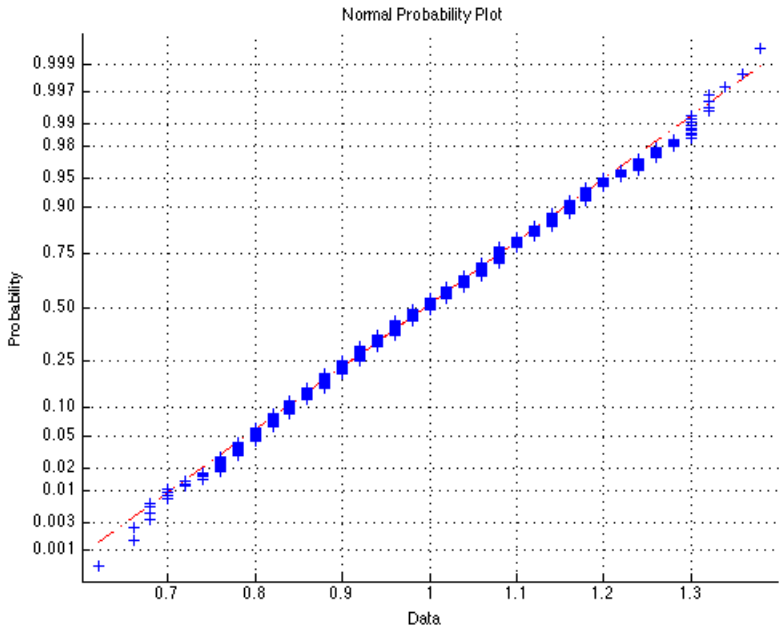
```
n=[2 5 10 20 50];
number_of_sizes=length(n);
nSample = 1000;
N = 5;
p = 0.2;
for i=1:number_of_sizes
          bin_sample_mean = mean(binornd(N,p,n(i),nSample));
          samplesize_string=num2str(n(i));
          figure
          hist(bin_sample_mean);
          xlabel('Gjennomsnitt');
          ylabel('Frekvens');
```



Figur 18: Histogram av gjennomsnittsverdiene regnet fra 1000 utvalg av størrelse 100 fra normalfordelingen med forventning 5 og standardavvik 2



Figur 19: Normal kvantil-kvantil plott av gjennomsnittsverdiene regnet fra 1000 utvalg av størrelse 100 fra normalfordelingen med forventning 5 og standardavvik 2



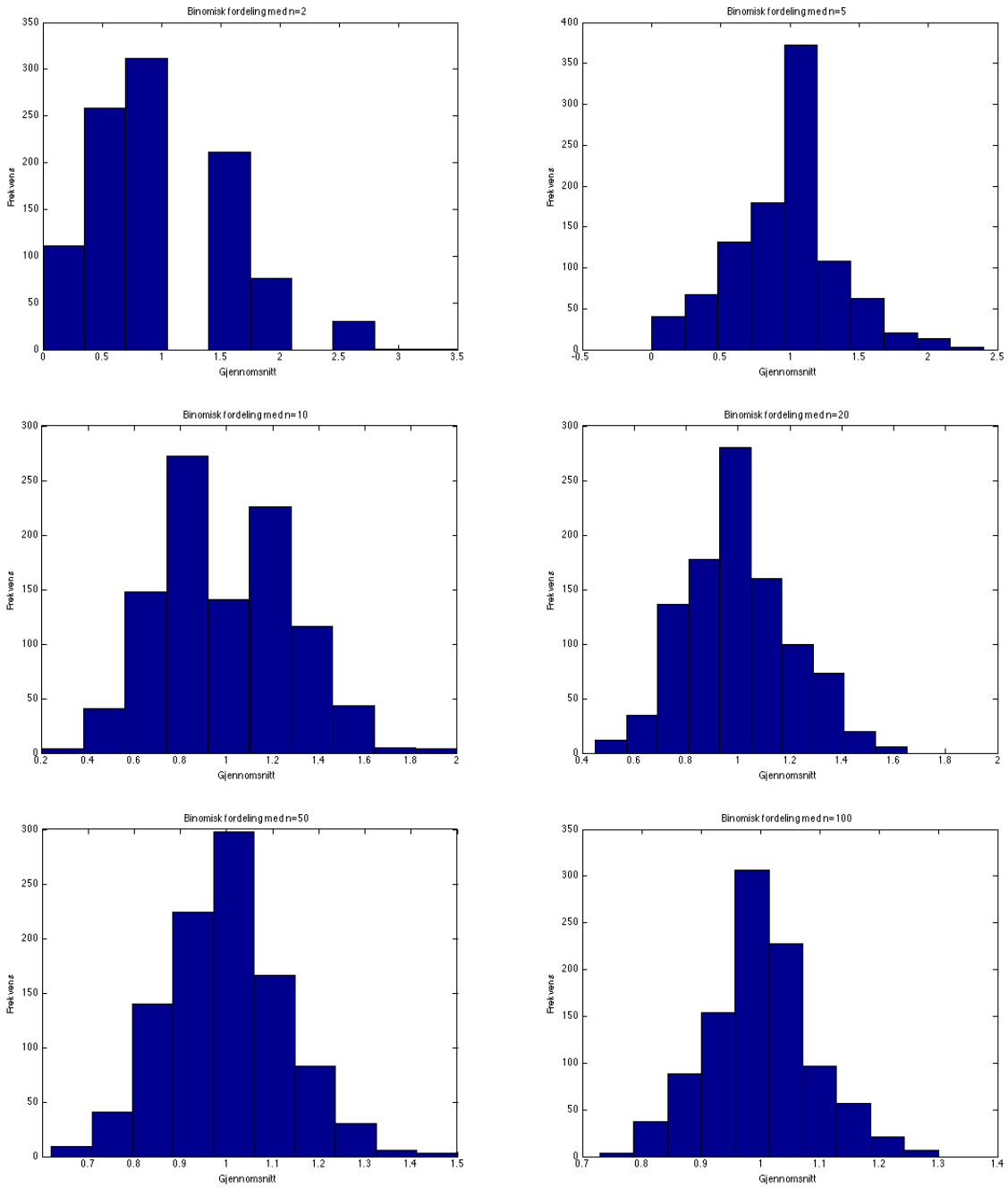
Figur 20: Normal probability plot for a sample of size 50 from the Bin(5,0.2) distribution.

```
            title(['Binomisk fordeling med n=',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 21 at de ligner på en normalfordeling allerede ved utvalgsstørrelse $n = 20$. Vi vet fra sentralgrenseteoremet at hvis utvalgsstÃ¸rrelsen er stor nok kan vi tilnærme fordelingen med en normalfordeling. Vårt resultat her viser at den binomiske fordelingen kan tilnærmes godt med en normalfordeling for utvalgsstørrelser så små som 20.

```
R = mean(binornd(5,0.2,50,1000))
normplot(mean(R))
```



Figur 21: Gjennomsnittsverdier for 1000 utvalg fra binomisk fordeling med $p = 0.2, N = 5$, utvalgsstÃ¸rrelser $n = 2, 5, 10, 20, 50, 100$