Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

TMA4240 Statistikk
Høst 2017

Anbefalt øving 8
Løsningsskisse

Oppgave 1

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

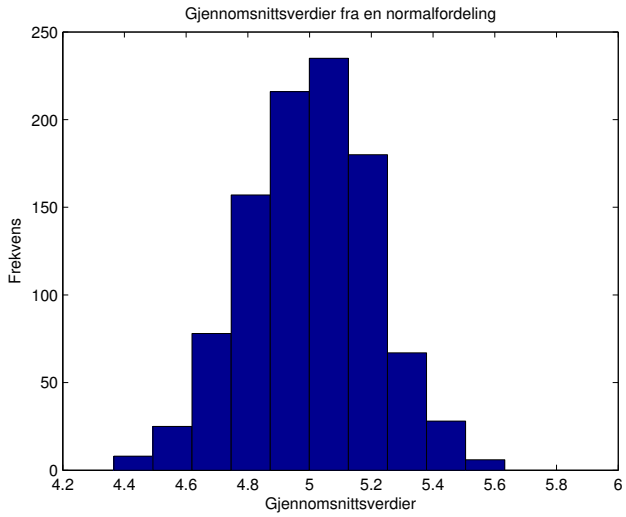
```
sample_size=100;
number_of_samples=1000;
mu=5; %forventning
sigma=2; %standardavvik
sample_matrix=normrnd(mu,sigma,sample_size,number_of_samples);
```

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:

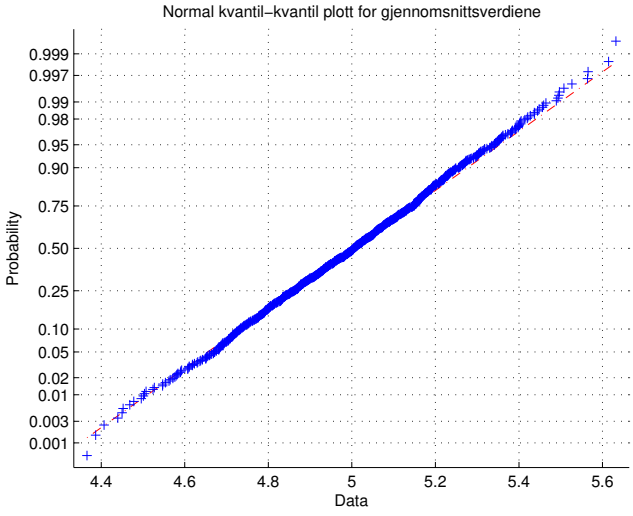
```
sample_matrix_mean=mean(sample_matrix);
hist(sample_matrix_mean);
xlabel('Gjennomsnittsverdier');
ylabel('Frekvens');
title('Gjennomsnittsverdier fra en normalfordeling');
figure
normplot(sample_matrix_mean);
title('Normal kvantil-kvantil plott for gjennomsnittsverdiene');
```

Fra Figur 3 ser vi at gjennomsnittsverdiene minner om en normalfordeling og dette støttes av kvantil-kvantil plottet i Figur 2. 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:



Figur 1: Histogram av gjennomsnittsverdiene regnet fra 1000 utvalg av størrelse 100 fra normalfordelingen med forventning 5 og standardavvik 2



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

```
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');
            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 4 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))
```

Oppgave 2

a) Variansen til utvalgsgjennomsnittet er

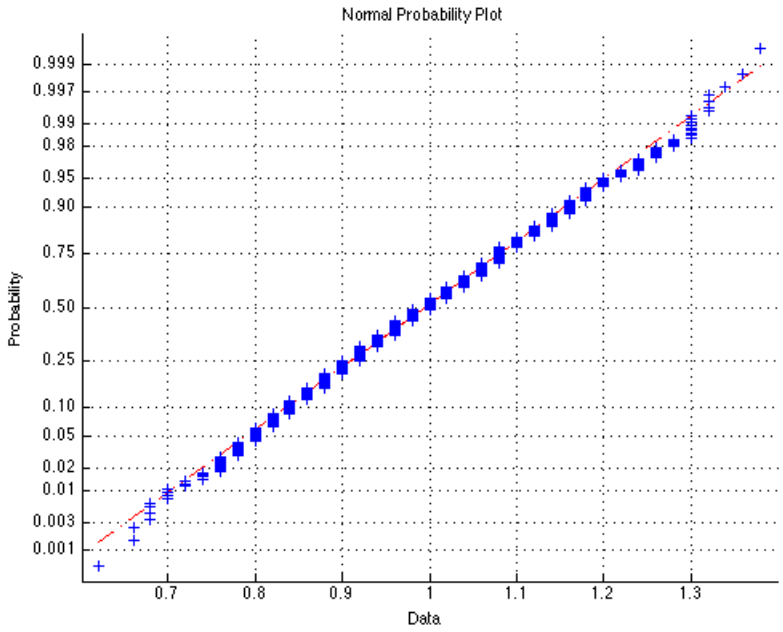
$$\begin{aligned}\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.\end{aligned}$$

Sannsynlighetstetthetsfunksjonen til normalfordelingen er gitt på s. 25 i *Tabeller og formler i statistikk* som

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right),$$

slik at vi har

$$f(\mu) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\cdot\frac{0}{\sigma^2}\right) = \frac{1}{\sqrt{2\pi}\sigma}e^0 = \frac{1}{\sqrt{2\pi}\sigma}.$$



Figur 3: Normalkvantilplott av et utvalg med 50 datapunkter trukket fra Bin(5,0.2)-fordelingen.

Dette gir at

$$\mathrm{Var}(\tilde{X}) = \frac{1}{4n\big(f(\mu)\big)^2} = \frac{1}{4n\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^2} = \frac{\pi\sigma^2}{2n} = \frac{\pi}{2}\mathrm{Var}(\bar{X}),$$

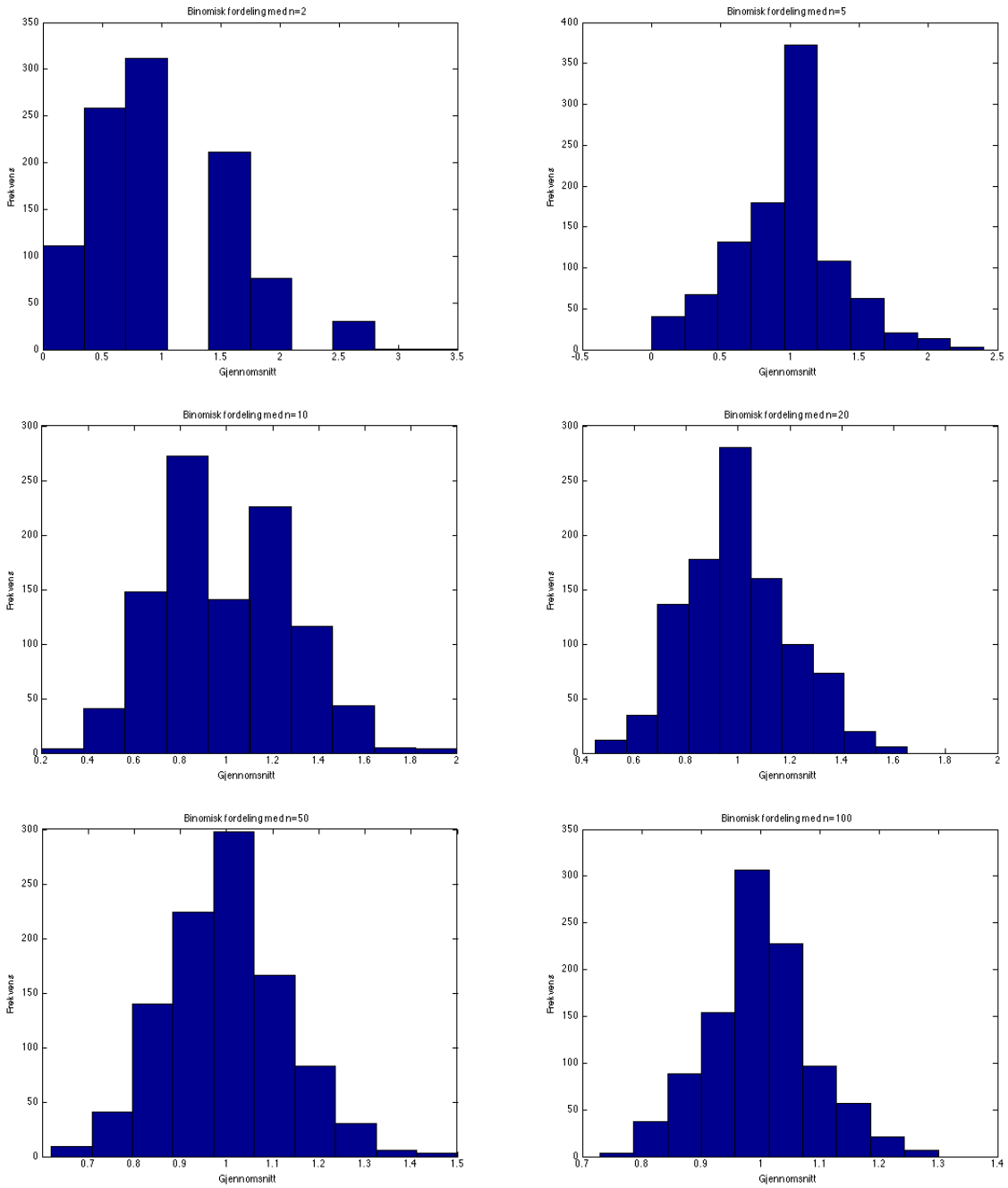
hvilket skulle vises.

Når vi skal velge mellom to estimatorer som begge er forventningsrette, velger vi alltid den med minst varians. Siden $\frac{\pi}{2} \approx 1.57 > 1$ har vi $\mathrm{Var}(\tilde{X}) > \mathrm{Var}(\bar{X})$, som betyr at vi foretrekker å bruke $\bar{X}$ som estimator for μ.

b) På grunn av de to tydelige outlierne på oppsiden, kommer medianen $\tilde{X}$ til å være mindre enn utvalgsgjennomsnittet $\bar{X}$ (for disse dataene er $\tilde{X} = 171.0$ mens $\bar{X} = 175.3$).

Vi har antatt at rekruttenes høyder er normalfordelte. Utfra histogrammet ser det ut til at gjennomsnittet ligger rundt 170 cm. I så fall er sannsynligheten for at to av de tretti datapunktene er større enn 235 cm neglisjerbar, så de ekstreme verdiene til disse to datapunktene skyldes antakelig en feil hos rekrutten som fylte inn dataene i regnearket – ikke spesielt usannsynlig, gitt det gulnede papiret og falmede blekket. Siden utvalgsgjennomsnittet er følsomt for outliere, mens utvalgsmedianen ikke er det, gir medianen et bedre estimat enn gjennomsnittet i dette tilfellet.

Anmerkning vedrørende dataene
Datasettet i denne oppgaven er naturligvis fiktivt. Histogrammet er laget for 28 datapunkt trukket tilfeldig fra en normalfordeling med forventningsverdi 166 cm (litt lavere



Figur 4: Gjennomsnittsverdier for 1000 utvalg fra binomisk fordeling med $p = 0.2, N = 5$, utvalgsstÃ¸rrelser $n = 2, 5, 10, 20, 50, 100$

enn gjennomsnittshøyden for 1878, som er 169.5 cm) og standardavvik 7 cm, og med to outliere på 239 cm og 251 cm (høyden til verdens høyeste mann). Når $X \sim N(166, 7^2)$ *så er* $P(X \geq 239) = 9 \cdot 10^{-26}$.

Oppgave 3

a) For å regne ut $P(L'|A_2)$ benytter vi regelen for sannsynlighet for komplementære hendelser:

$$\begin{aligned} P(L'|A_2) + P(L|A_2) &= 1 \\ P(L'|A_2) &= = 1 - P(L|A_2) = 1 - 0.2 = \underline{\underline{0.8}} \end{aligned}$$

For å regne ut $P(L)$ bruker vi setningen om total sannsynlighet. Vi vet at A_1, A_2, A_3 er en partisjon av utfallsrommet (det ser vi lett av venndiagrammet).

$$\begin{aligned} P(L) &= P(L \cap A_1) + P(L \cap A_2) + P(L \cap A_3) \\ &= P(L|A_1) \cdot P(A_1) + P(L|A_2) \cdot P(A_2) + P(L|A_3) \cdot P(A_3) \\ &= 0.05 \cdot 0.1 + 0.2 \cdot 0.4 + 0.6 \cdot 0.5 = \underline{\underline{0.385}} \end{aligned}$$

b) Betingelser for at X er binomisk fordelt:

- Vi spør n personer.
- For hver person registerer vi om personen lyver eller ikke lyver (to komplmentære hendelser).
- Sannsynligheten for at en tilfeldig valgt person lyver er p, og denne er den samme for alle de n personene vi spør.
- De n personene vi spør svarer uavhengig av hverandre (n uavhengige forsøk).

Under disse 4 betingelsene er X="antall personer som lyver" binomisk fordelt med parametere n og p. Dermed er sannsynlighetsfordelingen til X gitt ved punktsannsynligheten $f(x)$,

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, ..., n$$

Vi vet at da er forventningen til X $E(X) = np$ og variansen $\text{Var}(X) = np(1-p)$.

Videre: vi har at $p = 0.2$, og $n = 20$. $P(X = 4)$ finner vi ved å sette inn $X = 4$ i punktsannsynligheten $f(x)$ over.

$$P(X = 4) = f(4) = \binom{20}{4} 0.2^4 (1 - 0.2)^{20-4} = \underline{\underline{0.218}}$$

Det er også mulig å finne $P(X = 4)$ ved tabelloppslag (s 17 i formelsamlingen),

$$P(X = 4) = P(X \leq 4) - P(X \leq 3) = 0.630 - 0.411 = 0.219$$

Sannsynligheten $P[(X \leq 2) \cup (X > 5)]$ finner vi enklest ved tabelloppslag (s 17 i formelsamlingen),

$$\begin{aligned} P[(X \leq 2) \cup (X > 5) = P(X \leq 2) + P(X > 5) &= (X \leq 2) + 1 - P(X \leq 5) \\ &= 0.206 + 1 - 0.804 = \underline{\underline{0.402}} \end{aligned}$$

c) Nå er p ukjent.

Først forventning:

$$\begin{aligned} E(\hat{p}) &= E(\frac{X}{n}) = \frac{1}{n}E(X) = \frac{1}{n}np = p \\ E(p^*) &= E(\frac{X}{n-1}) = \frac{1}{n-1}E(X) = \frac{1}{n-1}np = \frac{n}{n-1}p \end{aligned}$$

Vi ser videre på varians:

$$\begin{aligned} \text{Var}(\hat{p}) &= \text{Var}(\frac{X}{n}) = \frac{1}{n^2}\text{Var}(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n} \\ \text{Var}(p^*) &= \text{Var}(\frac{X}{n-1}) = \frac{1}{(n-1)^2}\text{Var}(X) = \frac{1}{(n-1)^2}np(1-p) = \frac{np(1-p)}{(n-1)^2} \end{aligned}$$

En god estimator $\hat{p}$ er en estimator som er

- forventningsrett, dvs. $E(\hat{p}) = p$, og
- har liten varians, dvs. $\text{Var}(\hat{p})$ er liten.

Vi liker veldig godt hvis variansen minker når antall observasjoner som estimatoren er basert på øker.

Sammenligner vi to estimatorer som begge er forventningsrette velger vi estimatoren med minst varians. Sammenlinger vi to estimatorer der kun den ene er forventningsrett, velger vi gjerne den estimatoren som er forventningsrett (ofte sjekker vi også at det ikke er veldig stor forskjell på variansene).

For å velge mellom $\hat{p}$ og p^* ser vi på uttrykkene for forventning og varians til begge estimatorene.

Vi ser at $\hat{p}$ er forventningsrett, men det er ikke p^*. I prinsippet kan vi stoppe her og konkluere med at vi foretrekker den forventningsrette estimatoren $\hat{p}$. Men, det kan være fint å sjekke at det ikke er stor forskjell på variansen til de to estimatorene (hva hvis den ene hadde hatt to ganger så stor varians?).

Vi ser at $\text{Var}(\hat{p}) = (\frac{(n-1)}{n})^2\text{Var}(p^*)$, dvs. $\text{Var}(\hat{p}) < \text{Var}(p^*)$ med en faktor $(\frac{n-1}{n})^2$ i forskjell. For $n = 20$ er denne faktoren $(\frac{19}{20})^2 = 0.95^2 = 0.9$, dvs. $\text{Var}(\hat{p}) = 0.9 \cdot \text{Var}(p^*)$. Dermed har estimatoren $\text{Var}(\hat{p})$ både minst varians og er forventningsrett. Vi velger derfor estimatoren $\hat{p}$.

Kommentarer: Asymptotisk (når $n \to \infty$) vil de to estimatorene være like gode. Vi har i vårt pensum ikke snakket om begrepet konsistente estimatorer, men begge disse estimatorene er konsistente.

Oppgave 4

a) Setning om forventning til funksjoner av stokastiske variable gir at

$$\begin{aligned}
E(\sqrt{Y}) &= \int_0^\infty y^{1/2} f(y) dy \\
&= \int_0^\infty y^{1/2} \frac{1}{2^{\frac{v}{2}}\Gamma(\frac{v}{2})} y^{\frac{v}{2}-1} e^{-\frac{y}{2}} dy \\
&= \int_0^\infty \frac{1}{2^{\frac{v}{2}}\Gamma(\frac{v}{2})} y^{\frac{v+1}{2}-1} e^{-\frac{y}{2}} dy \\
&= \frac{2^{\frac{v+1}{2}}\Gamma(\frac{v+1}{2})}{2^{\frac{v}{2}}\Gamma(\frac{v}{2})} \int_0^\infty \frac{1}{2^{\frac{v+1}{2}}\Gamma(\frac{v+1}{2})} y^{\frac{v+1}{2}-1} e^{-\frac{y}{2}} dy \\
&= \frac{\sqrt{2}\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})}
\end{aligned}$$

siden integranden i nest siste uttrykk ovenfor er en sannsynlighetstetthet (til en kji-kvadratfordelt variabel med $v+1$ frihetsgrader).

b) Bruker vi resultatet i forrige punkt med $v = n-1$ følger det at

$$E\sqrt{\frac{S^2(n-1)}{\sigma^2}} = \frac{\sqrt{n-1}}{\sigma} ES = \frac{\sqrt{2}\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}.$$

Altså er

$$ES = \frac{\sigma}{\sqrt{n-1}} \cdot \frac{\sqrt{2}\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}$$

slik at S ikke er forventningsrett for σ. En forventingsfeilkorrigert, forventningsrett estimator av σ er dermed

$$\hat{\sigma} = S\sqrt{n-1} \cdot \frac{\Gamma(\frac{n-1}{2})}{\sqrt{2}\Gamma(\frac{n}{2})} = \frac{\Gamma(\frac{n-1}{2})}{\sqrt{2}\Gamma(\frac{n}{2})} \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

På tilsvarende måte som i punkt a) kan en medianrett estimator for σ utledes med utgangspunkt i samme pivotale stÃ¸rrelse. Vi vet at

$$P\left(\frac{S^2(n-1)}{\sigma^2} < \chi^2_{1/2,n-1}\right) = 1/2.$$

Omskriving av ulikheten gir at

$$P\left(S\sqrt{\frac{n-1}{\chi^2_{1/2,n-1}}} < \sigma\right) = 1/2,$$

som i følge definisjon av medianretthet betyr at

$$\tilde{\sigma} = S\sqrt{\frac{n-1}{\chi^2_{1/2,n-1}}} = \sqrt{\frac{1}{\chi^2_{1/2,n-1}} \sum_{i=1}^n (X_i - \bar{X})^2}$$

er medianrett for σ.