STK1100 Oblig 2

Daniel Heinesen, daniehei

3. mai 2017

Oppgave 1)

a)

Vi har en simultan sannsynlighetsfordelig

$$f(x,y) = \begin{cases} k(x-y) & 0 \leq y \leq x \leq 1 \\ 0 & \text{ellers} \end{cases} \tag{1}$$

for å finne k kan vi bruke at fordeligen må være normalisert. Dette betyr at

$$I = \int_0^1 \int_0^x k(x-y)dydx = 1 \tag{2}$$

Vi kan regne dette ut:

$$\begin{aligned} I = \int_0^1 \int_0^x k(x-y)dydx &= k\int_0^1 \left[xy - \frac{1}{2}y^2\right]_0^x dx \\ &= k\int_0^1 \frac{1}{2}x^2 dx = k\frac{1}{6}x^3\Big|_0^1 = \frac{k}{6} = 1 \end{aligned}$$

Dette gir oss at

$$\underline{\underline{k = 6}} \tag{3}$$

b)

Vi ønsker å finne $P(2Y \leq X)$. Dette er sannsynligheten at om vi velger en tilfeldig x og en tilfeldig y, så er $2Y \leq X \Leftrightarrow Y \leq X/2$. Vi finner sannsynligheten ved

$$\begin{aligned} P(2Y \leq X) = \int_0^1 \int_0^{x/2} k(x-y)dydx &= k\int_0^1 \left(xy - \frac{1}{2}y^2\right)\Big|_0^{x/2} dx = k\int_0^{x/2} \frac{1}{2}x^2 - \frac{1}{8}x^2 dx \\ &= k\frac{1}{8}x^3\Big|_0^1 = \frac{k}{8} \end{aligned}$$

Setter vi inn verdien for k det endelig svaret:

$$\underline{\underline{P(2Y \leq X) = \frac{3}{4}}} \tag{4}$$

c)

Vi ønsker så å finne den marginale sannsynlighetsfordeligen for X. Vi ser først på hvor $0 \leq y \leq x \leq 1$. For å få den marginale sannsynlighetsfordeligen må vi integrere sannsynlighetsfordeligen over alle mulige verdier av y, som er $0 \leq y \leq x$, vi får da

$$f_X(x) = \int_0^x k(x-y)dy = k\left(xy - \frac{1}{2}y^2\right)\Big|_0^x = \frac{k}{2}x^2 = \underline{3x^2} \tag{5}$$

For vi ser nå på de øvrige funksjonene. Her vil $f(x,y) = 0$, som fører til at også $f_X(x) = 0$. Så vi ender opp med at

$$f_X(x) = \begin{cases} 3x^2 & 0 \leq x \leq 1 \\ 0 & \text{ellers} \end{cases} \tag{6}$$

d)

Vi ønsker så å finne den marginale sannsynlighetsfordeligen for Y. Vi starter også her med å se på tilfellene hvor $0 \leq y \leq x \leq 1$. Vi ønsker å integrere $f(x,y)$ over alle de mulige verdiene av x, hvilket er $y \leq x \leq 1$:

$$f_Y(x) = \int_y^1 k(x-y)dy = k\left(\frac{1}{2}x^2 - yx\right)\Big|_y^1 = k\left(\frac{1}{2}y^2 - y + \frac{1}{2}\right) = \underline{3y^2 - 6y + 3} \tag{7}$$

Akkurat som i deloppgaven over vil $f(x,y) = 0$ for de øvrige tilfellen, noe som gjør at $f_Y(x) = 0$ her. Så vi ender opp med:

$$f_X(x) = \begin{cases} 3y^2 - 6y + 3 & 0 \leq y \leq 1 \\ 0 & \text{ellers} \end{cases} \tag{8}$$

e)

For å sjekke om X og Y er uavhenginge kan vi gange samme de marginale sannsynlighetsfordelingene og se om vi før den orginale sannsynlighetsfordelingen (1):

$$f_X(x) \cdot f_Y(y) = 3x^2 \cdot (3y^2 - 6y + 3) = 9x^2y^2 - 6x^2y + 9x^2 \neq f(x,y) \tag{9}$$

Siden $f_X(x) \cdot f_Y(y) \neq f(x,y)$ er X og Y **ikke** uavhenginge.

Oppgave 2)

Koden finnes som opp2b.py. Denne koden vil lage alle plottene som er brukt i denne oppgaven.(Alle plotten vil lages og vises når man kjører programmet, så vær forberedt på den når den som retter dette kjører programmet)

a)

Vi har flere identiske stokatiske variabler X_i som følger:

$$E(X_i) = \mu, \qquad V(X_i) = \sigma^2 \tag{10}$$

Vi innfører så at

$$\overline{X}_i = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{11}$$

Vi har også det standariserte gjennomsnittet

$$Z_n = \frac{\overline{X}_i - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\overline{X}_i - \mu}{\sigma} \tag{12}$$

Vi kommer til å bruke at for forventingsverdien til flere uavhengige stokastiske variabler er

$$E\left(a + \sum_{i=1}^{n} b_i X_i\right) = a + \sum_{i=1}^{n} b_i E(X_i) \tag{13}$$

Og for variansen er gjelder

$$V\left(a + \sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i^2 V(X_i) \tag{14}$$

Vi kan fra dette se at for de uavhengige stokastiske variablene X_i gjelder:

$$E(\overline{X}_i) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \underline{\underline{\mu}} \tag{15}$$

og

$$V(\overline{X}_i) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2} n\sigma^2 = \underline{\underline{\frac{\sigma^2}{n}}} \tag{16}$$

Og for det standariserte gjennomsnittet gjelder

$$E(Z_n) = E\left(\sqrt{n}\frac{\overline{X}_i - \mu}{\sigma}\right) = \sqrt{n}\left(\frac{E(\overline{X}_i)}{\sigma} - \frac{\mu}{\sigma}\right) = \sqrt{n}\left(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}\right) = \underline{\underline{0}} \tag{17}$$

og

$$V(Z_n) = V\left(\sqrt{n}\frac{\overline{X}_i - \mu}{\sigma}\right) = n\frac{V(\overline{X}_i)}{\sigma^2} = n\frac{\sigma^2/n}{\sigma^2} = \underline{\underline{1}} \tag{18}$$

b)

Vi skal her bestemme forventingsverdien og variansen til de 3 følgende fordelingene:

Uniform Fordeling:

Vi har en uniform fordeling:

$$f(x) = \begin{cases} \frac{1}{2} & -1 \leq x \leq 1 \\ 0 & ellers \end{cases} \tag{19}$$

Vi kan regne ut fordelingen:

$$\mu = E(X_i) = \int_{-1}^{1} x f(x) dx = \int_{-1}^{1} x\frac{1}{2} dx = \underline{0} \tag{20}$$

Siden integranden er en 'odd' funksjon. For å regne ut variansen regner vi først ut:

$$E(X_i^2) = \int_{-1}^{1} x^2 f(x) dx = \int_{-1}^{1} \frac{1}{2} x^2 dx = \frac{1}{6} x^3 \Big|_{-1}^{1} = \frac{1}{3} \tag{21}$$

Vi har da at variansen er

$$\sigma^2 = V(X_i) = E(X_i^2) - E(X_i)^2 = \underline{\underline{\frac{1}{3}}} \tag{22}$$

Gammafordeling:

Vi har gammafordelingen:

$$f(x) = \begin{cases} \frac{1}{\sqrt{\pi x}} e^{-x} & x > 0 \\ 0 & ellers \end{cases} \tag{23}$$

Dette er en gammafordeling med $\alpha = 1/2$ og $\beta = 1$. Vi vet hva definisjonene på forventingsverdien og variansen er for gammafordelinger, så vi kan fort finne at for denne fordelingen er:

$$\mu = \alpha\beta = \underline{\underline{\frac{1}{2}}} \tag{24}$$

og

$$\sigma^2 = \alpha\beta^2 = \underline{\underline{\frac{1}{2}}} \tag{25}$$

Bernoullifordeling:

Vi har en Bernoullifordeling med punktfordelingen $p(x) = P(X = x)$, med en $p = 0.75$, som gir at $p(1) = 0.75$ og $p(0) = 0.25$. Vi vet også de generelle formelene for forventingsverdien og variansen til Bernoullifordelinger. Vi finner så at

$$\mu = p = \underline{\underline{0.75}} \tag{26}$$

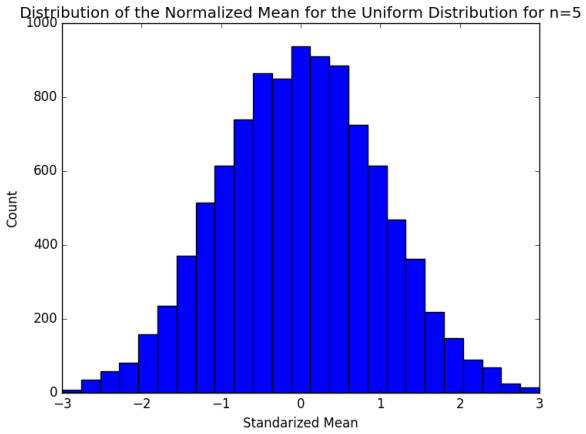
og

$$\sigma^2 = p(1-p) = \frac{1}{4} \cdot \frac{3}{4} = \underline{\underline{\frac{3}{16}}} \tag{27}$$

c)

Vi forventer å se at det normerte histogrammet av observasjonene nærmer seg sannsynlighetstetthetene til Z_n blir større og større. Teller vi opp antall observasjoner med en viss verdi/innen for et vist intervall vil forvente at jo høyere sannsynlighet for å finne denne observasjonen, jo flere tilfeller vil vi observere, og vis versa. Med nok antall observasjoner vil observasjonen alltid legge seg etter fordlingen til Z_n. Siden fordelingen til Z_n er normalisert, så må vi også normere histogrammet for at det skal likne på fordelingen til Z_n.

d)



Figur 1: Histogram for en uniform fordeling med $n = 5$

Vi ser over et histogram med de standariserte gjennomsnittene for en uniform fordeling. Siden vi regnet ut standardiserte gjennomsnitt av observasjoner av fordelingene, så vil vi etter sentralgrenseteoremet forvente at Z_n og dermed histogrammet skal gå mot en normalfordeling. Vi kan se at histogrammet over er veldig nært en normalfordeling. Rett rundt $\mu = 0$ er det noen ekstra topper, men ellers er fordelingen ganske normalfordelt.

e)

Verdiene under finnes i utskrifen til python-programmet.

Tabell for sannsynlighetene til en standard nomalfordelt variable:

Interval	Sannsynlighet
$-\infty$ til -2.5	0.0062
-2.5 til -2.0	0.0165
-2.0 til -1.5	0.0441
-1.5 til -1.0	0.0919
-1.0 til -0.5	0.1499
-0.5 til 0.0	0.1915
0.0 til 0.5	0.1915
0.5 til 1.0	0.1499
1.0 til 1.5	0.0919
1.5 til 2.0	0.0441
2.0 til 2.5	0.0165
2.5 til ∞	0.0062

Tabell 1: Sannsynlighetene til en standard nomalfordelt variable

f)

På samme måte som vi fant verdiene for en standard nomalfordelt variable, kan vi finne verdiene for vår fordelinge av det standardiserte gjennomsnittet av den uniforme fordelingen:

Interval	Sannsynlighet
$-\infty$ til -2.5	0.0044
-2.5 til -2.0	0.0171
-2.0 til -1.5	0.0457
-1.5 til -1.0	0.0926
-1.0 til -0.5	0.1486
-0.5 til 0.0	0.1905
0.0 til 0.5	0.1931
0.5 til 1.0	0.1494
1.0 til 1.5	0.0901
1.5 til 2.0	0.0469
2.0 til 2.5	0.0175
2.5 til ∞	0.0041

Tabell 2: Sannsynlighetene til en unifrom fordeling med $n = 5$

Vi kan se at disse verdiene er veldig nære de vi fant i deloppgaven over 1. Vi kan finne den relative forskjellen mellom fordelingen:

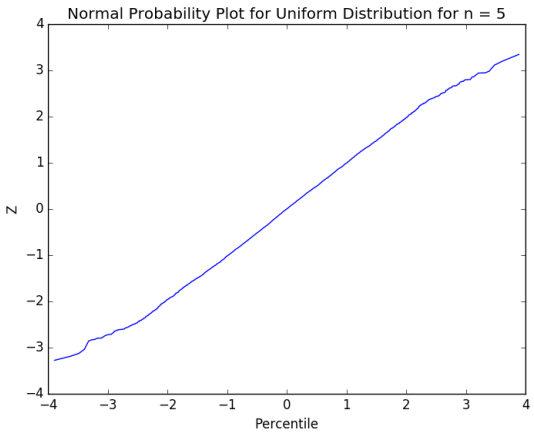
[0.19480363 0.03872119 0.02594205 0.05609207 0.02146828 0.04263218 0.02317707 0.01055685 0.01036437 0.04215145 0.00359926 0.24311541]

Interval	Relativ forskjell
$-\infty$ til -2.5	0.195
-2.5 til -2.0	0.039
-2.0 til -1.5	0.026
-1.5 til -1.0	0.056
-1.0 til -0.5	0.022
-0.5 til 0.0	0.043
0.0 til 0.5	0.023
0.5 til 1.0	0.011
1.0 til 1.5	0.01
1.5 til 2.0	0.042
2.0 til 2.5	0.004
2.5 til ∞	0.243

Tabell 3: Forskjellen til en unifrom fordeling med $n = 5$

Vi ser her at forskjellen mellom standard normalfordelingen og vår fordeling er svært liten, som vi forventet ut i fra histogrammet i 1.

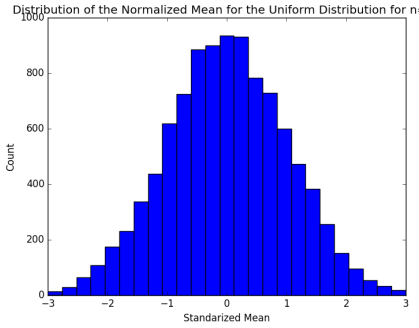
Vi kan så sjekke hvor standard normalfordelt det standariserte gjennomsnittet er ved hjelp av et normalsannsynlighetsplot. For den uniforme fordelingen med $n = 5$ blir dette:



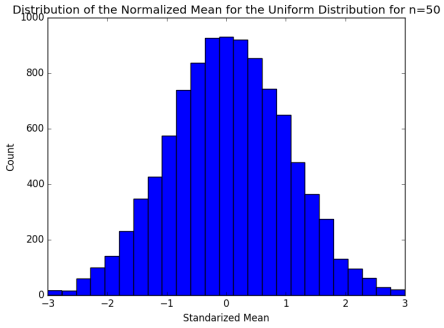
Figur 2: Normalsannsynlighetsplot for uniform fordeling med $n = 5$

Om vår fordeling hadde vært helt standardnormalfordelt ville vi forventet en helt rett linje. Vi kan se at fordelingen er ganske standard normalfordelt, men vi ser fra avbøyningene i endene at den ikke er helt perfekt. En av grunnene til dette kan være fordi vi har for så datapunkter, så vi kan ende opp med en litt skjev fordeling, som forutsaker avviket fra den rette linjen.

g)



(a) Histogram for en uniform fordeling med $n = 15$



(b) Histogram for en uniform fordeling med $n = 50$

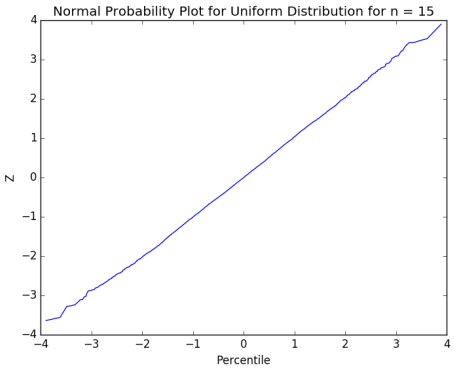
Vi kan se at vi nå får noe som ser mye mer ut som en standard normalfordeling. Dette er det vi forventer av flere verdier vår vi regner ut det standariserte gjennomsnittet. Ser vi på $n = 50$ ser vi at vi har en fordeling som er svært like en normalfordelingen. Vi kan så se på de relative frekvensene til fordelingen:

[0.1786997 0.05196549 0.03988167 0.03536333 0.03524289 0.01168657 0.02696331 0.02480423 0.07351214 0.04183053 0.05801126 0.03065136]

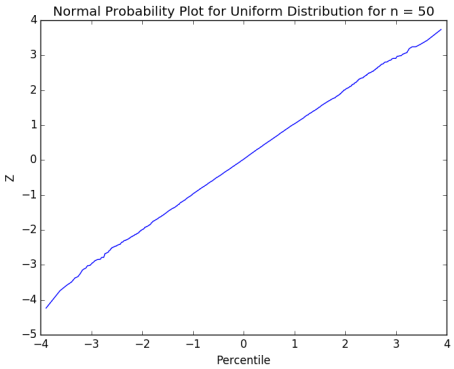
Interval	Sannsynlighet $n = 15$	Forskjell $n = 15$	Sannsynlighet $n = 50$	Forskjell $n = 50$
$-\infty$ til -2.5	0.0055	0.082	0.0065	0.179
-2.5 til -2.0	0.0174	0.112	0.0176	0.052
-2.0 til -1.5	0.0439	0.046	0.0424	0.04
-1.5 til -1.0	0.0928	0.0136	0.09	0.035
-1.0 til -0.5	0.1484	0.004	0.1499	0.035
-0.5 til 0.0	0.19	0.009	0.1933	0.012
0.0 til 0.5	0.1849	0.026	0.1963	0.027
0.5 til 1.0	0.1507	0.017	0.146	0.025
1.0 til 1.5	0.0957	0.038	0.0943	0.074
1.5 til 2.0	0.047	0.021	0.0417	0.042
2.0 til 2.5	0.0188	0.076	0.0171	0.058
2.5 til ∞	0.0049	0.134	0.0049	0.031

Tabell 4: Forskjellen til en unifrom fordeling med $n = 5$

Som vi forventet er det veldig liten forskjell mellom den relative frekvensen til vår fordeling og standard normalfordelingen. Selv om det er histogrammet til $n = 50$ som ser mest normalfordelt ut, så kan vi se fra den relative forskjellen at $n = 15$ er noe nærmere normalfordlingen enn $n = 50$, men ikke med mye. Vi ser noe av det samme om vi ser på normalsannsynlighetsplottene:



(a) Normalsannsynlighetsplot for uniform fordeling med $n = 15$

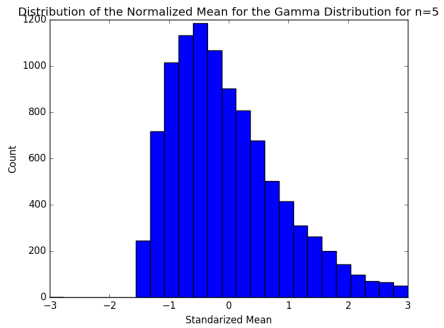


(b) Normalsannsynlighetsplot for uniform fordeling med $n = 50$

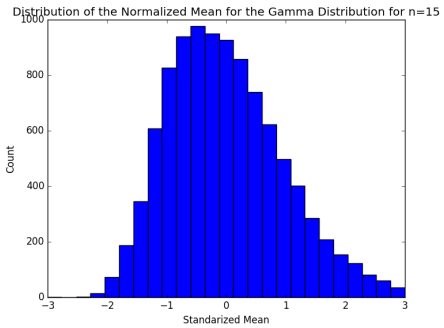
For begge verdiene av n har vi nesten en rett linje. Så vi kan se at vi ikke trenger mange datapunkter for det standardiserte gjennomsnittet før vi nesten har en standard normalfordeling.

h)

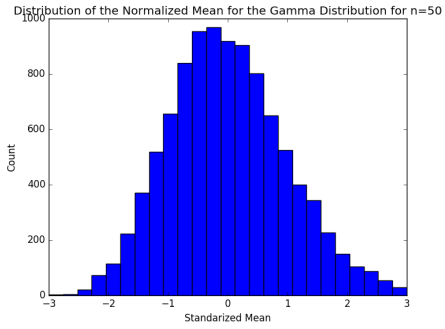
Vi skal nå gjøre det samme vi gjorde over bare med en gammafordeling. Først ser vi på histogrammene:



(a) Histogram for en gammafordeling med $n = 5$



(b) Histogram for en gammafordeling med $n = 15$



(c) Histogram for en gammafordeling med $n = 50$

Vi kan se at det standariserte gjennomsnitte til gammafordelingen ser mye mindre normalfordelt ut. For de lave verdiene av n har vi en ganske høy positiv skjevhet. For $n = 50$ begynner vi å se en likhet med normalfordelingen, men det er fortsatt noe skjevhet. Vi kan så se på forskjellen på

sannsynligheten til standardnormalfordelingen og de relative frekvensene. Siden det er så vanvittig mange tall, som tar veldig lang tid å skrive inn i tabeller, så gir jeg heller fra nå at utskrifen fra programmet mitt. Verdien er gitt i samme rekkefølge som intervallene oppgitt i oppgaven:

Relative frekvens for n=5

```
[ 0.      0.      0.0015  0.1239  0.2356  0.2201  0.1685  0.1035  0.0646
  0.0369  0.0191  0.0263]
```

Relativ forskjell fra standard normalfordelingn for n=5

```
[ 1.          1.          0.96595325  0.34896709  0.57190024  0.1495726
  0.11993192  0.30945808  0.29666446  0.16244996  0.15474372  3.23533292]
```

Relative frekvens for n=15

```
[ 0.      0.0022  0.032   0.115   0.1909  0.2016  0.172   0.1269  0.0793
  0.0402  0.0218  0.0181]
```

Relativ forskjell fra standard normalfordelingn for n=15

```
[ 1.          0.86699287  0.27366935  0.25206792  0.2736662   0.05294792
  0.10165158  0.15333556  0.13661751  0.08754712  0.31797975  1.91481087]
```

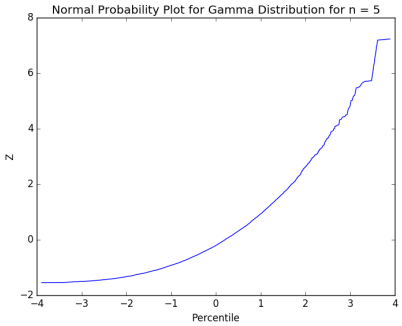
Relative frekvens for n=50

```
[ 0.0008  0.0105  0.0415  0.1014  0.1678  0.1977  0.1847  0.1342  0.0848
  0.0408  0.0221  0.0137]
```

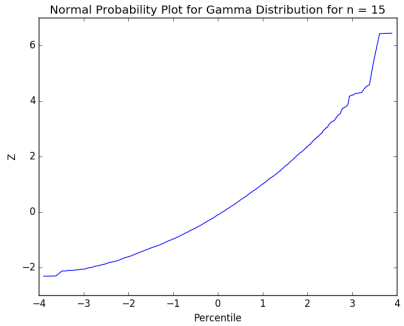
Relativ forskjell fra standard normalfordelingn for n=50

```
[ 0.87116858  0.36519324  0.05803993  0.10399728  0.11954525  0.03257839
  0.03532004  0.10463068  0.07673601  0.07392842  0.33611708  1.20623806]
```

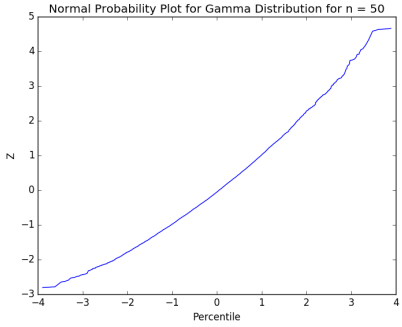
Her ser vi akkurat det samme som vi så på histogrammene; for de to lave verdiene av n har vi langt fra normalfordelingen. Ute i kantene har vi opp til over 300% forskjell mellom fordelingen og normalfordelingen. For $n = 50$ nærmer vi oss, men på kantene er det fortsatt rundt 100% forskjell, noe vi ikke var i nærheten av med den uniforme fordelingen. Vi ser på slutt på normalsannsynlighetsplottene:



(a) Normalsannsynlighetsplottet for en gammafordeling med $n = 5$



(b) Normalsannsynlighetsplottet for en gammafordeling med $n = 15$

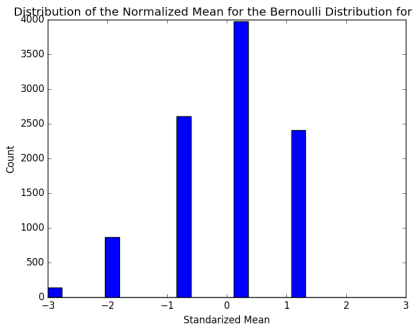


(c) Normalsannsynlighetsplottet for en gammafordeling med $n = 50$

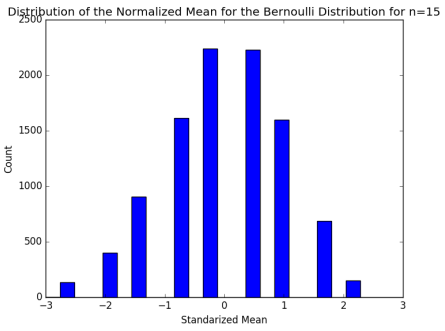
Som forventet er vi ikke i nærheten av å få rette linjer. Jo høyere n er, jo rettere blir linjen, men selv for $n = 50$ er vi ganske langt unna en rett linje. Vi ser at linjene er bøyd oppover, dette er fordi fordelingene hadde positiv skjevhet. Dette viser at det standariserte gjennomsnittet til en gammafordelingen går saktere mot standardnormalfordelingen.

i)

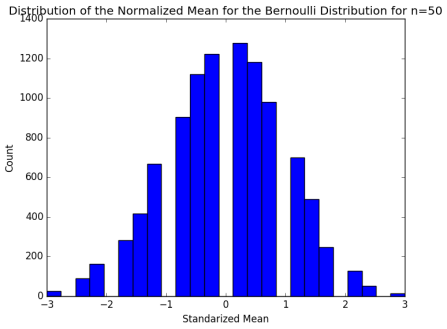
Vi skal nå se på de samme figurene for en Bernoullifordelinge. Vi starter med histogrammene:



(a) Histogram for en bernoullifordeling med $n = 5$



(b) Histogram for en bernoullifordeling med $n = 15$



(c) Histogram for en bernoullifordeling med $n = 50$

Siden Bernoullifordelingen gir et binært utfall, så er det bare en diskret mengde standardiserte gjennomsnitt det er mulig å oppnå for en gitt n. Dette kan vi se ut i fra histogrammene over. Det er derfor ganske vanskelig å se hva slags fordeling vi får. Men vi kan se at det har likheter med normalfordelingen, men har en negativ skjevhet(en hale mot venstre). For $n = 50$ har vi fått så mange punkter at det er mulig å se at fordelingen går mot standardnormalfordelingen, selv om vi fortsatt kan se noe negativ skjevhet. Det er ikke rart at vi ser en negativ skjevhet, siden Bernoullifordelingen er veldig skjev, siden et utfall har $p = 0.75$, mens det andre bare har $p = 0.25$. La oss så se på de relative frekvensene:

Relative frekvens for n=5

```
[ 0.0145  0.      0.0867  0.      0.261   0.      0.3973  0.      0.2405
  0.      0.      0.    ]
```

Relativ forskjell fra standard normalfordelingn for n=5

```
[ 1.33506948  1.          0.96790212  1.          0.74136657  1.          1.0750804
  1.          1.61845508  1.          1.          1.        ]
```

Relative frekvens for n=15

```
[ 0.0177  0.      0.04    0.0908  0.1615  0.2238  0.2228  0.      0.1598
  0.0687  0.0149  0.    ]
```

Relativ forskjell fra standard normalfordelingn for n=15

```
[ 1.85039516  1.          0.09208668  0.01141072  0.07751226  0.16889754
  0.16367458  1.          0.73983003  0.55934112  0.09917898  1.        ]
```

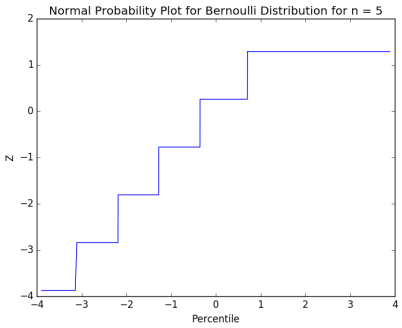
Relative frekvens for n=50

```
[ 0.0056  0.0253  0.0283  0.1087  0.0905  0.2341  0.2459  0.098   0.119
  0.0249  0.0179  0.0018]
```

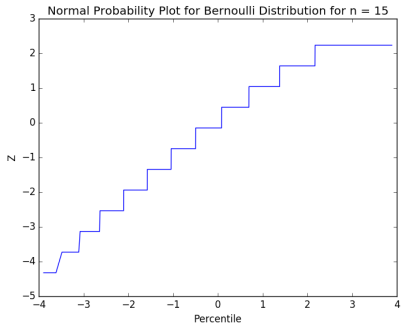
Relativ forskjell fra standard normalfordelingn for n=50

```
[ 0.09818006  0.529582    0.35765133  0.18347637  0.39619282  0.22269399
  0.28432487  0.34615355  0.29561811  0.43482396  0.08219438  0.71012931]
```

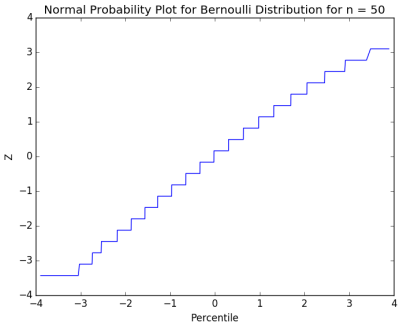
Igjen mangle er det data for mange av punktene der vi ikke har fått noe standardisert gjennomsnitt, og som vi da forventer, så er den relative forskjell her. Men for $n = 50$ har vi fått mer data, men likevel har vi en ganske stor relativ forskjell, generelt $20 - 70\%$. Vi kan så se på normalsannsynlighetsplottene:



(a) Normalsannsynlighetsplottet for en Bernoullifordeling med $n = 5$



(b) Normalsannsynlighetsplottet for en Bernoullifordeling med $n = 15$



(c) Normalsannsynlighetsplottet for en Bernoullifordeling med $n = 50$

Igjen, siden vi bare kan få noen diskrete verdier for det standardiserte gjennomsnittet, så får vi en hakkete kurve. Selv om det er vanskelig å få en trend fra de lave verdiene av n, så ser vi at vi får en nogenlund rett kurve etterhvert. Men den bøyer seg litt nedover, som reflekterer den negative skjevheten. Vi kan se det standardiserte gjennomsnittet til en Bernoullifordelingen har en moderat likehet med standardnormalfordelingen, selv om den selv for store n er fordelingen litt skjev grunnet verdien av p for Bernoullifordelingen.

j)

Som vi har sett i de forrige deloppgavene går de standardiserte gjennomsnitten til de forskjellige fordelingene alle mot standardnormalfordelingen, men med forskjellige 'hastigheter'.

Av de fordelingen vi har sett på, så er det den uniforme fordelingen som ligner mest på normalfordelingen. Jeg tror det har å gjøre med at den uniforme fordelingen har sannsynlighetsfordelingen symmetrisk rundt gjennomsnittet, akkurat som normalfordelingen. Mens Bernoulli- og gammafordelingen har en, forholdsvis, negativ og positiv skjevhet om gjennomsnittet. Dette gjør at vi trenger høyere verdier av n for at disse fordelingen skal ligne på normalfordelingen.

Om denne hypotesen er korrekt har vi for få fordelinger til å si konkret. Men det vi hvertfall kan se ut fra oppgaven er at den uniforme fordelingen nærmer seg normalfordelingen raskere enn Bernoullifordelingen, som nærmer seg raskere enn gammafordelingen.