Gennemsnit og normalfordeling illustreret med terningkast, simulering og SLUMP()

John Andersen, Læreruddannelsen i Aarhus, VIA

Et kast med 10 terninger gav følgende udfald

Fig. 1 Result of rolling 10 dices

Summen af øjnene divideret med 10 giver gennemsnittet

$$\frac{1+1+3+6+1+5+5+5+1+6}{10} = \frac{17}{5} = 3.4$$

Kaster man mange gange med de 10 terninger får man en masse gennemsnit der alle vil tilhøre talmængden

$$\{1, 1.1, 1.2, 1.3, ..., 5.8, 5.9, 6\}$$

Ved hjælp af Excel kan vi undersøge om der er et mønster i disse gennemsnit.

Simulering af 100 kast med 10 terninger

Fig. 2 viser resultatet af 100 simuleringer af gennemsnit ved kast med 10 terninger.

	A	B	C	D	E	F	G	H	I	J
1	Gennemsnit af 100 kast med 10 terninger									
2	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6
3	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6
4	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6
5	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6
6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6
7	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6
8	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6
9	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6
10	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6
11	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6

Fig. 2 Skærmbillede fra Excel

Formlen i hver celle er

=(SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6))/10

Den er skrevet ind i A2 og derpå kopieret til de øvrige. Under indskrivninger og også brugt kopiering så skrivearbejdet var ikke så stort som man måske tror når man ser på formlen. SLUMPMELLEM-funktionen er indrettet sådan at hver udgave vi skriver af den fungerer uafhængigt af de øvrige. Det er det der gør at denne formel kan bruges til en simulering som denne.

Fig. 3 viser formler til optælling af de simulerede gennemsnit

K	L	M	N
Antal kast	**Mulige værdier**	**Hyppigheder**	**Frekvenser**
100	1	0	0
	1,1	0	0
	1,2	0	0
	1,3	0	0
	1,4	0	0
	1,5	0	0
	1,6	0	0
	1,7	0	0
	1,8	0	0
	1,9	0	0
	2	0	0
	2,1	1	0,01
	2,2	1	0,01
	2,3	0	0
	2,4	1	0,01
	2,5	1	0,01
	2,6	5	0,05
	2,7	2	0,02
	2,8	5	0,05
	2,9	7	0,07
	3	8	0,08

K	L	M	N
Antal kast	**Mulige værdier**	**Hyppigheder**	**Frekvenser**
=TÆL(A2:J11)	1	=TÆL.HVIS($A$2:$J$11;L2)	=M2/$K$2
	1,1	=TÆL.HVIS($A$2:$J$11;L3)	=M3/$K$2
	1,2	=TÆL.HVIS($A$2:$J$11;L4)	=M4/$K$2
	1,3	=TÆL.HVIS($A$2:$J$11;L5)	=M5/$K$2
	1,4	=TÆL.HVIS($A$2:$J$11;L6)	=M6/$K$2
	1,5	=TÆL.HVIS($A$2:$J$11;L7)	=M7/$K$2
	1,6	=TÆL.HVIS($A$2:$J$11;L8)	=M8/$K$2
	1,7	=TÆL.HVIS($A$2:$J$11;L9)	=M9/$K$2
	1,8	=TÆL.HVIS($A$2:$J$11;L10)	=M10/$K$2
	1,9	=TÆL.HVIS($A$2:$J$11;L11)	=M11/$K$2
	2	=TÆL.HVIS($A$2:$J$11;L12)	=M12/$K$2
	2,1	=TÆL.HVIS($A$2:$J$11;L13)	=M13/$K$2
	2,2	=TÆL.HVIS($A$2:$J$11;L14)	=M14/$K$2
	2,3	=TÆL.HVIS($A$2:$J$11;L15)	=M15/$K$2
	2,4	=TÆL.HVIS($A$2:$J$11;L16)	=M16/$K$2
	2,5	=TÆL.HVIS($A$2:$J$11;L17)	=M17/$K$2
	2,6	=TÆL.HVIS($A$2:$J$11;L18)	=M18/$K$2
	2,7	=TÆL.HVIS($A$2:$J$11;L19)	=M19/$K$2
	2,8	=TÆL.HVIS($A$2:$J$11;L20)	=M20/$K$2
	2,9	=TÆL.HVIS($A$2:$J$11;L21)	=M21/$K$2
	3	=TÆL.HVIS($A$2:$J$11;L22)	=M22/$K$2

Fig. 3 Skærmbilleder der viser hvilke formler der er brugt til at optælle og beregne frekvenser af de simulerede gennemsnit

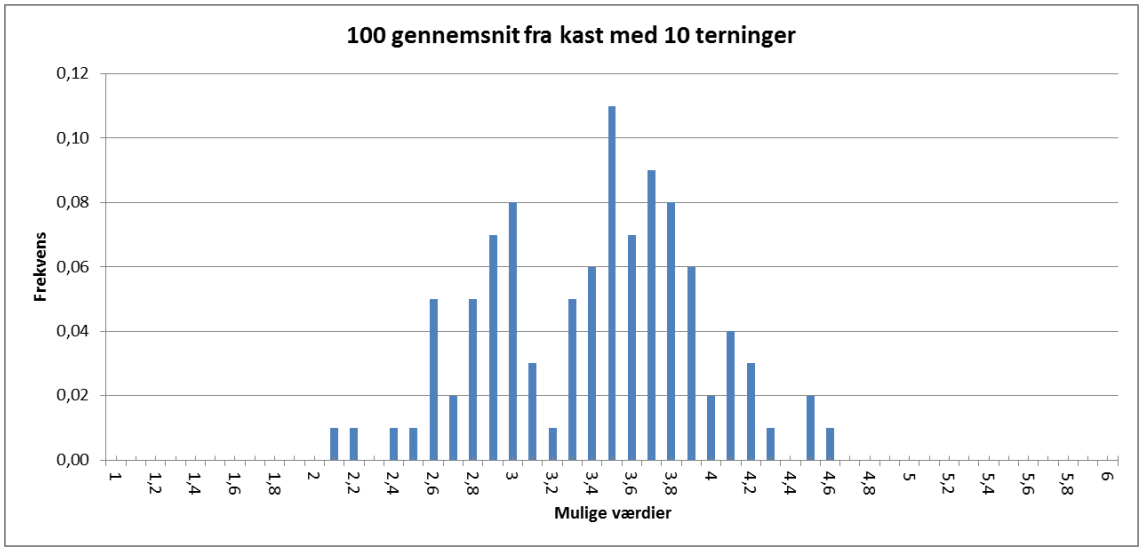

Fig. 4 Pindediagram der visualiserer statistikken. Denne diagramtyper er valgt fordi vi kender de mulige, adskilte (diskrete) værdier af gennemsnittene.

Så snart regnearket er udformet kan vi se resultatet af en ny simulering af 100 kast ved at trykke på F9-tasten (beregn). Det kan du gøre en masse gange og se at gennemsnittene ganske vist fordeler sig ret så tilfældigt, men alligevel er der en tendens til at de samler sig i midten.

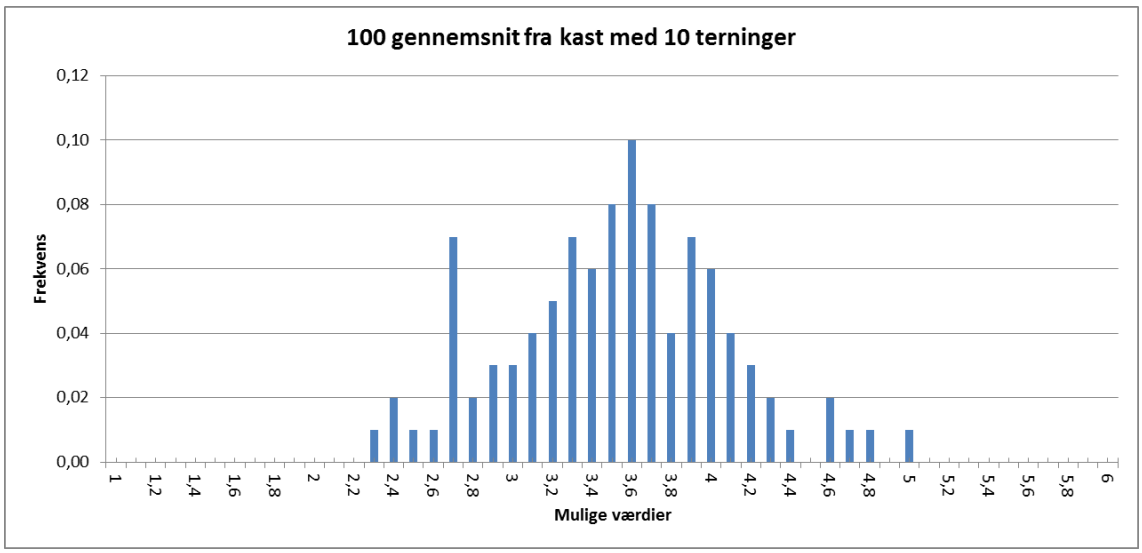

Fig. 5 100 nye kast med de 10 terninger

Med computer og regneark kan man håndtere store datamængder så lad os se hvad der sker, hvis vi kaster de 10 terninger nogle flere gange. Det er de samme formler der skal bruges. SLUMPMELLEM-formlerne skal blot kopieres til et større område og optællings området skal justeres så det passer dertil. På de næste to figurer ser de statistikker over 5000 kast med 10 terninger.

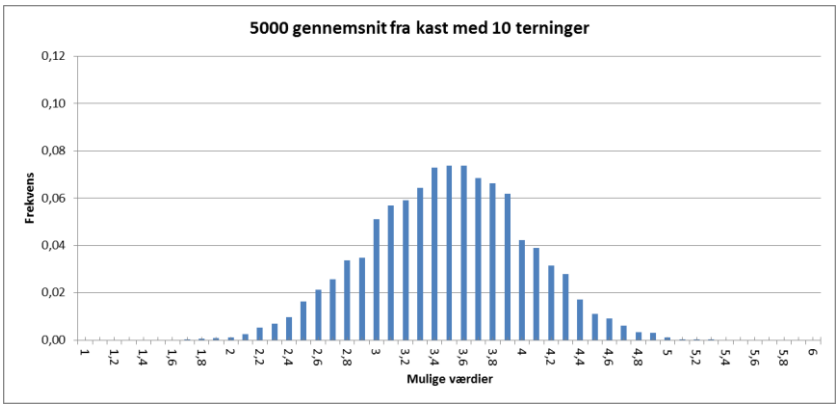

Fig. 6 Pindediagram over 5000 kast med 10 terninger

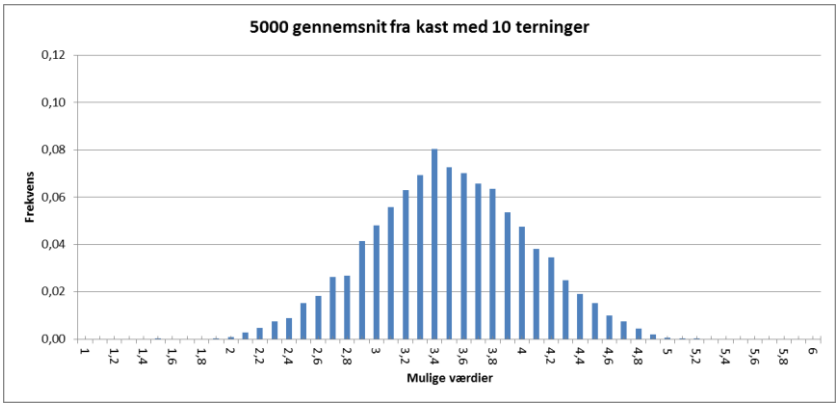

Fig. 7 Ved at trykke på F9 får man 5000 nye kast.

Vi kan nu se et mønster der ser mere regelmæssigt ud, og man får en ide om, hvad der kan forventes - eller ikke forventes. Det vil nok ikke være klogt at sætte en formue på gennemsnit tæt ved 1 eller 6.

Nu kan det være spændende at se hvad der sker hvis vi kaster endnu flere gange. Hvad med 100.000? Kan det lade sig gøre? Er regnearket og computerens hukommelse stor nok?

Kører jeg helt ned i bunden af mit regneark kan jeg se at der er $2^{20} = 1048576$ rækker og $2^{14} = 16384$ søjler at gøre godt med, så lad os prøve at se hvad der sker:

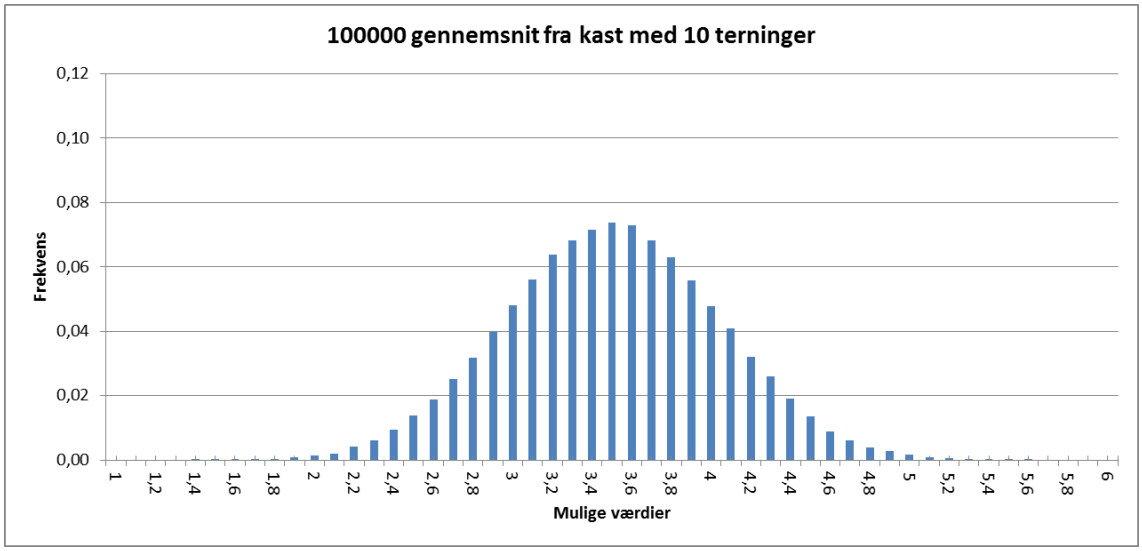

Fig. 8 En smuk "klokkeformet" fordeling af pinde for de 100000 kast.

Et tryk på F9 giver et nyt billede (Fig. 9) der til forveksling ligner billedet på Fig. 8. og gentagelser resulterer i det samme.

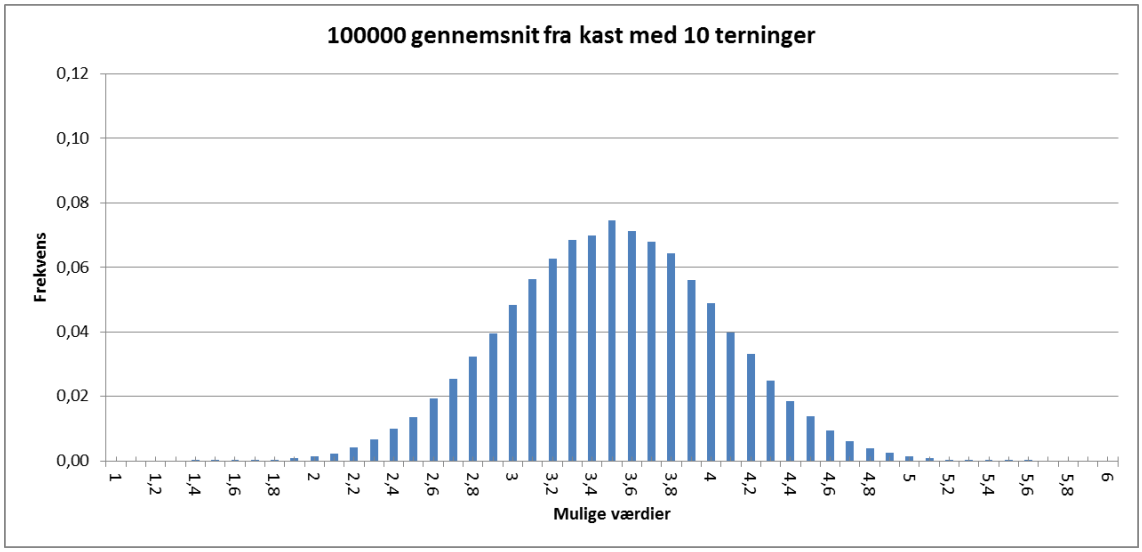

Fig. 9 En ny simulering af 100000 kast med 10 terninger giver gennemsnit, der er fordelt næsten som på Fig. 8.

Forskellen fra regnearket der blev brugt til Fig. 3 er at området der simulerers i og optælles over er noget større:

K	L	M	N
Antal kast	**Mulige værdier**	**Hyppigheder**	**Frekvenser**
=TÆL(A2:J10001)	1	=TÆL.HVIS($A$2:$J$10001;L2)	=M2/$K$2
	1,1	=TÆL.HVIS($A$2:$J$10001;L3)	=M3/$K$2
	1,2	=TÆL.HVIS($A$2:$J$10001;L4)	=M4/$K$2
	1,3	=TÆL.HVIS($A$2:$J$10001;L5)	=M5/$K$2
	1,4	=TÆL.HVIS($A$2:$J$10001;L6)	=M6/$K$2
	1,5	=TÆL.HVIS($A$2:$J$10001;L7)	=M7/$K$2
	1,6	=TÆL.HVIS($A$2:$J$10001;L8)	=M8/$K$2
	1,7	=TÆL.HVIS($A$2:$J$10001;L9)	=M9/$K$2

Fig. 10 Del af formlerne der bliver brugt til optællingen af de 100000 kast med de 10 terninger.

Den regelmæssige form af pindediagrammet på Fig. 9 antyder at der måske er en eller anden formel for en kurve der kan beskrive formen af det samlede billede af pindene - altså f.eks. afgrænsningen opad. Det gør der - jeg havde nok slet ikke skrevet dette kapitel hvis ikke det var fordi jeg på forhånd vidste det. Det er mange år siden at den blev fundet. Det drejer sig om normalfordelingen eller Gaussfordelingen som den også kaldes.

Baggrunden for dette kapitel er et ønske om at give et konkret eksempel på hvad konklusionen i den såkaldte centrale grænseværdisætning siger.

Jeg vil ikke gå ind i mere formulerings- og bevistekniske detaljer omkring denne emnekreds men holde mig til konkrete eksempler der er til at tage og føle på - i det omfang man kan sige at computersimuleringer er til at tage og føle på. Hvis du vil lidt mere ind i detaljeret teoretisk baggrund så se f.eks. [1] og [2].

Grænsefordelingen af gennemsnittene ved kast med 10 terninger

Hvad skal man se efter? Den centrale grænseværdisætning fortæller at gennemsnittene tilnærmelsesvis (jo flere kast desto bedre) normalfordelt med middelværdi m og varians $s^2/10$, hvor m og s^2 er middelværdi og varians for et kast med en terning. For at minde om definitioner på disse størrelser vælger jeg her at beregne dem direkte med regnearket (Se Fig. 11 & 12):

	A	B	C	D	E
1	**Antal øjne**	**Sandsynlighed**	**Middelværdi-beregning**	**Variansberegning**	
2	**k**	**P(D = k)**	**k·P(D = k)**	**$(k-m)^2 \cdot P(D = k)$**	
3	1	1/6	0,1667	1,0417	
4	2	1/6	0,3333	0,3750	
5	3	1/6	0,5000	0,0417	
6	4	1/6	0,6667	0,0417	
7	5	1/6	0,8333	0,3750	
8	6	1/6	1,0000	1,0417	
9			**Middelværdi**	**Varians**	**Spredning**
10			**$\Sigma k \cdot P(D = k)$**	**$\Sigma(k-m)^2 \cdot P(D = k)$**	
11			**m**	**s^2**	**s**
12			3,5000	2,9167	1,7078

Fig. 11 Beregning af middelværdi, varians og spredning for en terning.

	A	B	C	D	E
1	**Antal øjne**	**Sandsynlighed**	**Middelværdi-beregning**	**Variansberegning**	
2	**k**	**P(D = k)**	**k·P(D = k)**	**$(k-m)^2 \cdot P(D = k)$**	
3	1	=1/6	=A3*B3	=(A3-$C$12)^2*B3	
4	2	=1/6	=A4*B4	=(A4-$C$12)^2*B4	
5	3	=1/6	=A5*B5	=(A5-$C$12)^2*B5	
6	4	=1/6	=A6*B6	=(A6-$C$12)^2*B6	
7	5	=1/6	=A7*B7	=(A7-$C$12)^2*B7	
8	6	=1/6	=A8*B8	=(A8-$C$12)^2*B8	
9			**Middelværdi**	**Varians**	**Spredning**
10			**$\Sigma k \cdot P(D = k)$**	**$\Sigma(k-m)^2 \cdot P(D = k)$**	
11			**m**	**s^2**	**s**
12			=SUM(C3:C8)	=SUM(D3:D8)	=KVROD(D12)

Fig. 12 Formler til beregningerne i Fig. 11.

Da normalfordelingen er en kontinuert fordeling kan vi ikke direkte sammenligne dens sandsynlighedstæthed med pindediagrammer som på Fig. 9.

Hvis vi betragter de 100000 gennemsnit som ombservationer af kontinuerte (grupperede) observationer kan vi derimod lave en grafisk fremstilling som direkte kan sammenlignes med de teoretiske grafer fra normalfordelingen. Det gør vi ved at fremstille histogram og sumkurve for de 100000 gennemsnit baseret på en eller anden intervalinddeling.

Da vores talmateriale består at 100000 observationer er der rigeligt mange til at vi kan få det mere detaljeret billede ved at vælge smallere grupperingsintervaller. På Fig. 14 & 15 er brugt inddelingen 1,05<1,25<1,45<.....<5,85 for at undgå af delepunkterne rammer oven i de mulige værdier af gennemsnittene (der jo ikke er ægte kontinuerte observationer jvnf. bemærkningen lige efter Fig. 1)

Vi ser et næsten perfekt sammenfald med normalfordelingen:

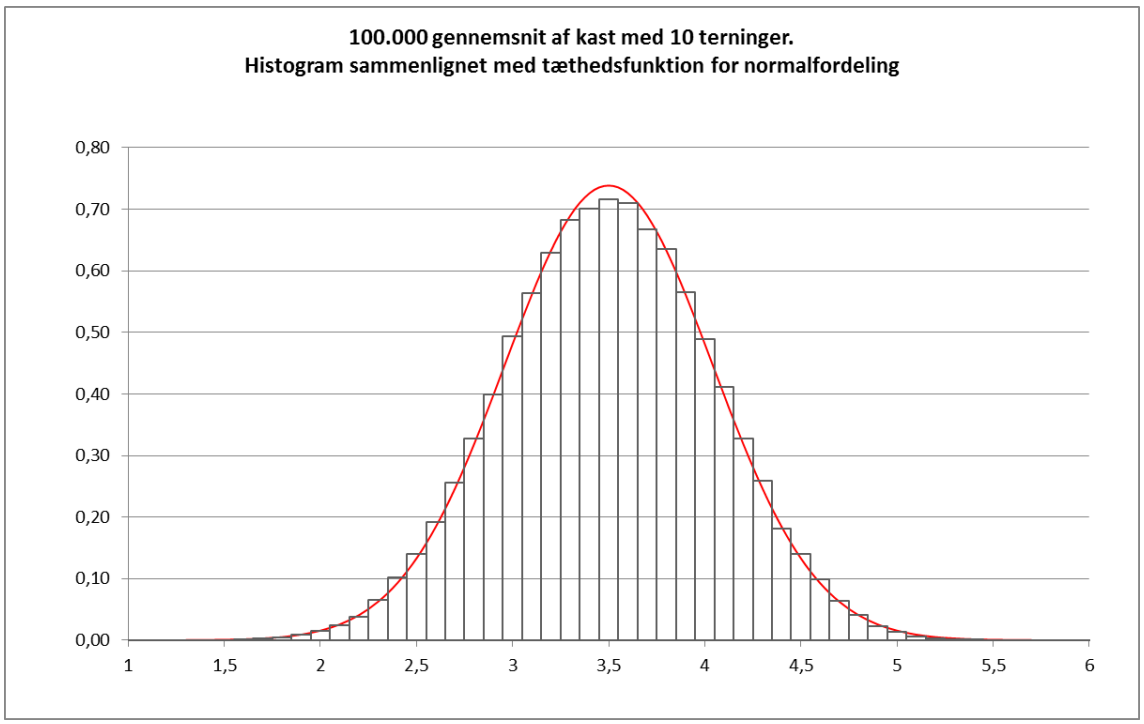

Fig. 13 Histogram sammenlignet med sandsynlighedstæthed for normalfordeling med middelværdi 3,5 spredning $^{1,7078}/_{\sqrt{10}} = 0{,}54$

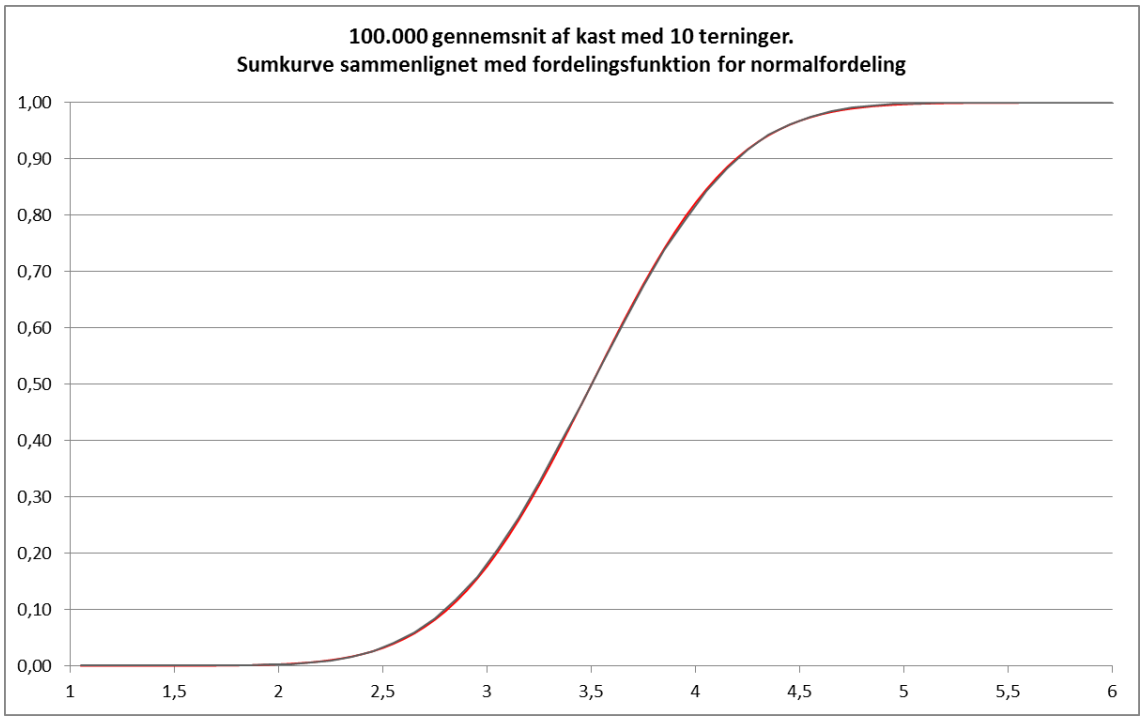

Fig. 14 Sumkurve sammenlignet med fordelingsfunktion for normalfordeling med samme parametre som på Fig. 13. Der er meget fint sammenfald og man skal zoome ind for at se forskellene - se Fig. 15

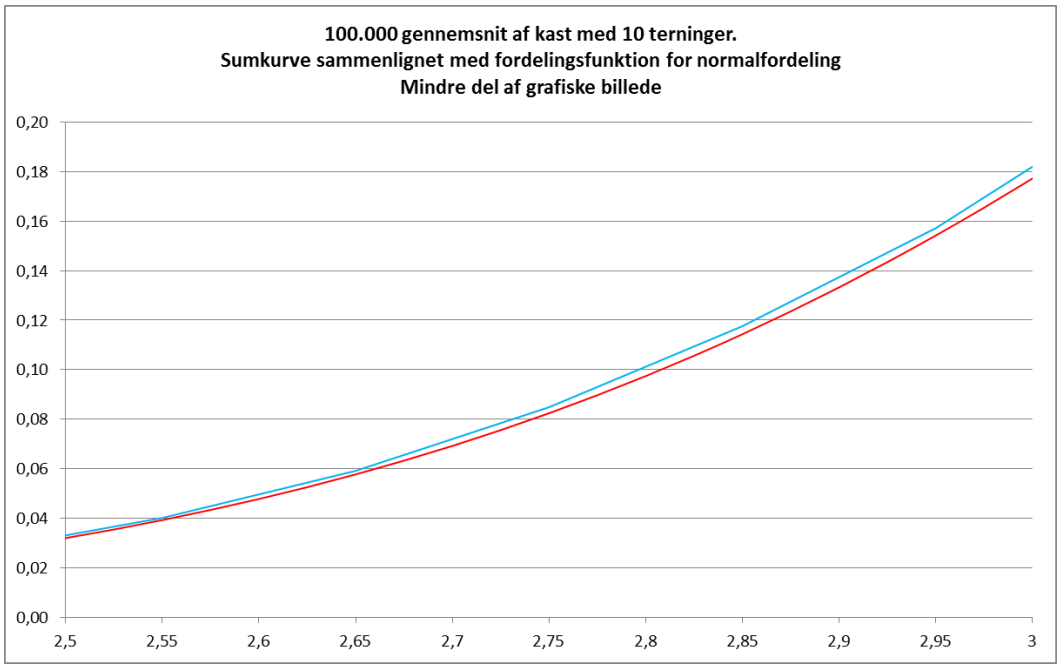

Fig. 15 Sumkurve sammenlignet med fordelingsfunktion for normalfordeling med samme parametre som på Fig. 13. Her er der zoomet ind på en del af graferne i Fig. 14 og man kan se at der er forskel (blå graf er sumkurve, rød graf er fordelingsfunktion).

Fordelingen af gennemsnittene ved kast med 100 terninger

I næste omgang vil vi prøve at kaste 100 terninger og beregne deres gennemsnitlige øjental. Det gøre med et regneark, der næsten er en kopi af regnearket i Fig. 3. Formlen på Fig. 17 til beregning af gennemsnittet indeholder 100 kopier af SLUMPMELLEM(1;6) og det kan umiddelbart lyde som noget af en tålmodighedsprøve at indtaste den. Men skriver man den f.eks. i Word bliver det en overkommelig opgave hvis man bruger "kopier" og "sæt ind" nogle et passende antal gange. Den færdige formel kopieres fra Word over i den første celle i regnearket. Derpå kopieres den ved træk med musen til de celler der skal indgå i simuleringen.

```
=(SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;
6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1
;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(
1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM
(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLE
M(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELL
EM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMEL
LEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPME
LLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPM
ELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMP
MELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUM
PMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLU
MPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SL
UMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+S
LUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+
SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)
+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6
)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;
6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1
;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(
1;6)+SLUMPMELLEM(1;6)])/100
```

Fig. 16 Denne lange formel beregner gennemsnitligt øjental for et kast med 100 terninger. Når man bruger diverse kopieringsmuligheder er der ikke megen forskel på at arbejde med lange eller korte formler.

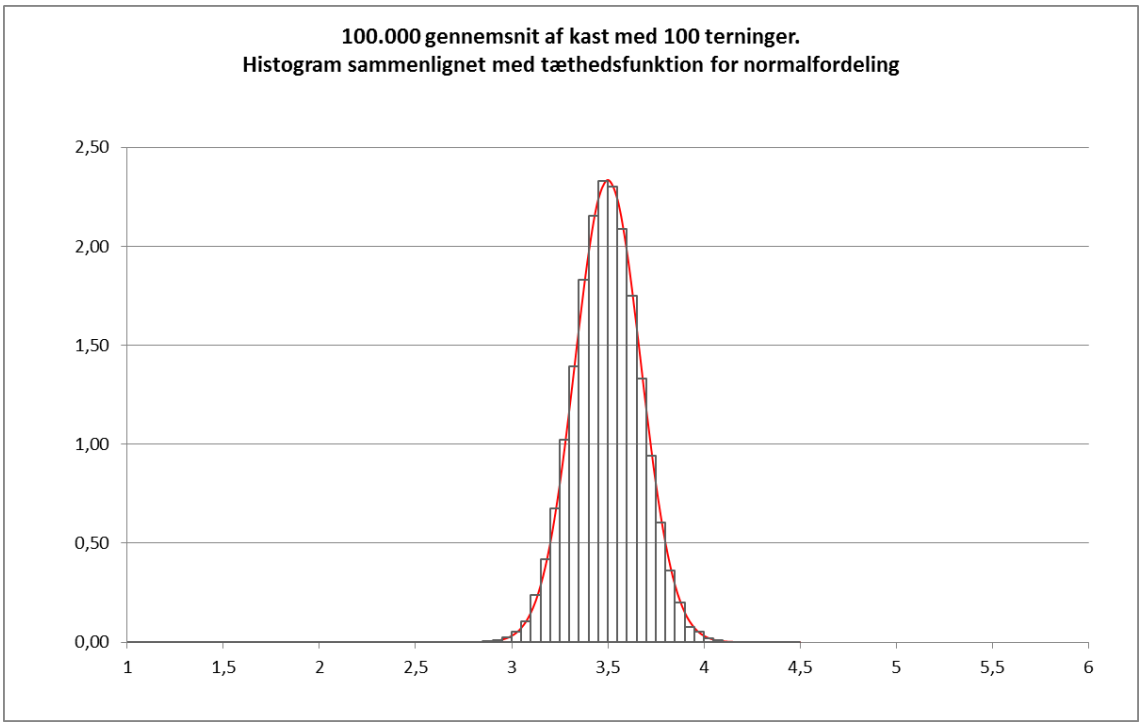

Fig. 157 Histogram sammenlignet med sandsynlighedstæthed for normalfordeling med middelværdi 3,5 spredning $^{1,7078}/_{\sqrt{100}} = 0,1708$

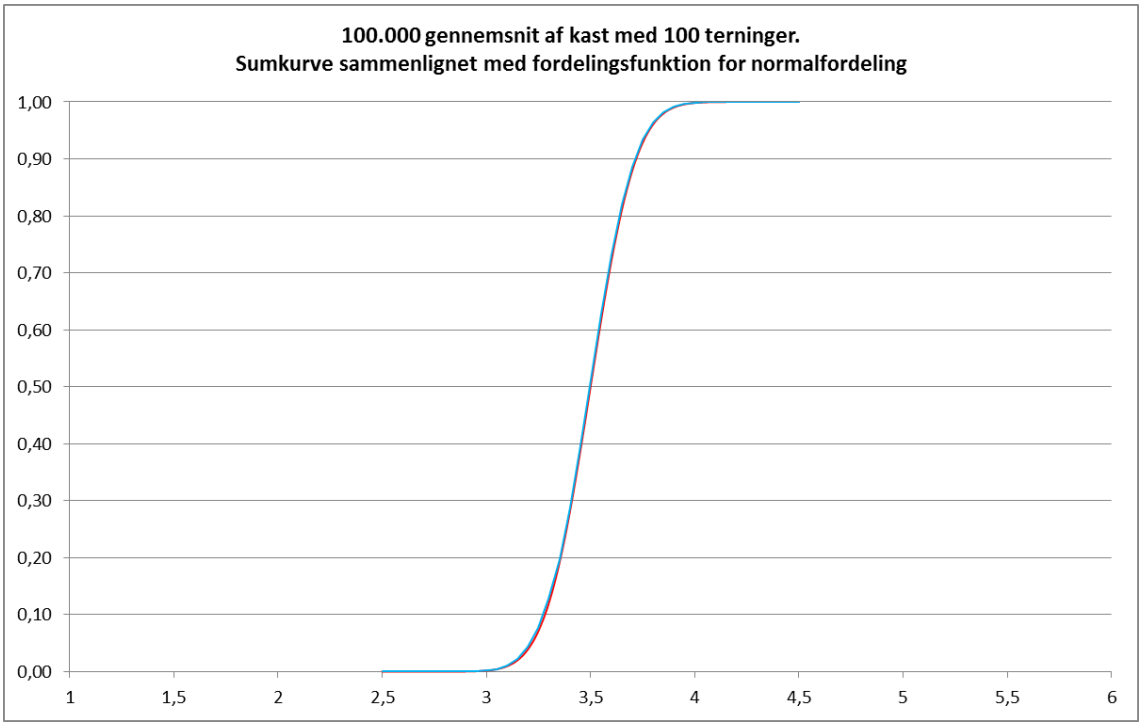

Fig. 168 Sumkurve sammenlignet med fordelingsfunktion for normalfordeling med samme parametre som på Fig. 17.

Sammenfatning vedrørende terningkastene

Vi har set forskellige muligheder for at variere disse simuleringer:

På Fig. 4 - Fig. 9 varieres antallet af kast men der kastes hver gang 10 terninger. Jo flere kast desto mere regelmæssigt mønster. Det er store tals lov der illustreres der.

På Fig 13. - 18. er det antallet af terninger der varieres. Læg bl.a. andet mærke til at spredningen af de observerede gennemsnit bliver mindre desto flere terninger der kastes. Sammenlign f.eks. Fig. 13. og 17. Det er måske den fornemmelse de fleste har: Når man laver flere målinger (her flere terninger) til beregning af et gennemsnit så får man den forventede gennemsnitsværdi bestemt med større nøjagtighed.

Det som den centrale grænseværdisætning fortæller os er bl.a. at når antallet af terninger øges med en faktor k så formindskes spredningen med en faktor $1/\sqrt{k}$, f.eks. fra Fig. 13 til Fig. 17. formindskes spredningen med en faktor $1/\sqrt{10} \approx 1/3$. Tjek selv på de to figurer: Sammenlign bredderne af de klokkeformede områder.

Der er også den mulighed at man kan variere finheden af intervalinddelingerne når vi grupperer data. Men så går noget tilsyneladende galt. Eller rettere sagt: Det går ikke galt, men det bliver afsløret at gennemsnittene af terningerne ikke er kontinuerte variable men derimod diskrete variable:

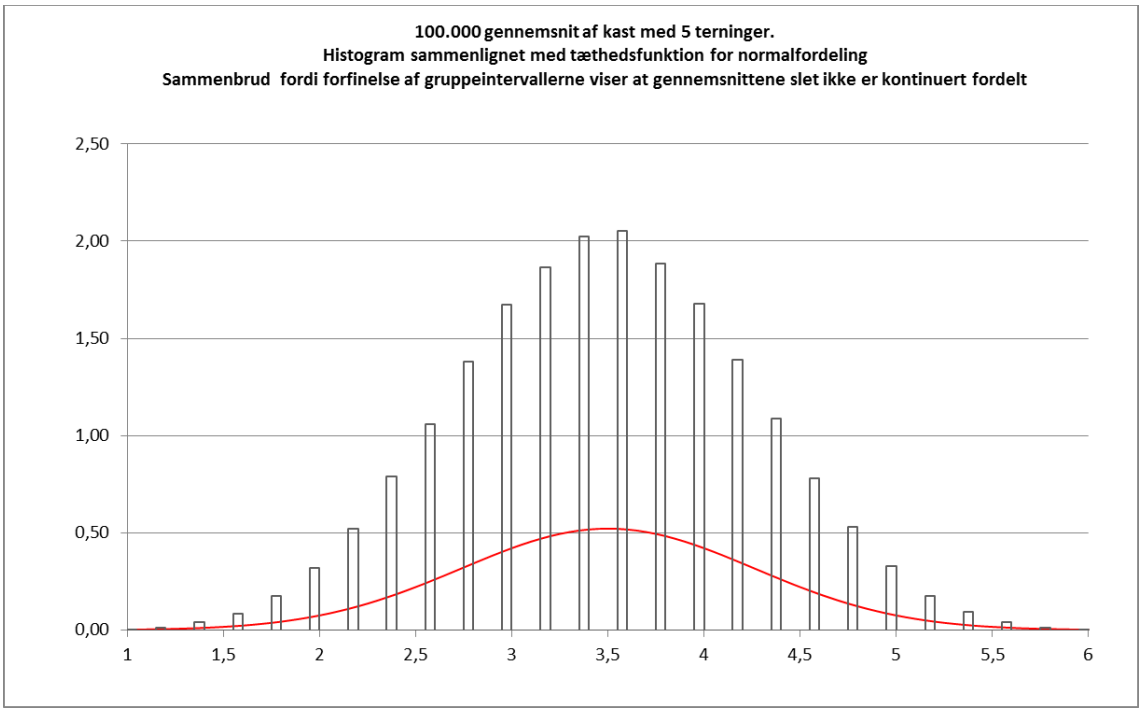

Fig 19. Gennemsnittene af terningerne afsløret: De er ikke kontinuerte variable.

Hvad gør jeg så nu?

Det her blev afsløret fordi jeg gik ned i antallet af terninger. Så blev springene mellem gennemsnitsværdierne efterhånden så store at det blev afsløret ved forfinelse af gruppeintervallerne.

Der er ellers endnu et aspekt ved den centrale grænseværdisætning jeg gerne vil vise og det kræver at jeg kan komme helt ned på gennemsnit af to variable. Fordi terninger producerer diskrete variable skifter jeg nu terningerne, der jo giver hele tal ud mellem 1 og 6 begge inklusive, så vælger jeg at erstatte terningerne med variable der leverer tilfældige tal, jævnt fordelt i intervallet [1;6].

Den indbyggede funktion SLUMP() giver tilfældige tal jævnt fordelt i intervallet [0;1]. Derfor kan jeg bruge funktionen 1 + 5*SLUMP() til at skaffe mig tilfældige tal i intervallet [1;6] og så udskifte SLUMPMELLEM(1;6) i alle regneark der tidligere er blevet brugt med denne nye funktion.

Middelværdien af denne nye variabel er 3,5 og dens varians er 31,25/15 og dermed er dens spredning 1,4433. Den centrale grænseværdisætning siger nu at hvis man tager gennemsnittet af n af hinanden uafhængige udgaver af denne variabel så vil dette gennemsnit være approksimativt normalfordelt med middelværdi 3,5 og spredning $^{1,4434}/_{\sqrt{n}}$. Nedenfor vises nu en billedserier der illustrerer dette. En mere formaliseret og stringent formalisering ligger uden for denne fremstillings intentioner.

Vær opmærksom på at ved histogrammerne ændres skaleringen på 2. aksen. Men bortset fra dette så viser figurerne at fordelingen er ret langt fra en normalfordeling når antallet der tages gennemsnit er så lille som 2 og 3 men derefter kommer den til at ligne mere og mere.

Jeg stopper her og lader billederne tale for sig selv.

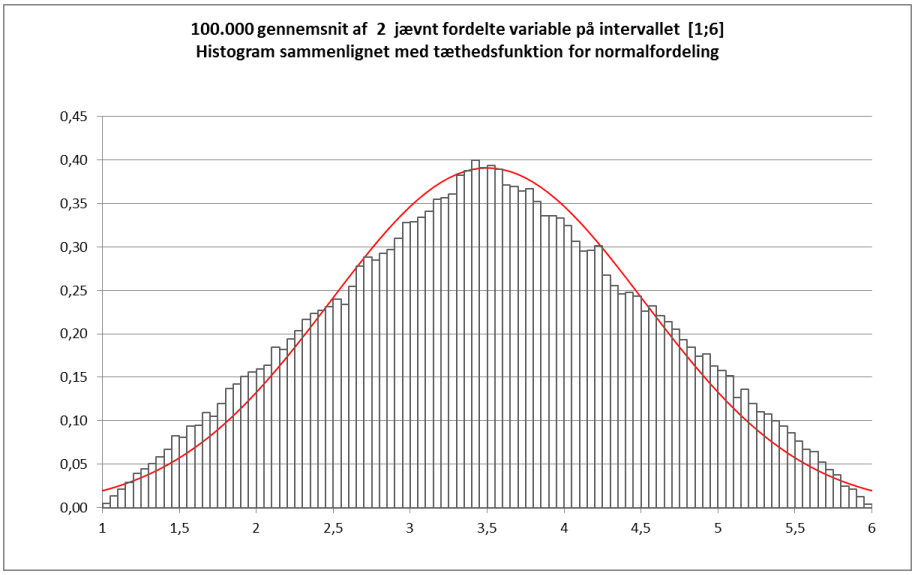

Fig 19

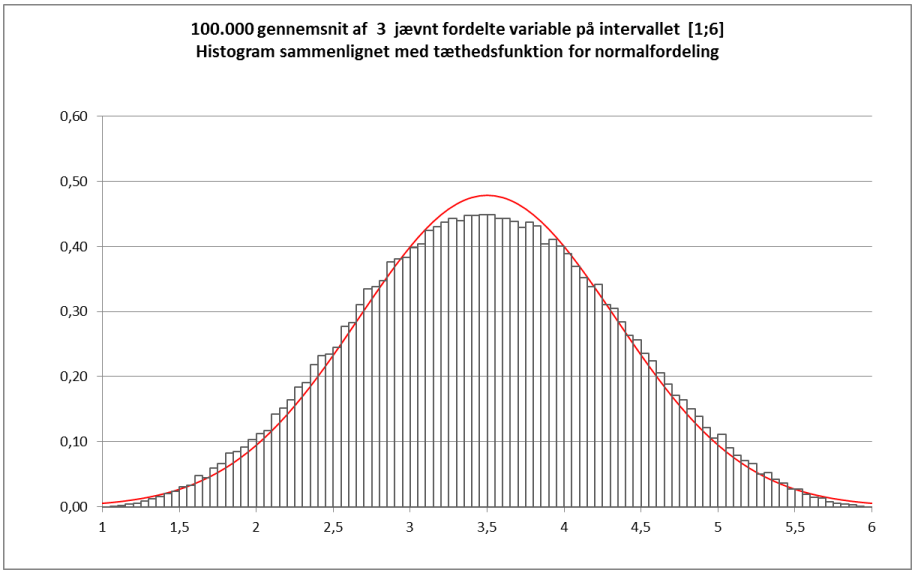

Fig 20

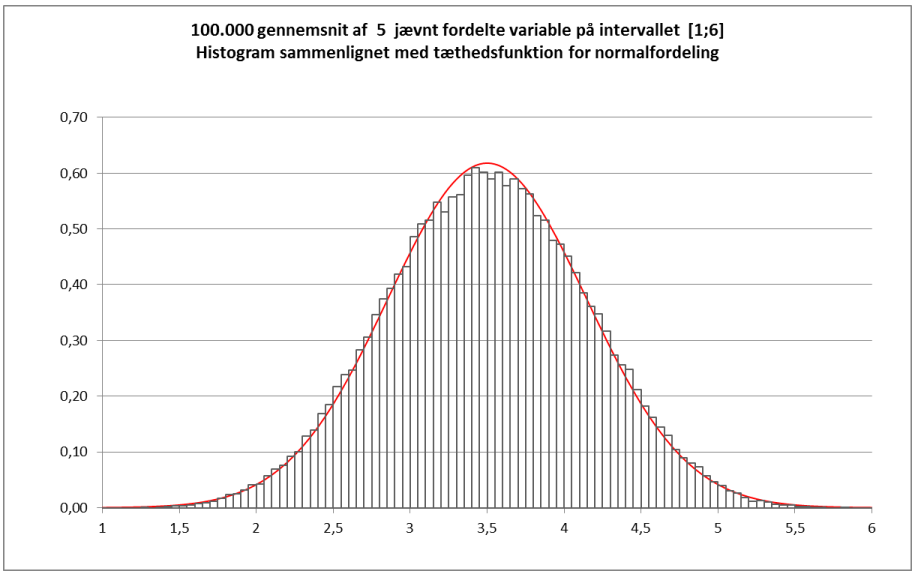

Fig 21

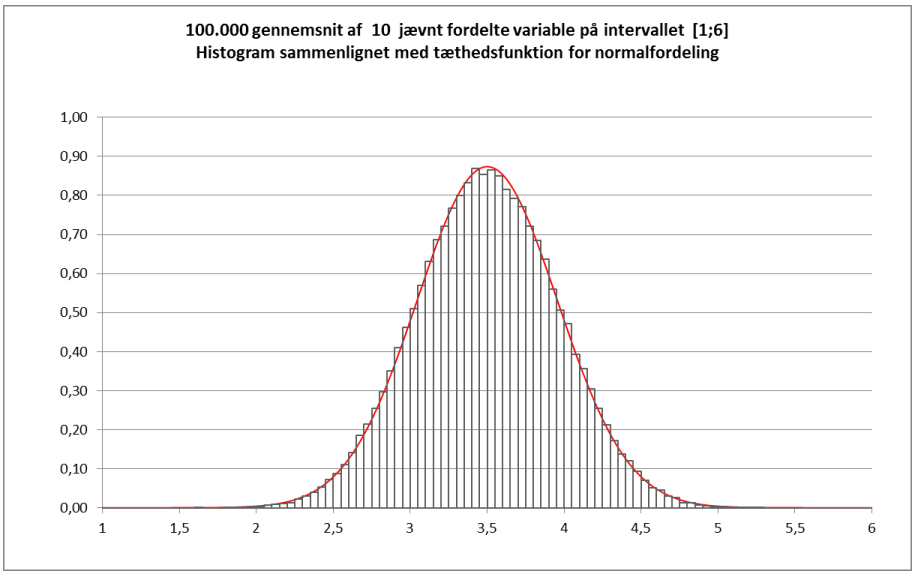

Fig 22

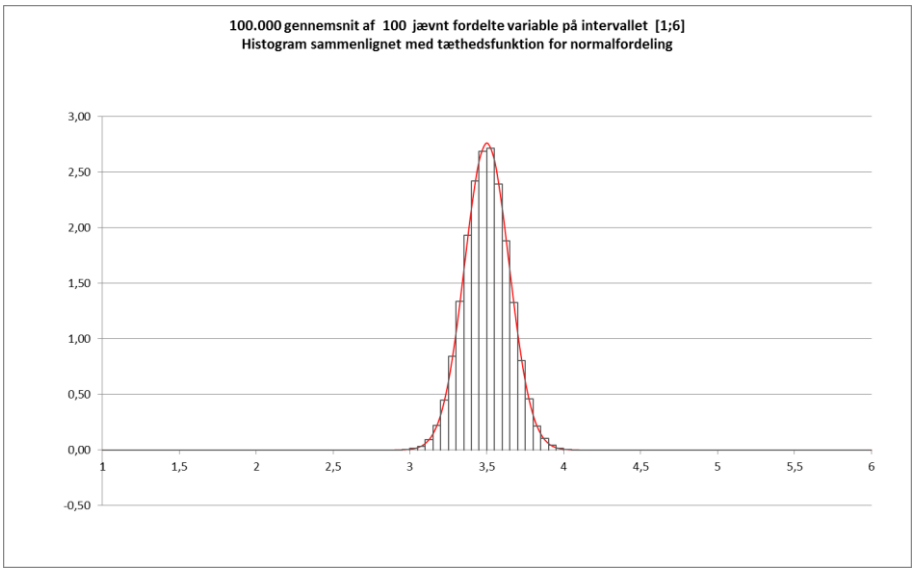

Fig 23

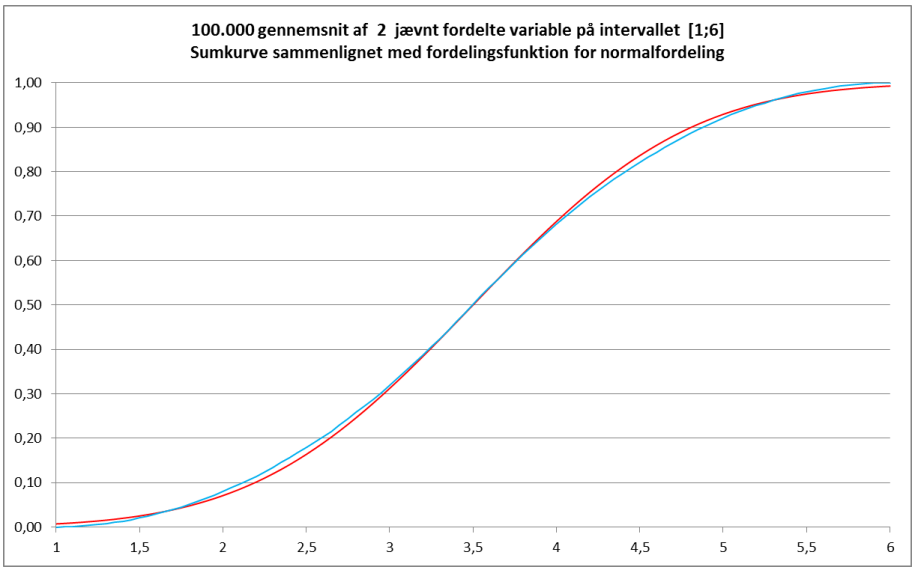

Fig 24

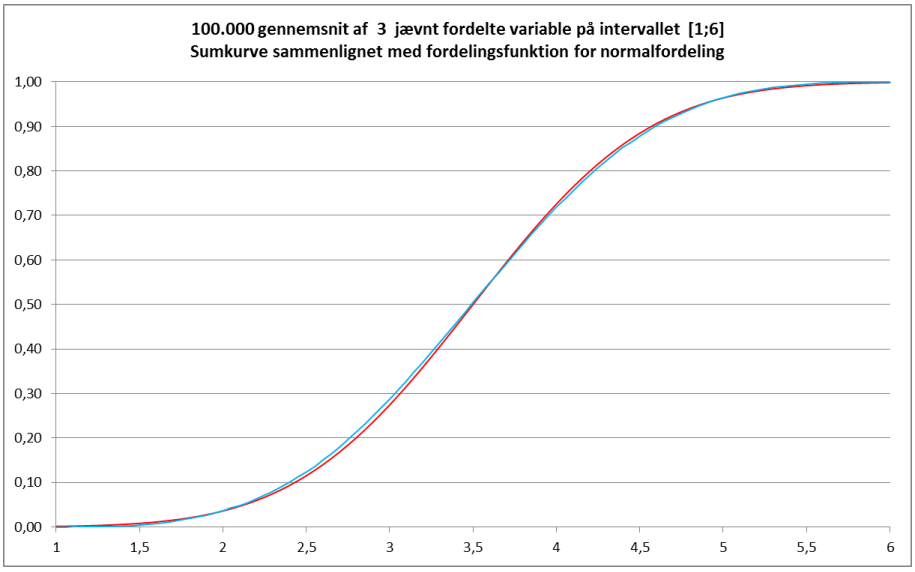

Fig 25

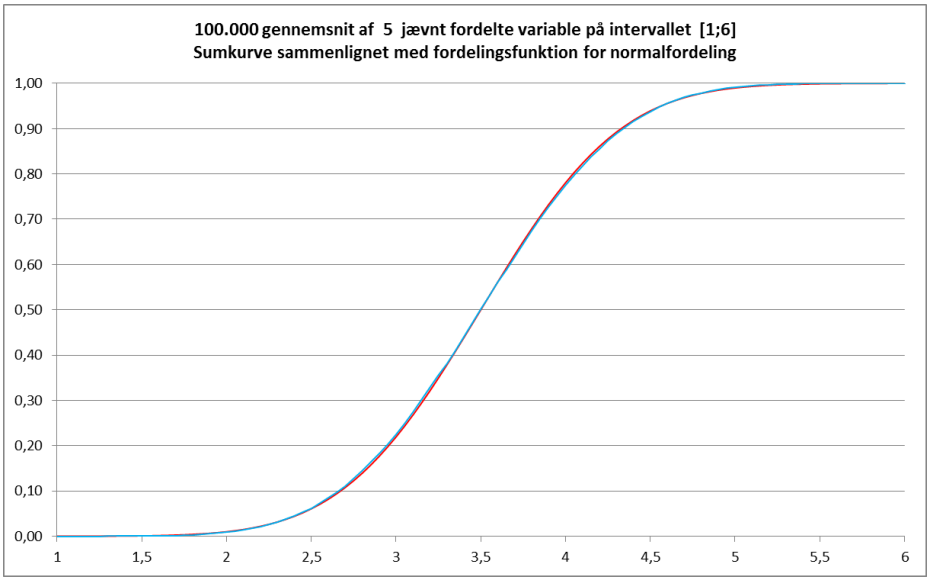

Fig 26

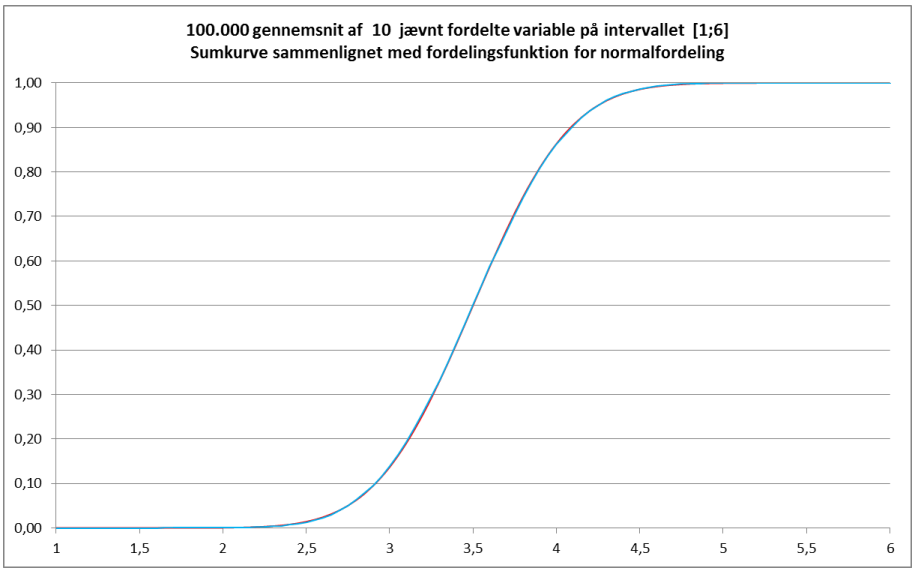

Fig 27

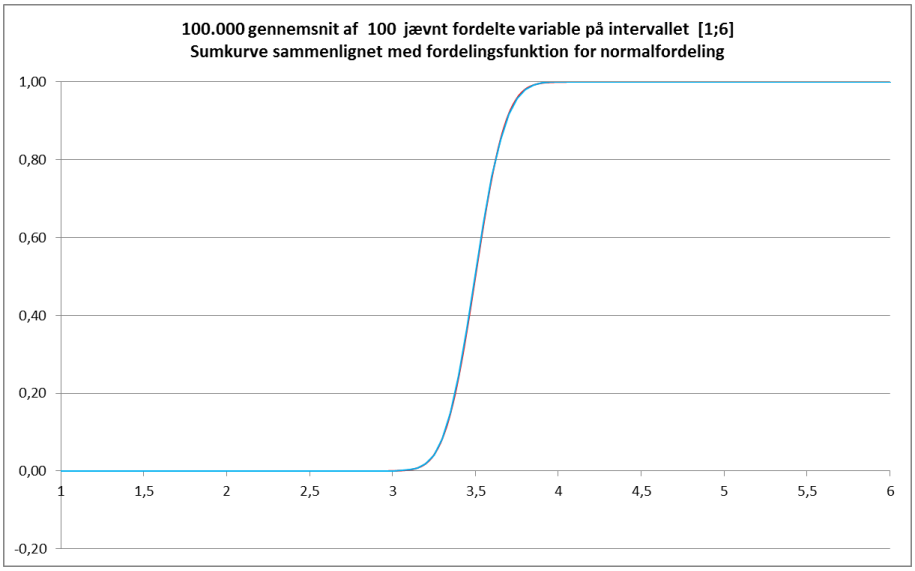

Fig 28

References

[1] http://en.wikipedia.org/wiki/Central_limit_theorem (August 31, 2011)

[2] http://mathworld.wolfram.com/CentralLimitTheorem.html (August 31, 2011)