Høgskolen i Gjøviks notatserie, 2001 nr 5

5 Java-applet's for faget Statistikk

Tor Slind
Avdeling for Teknologi

Gjøvik 2001
ISSN 1501-3162

Sammendrag

Dette notatet beskriver 5 JAVA-applets som demonstrerer noen grunnleggende metoder som brukes i faget Statistikk. De 5 programmene er tilgjengelig for studentene via fagets Internett-hjemmeside. 6 andre JAVA-program er beskrevet tidligere (Slind, T., 2000).

Den som bruker programmmene kan lese inn sine egne data og beregningsvalg. Deretter vises resultatene i et grafisk vindu og som tallverdier. For hver applet er det en web-side med en brukerveiledning og en kort forklaring av hva som blir beregnet.

Metodene som demonstreres er:

Simulering :	Sannsynlighet som grenseverdi for relativ frekvens(Terningkast)
	Histogram: Simulering av data fra kontinuerlige fordelinger
	Minste kvadraters tilpasning for forskjellige r-verdier
	Simulering av sentralgrenseprinsippet
Beregninger:	Kvantiler beregnet fra innlest sannsynlighet. Forskj. fordelinger.

Innholdsfortegnelse

1. Innledning

Det er laget 5 interaktive JAVA-applets som kan vises og brukes ved hjelp av nettlesere som Netscape, Opera eller Explorer via en html-fil. Programmene kan også startes direkte fra DOS-vinduet med programmet *appletviewer*. De 6 programmene kan i vårsemestret brukes via hjemmesiden til faget Statistikk ved HIG :
http://www.hig.no/at/realfag/statistikk/Stat_ing

2. Statistiske metoder

De fem JAVA-programmene som beskrives i dette notatet er basert på en del enkle statistiske begreper og metoder (Daly & al, 1995). Hensikten med programmene er å visualisere de statistiske begrepene ved at brukeren kan velge/lese inn noen parametre og at resultatet presenteres både grafisk og numerisk. Det er skrevet noen enkle matematiske funksjoner som ikke finnes i et standard JAVA matematikk-bibliotek (Sun Microsystems, 2000).

2.1 Tilfeldig fordelte tall

I programmene som simulerer forskjellige forsøk brukes tilfeldig fordelte tall som skal følge en bestemt statistisk sannsynlighetsfordeling f(x). For å generere en slik rekke av tall brukes den kumulative fordelingen F(x) sammen med en tilfeldig-tall-generator som gir uniformt fordelte tall. Denne metoden er beskrevet tidligere (Slind, T., 2000).

2.2 Sannsynlighet som grenseverdi for relativ frekvens

Sannsynlighetsverdier må ofte bestemmes fra resultatene fra en forsøkserie med et stort antall enkeltforsøk. Programmet *Freq_simul* simulerer resultatet av en forsøksserie med terningkast. Brukeren velger antall kast. Fra forsøksserien telles fortløpende opp antall 6'ere, og en beregner den relative frekvensen av 6'ere. Dette vises i det grafiske vinduet, og en kan se hvordan den relative frekvensen endrer seg i løpet av forsøksserien. Vinduet viser også den teoretiske sannsynligheten 1/6.

2.3 Histogram: Simulering av data fra kontinuerlige fordelinger

Selv om en har forsøksdata fra en kjent sannsynlighetsfordeling kan det være vanskelig å gjenkjenne denne fordelingen fra et histogram. Ved denne simuleringen kan brukeren velge fordeling og antall dataserier. Ver dataserier inneholder 25 verdier. Verdiene for de 10 første seriene vises på tallinjen sammen med middelverdi $\bar{x}$ og standardavvik s. I tillegg vises fordelingens forventningsverdi μ og standardavvik σ. Ved å øke antallet kan en se at histogrammet etter hvert nærmer seg den teoretiske sannsynlighetstettheten. En kan også velge å få tegnet opp den kumulative fordelingen. Brukeren kan velge mellom følgende fordelinger:

Uniform fordeling
Eksponentialfordelingen
Normalfordelingen
Weibullfordelingen
Rayleighfordelingen
χ^2-fordelingen
Fisherfordelingen

2.4 Minste kvadraters metode: Sammenheng mellom r-verdi og linjene a+bx

Minste kvadraters metode brukes for å tilpasse en rett linje y = a + bx til et sett av måledata (Slind, T., 2000). Selv om en regner ut r-verdien får en i første omgang ikke noe inntrykk av hvordan spredningen av y-verdiene er knyttet til spredningen av rette linjer a + bx rundt den ukjente, feilfrie linjen $Y = \alpha + \beta x$. I programmet *Lsq_sim* simuleres 5 datasett hvert på 25 x/y-par. Brukeren kan velge mellom 3 forskjellige r-verdier som brukes til å bestemme spredningen rundt den rette linjen Y = 5.0 + 0.5x. De simulerte datasettene presenteres grafisk i et spredningsdiagram sammen med de tilhørende 5 rette linjene og den feilfrie linjen 5.0 + 0.5x.

2.5 Simulering av sentralgrenseprinsippet

Sentralgrenseprinsippet fører til at en hvis en har middelverdien av mer enn 25 måleverdier fra den samme fordelingen så vil middelverdiene være tilnærmet normalfordelt. Med programmet *Sentral* kan brukeren velge å simulere et stort antall måleserier hver på 25 måleverdier. Resultatet av simuleringen vises som 2 histogram og som 2 tetthetsgrafer. Det ene histogrammet viser alle de simulerte verdiene (alle fra den samme valgte fordelingen), det andre histogrammet viser middelverdiene til hver måleserie. Sammen med det andre histogrammet vises også normalfordelingen. Brukeren kan velge mellom følgende fordelinger:
Uniform fordeling
Eksponentialfordelingen
Normalfordelingen

2.6 Beregning av kvantiler.

Når en skal estimere parametre i Normalfordelingen ut fra måleverdier brukes ofte intervall-estimater. Ut fra en valgt sannsynlighet for at intervallet skal inneholde den sanne parameter-verdien må en bestemme kvantil-verdier enten fra Normalfordelingen eller Students t-fordeling. En tilsvarende utregning må en gjøre ved hypotese-testing. Programmet *Kvantil* lar brukeren velge en fordeling og en sannsynlighet *p*. En kan velge mellom å få 1 kvantil-verdi *z* fra den inverse av fordelingsfunksjonen p = F(z), eller 2 kvantilverdier når *p* er plassert i sentrum av fordelingen med sanns. (1-*p*)/2 i hver ende. Dette vises også grafisk. Brukeren kan velge mellom følgende fordelinger:
Normalfordelingen
Students t-fordeling
χ^2-fordelingen
Fisherfordelingen

3. Programbeskrivelser

Dette kapitlet gir en kort beskrivelse av programstrukturen og de forskjellige klassene og metodene.

3.1 Sannsynlighet som grenseverdi for relativ frekvens

Klasser:	Metoder
Freq_simul	*init*() – definerer vinduets utseende
(Applet)	*aktiv*()- beregner relativ frekvens for 6'ere, kast med 1 terning
	actionPerformed() – leser inn # terningkast
	paint()- tegner grafisk vindu

3.2 Histogram: Simulering av data fra kontinuerlige fordelinger

Klasser:	Metoder
Kont_ford	*init*() – definerer vinduets utseende
(Applet)	*itemStateChanged*()- leser inn valgt fordeling
	actionPerformed() – leser inn # forsøksserier
	run() – finner middelverdier, std.avvik og histogram, beregner punkter på tetthetskurven f(x) kaller metoden *hist* og klassene *Pkum*, *Stat* og *Tilfeldig_3p*
	hist() – sorterer innleste verdier til et histogram
	paint()- tegner grafisk vindu
Pkum	*ny_p*() - beregner sannsynligheter for noen kontinuerlige fordelinger kaller klassene *Simpson* og *Gamma*
	ny_f() - regner ut tetthetsverdier $f(x)$ for noen kont. fordelinger kaller klassen *Gamma*
Simpson	*flate_num*()- numerisk beregning av arealet under en gitt funksjon
Gamma	*g*() - beregner Γ-funksjonen for $n = 1,2,3,..$ og $m = 1/2, 3/2, 5/2,..$
Stat	*gj_snitt*() - beregner empirisk middelverdi
	std_avvik() - beregner empirisk standardavvik kaller metoden *gj_snitt*
	norm_z() - normaliserer data til en standard normalfordeling kaller metodene *gj_snitt* og *std_avvik*
	konf_int() - regner ut 95 % konf.intervall for $N(m, s)$ med ukjent s
Tilfeldig_3p	*f_pros*()- sorterer n tilfeldig tall som klassefrekvenser kaller metodene *F_to_f_bin, F_to_f_dice, F_to_f_exp, F_to_f_uni* og *F_to_f_wei*
	F_to_f_bin()- inverterer en fordelingsverdi F til en kvantil x_q for den binomiske fordelingen
	F_to_f_dice()- inverterer en fordelingsverdi F til en kvantil x_q for sanns.fordelingen til sum øyne/2 terningkast
	F_to_f_exp()- inverterer en fordelingsverdi F til en kvantil x_q

for eksponentialfordelingen
F_to_f_uni()- inverterer en fordelingsverdi *F* til en kvantil x_q
for den uniforme fordelingen
F_to_f_wei()- inverterer en fordelingsverdi *F* til en kvantil x_q
for Weibullfordelingen med 3 parametre
bin() – beregner binomialkoeffisienter

3.3 Minste kvadraters metode: Sammenheng mellom r-verdi og linjene a+bx

Klasser:	Metoder
Lsq_sim	*init*() – definerer vinduets utseende
(Applet)	*aktiv*()- genererer tilfeldige verdier, 5 datasett a 25 verdier Finner ligningen for 5 rette linjer vha. minstekvadraters metode kaller klassen *Stat_2D*
	actionPerformed() – leser inn valgt r-verdi
	paint()- tegner grafisk vindu
Stat_2D	*gj_snitt*() - beregner empirisk middelverdi
	std_avvik() – finner summen av kvadrerte avvik kaller metoden *gj_snitt*
	lin_reg() – beregner koeffisienter knyttet til minstekvadraters lineær kurvetilpasning kaller metodene *gj_snitt*
	konf_int2D() - regner ut std.avvik i y-retning og t-kvantiler for et 95 % konf.intervall kaller metoden *lin_reg*

3.4 Simulering av sentralgrenseprinsippet

Klasser:	Metoder
Sentral	*init*() – definerer vinduets utseende
(Applet)	*itemStateChanged*()- leser inn valgt fordeling
	actionPerformed() – leser inn # forsøk
	run()- beregner tilfeldige tall fra riktig fordeling beregner middelverdier som sorteres i et histogram finner punkter på tetthetskurven kaller metoden *hist* og klassene *Pkum*, *Stat* og *Tilfeldig*
	hist() - sorterer innleste verdier til et histogram
	paint()- tegner grafisk vindu

3.5 Beregning av kvantiler.

Klasser:	Metoder
Kvantil	*init*() – definerer vinduets utseende
(Applet)	*itemStateChanged*()- leser inn valgt fordeling
	actionPerformed() – leser inn fordelingsparametre og sannsynlighet sjekk på tillatte tall-verdier

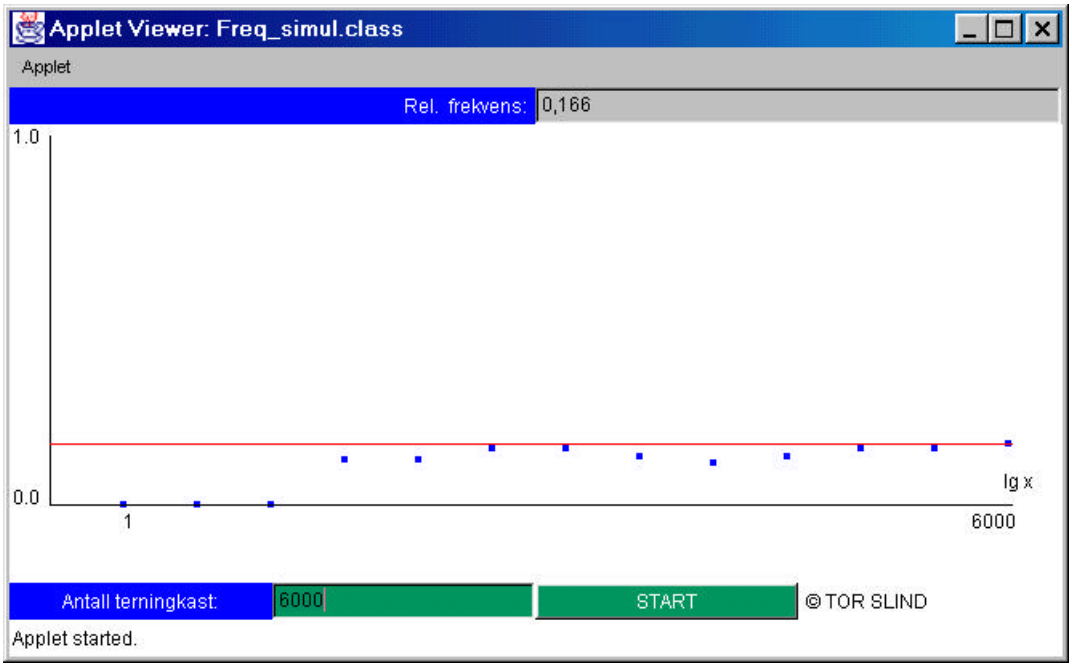
z-frakt()- beregner kvantilen fra innlest sanns.verdi
skriver ut resultater og beregner verdier for grafisk vindu
alfa-frakt()- beregner øvre/nedre kvantil fra innlest sanns.verdi
skriver ut resultater og beregner verdier for grafisk vindu
paint()- tegner grafisk vindu

4. Eksempler på bruk av programmene

Dette kapitlet viser eksempler på bruk av programmene i form av et sett med inngangs-data til programmet og resultatet slik det vises i det grafiske vinduet. For brukeren følger det med en web-side med brukerveiledning og en kort forklaring av beregningene.

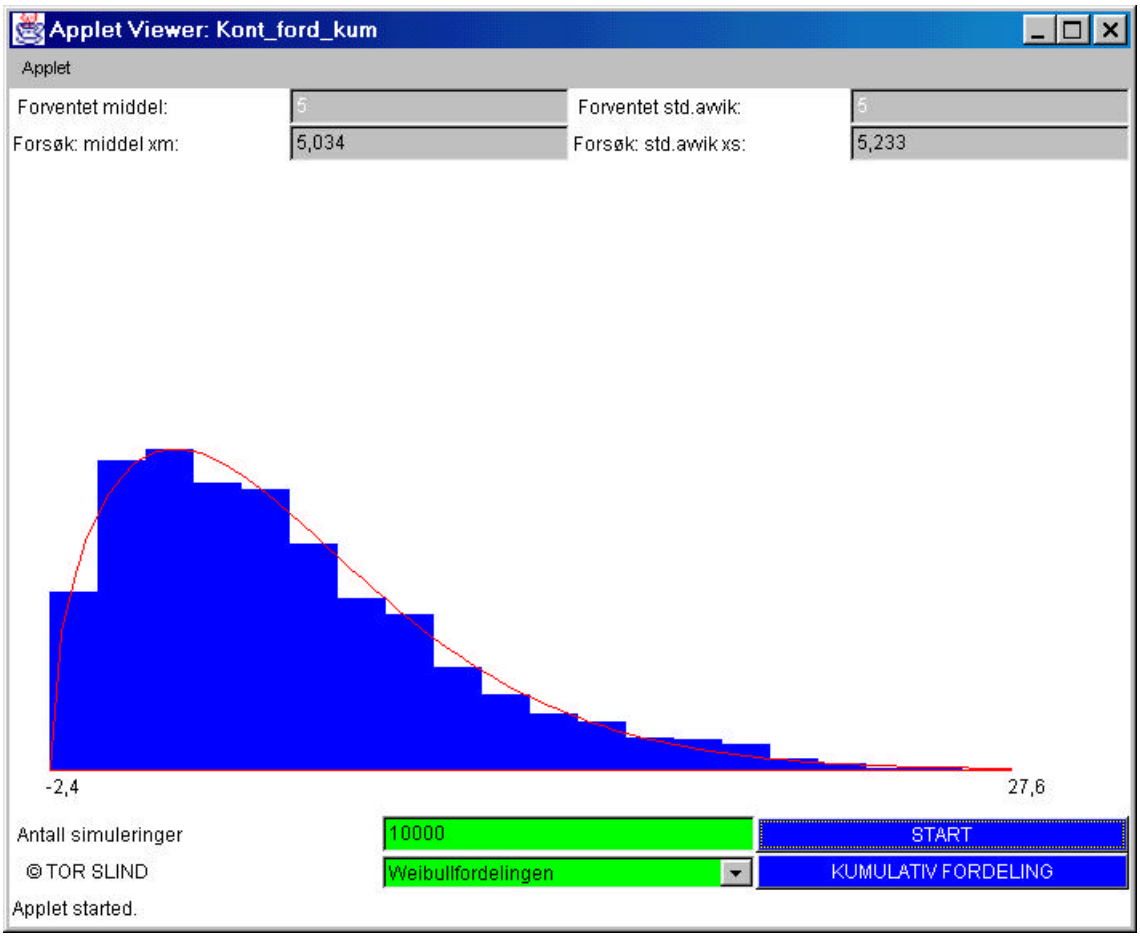
4.1 Sannsynlighet som grenseverdi for relativ frekvens

Velger å simulere 6000 terningkast. Relativ frekvens for 6'ere er vist i figuren. Etter hvert som antall kast øker vil den relative frekvensen nærme seg den teoretiske verdien 1/6.



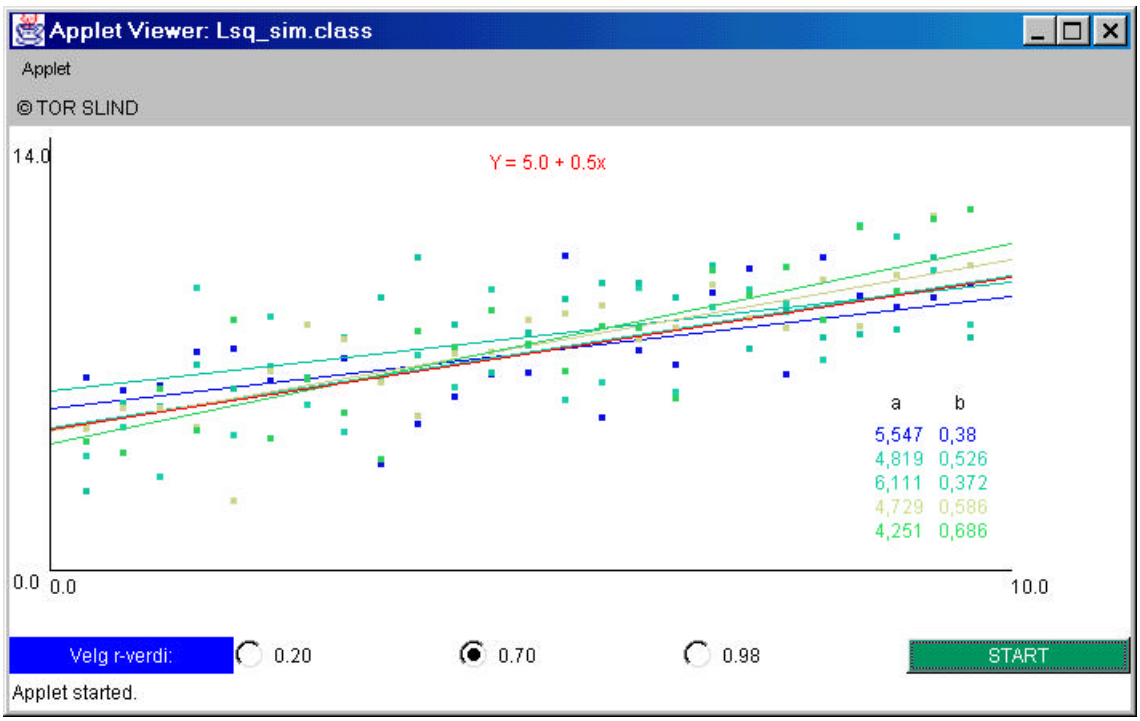
4.2 Histogram: Simulering av data fra kontinuerlige fordelinger

Det grafiske vinduet viser et histogram gitt av simulerte data fra 10000 måleserier. En har valgt en Weibullfordeling med forventningsverdi $\mu = 5$ og standardavvik $\sigma = 5$. Den teoretiske tetthetsfunksjonen er tegnet opp sammen med histogrammet. Brukeren har også mulighet til å se den kumulative fordelingen sammen med et kumulativt histogram.



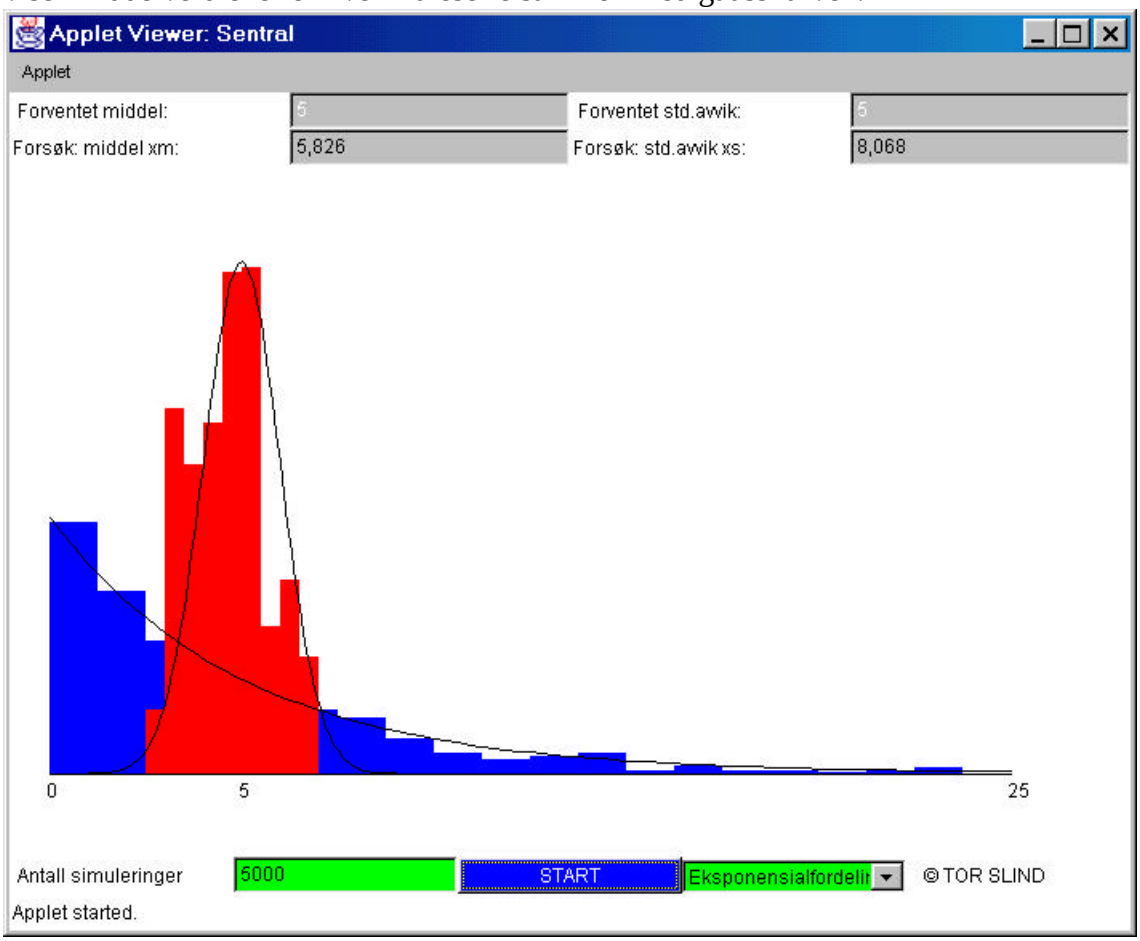
4.3 Minste kvadraters metode: Sammenheng mellom r-verdi og linjene a+bx

Det grafiske bildet viser 5 datasett besert på r = 0.7 for spredningen rundt linjen Y= 5+0.5x.



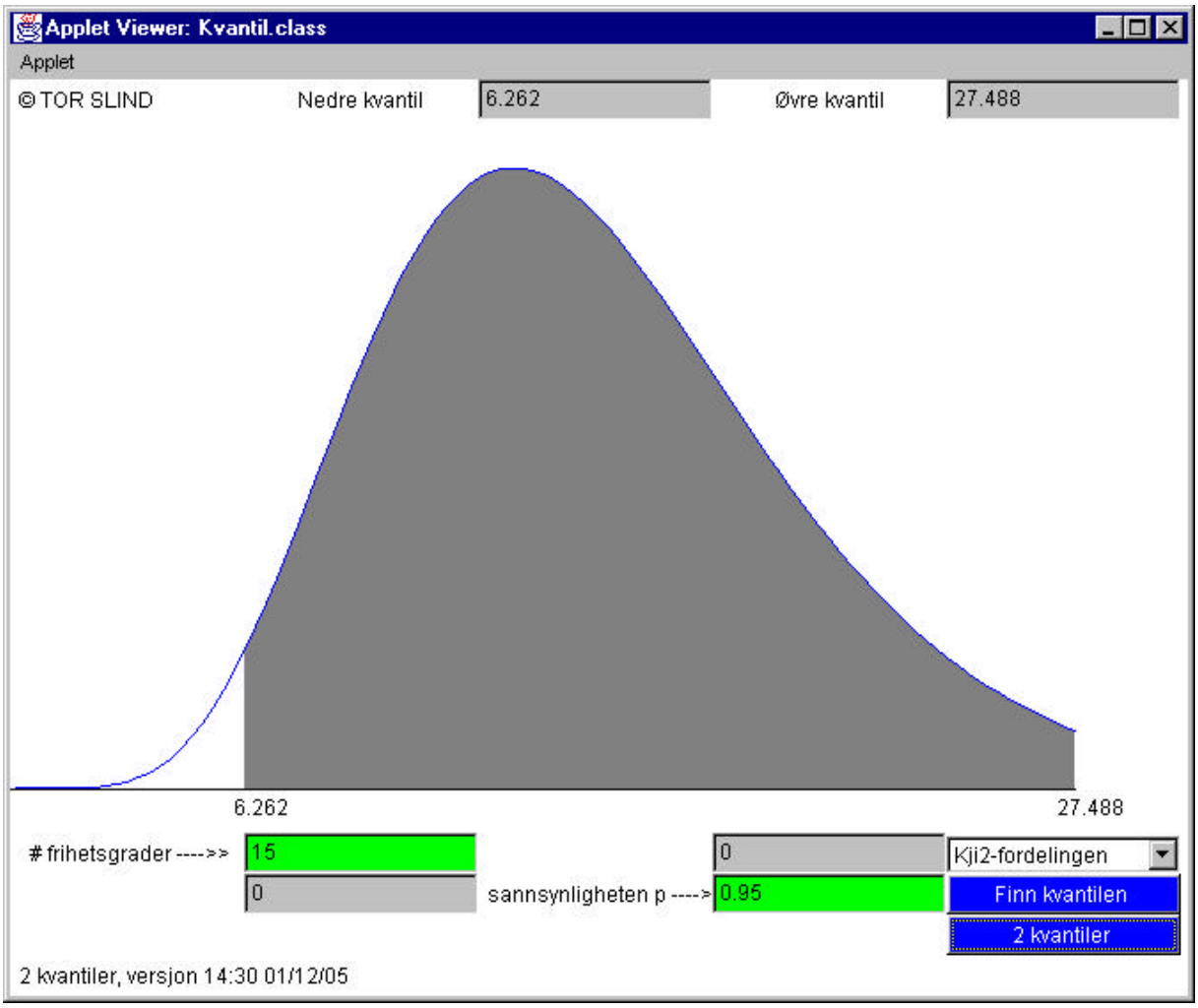
4.4 Simulering av sentralgrenseprinsippet

Her har en valgt 5000 måleserier der dataene er simulert fra en eksponentialfordeling. Det blå histogrammet viser dataene sammen med den teoretiske fordelingen. Det røde histogrammet viser middelverdiene for hver måleserie sammen med gausskurven.



4.5 Beregning av kvantiler.

Ved bergning av kvantil-verdier har en valgt p = 0.95 for χ^2-fordelingen med 15 frihetsgrader. Dessuten er det valgt å finne øvre og nedre kvantil nå de 95 %'ene er plassert sentralt, mens de resterende 5 % er fordelt med 2.5 % i øvre og 2.5 % nedre ende.



5. Referanser

Daly & al, 1995: "Elements of Statistics", Addison-Wesley Pub. Co.,1995, ISBN 0-201-42278-6

Slind, T., 2000: Noen Java-applet's for faget Statistikk, Høgskolen i Gjøviks notatserie, 2000 nr. 2, ISSN 1501-3162

Sun Microsystems, 2000: JAVA 2 Platform (http://java.sun.com/j2se/)