

TILASTOLLISET MENETELMÄT TÄHTITIETEESSÄ

IDL-harjoitus 3, Heikki Salo 27.3.2008

1. Luennoilla on esitetty yhtälö ('error propagation equation'), jonka avulla voidaan arvioida satunnaissuureista johdettujen suureiden varianssia. Oletetaan, että satunnaismuuttujat X ja Y noudattavat jakaumia, joilla on varianssit σ_x^2 ja σ_y^2 . Mikäli x ja y ovat riippumattomia, niin silloin funktion $f = f(x, y)$ varianssia voidaan approksimoida

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2$$

Tutkitaan miten hyvin tämä approksimaatio pätee funktioille $f(x, y) = xy$ ja $f(x, y) = x/y$.

Käytännössä: valitse esim. $\bar{x} = \bar{y} = 1$, ja $\sigma_x = \sigma_y$, ja tutki erilaisia suhteita $\sigma_x/\bar{x} = 0, \dots, 0.8$, käyttäen sekä Gaussista että tasaista jakaumaa x :lle ja y :lle.

Esim $f(x, y) = xy \Rightarrow \frac{\partial f}{\partial x} = y$ ja $\frac{\partial f}{\partial y} = x$

Siten $\sigma_f^2 = y^2 \sigma_x^2 + x^2 \sigma_y^2 \Rightarrow (\sigma_f/f)^2 = (\sigma_x/x)^2 + (\sigma_y/y)^2$

Oletetaan x ja y noudattaa Gaussista jakaumaa ja luodaan satunnaisotokset X_i ja Y_i . Tarkistetaan että eo. approksimaatio pätee kun hajonnat pieninä. Osittaisderivaatat lasketaan keskiarvojen $y = \bar{y}$ ja $x = \bar{x}$ kohdalla.

```
x=1.+ randomn(seed,10000)*.1
y=2. + randomn(seed,10000)*.05
z=x*y

IDL> print,stdev(z)/mean(z)
0.141812
IDL> print,sqrt((stdev(x)/mean(x))^2+(stdev(y)/mean(y))^2)
0.140881
```

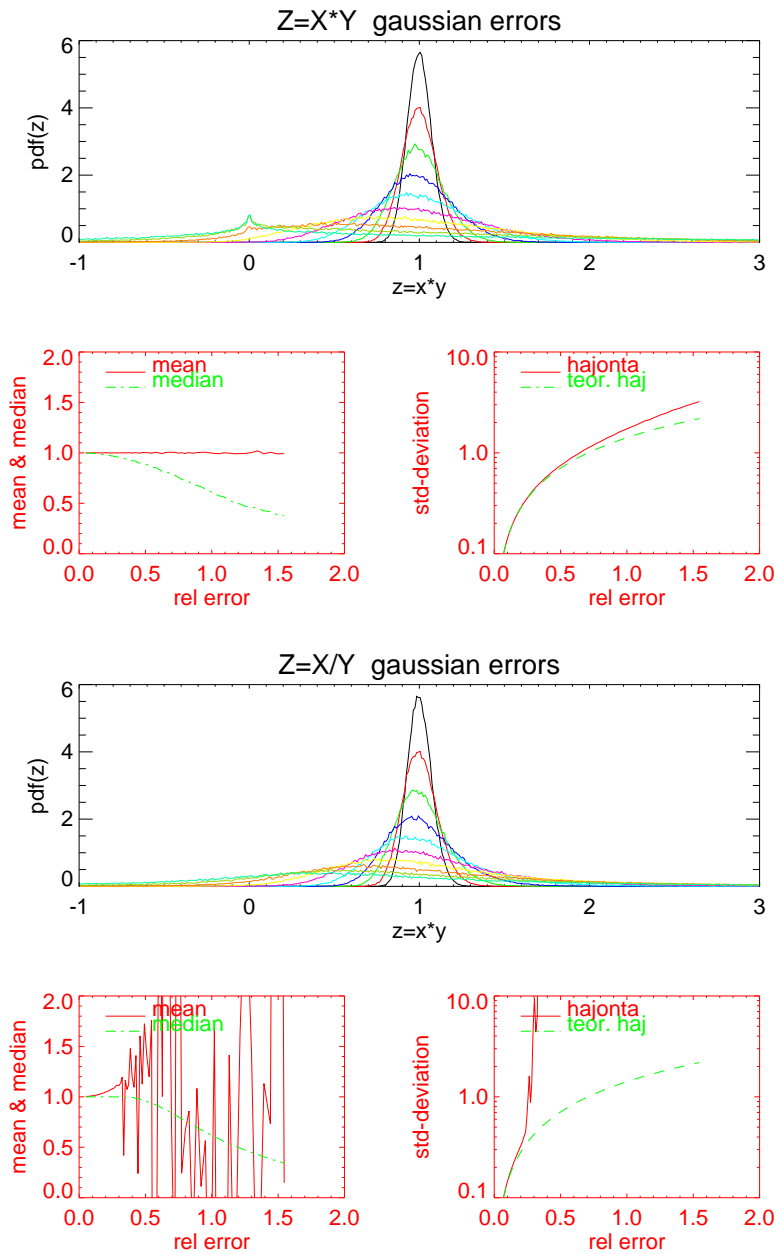
Entä kun virheet (hajonnat) ovat isommat? Esim. 50%:

```
y=2. + randomn(seed,10000)*1.
x=1.+ randomn(seed,10000)*.5
z=x*y

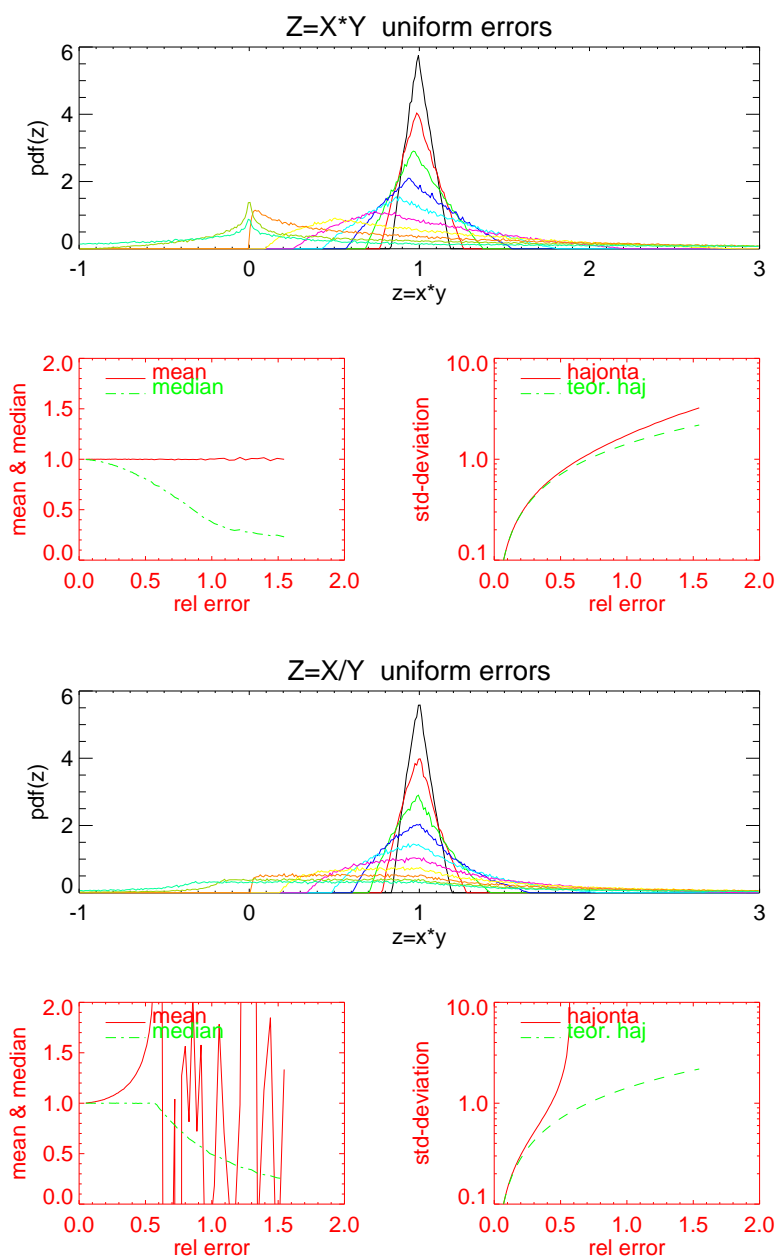
IDL> print,stdev(z)/mean(z)
0.742651
IDL> print,sqrt((stdev(x)/mean(x))^2+(stdev(y)/mean(y))^2)
0.702376
```

Edellä olevan perusteella aproksimaatio pätee aika suurillekin virheille.

Esimerkki-ohjelma `idlharj3_error_propagation.pro` tutkii tulon xy ja osamäärän x/y virheiden käyttäytymistä, kun suureiden x ja y suhteellinen virhe on 0.05 - 1.54 (100 logaritmisesti valittua arvoa). Hajonnan ja keskiarvon lisäksi se tulostaa myös suureiden jakaumat



Kuten edellä, paitsi että x ja y tasaisesti jakautuneet.



```

;-----
program0='idlharj3_error_propagation' & ps=-1
;-----

mu_x=1.d0 & mu_y=1.d0
reltab=0.05*2^(findgen(100)/20.)

for ica=1,4 do begin
  if(ica eq 1) then title='Z=X*Y gaussian errors'
  if(ica eq 2) then title='Z=X/Y gaussian errors'
  if(ica eq 3) then title='Z=X*Y uniform errors'
  if(ica eq 4) then title='Z=X/Y uniform errors'
  if(ica eq 1) then program=program0+'_a'
  if(ica eq 2) then program=program0+'_b'
  if(ica eq 3) then program=program0+'_c'
  if(ica eq 4) then program=program0+'_d'

  psdirect,program,ps,color=1
  nwin & !p.multi=[0,1,2]
  meantab=reltab*0. & medtab=reltab*0. & sttab=reltab*0.

  for i=0,n_elements(reltab)-1 do begin
    relx=reltab(i)
    sigma_x=mu_x*relx
    rely=relx
    sigma_y=mu_y*rely

    N=1000001
    if(ica eq 1 or ica eq 2) then begin
      x=randomn(seed,n)*sigma_x+mu_x
      y=randomn(seed,n)*sigma_y+mu_y
    endif
    if(ica eq 3 or ica eq 4) then begin
      x=(randomu(seed,n)-0.5)*sqrt(12.)*sigma_x+mu_x
      y=(randomu(seed,n)-0.5)*sqrt(12.)*sigma_y+mu_y
    endif
    if(ica eq 1 or ica eq 3) then z=x*y
    if(ica eq 2 or ica eq 4) then z=x/y

    meantab(i)=mean(z)
    medtab(i)=median(z)
    sttab(i)=stdev(z)

    if(i mod 10 eq 0) then begin
      histo_f,z,-1,3,.01,xx,yy
      if(i eq 0) then plot,xx,yy,yr=[0,6],xtitle='z=x*y',ytitle='pdf(z)',title=title
      if(i ne 0) then oplot,xx,yy,col=i*.1+1
    endif
  endfor

  !p.multi=[2,2,2]
  plot,reltab,meantab,yr=[0,2],xtitle='rel error',ytitle='mean & median',col=2
  oplot,reltab,medtab,lines=3,col=3
  label_data,0.1,0.9,['mean','median'],lines=[0,3],col=[2,3]

  plot,reltab,sttab,/ylog,yr=[0.1,10],xtitle='rel error',ytitle='std-deviation',col=2
  oplot,reltab,reltab*sqrt(2),lines=2,col=3
  label_data,0.1,0.9,['hajonta','teor. haj'],lines=[0,3],col=[2,3]

  psdirect,program,ps,color=1,/stop
endfor
end

```

2. Tutkitaan 'poikkeavien havaintojen' (outlier) vaikutusta jakauman leveyden arvioinnissa eri tilastollisia tunnuslukuja käytettäessä. Mikä seuraavista tunnusluvuista on 'vakain' mitta jakauman leveydelle (most robust), eli vähiten riippuvainen poikkeavista arvoista?

i) Neliöllisen poikkeaman keskiarvo

$$RMS = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

ii) keskipoikkeama

$$\overline{\Delta X} = \sqrt{\frac{\sum |X - \bar{X}|}{N}}$$

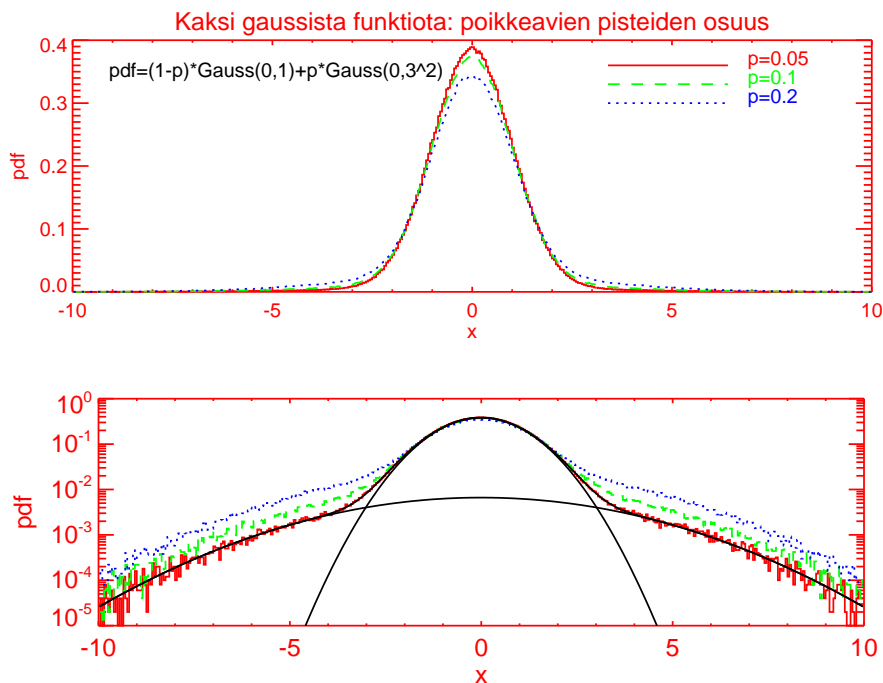
iii) keskimmäiset 50% arvoista sisältävän välin leveys c_{50}

Luodaan outlier-havaintoja sisältävä satunnaisjakauma summaamalla kaksi Gaussista tiheysjakaumaa:

$$f(x) = (1 - p)f_G(\mu, \sigma) + pf_G(\mu, 3\sigma)$$

Jakaumilla sama μ mutta hajonnat poikkeavat tekijällä 3. Parametri $p \ll 1$ mittaa poikkeavien havaintojen osuutta.

a) Plottaa eo. jakauma eri p :n arvoilla. Millainen koordinaatisto havainnollistaa parhaiten poikkeavia pisteitä (jakauman häntiä)?



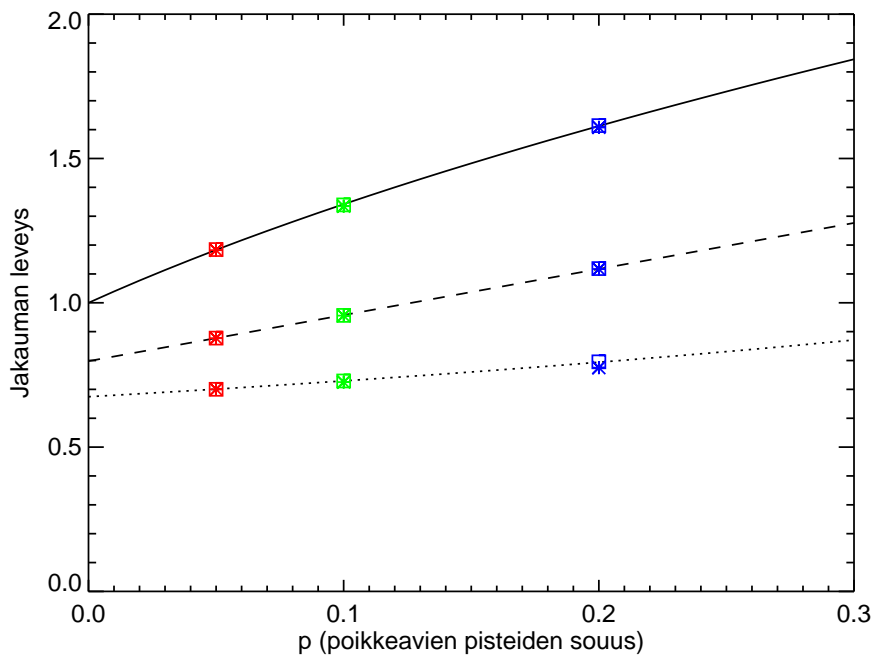
/home/heikki/STATI2008/EXERCISES/EXERCISE3/dlhar3_gaussian_a

heikki@canopus.fysik.yo.oulu.fi Thu Apr 3 10:41:59 2008

b) Totea RMS-poikkeaman riippuvuus p :stä, luomalla eo. jakauman mukaisia satunnaislukuja, ja laskemalla otoksien RMS. (teoreettinen tulos $RMS = \sigma \sqrt{1 + 8p}$)

c) Toista sama käyttäen keskipoikkeamaa. (teoreettinen tulos $\overline{\Delta X} = \sqrt{2/\pi}(1 + 2p)$)

d) Ja käyttäen keskimmäistä 50% pisteistä (teoreettinen ratkaisu saadaan numeerisesti yhtälöstä $(1 - p)2\Phi(C_{50}) + p2\Phi(C_{50}/3) - 0.5 = 0$, joka voidaan ratkaista esim. Newtonin iteraatiomenetelmällä.



/home/heikki/STATI2008/EXERCISES/EXERCISE3/ldharj3_gaussian_b

heikki@canopus.fysik.yo.oulu.fi Thu Apr 3 10:42:01 2008

```
;-----  
;ldharj3_gaussian.pro  
;Poikkeavien havaintojen vaikutus (2005/3.3.2008)  
  
;-----  
;aliohjelma jota kaytetaan c50-laskemiseen (inner quartile)  
;-----  
function newton_c50, c50  
  
;p:n arvo tuodaan aliohjelmaan common-lauseen kautta  
common newton_c50_common, pcom  
p=pcom  
  
;seuraavat lausekkeet antavat saman tuloksen  
;1) using IDL cumulative function of normalized gaussian  
func=(1.-p)*(gauss_pdf(c50)-gauss_pdf(-c50))+  
p*(gauss_pdf(c50/3.)-gauss_pdf(-c50/3.))  
-0.5
```

```

;2) error-funktion avolla
func=(1-p)*erf(c50/sqrt(2.))+ $
      p*erf(c50/sqrt(2.)/3.)$
      -0.5
return,func
end

;-----
;paaohjelma
;-----
common newton_c50_common, pcom

      program='idlharj3_gaussian'
      ps=0

;Havainnollista poikkeavien pisteiden vaikutus
;tutkimalla jakaumaa joka on kahden Gaussin funktion summa
;hajonnat sigma=1 ja sigma=3
;(varianssit sigma^2=1 ja 9)
;sigma=3 jakauman osuus (=poikkeavat pisteet) maaraytyy parametrasta p
; g1=1./sqrt(2!*pi)*exp(-0.5*x^2)
; g2=1./sqrt(2!*pi)/3.*exp(-0.5*x^2/3.^2)
; g=(1-p)*g1 + p*g2

;-----
;a) Luodaan jakaumat ja plotataan histogrammeina
;-----
      psdirect,program+'_a',ps,/color
      !p.multi=[0,1,2]
      nwin

;-----
;ensin p=0.05

      N=10000001          ; number of points
      p=0.05
      N1=(1.-p)*N
      N2=p*N
      X1=[randomn(seed,N1),3*randomn(seed,N2)]

;plot results with histo_f procedure:
; x1=minimum of tabulation
; x2=maximum of tabulation
; dx=binsize of tabulation
; xbin = centers of bins
; ybin = number of points in the bin
; /noscale keyword ->
;      histo_f returns yt =ybin
;      probability density pdf = yt/(dx*N)

      xmin=-10.
      xmax= 10.
      dx=0.05
      histo_f,X1,xmin,xmax,dx,xt1,yt1,/noscale
      pdf1=1.*yt1/(n_elements(x1)*dx)
      plot,xt1,pdf1,psym=10,xtitle='x',ytitle='pdf',col=2,$
          title='Kaksi gaussista funktiota: poikkeavien pisteiden osuus',chars=0.8

;-----
; samalla tavalla p=0.1 and p=0.2

      p=0.10

```

```

N1=(1.-p)*N
N2=p*N
X2=[randomn(seed,N1),3*randomn(seed,N2)]
histo_f,X2,xmin,xmax,dx,xt2,yt2,/noscale
pdf2=1.*yt2/(n_elements(x2)*dx)
oplot,xt2,pdf2,col=3,lines=2

p=0.20
N1=(1.-p)*N
N2=p*N
X3=[randomn(seed,N1),3*randomn(seed,N2)]
histo_f,X3,xmin,xmax,dx,xt3,yt3,/noscale
pdf3=1.*yt3/(n_elements(x3)*dx)
oplot,xt3,pdf3,col=4,lines=1

xyouts,0.15,.9,'pdf=(1-p)*Gauss(0,1)+p*Gauss(0,3^2)',/normal,chars=.8
label_data,0.67,.9,['p=0.05','p=0.1','p=0.2'],lines=[0,2,1],col=[2,3,4],size=.8

;-----
;plotataan sama logaritmisessa y-skaalassa
;+ teoreettinen jakauma p=0.05
  nwin
  plot,xt1,pdf1,psym=10,/ylog,yr=[1d-5,1],col=2,xtitle='x',ytitle='pdf'
  oplot,xt2,pdf2,psym=10,col=3,lines=2
  oplot,xt3,pdf3,psym=10,col=4,lines=1

;teoreettinenäp=0.05
  xarg=xt1
  p=0.05
  g1=1./sqrt(2*pi)*exp(-0.5*xarg^2)
  g2=1./sqrt(2*pi)/3.*exp(-0.5*xarg^2/3.^2)
  oplot,xarg,(1.-p)*g1
  oplot,xarg,p*g2
  oplot,xarg,(1.-p)*g1+p*g2,lines=0

!p.multi=0
psdirect,program+'_a',ps,/stop

;-----
; b,c,d)
; Verrataan miten erilaiset jakauman leveyden mitat
; RMS
; keskipoikkeama
; keskimmaiset 50% (C50 inner quartile width)
; riippuvat poikkeavien pisteiden maarasta
;-----
psdirect,program+'_b',ps,/col

;-----
;Analyttiset lausekkeet arvoille p 0 - 0.3
;-----
  parg=findgen(31)*.01

;root-mean-square: integral of x^2g(x)
  RMS=sqrt((1.-parg)+parg*3.^2)

;mean (absolute) deviation: integral of abs(x)g(x)
  MAD=2./sqrt(2.*pi)*((1.-parg)*1. +parg/3.*9.)

;interquartile width c50: integral from -c50 to c50 g(x) equals 0.5
;need to solve numerically (G is the cumulative distribution of gaussian)

```



```

; (1-p)*(G1(c50)-G1(-c50))+p*(G2(c50)-G2(-c50))=0.5

;use Newton's method to find the root of
; f(c50;p)=(1-p)*(G1(c50)-G1(-c50))+p*(G2(c50)-G2(-c50))-0.5
;this function f(c50;p) is returned by
;the function-procedure newton_c50(c50)
; p is passed to the function via a common block

c50=parg*0
for i=0,n_elements(parg)-1 do begin
  pcom=parg(i) ;passed via common to newton_c50
  p_ini=0. ;initial guess for iteration
  c50(i)=newton(p_ini,'newton_c50')
endfor

nwin
plot,parg,rms,lines=0,xtitle='p (poikkeavien pisteiden sousus)',ytitle='Jakauman leveys'
oplot,parg,mad,lines=2
oplot,parg,c50,lines=1

;-----
;Numeerisesti luotujen jakaumien avulla
;-----;-----;-----
;for p=0.05,0.1,0.2

for ip=1,3 do begin
  if(ip eq 1) then begin
    p=0.05
    pdf=pdf1
    x=x1
    xt=xt1
  endif
  if(ip eq 2) then begin
    p=0.1
    pdf=pdf2
    x=x2
    xt=xt2
  endif
  if(ip eq 3) then begin
    p=0.2
    pdf=pdf3
    x=x3
    xt=xt3
  endif
endif

;Luodun otoksen perusteella:
rms=sqrt(mean(x^2))
mad=mean(abs(x))
x=x(sort(x))
c50=0.5*(x(N*0.75)-x(N*0.25))

plots,[p,rms],psym=6,col=ip+1
plots,[p,mad],psym=6,col=ip+1
plots,[p,c50],psym=6,col=ip+1

;taulukoidun tiheysfunktion avulla

rms=sqrt(mean(xt^2*pdf)/mean(pdf))
mad=mean(abs(xt)*pdf)/mean(pdf)
sum=pdf*0.
for ii=1,n_elements(xt)-1 do begin

```

```

        sum(ii)=sum(ii-1)+pdf(ii)*dx
    endfor

    ind=where(sum gt 0.25)
    ind25=ind(0)
;0.25 attained between xt(ind25) and xt(ind25-1)
;weighted mean
    c50m=(sum(ind25-1)*xt(ind25)+sum(ind25)*xt(ind25-1))/$
        (sum(ind25)+sum(ind25-1))+0.5*dx

    ind=where(sum gt 0.75)
    ind75=ind(0)
;0.75 attained between xt(ind75) and xt(ind75-1)
;weighted mean
    c50p=(sum(ind75-1)*xt(ind75)+sum(ind75)*xt(ind75-1))/$
        (sum(ind75)+sum(ind75-1))+0.5*dx

    print,c50m,c50p
    c50=(c50p-c50m)/2.
    plots,[p,rms],psym=2,col=ip+1
    plots,[p,md],psym=2,col=ip+1
    plots,[p,c50],psym=2,col=ip+1

endfor

label_data,0.1,.9,['RMS','MD','C50'],lines=[0,2,1]

psdirect,program+'_b',ps,/stop

end

```

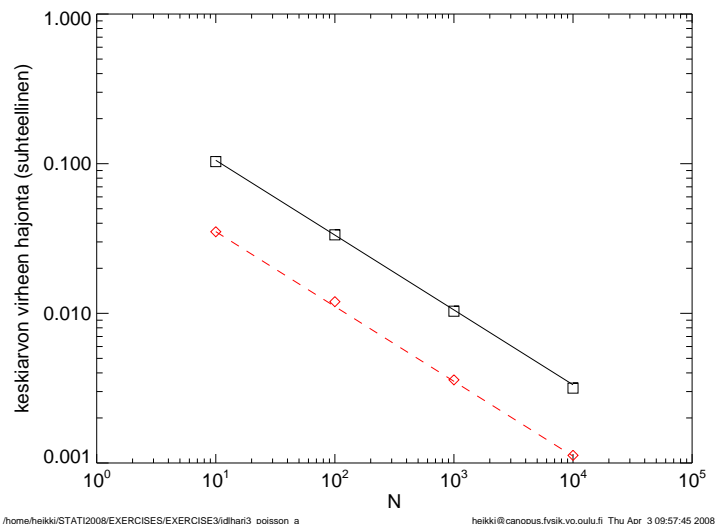
3. Luo Poisson tiheysjakaumaa noudattavia satunnaismuuttujia, siten että teoreettinen keskiarvo $\mu = 9$ ja $\mu = 81$. Tutki $N_{\text{sample}} = 10$ ja $N_{\text{sample}} = 100$ luvun muodostamien otoksien keskiarvoa \bar{X} ja otoshajontaa σ_s .

a) Totea, että otoskeskiarvo \bar{X} approksimoi jakauman keskiarvoa μ , ja että \bar{X} :n hajonta on yhtäpitävä teoreettisen arvon $\sigma_{\bar{X}} = \sigma / \sqrt{N_{\text{sample}}}$ kanssa, jossa $\sigma = \sqrt{\mu}$ on perusjakauman hajonta. Suhteellinen virhe

$$\frac{\sigma_{\bar{X}}}{\mu} = \frac{1}{\sqrt{\mu N_{\text{sample}}}}$$

```
IDL> x=randomu(seed,10000,poisson=10)
IDL> print,mean(x),10.
      9.97360      10.00000
IDL> print,stdev(x),sqrt(10)
      3.17100      3.16228
IDL> print,sqrt(mean(x)),sqrt(10)
      3.15987      3.16228
```

Tutkitaan otoskeskiarvon suhteellista virhettä eri N arvoilla (**idlharj3_poisson.pro**). Käytetään kuten pitaakin: Suhteellinen virhe verrannollinen $1/\sqrt{N}$ ja $1/\sqrt{\mu}$. Kuvassa neliot $\mu=9$, vinoneliot $\mu=81$; jalkimmaisessa tapauksessa virhe 3 kertaa pienempi.

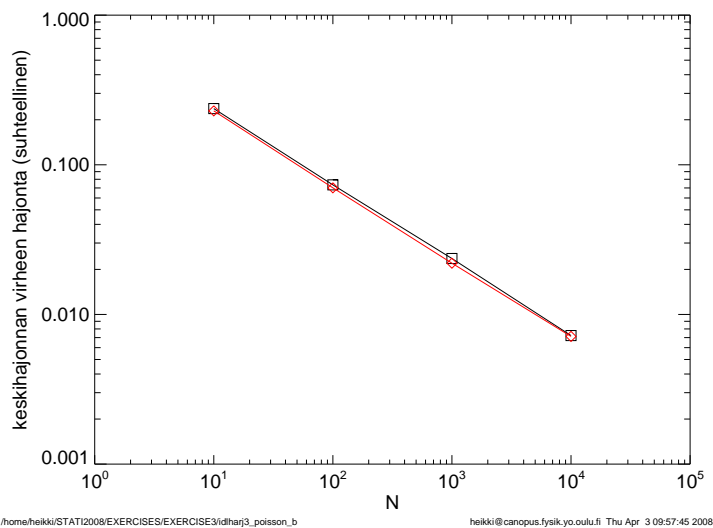


b) Miten hyvin otoshajonta

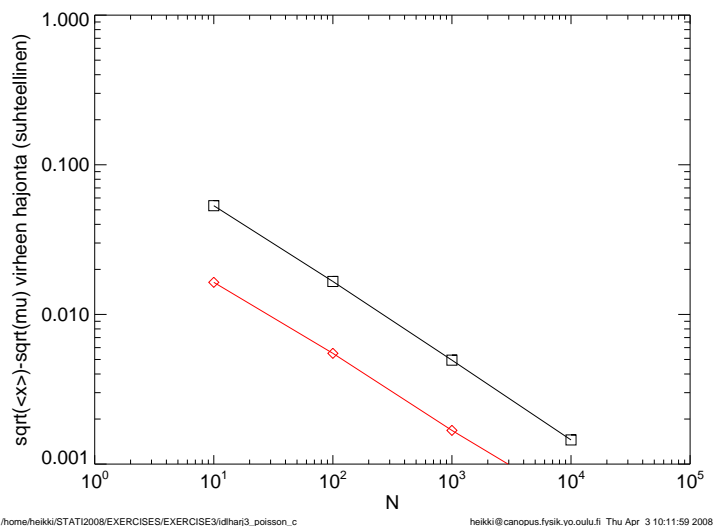
$$\sigma_s = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}}$$

ja otoskeskiarvon neliöjuuri $\sqrt{\bar{X}}$ approksimoivat jakauman teoreettista hajontaa $\sqrt{\mu}$?

Otoshajonnan suhteellinen virhe eri N arvoilla (**idharj3_poisson.pro**). Huom ei riipu μ 'sta. Kuvassa neliöt $\mu=9$, vinoneliöt $\mu=81$; virhe sama ja verrannollinen $1/\sqrt{N}$.



Erotuksen $\sqrt{\bar{X}} - \sqrt{\mu}$ suhteellinen virhe eri N arvoilla:



```

;-----
program='idlharj3_poisson' & ps=0
;-----

;-----
;Poisson prosessi mu=9 ja mu=81 tapausta/aikayksikko
;Jakauman teoreettinen keskiarvo mu, keskihajonta sqrt(mu)

;Arvioidaan otoskeskiarvon ja otoshajonnan
;poikkeamat teoreettisista arvoista eri otoskoolla Nsample
;kayetaan Nsample=10,100,1000,10000

;kuten aiemmin, arvioidaan poikkeamien kayttaytymista
;muodostamalla M:n otoksen otos ja laskemalla niiden otoshajonnat
;M=100
;-----

nsample=[10,100,1000,10000]
M=100

;talletetaan dev_sample_mean = hajonta (<x>-mu)
;talletetaan dev_sample_haj = hajonta (haj(x)-\sqrt(mu))
;talletetaan dev_sample_sgrt = hajonta (sqrt(<x> -\sqrt(mu))
dev_sample_mean=fltarr(m)
dev_sample_haj=fltarr(m)
dev_sample_sqrt=fltarr(m)

;-----
mu=9.

for i=0l,n_elements(nsample)-1 do begin
  n=nsample(i)
  meanit=fltarr(m)
  hajot=fltarr(m)
  sqrtt=fltarr(m)
  for j=0,m-1 do begin
    x=randomu(seed,n,poisson=mu)
    meanit(j)=mean(x)-mu
    HAJOT(j)=STDEV(x)-SQRT(mu)
    sqrtt(j)=sqrt(mean(x))-SQRT(mu)
  endfor
  dev_sample_mean(i)=stdev(meanit)
  dev_sample_haj(i)=stdev(hajot)
  dev_sample_sqrt(i)=stdev(sqrtt)
endfor
dev_sample_mean_9=dev_sample_mean
dev_sample_haj_9=dev_sample_haj
dev_sample_sqrt_9=dev_sample_sqrt

;-----
mu=81.

for i=0l,n_elements(nsample)-1 do begin
  n=nsample(i)
  meanit=fltarr(m)
  hajot=fltarr(m)
  sqrtt=fltarr(m)
  for j=0,m-1 do begin
    x=randomu(seed,n,poisson=mu)
    meanit(j)=mean(x)-mu
    HAJOT(j)=STDEV(x)-SQRT(mu)

```

```

        sqrtt(j)=sqrt(mean(x))-SQRT(mu)
    endfor
    dev_sample_mean(i)=stdev(meanit)
    dev_sample_haj(i)=stdev(hajot)
    dev_sample_sqrt(i)=stdev(sqrtt)
endfor
dev_sample_mean_81=dev_sample_mean
dev_sample_haj_81=dev_sample_haj
dev_sample_sqrt_81=dev_sample_sqrt

;-----
;poikkeamat teoreettisesta keskiarvosta:

psdirect,program+'_a',ps,/color
nwin
plot,nsample,dev_sample_mean_9/9.,/xlog,/ylog,xtitle='N',$
    ytitle='keskiarvon virheen hajonta (suhteellinen)',psym=6,xr=[1,100000]
oplot,nsample,sqrt(9./nsample)/9.

oplot,nsample,dev_sample_mean_81/81.,psym=4,col=2
oplot,nsample,sqrt(81./nsample)/81.,lines=2,col=2
psdirect,program+'_a',ps,/color,/stop

;-----
;poikkeamat teoreettisesta hajonnasta

psdirect,program+'_b',ps,/color
nwin
plot,nsample,dev_sample_haj_9/sqrt(9),/xlog,/ylog,xtitle='N',$
    ytitle='keskihajonnan virheen hajonta (suhteellinen)',psym=-6,xr=[1,100000]
oplot,nsample,dev_sample_haj_81/sqrt(81),psym=-4,col=2
psdirect,program+'_b',ps,/color,/stop

;-----
;poikkeamat sqrt(mean(x)) -sqrt(mu)

psdirect,program+'_c',ps,/color
nwin
plot,nsample,dev_sample_sqrt_9/sqrt(9),/xlog,/ylog,xtitle='N',$
    ytitle='sqrt(<x>)-sqrt(mu) virheen hajonta (suhteellinen)',psym=-6,xr=[1,100000],yr=[0.001,1]
oplot,nsample,dev_sample_sqrt_81/sqrt(81),psym=-4,col=2
psdirect,program+'_c',ps,/color,/stop

end

```

4. Studentin t-jakauma kuvaa Gaussisesta jakaumasta otetun otoksen otoskeskiarvon poikkeamaa teoreettisesta keskiarvosta: suure

$$\frac{\bar{X} - \mu}{\sigma_s / \sqrt{N}}$$

noudattaa t-jakaumaa vapaus-asteella $df=N-1$. Tässä \bar{X} ja σ_s ovat N:n kappaleen otoksen keskiarvo ja otoshajonta. (HUOM: jos σ_s sijasta käytettäisiin jakauman teoreettista varianssia σ niin suure olisi Gaussisesti jakaantunut).

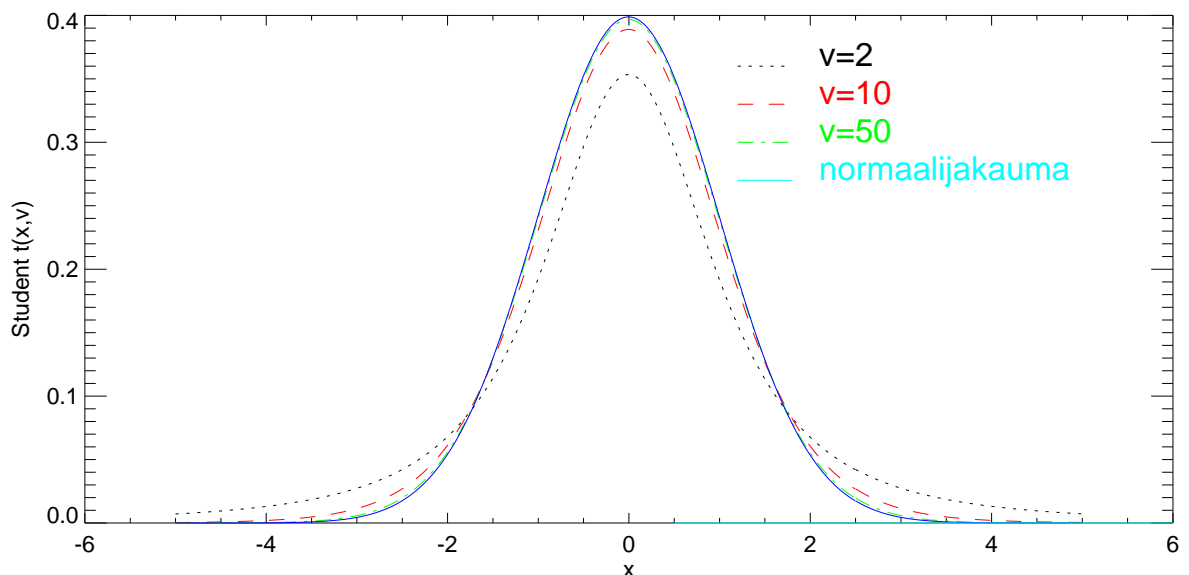
a) Piirrä Student'n t-jakaumia eri vapausasteilla df . Vertaa luennolla (WJ Table 2.1) annettua kaavaa ja IDL:n `T_pdf`-funktioita.

Oheisessa kuvassa (tehty ohjelmalla **idharj3_student_plot.pro**) vapausasteiden määrää merkitty suurella v

Teoreettinen tiheysfunktio laskettu kaavalla:

$$f(x, v) = \Gamma\left(\frac{v+1}{2}\right) \frac{(1 + x^2/v)^{-(v+1)/2}}{\sqrt{\pi v} \Gamma(v/2)}$$

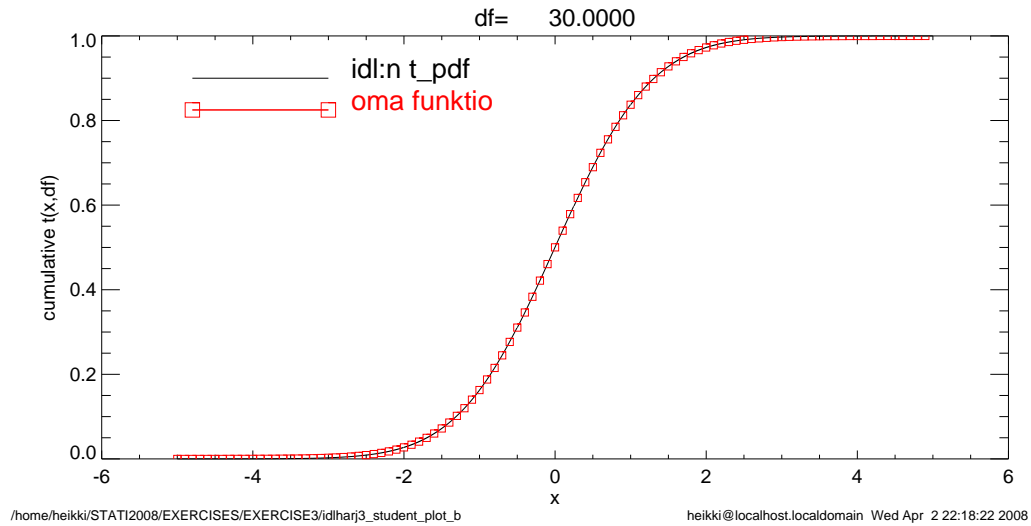
Huomaa miten jakauma lähestyy nopeasti Gaussista jakaumaa kun vapausasteiden luku kasvaa. Pienillä vapausasteiden määrällä Student'n t-jakauma on hyvä malli jakaumalle jossa on Gaussista jakaumaa voimakkaammat hännät (paljon poikkeavia 'outlier' pisteitä).



/home/heikki/STATI2008/EXERCISES/EXERCISE3/idharj3_student_plot_a

heikki@localhost.localdomain Wed Apr 2 22:18:22 2008

Esimerkki Student'n t-jakaumaa vastaavaavasta kertymäfunktioista (v=50): Verrattu idl:n funktiota ja eo kaavasta yksinkertaisella numeerisella integroinnilla saatua tulosta:



```

;-----
;student'n t-jakauman tiheysfunktion palauttava aliohjelma
;-----
function t_pdf_nocum,x,df
  pdf=gamma((df+1.)/2.)*(1.+x^2/df)^(-(df+1.)/2.)/sqrt(!pi*df)/gamma(df/2.)
  return,pdf
end

;paaohjelma
;-----
; a1) plot t-jakauma distribution with different degrees of freedom
;-----
program='idlharj3_student_plot'
ps=0
psdirect,program+'_a',ps,/color,xsize=16,ysize=8
col1=1 & if !d.name eq 'PS' then col1=0 ;(valkoinen mustaksi PS-tulostuksessa)

nwin
!p.multi=1

x=findgen(1000)*.01-5.
y2=t_pdf_nocum(x,2.)
y10=t_pdf_nocum(x,10.)
y50=t_pdf_nocum(x,50.)

plot,x,y2,xtitle='x',ytitle='Student t(x,v)',xr=[-5,5],lines=1,col=col1
oplot,x,y10,col=2,lines=2
oplot,x,y50,col=3,lines=3
oplot,x,1./sqrt(2*!pi)*exp(-0.5*x^2),col=4

label_data,0.6,0.9,['v=2','v=10','v=50','normaalijakauma'],col=[col1,2,3,5],lines=[1,2,3,0] ,len=0.05

```



```

;-----
;plotataan vielä samaan kuvaan jakauma, joka saadaan
;laskemalla normaali-jakaumasta otettujen N=50 luvun otoksien
;lukujen nelioiden summaa. Tman pitäisi olla  $\chi^2$  jakautunut
;vapausasteella 50
  m=10000
  n=50
  x2sum=findgen(m)
  for i=01,m-1 do begin
    x2sum(i)=total(randomn(seed,n)^2)
  endfor
  histo_f,x2sum,0,100,1,xx,yy
  oplot,xx,yy,psym=10,col=5
  psdirect,program+'_a',ps,/stop

;-----
;verrataan idl t_pdf - funktioon joka antaa cumulaatiivisen jakauman
;-----
  psdirect,program+'_b',ps,/color,xsize=16,ysize=8

  x=findgen(10000)*.001-5.
  df=30.

;idl-funktio
  fcumu=t_pdf(x,df)
  plot,x,fcumu,xtitle='x',ytitle='cumulative t(x,df)',title='df='+string(df)

;oma funktioaliohjelma palauttaa tiheysfunktion --> cumulative
  pdf=t_pdf_nocum(x,df)
  dx=x(1)-x(0)
  cpdf=total(pdf,/cumu)*dx

;plotataan joka sadas
  index=lindgen(n_elements(x)/100.)*100.
  oplot,x(index),cpdf(index),psym=6,syms=.5,col=2
;teksti
  label_data,0.1,.9,['idl:n t_pdf','oma funktio'],psym=[0,-6],col=[0,2],lines=[0,-1]

  psdirect,program+'_b',ps,/stop
end

```

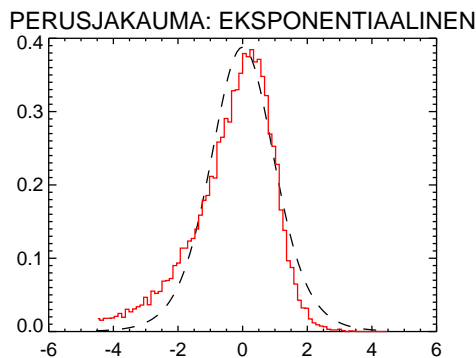
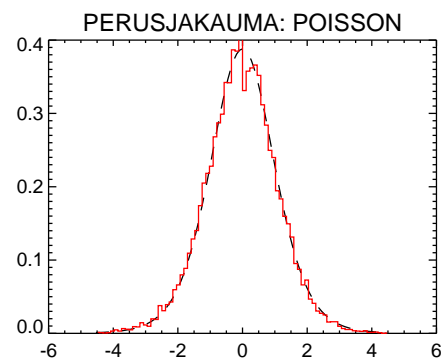
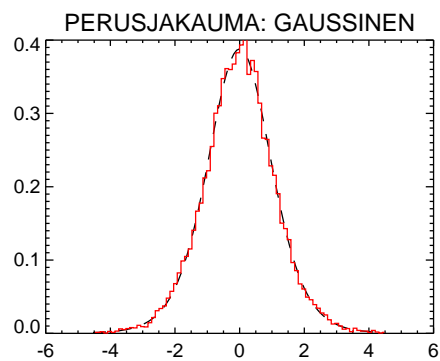
b) Verifioi eo. otoskeskiarvon jakaumaa kuvaava kaava luomalla Gaussisesti jakaantuneita suureita. Käytä esim. $N=10$ ja $N=100$ otoskokoja. Muodosta histogrammit $M=100000$ otoksen perusteella. (t-jakauman pitäisi olla eksakti)

c) Toista sama luomalla otoksia Poisson-jakaantuneista suureista. Päteekö approksimaatio yhä?

d) Toista sama käyttäen eksponentiaalisesti jakaantuneita suureita. Entä nyt?

Esimerkkiohjelma **idharj3_student_otos.pro** vertaa otosvarianssin jakaumaa eri perusjakaumille.

Ohessa $N=10$ luvun otoksia (toistettu $M=20000$ kertaa)

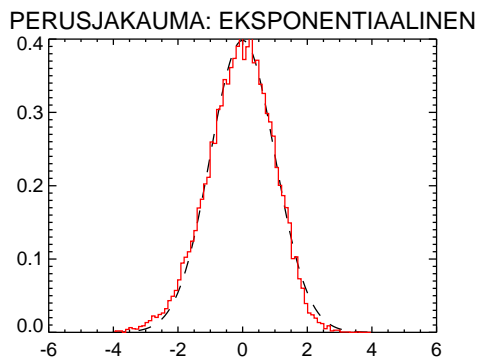
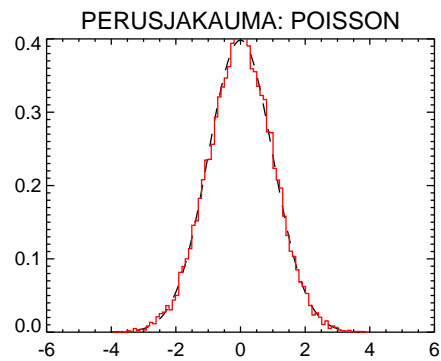
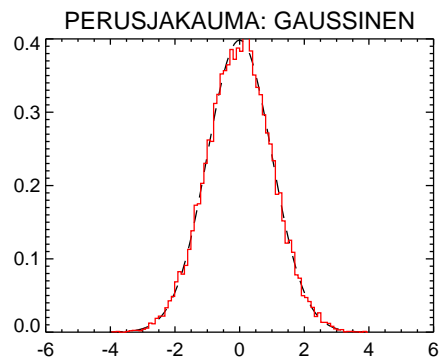


Suureen $(\bar{X} - \mu) / (\sigma_s / \sqrt{N})$ jakauma
 $t(x, v)$ (katkoviiva)
 $N = 10$

/home/heikki/STAT2008/EXERCISES/EXERCISE3/idharj3_student_otos

heikki@localhost.localdomain Wed Apr 2 22:59:41 2008

Ohessa $N=200$ luvun otoksia (toistettu $M=20000$ kertaa)



Suureen $(\langle X \rangle - \mu) / (\sigma_s / \sqrt{N})$ jakauma
 $t(x, v)$ (katkoviiva)
 $N = 200$

/home/heikki/STATI2008/EXERCISES/EXERCISE3/idlharj3_student_otos

heikki@localhost.localdomain Wed Apr 2 22:58:53 2008

```

;-----
program='idlharj3_student_otos' & ps=0
;-----
psdirect,program,ps,/color

;Otetaan N-luvun otos
; Gaussisesta      ica=1
; Poisson-jakaumasta ica=2
; eksponentiaalisesta ica=3
;ja lasketaan otoskeskiarvon poikkeama teoreettisesta
;( $\langle X \rangle - \mu$ ) / ( $\sigma_s / \sqrt{N}$ )
;toistetaan M kertaa --> saadaan otospoikkeaman jakauma
;verrataan teoreettiseen Student'n t-jakauman avulla lausuttuun
;(patee Gaussiselle)

N=10
m=200001
otos_dev_tab=fltarr(m)

nwin
!p.multi=[0,2,2]
!p.charsize=1.
if(!d.name eq 'PS') then !p.charsize=0.7

```

```

;oletettu teoreettinen keskiarvo
;Ei vaikuta tuloksiin
mu=100.

for ica=1,3 do begin
  for i=01,m-1 do begin
    if(ica eq 1) then x=randomn(seed,n)+mu
    if(ica eq 2) then x=randomn(seed,n,poisson=mu)
    if(ica eq 3) then x=-alog(randomu(seed,n))*mu
    if(ica eq 1) then title='PERUSJAKAUMA: GAUSSINEN'
    if(ica eq 2) then title='PERUSJAKAUMA: POISSON'
    if(ica eq 3) then title='PERUSJAKAUMA: EKSPONENTIAALINEN'
    otos_dev_tab(i)=(mean(x)-mu)/(stdev(x)/sqrt(N))
  endfor
endfor

;jakauman teoreettinen keskikohta ja keskihajonta -> plottausrajat
df=N-1.d0
xm=0.
xsig=sqrt(df/(df-2))
print,stdev(otos_dev_tab),xsig
x1=xm-4*xsig
x2=xm+4*xsig
dx=xsig/10.

;teoreettinen jakauma
xtab=x1+(x2-x1)*dindgen(101.)/100.
pdftab=gamma((df+1.)/2.)*(1.d0+xtab^2/df)^(-(df+1.)/2.)/sqrt(!dpi*df)/gamma(df/2.)
plot,xtab,pdftab,lines=2,title=title
;havaittu
histo_f,otos_dev_tab,x1,x2,dx,xx,yy
oplot,xx,yy,psym=10,col=2
endfor

;tyhja tila teksteja varten = tyhja plotti (0,0) - (1,1)
plot,lindgen(2),lindgen(2),xs=15,ys=15,/nodata
xyouts,0.1,.8,'Suureen (<X>-mu)/(sig_s/sqrt(N)) jakauma',col=2
xyouts,0.1,.7,'t(x,v) (katkoviiva)'
xyouts,0.1,.6,'N='+string(N)

psdirect,program,ps,/color,/stop
!p.multi=0

end

```

5. χ^2 jakauma kuvaa normeeratusta normaalijakaumasta otettujen lukujen X_i neliöiden summan jakaumaa: suure $\sum_{i=1}^N X_i^2$ noudattaa χ^2 jakaumaa vapausasteella $df=N$.

Jos otetaan N satunnaismuuttujaa Gaussisesta jakaumasta, jolla hajonta σ , tällöin suure

$$(N-1) \frac{\sigma_s^2}{\sigma^2}$$

noudattaa χ^2 jakaumaa vapausasteella $df=N-1$

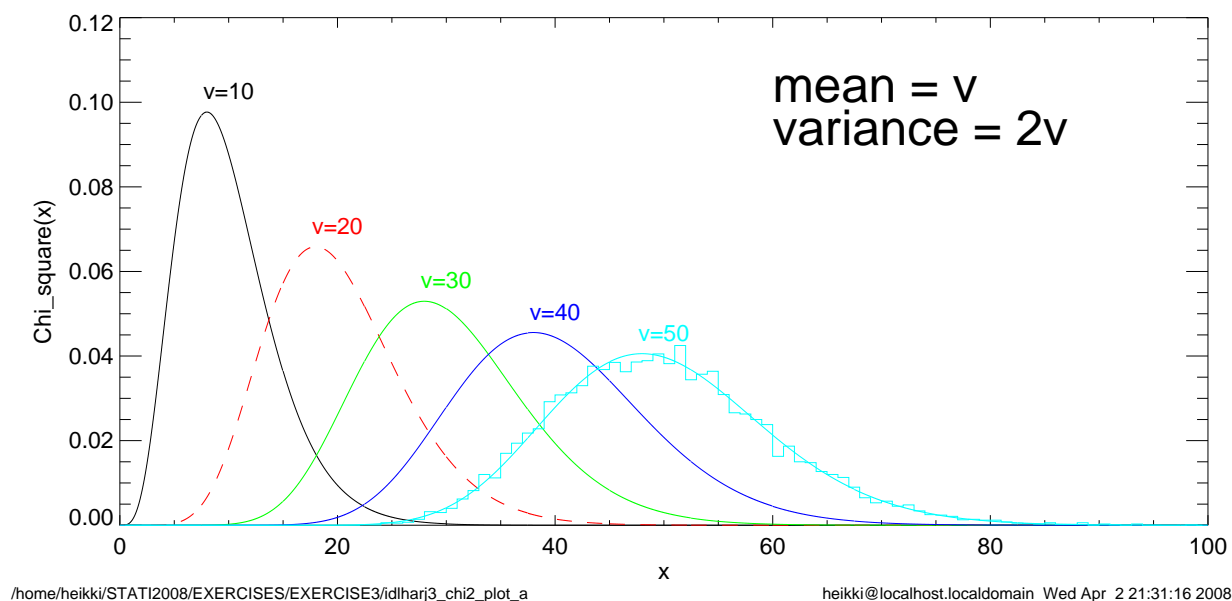
a) Piirrä χ^2 -jakaumia eri vapausasteilla. Vertaa luennolla (WJ Table 2.1) annettua kaavaa ja IDL:n `chisqr_pdf`-funktia.

Oheisessa kuvassa (tehty ohjelmalla `idlharj3_chi2_plot.pro`) vapausasteiden määrää merkitty suureella v

Teoreettinen tiheysfunktio laskettu kaavalla:

$$f(x, v) = \frac{2^{-v/2}}{\Gamma(v/2)} x^{v/2-1} e^{-x/2}$$

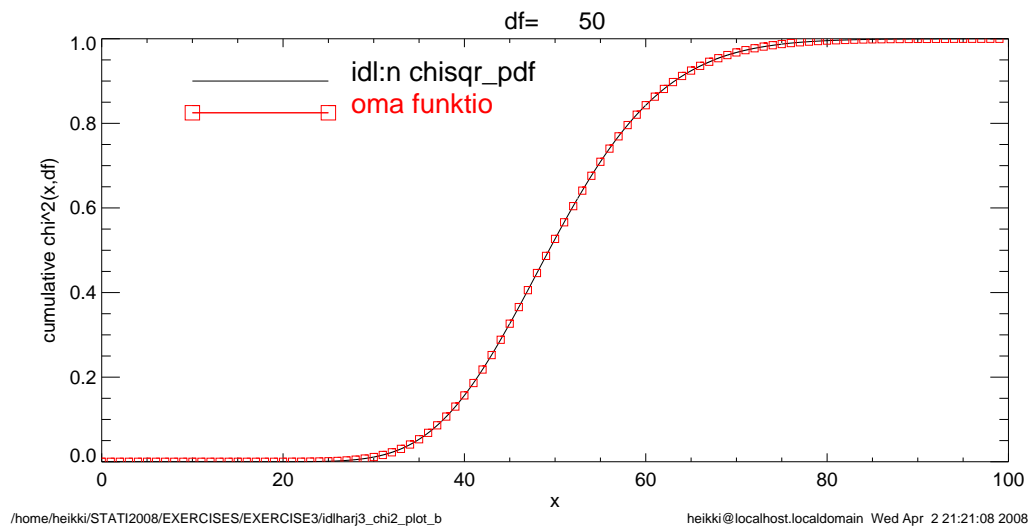
Vapausasteella $v=50$ on piirretty myös histogrammi, joka on saatu tutkimalla $N=50$ gaussisesti jakaantuneen luvun neliöiden summia ($M=10000$ otos-otosta)



Esimerkki χ^2 -jakaumaa vastaavaavasta kertymäfunktioista ($v=50$): Verrattu idl:n funktiota ja eo kaavasta yksinkertaisella numeerisella integroinnilla saatua tulosta:

$$F(x, v) = \int_0^x \frac{2^{-v/2}}{\Gamma(v/2)} x^{v/2-1} e^{-x/2} dx \approx \Delta x \sum_{x_i \leq x} \frac{2^{-v/2}}{\Gamma(v/2)} x_i^{v/2-1} e^{-x_i/2}$$

jossa x_i on tasaisin välein Δx valituja pisteitä, joissa tiheysfunktion arvo lasketaan (HUOM: on olemassa myös paljon tehokkaampia tapoja numeeristen integraalien laskemiseen)



```

;-----
;chi^2 tiheysfunktion palauttava aliohjelma
;-----
function chisqr_pdf_nocum,x,df
pdf=2.^(-df/2.)/gamma(df/2.)*x^(df/2.-1.)*exp(-x/2.)
return,pdf
end

;paaohjelma
;-----
; a1) plot chi-square distribution with different degrees of freedom
;-----
program='idlharj3_chi2_plot'
ps=0
psdirect,program+'_a',ps,/color,xsize=16,ysize=8
nwin
!p.multi=1

x=findgen(10000)*.01
y10=chisqr_pdf_nocum(x,10.)
y20=chisqr_pdf_nocum(x,20.)
y30=chisqr_pdf_nocum(x,30.)
y40=chisqr_pdf_nocum(x,40.)
y50=chisqr_pdf_nocum(x,50.)

```

```

plot,x,y10,xtitle='x',ytitle='Chi_square(x)',yr=[0,.12]
oplot,x,y20,col=2,lines=2
oplot,x,y30,col=3
oplot,x,y40,col=4
oplot,x,y50,col=5

dy=0.003
xyouts,ali=.5,10,max(y10)+dy,'v=10'
xyouts,ali=.5,20,max(y20)+dy,'v=20',col=2
xyouts,ali=.5,30,max(y30)+dy,'v=30',col=3
xyouts,ali=.5,40,max(y40)+dy,'v=40',col=4
xyouts,ali=.5,50,max(y50)+dy,'v=50',col=5

xyouts,60,.10,'mean = v',chars=1.5
xyouts,60,.09,'variance = 2v',chars=1.5

;-----
;plotataan vielä samaan kuvaan jakauma, joka saadaan
;laskemalla normaali jakaumasta otettujen N=50 luvun otoksien
;lukujen nelioiden summaa. Tman pitäisi olla chi^2 jakautunut
;vapausasteella 50
m=10000
n=50
x2sum=findgen(m)
for i=01,m-1 do begin
    x2sum(i)=total(randomn(seed,n)^2)
endfor
histo_f,x2sum,0,100,1,xx,yy
oplot,xx,yy,psym=10,col=5
psdirect,program+'_a',ps,/stop

;-----
;verrataan idl chisqr_pdf - funktioon joka antaa cumulaatiivisen jakauman
;-----
psdirect,program+'_b',ps,/color,xsize=16,ysize=8

x=findgen(10000)*.01
df=50

;idl-funktio
fcumu=chisqr_pdf(x,df)
plot,x,fcumu,xtitle='x',ytitle='cumulative chi^2(x,df)',title='df='+string(df)

;oma funktioaliohjelma palauttaa tiheysfunktion --> cumulative
pdf=chisqr_pdf_nocum(x,df)
dx=x(1)-x(0)
cpdf=total(pdf,/cumu)*dx

;plotataan joka sadas
index=lindgen(n_elements(x)/100.)*100.
oplot,x(index),cpdf(index),psym=6,syms=.5,col=2
;teksti
label_data,0.1,.9,['idl:n chisqr_pdf','oma funktio'],psym=[0,-6],col=[0,2],lines=[0,-1]

psdirect,program+'_b',ps,/stop
end

```

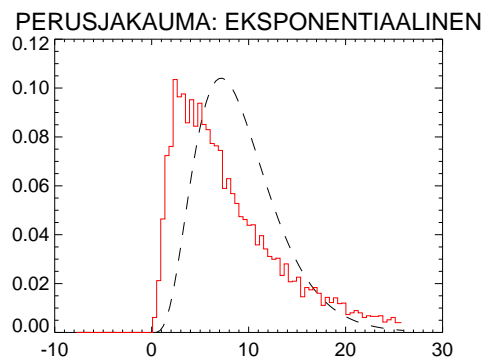
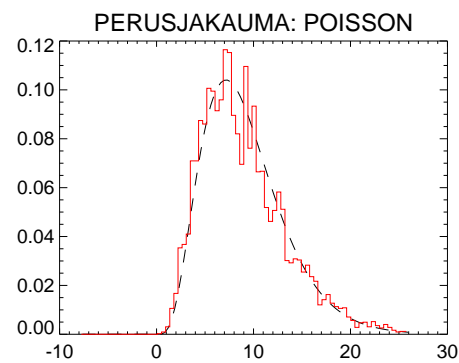
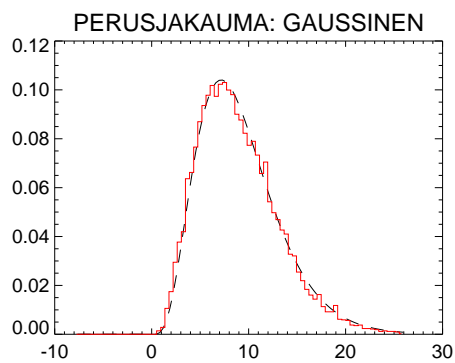
b) Verifioi otosvarianssin jakauma eri N arvoilla, käyttäen Gaussisesta jakaumasta otettuja muuttujia.

c) Toista Poisson-jakaantuneille muuttujille: miten hyvin otosvarianssin jakauma noudattaa eo. kaavaa?

d) Ja eksponentiaalisesti jakaantuneille muuttujille: miten hyvin otosvarianssin jakauma noudattaa eo. kaavaa?

Esimerkkiohjelma **idlharj3_chi2_otos.pro** vertaa otosvarianssin jakaumaa eri perusjakau-
mille.

Ohessa $N=10$ luvun otoksia (toistettu $M=10000$ kertaa)

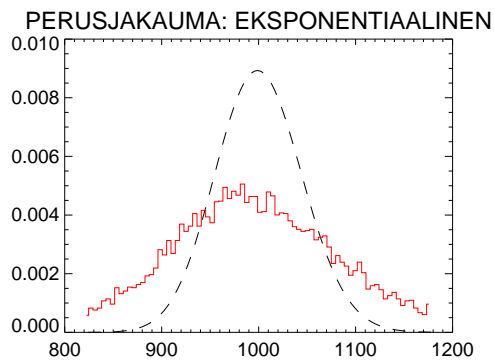
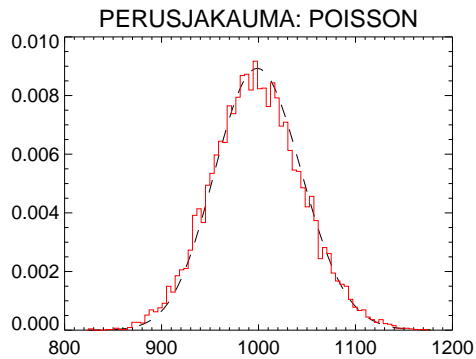
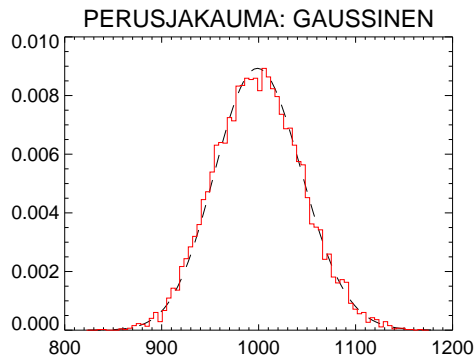


Suureen $(N-1) \cdot (\text{sig}_s / \text{sig})^2$ jakauma
 $\chi^2(df=N-1)$ (katkoviiva)
 $N= 10$

/home/heikki/STATI2008/EXERCISES/EXERCISE3/idlharj3_chi2_otos

heikki@localhost.localdomain Wed Apr 2 20:50:32 2008

Ohessa N=1000 luvun otoksia (toistettu M=10000 kertaa)



Suureen $(N-1) \cdot (\text{sig}_s / \text{sig})^2$ jakauma
 $\chi^2(df=N-1)$ (katkoviiva)
 N= 1000

/home/heikki/STATI2008/EXERCISES/EXERCISE3/ldharj3_chi2_otos

heikki@localhost.localdomain Wed Apr 2 20:48:31 2008

```

;-----
program='ldharj3_chi2_otos' & ps=0
;-----

psdirect,program,ps,/color

;Otetaan N-luvun otos ja lasketaan varianssi
; Gaussisesta      ica=1
; Poisson-jakaumasta ica=2
; eksponentiaalisesta ica=3
;toistetaan M kertaa --> saadaan otosvariانسsin jakauma
;verrataan teoreettiseen chi^2 avulla lausuttuun (patee gaussiselle)

N=10
m=100001
otos_sig_tab=fltarr(m)

;perusjoukon teoreettinen hajonta:
;poisson      sig_teo=sqrt(lambda)
;exponential  sig_teo=meanx

sig_teo=2.

```

```

nwin
!p.multi=[0,2,2]
!p.charsize=1.
if(!d.name eq 'PS') then !p.charsize=0.7

for ica=1,3 do begin
  for i=01,m-1 do begin
    if(ica eq 1) then x=randomn(seed,n)*sig_teo
    if(ica eq 2) then x=randomn(seed,n,poisson=sig_teo^2)
    if(ica eq 3) then x=-alog(randomu(seed,n))*sig_teo
    if(ica eq 1) then title='PERUSJAKAUMA: GAUSSINEN'
    if(ica eq 2) then title='PERUSJAKAUMA: POISSON'
    if(ica eq 3) then title='PERUSJAKAUMA: EKSPONENTIAALINEN'
    otos_sig_tab(i)=stdev(x)
  endfor

;jakauman teoreettinen keskikohta ja keskihajonta -> plottausrajat
  xm=n-1
  xsig=sqrt(2.*(n-1))
  print,stdev(otos_sig_tab^2/sig_teo^2*(n-1)),xsig
  x1=xm-4*xsig
  x2=xm+4*xsig
  dx=xsig/10.

;teoreettinen jakauma
  df=n-1.
  xtab=x1+(x2-x1)*findgen(101.)/100.
  pdf=2.^(-(df/2.)/gamma(df/2.)*xtab^(df/2.d0-1.)*exp(-xtab/2.)
  plot,xtab,pdftab,lines=2,title=title
;havaittu
  histo_f,otos_sig_tab^2/sig_teo^2*(n-1.),x1,x2,dx,xx,yy
  oplot,xx,yy,psym=10,col=2
endfor

;tyhja tila teksteja varten = tyhja plotti (0,0) - (1,1)
plot,lindgen(2),lindgen(2),xs=15,ys=15,/nodata
xyouts,0.1,.8,'Suureen (N-1)*(sig_s/sig)^2 jakauma',col=2
xyouts,0.1,.7,'chi^2(df=N-1) (katkoviiva)'
xyouts,0.1,.6,'N='+string(N)

psdirect,program,ps,/color,/stop
!p.multi=0

end

```