

TILASTOLLISET MENETELMÄT TÄHTITIEESSÄ

IDL-harjoitus 4, Heikki Salo 10.4.2008

1) Tutki Hubble(1936) havaitsemaa korrelaatiota galaksin punasiirtymän ja etääntymisnopeuden välillä

a) Lue 24 galaksin etäisyydet ja etääntymisnopeudet tiedostosta "4point1.dat" (WJ).

b) Testaa mahdollista korrelaatiota, käyttäen yhtä parametrista ja ainakin kahta ei-parametrista menetelmää:

- Pearsonin otoskorrelaatiokertoimen testaaminen
 - Spearmanin rank-korrelaatiokertoimen testaaminen
 - Kendall'n τ
-

Distance (Mpc)	speed (km/sec)
0.04	111.1
0.03	-83.3
0.19	97.2
0.25	27.8
0.27	-69.4
0.26	-208.3
0.42	819.4
0.50	819.4
0.50	958.3
0.63	666.7
0.79	777.8
0.89	194.4
0.89	430.6
0.88	888.9
0.91	1222.2
1.01	1736.1
1.10	1472.2
1.11	1166.7
1.42	1263.9
1.70	2111.1
2.02	1111.1
2.01	1611.1
2.02	1763.9
2.02	2250.0

Esimerkkiratkaisu: kts. `idlharj4_hubble.pro`

a) Datan lukeminen

Sisältää kaksi seliteriviä + 24 d,v paria seuraavilla riveillä

```
;use WJ hubble-data in 4point1.dat
;Hubble(1936): recession velocity vs galaxy distance

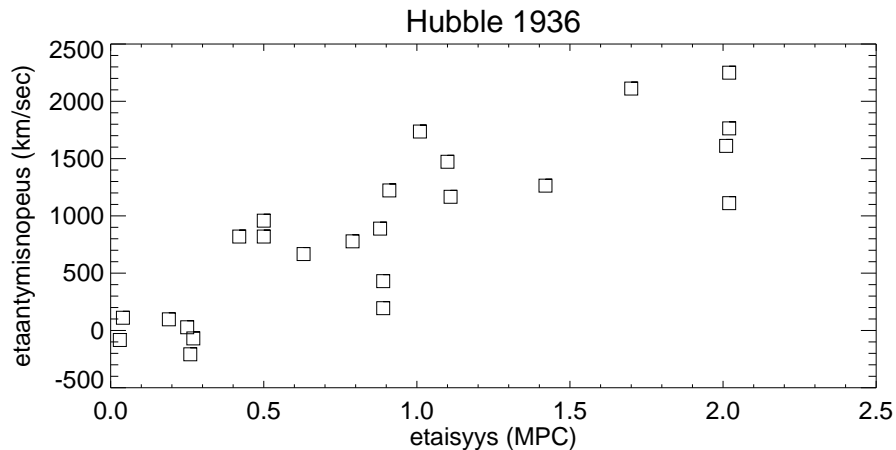
;varataan tilaa
d=fltarr(24)
v=fltarr(24)

;avataan tiedosto laitenumeralle 1
close,1 & openr,1,'4point1.dat'

;luetaan seliterivit merkijoinoina
line='' & readf,1,line
line='' & readf,1,line

;ja dataa sisältävät
for i=0,23 do begin
    readf,1,dd,vv
    d(i)=dd & v(i)=vv
endfor
close,1

plot,d,v,xtitle='etaisyys (MPC)',ytitle='etaantymisnopeus (km/sec)',title='Hubble 1936',psym=6
```



/home/heikki/STATI2008/EXERCISES/EXERCISE4/idlharj4_hubble_0

heikki@ikiturso Thu Apr 10 08:46:38 2008

Selvästikin v ja d:n välillä riippuvuus.
Korrelaatiokerroin:

```
IDL> print,correlate(d,v)
0.837084
x=d-mean(d)
y=v-mean(v)
r_check=total(x*y)/sqrt(total(x^2)*total(y^2))
IDL> print,r_check
0.837085
```

b) Korrelaation testaaminen Fisher r-testillä

- Oletetaan että x, y ovat riippumattomia ja että niiden perusjoukko on Gaussisesti jakaantunut
- Lasketaan Pearsonin otoskorrelaatiokerroin r otokselle, jonka muodostaa N lukuparia
- Muodostetaan testisuure t , joka noudattaa Student'n t -jakaumaa vapausasteella $N-2$
- Lasketaan mikä on todennäköisyys että voidaan sattumalta saada havaitun kokoinen tai suurempi testisuure, vaikka korrelaatiota ei todellisuudessa olekaan ($\rho = 0$)

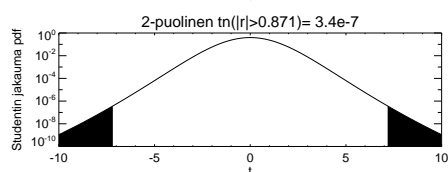
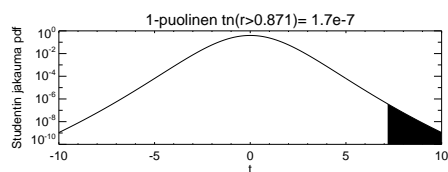
```
N=n_elements(d)
df=N-2
r=correlate(d,v)*1.d0
t=r*sqrt((N-2)/(1.-r^2))
; tarkistuksen vuoksi lasketaan r_s suoraan maaritelmasta
x=d-mean(d)
y=v-mean(v)
r_check=total(x*y)/sqrt(total(x^2)*total(y^2))

print, '-----'
print, 'Fisher r-testi otoskorrelaatiokertoimelle r_otos:'
print, '-----'
print, '  N= ', N, ' r_otos= ', r, ' tarkistus r= ', r_check
print, '  t= ', t
print, '1-puolinen testi'
print, '  H0: ei korrelaatiota'
print, '  H1: positiivinen korrelaatio'
print, '  prob(r>r_otos|H0)= ', 1-t_pdf(t,df)
print, '2-puolinen testi'
print, '  H0: ei korrelaatiota'
print, '  H1: positiivinen tai negatiivinen korrelaatio'
print, '  prob(|r|>|r_otos| |H0)= ', 2*(1-t_pdf(t,df))
print, '-----'
```

Tulostaa:

Fisher r-testi otoskorrelaatiokertoimelle r_otos:

```
N=          24 r_otos=          0.83708447 tarkistus r=          0.837085
t=          7.1768658
1-puolinen testi
  H0: ei korrelaatiota
  H1: positiivinen korrelaatio
  prob(r>r_otos|H0)=    1.7043772e-07
2-puolinen testi
  H0: ei korrelaatiota
  H1: positiivinen tai negatiivinen korrelaatio
  prob(|r|>|r_otos| |H0)=    3.4087543e-07
-----
```



homehaku1/STAT2008/EXERCISES/EXERCISE4/tdsp4_1_haku1_1

haku1@haku1: Thu Apr 10 08:46:38 2008

Rank-korrelaation testaaminen

Lasketaan x ja y järjestyslukujen ('rank') korrelaatio lukujen itsensä korrelaation sijaan. Riippumaton X ja Y jakaumista

```
;-----
;Test Spearman rank-correlation coefficient r_s
;non-parametric test
;-----
; H0: there is no correlation (rho=0)
; H1: there is positive correlation (rho>0)
; testisuure t noudattaa Student'in t-jakaumaa vapausasteella N-2
;      t=r_s*sqrt((N-2)/(1.-r_s^2))
; jossa N on otoksen koko
;      r_s on Spearman'n rank-korrelaatiokerroin

N=n_elements(d) & df=N-2
res=r_correlate(d,v)*1.d0 ; palauttaa r_s ja prob r_s 2-puolisessa testissa
r_s=res(0)
t=r_s*sqrt((N-2)/(1.-r_s^2))

;tarkistuksen vuoksi lasketaan r_s suoraan maaritelmaasta
x=sort(d)+1
y=sort(v)+1
x=x-mean(x)
y=y-mean(y)
r_s_check1=total(x*y)/sqrt(total(x^2)*total(y^2))
r_s_check2=1.-6*total((x-y)^2)/(N^3-N)
print,'-----'
print,'Spearman r_s rank-korrelaatiokertoimen testatus'
print,'-----'
print,' N= ',N,' r_s=',r_s,' tarkistus:',r_s_check1,r_s_check2
print,' ero tule yhtasuurien arvojen kasittelystä'
print,' t=',t
print,'1-puolinen testi'
print,' H0: ei korrelaatiota'
print,' H1: positiivinen korrelaatio'
print,' prob(r>r_s | H0)=' ,1-t_pdf(t,df)
print,'2-puolinen testi'
print,' H0: ei korrelaatiota'
print,' H1: positiivinen tai negatiivinen korrelaatio'
print,' prob(|r|>|r_s| |H0)=' ,2*(1-t_pdf(t,df))
print,'
print,'HUOM: IDL:n r_correlate plauttaa 2-puolisen testin todennakoisyyden'
print,'res=r_correlate(d,v) --> res=',res
print,'-----'
```

Tulostaa:

```
-----
Spearman r_s rank-korrelaatiokertoimen testatus
-----
N=          24 r_s=          0.87154412 tarkistus:          0.881739          0.881739
          ero tule yhtasuurien arvojen kasittelystä
t=          8.3372667
1-puolinen testi
H0: ei korrelaatiota
H1: positiivinen korrelaatio
prob(r>r_s | H0)= 1.4770860e-08
2-puolinen testi
H0: ei korrelaatiota
H1: positiivinen tai negatiivinen korrelaatio
prob(|r|>|r_s| |H0)= 2.9541720e-08

HUOM: IDL:n r_correlate plauttaa 2-puolisen testin todennakoisyyden
res=r_correlate(d,v) --> res=          0.87154412          2.9541710e-08
```

Rank-korrelaation testaaminen: Kendall τ

Toinen ei-parametrinen testi. Käytännössä ei suurta eroa Spearman'n rank-korrelaatioon nähden

```
-----
print, 'Testaa Kendallin rank-korrelaatiokerrointa'
res=r_correlate(d,v,/kendal)
print, 'res=r_correlate(d,v,/kendal) --> res=',res

res=      0.659358  6.37770e-06
```

Johtopäätös: kaikki kolme testiä viittaavat siihen että korrelaatio d ja v välillä on todellinen.

Huom. Korrelaatiokertoimen arvo riippuu siitä miten muuttujia transformoidaan:

;järjestyksellä ei väliä:

```
IDL> print,correlate(v,d)
0.837084
IDL> print,correlate(d,v)
0.837084
;eikä lineaarisilla muunnoksilla
```

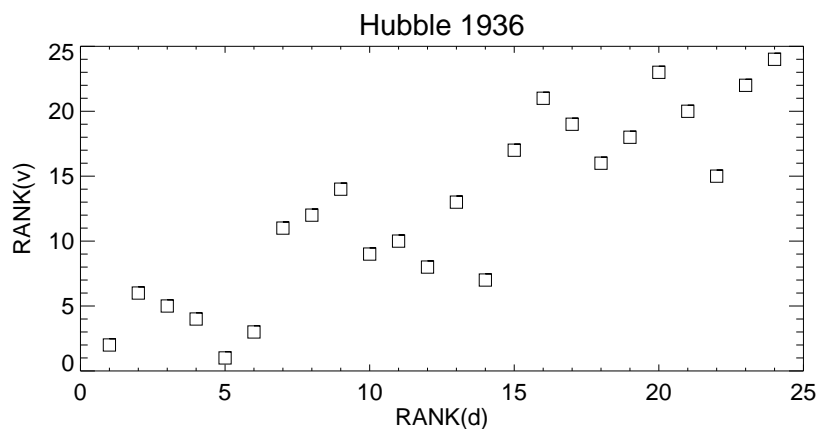
```
IDL> print,correlate(d,v*2)
0.837084
IDL> print,correlate(d-mean(d),v*2)
0.837084
```

;mutta ei-lineaarisilla muunnoksilla on

```
IDL> print,correlate(alog(d),v)
0.774086
IDL> print,correlate(alog(d),v^2)
0.650924
```

Rank-korrelaatiokertoimen arvo riippumaton muuttujien transformoinnista, kunhan se on monotoninen jolloin järjestys säilyy

```
IDL> print,r_correlate(d,v)
0.871544  2.95417e-08
IDL> print,r_correlate(alog(d),v)
0.871544  2.95417e-08
```



/home/heikki/STAT12008/EXERCISES/EXERCISE4/idiharj4_hubble_2

heikki@ikiturso Thu Apr 10 08:46:39 2008

2. Testataan miten hyvin Fisher r-testi toimii korreloimattomalle datalle

-Luo M kappaletta $N:n$ riippumattoman X, Y lukuparin otosta (X ja Y Gaussisesta jakaumasta).

-Laske kullekin otokselle korrelaatiokerroin r ja taulukoi $r:n$ jakauma

a) Valitse esim. $N = 25$ ja $M = 50000$. Arvioi saamasi r -arvojen avulla mikä on todennäköisyys että $r > 0.1$. Vertaa teoreettiseen arvioon.

b) Vertaa havaittua otoskorrelaatioiden jakaumaa teoreettiseen tiheysjakaumaan $f(r|\rho = 0$

- Esimerkkivastaukset ohjelmassa `idharj4_test_fisher_rtest.pro`

a) Todennäköisyyden $r > 0.1$ laskeminen:

Kokeellisesti: lasketaan niiden otosten suhteellinen osuus, joissa havaittu $r > 0.1$

```
N=25
M=50000
r_tab=fltarr(m) ;talletetaan otoskorrelaatiot tahan taulukkoon
for i=01,m-1 do begin
    x=randomn(seed,N)
    y=randomn(seed,N)
    r_tab(i)=correlate(x,y)
endfor
; Arvioidaan eo. kokeiden perusteella
ind=where(r_tab gt 0.1,count)
print,'kokeilun perusteella r> 0.1 todennakoisyys=',1.*count/M
```

Teoreettinen arvio:

Testisuure $t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$ noudattaa Studentin t -jakaumaa vapausasteella $df = N - 2$. Lasketaan mikä on $P(r > 0.1)$

```
;teoreettinen
r=0.1
df=N-2
t=r*sqrt(N-2)/sqrt(1.-r^2)
print,'t-jakaumasta teoreettinen todennkoisyys: ', 1.-t_pdf(t,df)
```

TULOS:

```
IDL> .run idharj4_test_fisher_rtest.pro
-----
Suureiden todellinen korrelaatio rho=0
pisteparien lkm N=      25 muodostetaan kaikkiaan M=      50000 otosta
-----
kokeilun perusteella r> 0.1 todennakoisyys=      0.312340
t-jakaumasta teoreettinen todennkoisyys:      0.317181
```

b) Todennäköisyysjakauman laskeminen. Tarkastellaan ensin testisuureen t jakauma ja sitten r :n jakaumaa

- Testisuure t kokeellisesti taulukoimalla:

```
; t=r*sqrt(N-2)/sqrt(1-r^2)
t_tab=r_tab*sqrt(N-2)/sqrt(1-r_tab^2)
histo_f,t_tab,-5,5,.1,xt,yt
plot,xt,yt,psym=10,xtitle='t_otos',ytitle='pdf(t_otos)', $
    title='otos testisuureiden t jakauma (rho=0), N='+string(N)
```

Verrataan teoreettiseen jakaumaan:

```
;Teoreettinen jakauma = Studentin jakauma vapausasteella df=N-2
df=N-2.d0
x=findgen(10001)/1001.*2-5
pdf=gamma((df+1.)/2.)*(1.+x^2/df)^(-(df+1.)/2.)/sqrt(!pi*df)/gamma(df/2.)
oplot,x,pdf,col=2
```

- Otokorrelaatio r kokeellisesti taulukoimalla:

```
histo_f,r_tab,-1,1,.05,xt,yt
plot,xt,yt,psym=10,xtitle='r_otos',ytitle='pdf(r_otos)', $
    title='otokorrelaatioiden r jakauma (rho=0), N='+string(N)
```

Verrataan teoreettiseen jakaumaan: lasketaan r :n jakauma luennoilla esitetyn jakaumien transformaatiokaavan avulla: tietyn tapahtuman todennäköisyys riippumaton käytetystä muuttujasta, eli

$$f(r)dr = g(t)dt \rightarrow f(r) = g(t)\frac{dt}{dr}$$

Nyt

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \rightarrow \frac{dt}{dr} = \frac{\sqrt{N-2}}{(1-r^2)^{1.5}}$$

Lähdetään edellä lasketusta jakaumasta $g(t)$ (=Student'n jakauma vapausasteella $N-2$)

\Rightarrow voidaan laskea haluttua r :n arvoa vastaava $f(r)$

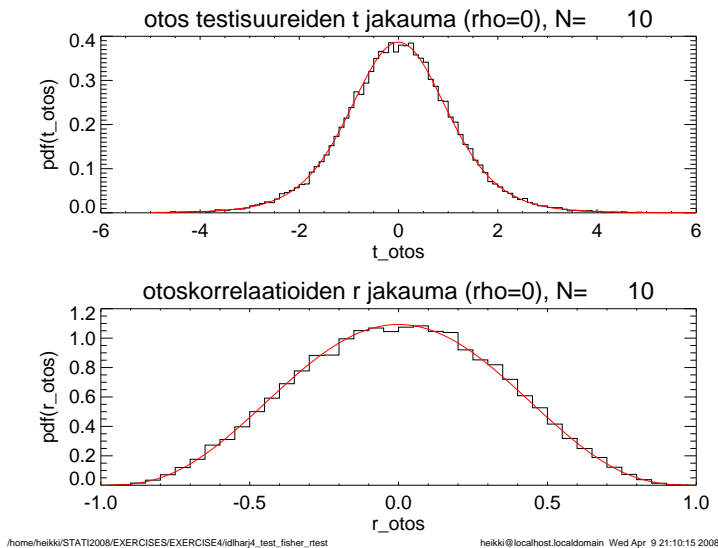
```
r=findgen(1001)/1001.*2-1
t=r*sqrt(N-2)/sqrt(1-r^2)
g=gamma((df+1.)/2.)*(1.+t^2/df)^(-(df+1.)/2.)/sqrt(!pi*df)/gamma(df/2.)
dtdr=sqrt(N-2)/(1-r^2)^1.5
pdf=g*dtdr
oplot,r,pdf,col=2
```

Tulokset seuraavalla sivulla

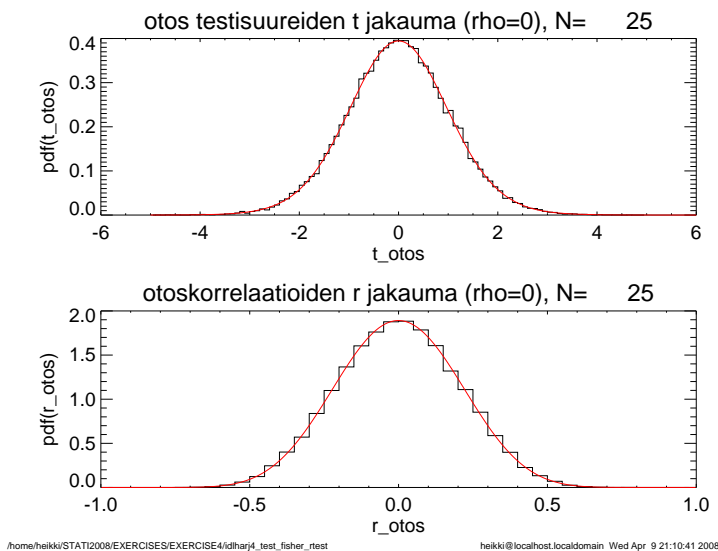
- Dataparien lukumäärä $N=10$

Histogrammi = kokeellinen jakauma

Käyrä = teoreettinen



- Dataparien lukumäärä $N=25$

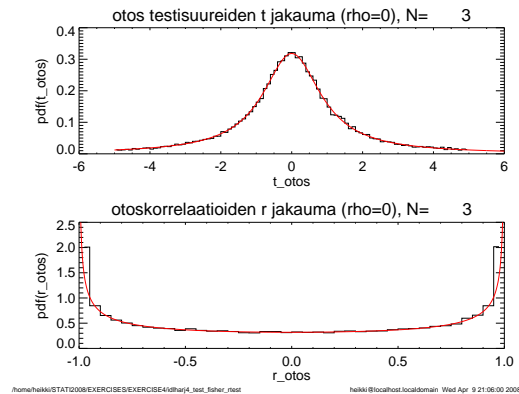


Nähdään että otoskorrelaatiokertoimen jakauma kapenee otoskoon pienetessä (hajonta verrannollinen $1/\sqrt{N}$)

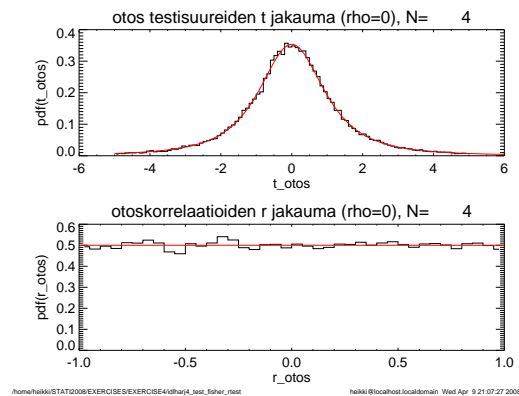
Huom: käytetty $M:n$ arvo määrää ainoastaan sen kuinka tarkasti kokeellinen jakauma saadaan määrättä

Pienen N otoksista laskettujen korrelaatiokertoimien jakaumat on sangen hupaisia:

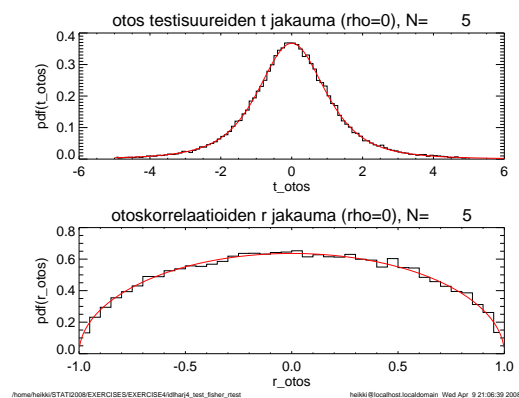
- $N=3 \Rightarrow$ kaikkein todennäköisintä saada voimakas positiivinen tai negatiivine korrelaatiokerroin, vaikkei todellista korrelaatiota olekaan.



- $N=4 \Rightarrow$ kaikki otoskorrelaatiokertoimen arvot yhtä todennäköisiä vaikkei todellista korrelaatiota olekaan.



- $N=5 \Rightarrow r = 0$:aa lähellä olevat arvot alkavat tulla todennäköisimmiksi



3. Tutkitaan korreloituneiden suureiden otoskorrelaatiokerrointa

-Luo M kappaletta N :n X, Y lukuparin otosta jotka noudattavat kahden muuttujan gaussista jakaumaa, todellinen korrelaatiokerroin ρ

Käytä X, Y parien luomiseen luennoilla annettua `bivariate_gaussian_f.pro` ohjelmaa.

-Laske kullekin otokselle otos-korrelaatiokerroin r ja taulukoi r :n jakauma

Valitse esim. $N = 25$ ja $M = 50000$ ja $\rho = 0.5$

a) Arvioi saamiesi r -arvojen avulla todennäköisyys, että $r > 0.6$. Vertaa teoreettiseen arvioon.

b) Vertaa havaittua otoskorrelaatioiden jakaumaa teoreettiseen tiheysjakaumaan $f(r|\rho = 0.5)$

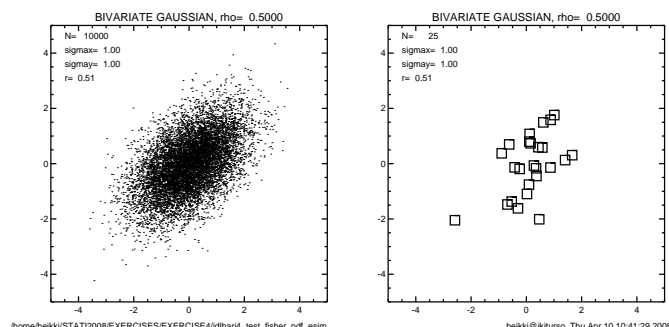
c) Entäpä jos X, Y noudattavatkin tasaista jakaumaa Gaussisen jakauman sijasta? Miten todennäköisyysjakauma muuttuu, mikä on todennäköisyys että $r > 0.6$.

Korreloituneiden dataparien luonti:

```
IDL> bivariate_gaussian_f
-----
bivariate_gaussian_f,x,y,N,rho,sx=sx,sy=sy,plot=plot,psym=psym
create N pairs x,y following a bivariate Gaussian pdf
with correlation coefficient rho, zero mean and unit variances
keywords:
  sx=sigmax, sy=sigmay specify standard deviations (defs=1)
  /plot -> plot y vs x
  plot=[nx,ny,widx,widy] -> plot 2-d histogram of pdf,
  together with theoretical pdf
EXAMPLE:
bivariate_gaussian_f,x,y,10000,0.5,/plot,sx=2
bivariate_gaussian_f,x,y,100000,0.5,plot=[100,100],sx=2
student=df -> use student t with df instead gaussian
uni -> unifrom instead of gaussian
-----
```

Esimerkki:

```
bivariate_gaussian_f,x,y,10000,0.5,/plot
bivariate_gaussian_f,x,y,25,0.5,/plot,psym=6
```



a) Kokeellinen jakauma kuten edellisessä esimerkissä

```
N=25
M=500001
rho=.5
nwin
r_tab=fltarr(m)
for i=01,m-1 do begin
    bivariate_gaussian_f,x,y,N,rho
    r_tab(i)=correlate(x,y)
endfor
;-----
;Esimerkki: mikä on todennakoisyys että otoskorrelaatio > 0.6
; kun todellinen korrelaatio rho= 0.5
; Arvioidaan eo. kokeiden perusteella
ind=where(r_tab gt 0.6,count)
print,'Suureiden todellinen korrelaatio rho=',rho
print,'kokeilun perusteella r> 0.6 todennakoisyys=',1.*count/M
```

Teoreettinen arvio: Voidaan laskea numeerisesti Fisher (1944) jakaumasta, joka on annettu luennoilla.

```
r=findgen(1001)/1000.*2.-1.
prob_r=(1.d0-rho^2)^(N/2.d0-0.5d0)*(1-r^2)^(N/2.d0-2.d0)/$
    (1.d0-rho*r)^(N-1.5d0)*(1.d0+1.d0/(N-0.5d0)*(1.d0+r*rho)/8.d0)
;normeerataan
prob_r=prob_r/(mean(prob_r)*2.)

;muodostetaan teoreettinen arvio r>0.6
ind=where(r gt 0.6)
print,'Fisher-jakauman perusteella r> 0.6 todennakoisyys=',$
    total(prob_r(ind))/total(prob_r)
```

Tulokset:

```
-----
Suureiden todellinen korrelaatio rho=      0.500000
-----
kokeilun perusteella r> 0.6 todennakoisyys=      0.263380
Fisher-jakauman perusteella r> 0.6 todennakoisyys=      0.26158947
-----
```

- Verrataan todennäköisyysjakaumia:

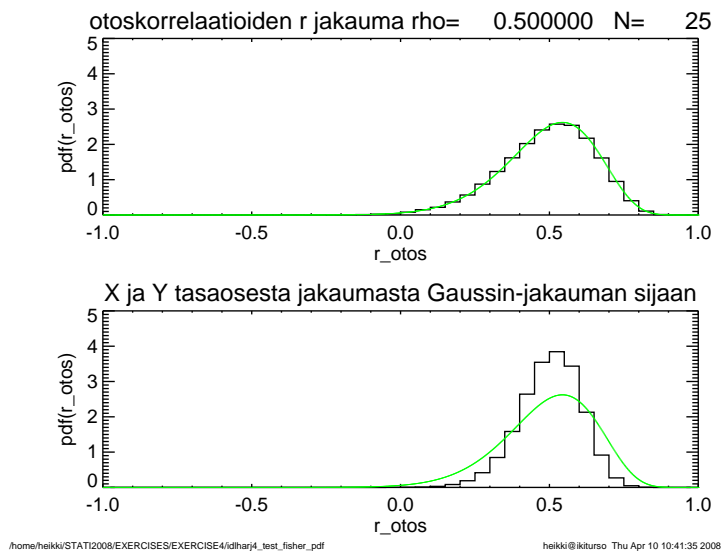
```

histo_f,r_tab,-1,1,.05,xt,yt
plot,xt,yt,psym=10,xtitle='r_otos',ytitle='pdf(r_otos)',yr=[0,5],$
    title='otoskorrelaatioiden r jakauma rho='+string(rho)+'    N='+string(N)

r=findgen(1001)/1000.*2.-1.
prob_r=(1.d0-rho^2)^(N/2.d0-0.5d0)*(1-r^2)^(N/2.d0-2.d0)/$
    (1.d0-rho*r)^(N-1.5d0)*(1.d0+1.d0/(N-0.5d0)*(1.d0+r*rho)/8.d0)
;normeerataan
prob_r=prob_r/(mean(prob_r)*2.)
oplot,r,prob_r,col=3

```

- Ylempi kuva: Gaussinen jakauma \Rightarrow havaittu ja teoreettinen hyvässä sopusoinnussa



- Alempi kuva (kohta c): X,Y otettu tasaisesta jakaumasta (ei päde enää)

```

for i=01,m-1 do begin
    bivariate_gaussian_f,x,y,N,rho,uni=1
    r_tab(i)=correlate(x,y)
endfor
histo_f,r_tab,-1,1,.05,xt,yt
plot,xt,yt,psym=10,xtitle='r_otos',ytitle='pdf(r_otos)',yr=[0,5],$
    title='X ja Y tasaisesta jakaumasta Gaussin-jakauman sijaan'
oplot,r,prob_r,col=3
ind=where(r_tab gt 0.6,count)
print,'-----'
print,'Tasainen jakauma Gaussisen jakauman sijaan:'
print,'kokeilun perusteella r> 0.6 todennakoisyys=',1.*count/M

```

kokeilun perusteella $r > 0.6$ todennakoisyys= 0.168420

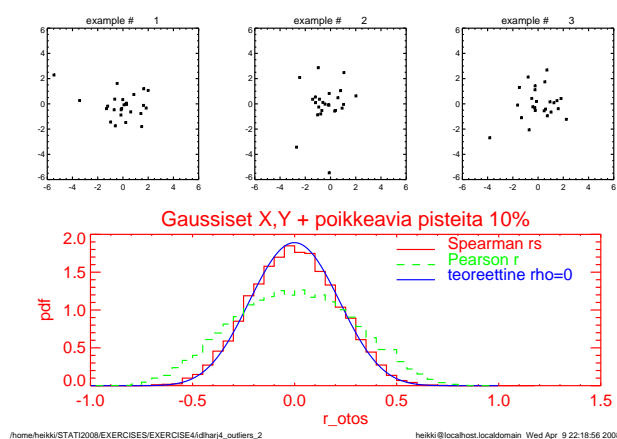
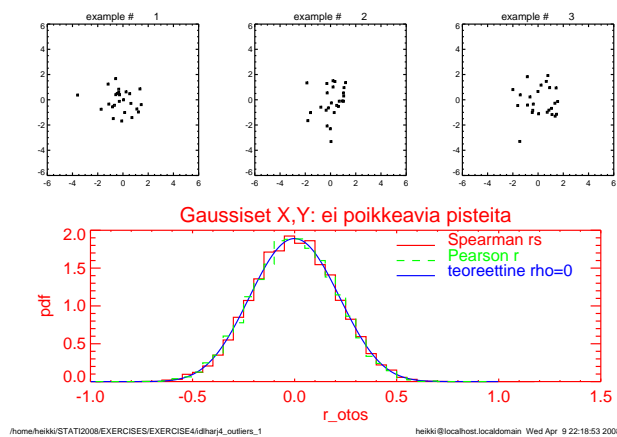
4. Tutkitaan poikkeavien pisteiden vaikutusta otoskorrelaatioon ja otos rank-korrelaatioon

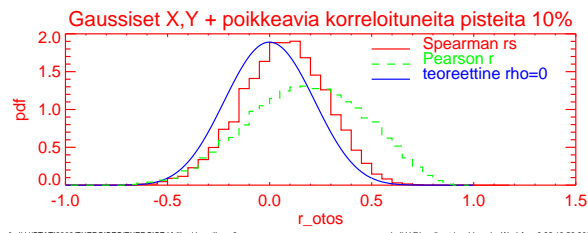
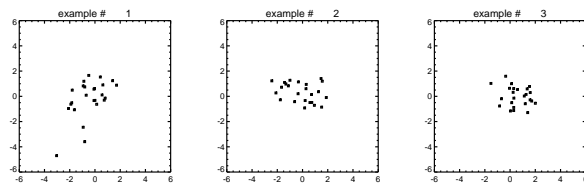
a) Luo M kappaletta $N:n$ X, Y lukuparin otosta jotka noudattavat kahden muuttujan gaussista jakaumaa, todellinen korrelaatiokerroin ρ_{00}

b) Lisää joukkoon 10 prosenttia poikkeavia pisteitä (riippumaton gaussinen jakauma, kolminkertainen hajonta)

c) Lisää joukkoon 10 prosenttia poikkeavia pisteitä (korreloitunut gaussinen jakauma ($\rho = 0$, kolminkertainen hajonta)

-Laske kullekin otokselle otoskorrelaatiokerroin r ja taulukoi $r:n$ jakauma. tee sama Spearman rank-korrelaatiokertoimelle r_s





/home/heikki/STAT2008/EXERCISES/EXERCISE4/dharj4_outliers_3

heikki@localhost.localdomain Wed Apr 9 22:18:58 2008