TILASTOLLISET MENETELMÄT TÄHTITIETEESSÄ
IDL-HARJOITUS I esimerkkivastaukset, Heikki Salo 21.2.2008

1) Binomijakauma

1.1) Tee IDL-ohjelma joka plottaa binomijakauman

$$\begin{aligned} \text{prob}(V) &= \binom{N}{V} p^V (1-p)^{N-V} \qquad V = 0, 1, ..., N \\ \binom{N}{V} &= \frac{N!}{V!(N-V)!} \end{aligned} \tag{1}$$

arvoille $N = 10, p = 0.5$, käyttäen suoraan kertomien avulla laskettuja binomitekijöitä. Lisää samaan kuvaan $p = 0.1$ ja $p = 0.9$.

1.2) Tee sama arvoille $p = 0.5$, $N = 10, 20, 40, 80, 160$ Käytä hyväksi relaatiota

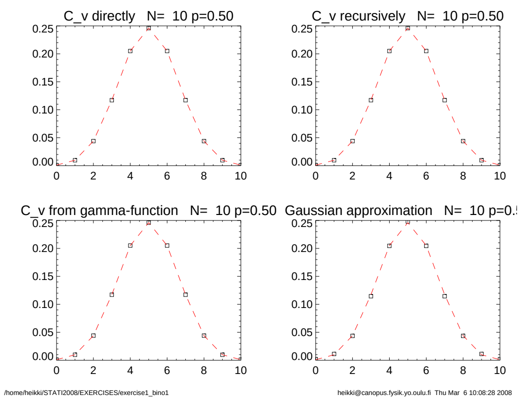
$$\begin{aligned} \log \frac{N!}{v!(N-v)!} &= \log N! - \log v! - \log(N-v)! \\ &= \log \Gamma(N+1) - \log \Gamma(v+1) - \log \Gamma(n-v+1) \end{aligned}$$

IDL-funktio `lngamma(x)` palauttaa lausekkeen $\log \Gamma(x)$

- ESIMERKKIOHJELMA: `exercise1_bino1.pro`

laskee binomijakauman 4 eri tavalla ja vertaa tata IDL:n kirjastofunktiolla saatuun:

1) lasketaan jokainen binomikerroin suoraan kertoman maaritelmasta
2) kaytetaan rekursiokaavaa binomikertoimille
3) Gamma-funktion avulla
4) Gaussin jakauman avulla approksimoituna



```
;--------------------------------------------------
;STATISTICAL METHODS IN ASTRONOMY
; EXERCISES I
; 21.02.2008 Heikki Salo
; exercise1_bino1.pro
;--------------------------------------------------

 program='exercise1_bino1'

;--------------------------------------------------
;calculate binomial probability prob(V;N,p)
;prob(V; N,p) = C_V p^V (1-p)^(N-v)
;
;           N!          N*(N-1)* ... *(N-V+1)
; C_V =  ---------   =  --------------------
;        V! (N-V)!      1*2* ....       * V

; the latter holds only when V>1
; C_0 = 1
;--------------------------------------------------

;--------------------------------------------------
;1. with a direct evaluation of binomial coefficients
;           N!
; C_V =  ---------
;        V! (N-V)!
;--------------------------------------------------
;
;--------------------------------------------------
;2. with binomial coefficients calculated recursively
;
; C_k = 1  for k=0
; C_(k+1)=C_k * (N-k)/(k+1)   k=1,2,..N-1
;
;--------------------------------------------------

;--------------------------------------------------
;3. Using the gamma-function
; log(C_v) = LNGAMMA(n+1)-LNGAMMA(v+1)-LNGAMMA(n-v+1)

;--------------------------------------------------

;--------------------------------------------------
;4. Approximating with a Gaussian functian
; Large N --> approaches Gaussian with
; mu= p*N
; sigma=sqrt(p*(1-p)*N)
; use gauss_pdf which returns cumulative density function of
; normalized gaussian (zero mean, unit standard deviation)

; subsitute  x1= (V-0.5- mu)/sigma
;            x2= (V+0.5- mu)/sigma
; prob(V)  = phi(x2)-phi(x1)   approximatively
;--------------------------------------------------

;-------------------------------------------------
;0. check by bino_pdf  (utilizing idl-library routine)
;-------------------------------------------------
;output to screen (or file)
 ps=0
 psdirect,program,ps,/color
;-------------------------------------------------
```

```
 !p.multi=[0,2,2]

 p=0.5d0
 N=10

;----------------------------------------------------------
;0. IDL-library routine

  vtab0=dindgen(n+1)
  bino_pdf,n,p,vtab0,ptab0

;------------------------------------------------------------
;1. directly
;calculate and store factorial
;------------------------------------------------------------

  vtab=dindgen(n+1)             ;(largest possible  n about 170)
  ptab1=vtab*0.

  fac=dindgen(n+1)
  fac(0)=1.
  for i=1,n do begin
     fac(i)=fac(i-1)*i
  endfor
  for v=0,n do begin
     c_v=fac(n)/fac(v)/fac(n-v)
     ptab1(v)=c_v*p^v*(1.-p)^(n-v)
  endfor

  nwin
  add='   N='+string(n,'(i4)')+' p='+string(p,'(f4.2)')
  plot,vtab,ptab1,title='C_v directly'+ add,psym=6,syms=.5
  oplot,vtab0,ptab0,col=2,lines=2

;------------------------------------------------------------
;2. calculate c_v recursively
;------------------------------------------------------------

  ptab2=vtab*0.
  c_v=1.d0
  for v=0,n do begin
     ptab2(v)=c_v*p^v*(1.-p)^(n-v)
;binomial factor for the NEXT v
     c_v=c_v*(n-v)/(v+1)
  endfor

  nwin
  add='   N='+string(n,'(i4)')+' p='+string(p,'(f4.2)')
  plot,vtab,ptab2,title='C_v recursively'+add,psym=6,syms=.5
  oplot,vtab0,ptab0,col=2,lines=2

;------------------------------------------------------------
;3. Use gamma-function
;------------------------------------------------------------

  ptab3=vtab*0.
  for v=0,n do begin
;remember to make in double precision!
     logfac = LNGAMMA(n+1.d0)-LNGAMMA(v+1.d0)-LNGAMMA(n-v+1.d0)
     ptab3(v)=exp(logfac)*p^v*(1.-p)^(n-v)
  endfor
```

```
  nwin
  add='   N='+string(n,'(i4)')+' p='+string(p,'(f4.2)')
  plot,vtab,ptab3,title='C_v from gamma-function'+add,psym=6,syms=.5
  oplot,vtab0,ptab0,col=2,lines=2

;-----------------------------------------------------------
;4. Use Gaussian approximation
;-----------------------------------------------------------

  ptab4=vtab*0.
  mu=p*n
  sigma=sqrt(p*(1.d0-p)*n)

  for v=0,n do begin
     x1= (V-0.5- mu)/sigma
     x2= (V+0.5- mu)/sigma
     ptab4(v)=gauss_pdf(x2)-gauss_pdf(x1)
  endfor

  nwin
  add='   N='+string(n,'(i4)')+' p='+string(p,'(f4.2)')
  plot,vtab,ptab4,title='Gaussian approximation'+add,psym=6,syms=.5
  oplot,vtab0,ptab0,col=2,lines=2

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

- IDL:ssa itsessaan on kirjastofunktio binomial joka palauttaa kumulatiivisen binomijakauman ($P(x \geq v)$.Tasta voidaan muodostaa tiheysfunktio ottamalla erotus:

$$P(x = v) = P(x \geq v) - P(x \geq v + 1)$$

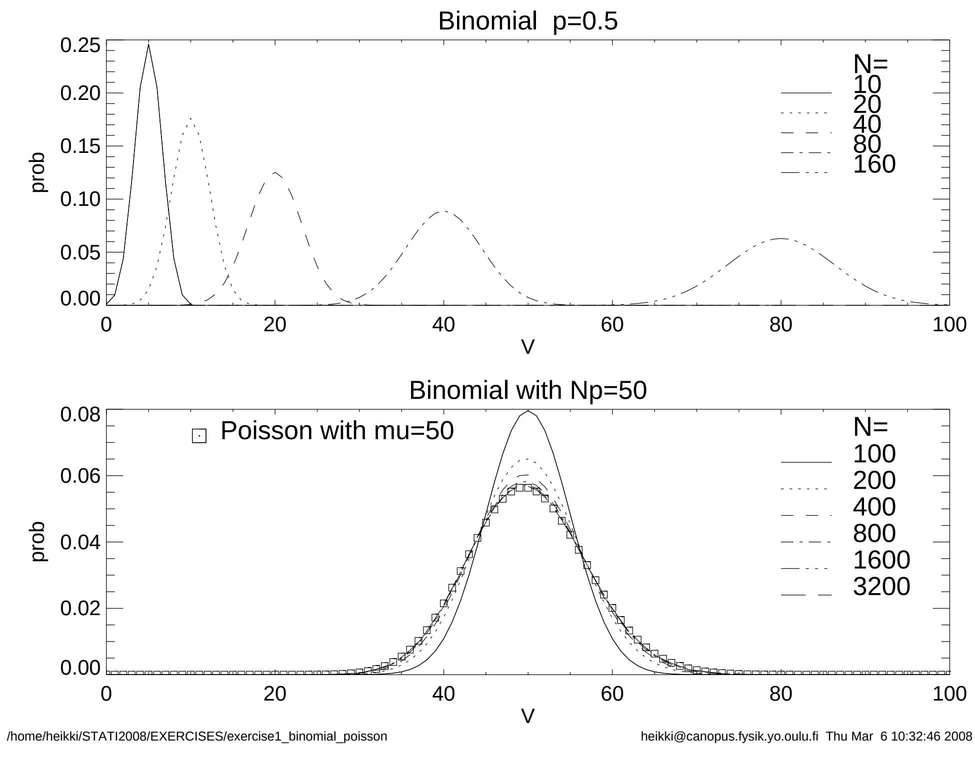
Tätä varten on tehty apuohjelma `bino_pdf.pro`

```
;----------------------------------
pro bino_pdf,n,p,vtab,ptab
;----------------------------------

if(n_params() le 0) then begin
   print,'calculate binomial probability using IDL-library function'
   print,'binomial for cumulative probability'
   print,'-----------------------------------------'
   print,"bino_pdf,n,p,vtab,ptab"
   print,'input= n,p -> returns  ptab(vtab)'
   return
endif

vtab=dindgen(n+1)
ptab=vtab*0.d0

for i=0,n do begin
   ptab(i)=binomial(i,n,p,/double)-binomial(i+1.d0,n,p,/double)
endfor
end
```

1.3) Laske keskeisen raja-arvolauseen perusteella tn V=1500 kun N=3000, p=0.5 (vastaus 0.014566503)
Tehty laskuharjoituksissa: totea IDL:n avulla

2. Poisson jakauma
- tuota luentojen sivulla 31 oleva kuva
- Osoita että binimijakauma lähestyy Poisson jakaumaa kun $N- > \infty$ ja nP äärellinen (luentojen sivun 29 kuva).

• <u>ESIMERKKIOHJELMA: `exercise1_binomial_poisson.pro`</u>
Binomijakauma lahestyy Poisson jakaumaa

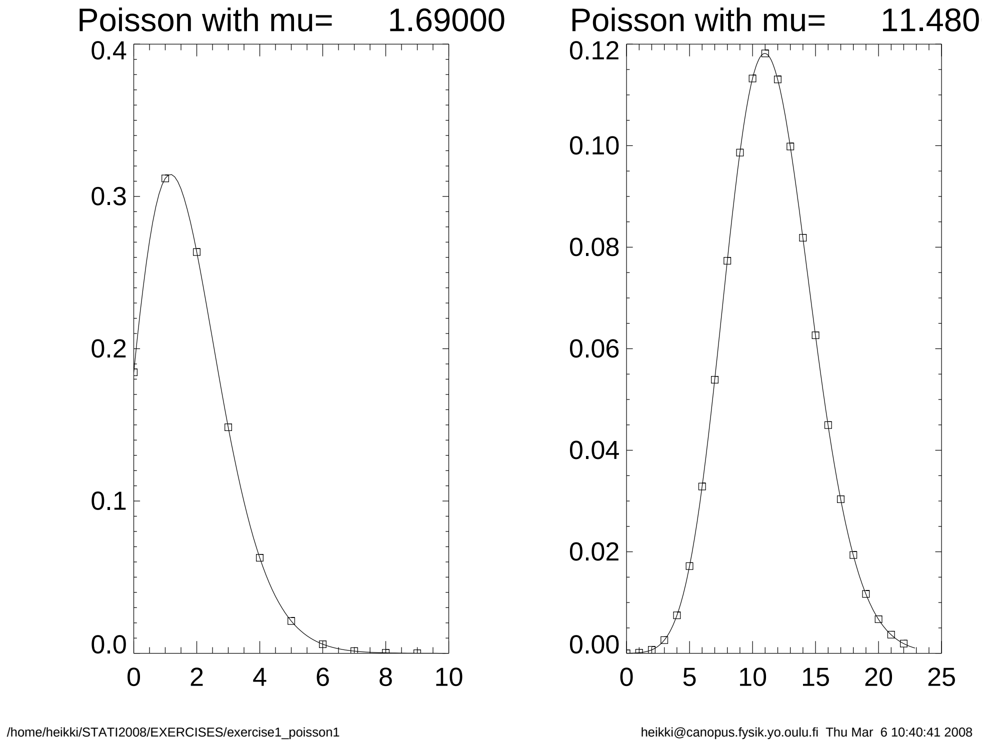


```
;---------------------------------------
program='exercise1_binomial_poisson'
;---------------------------------------

;----------------------------------------------------
;binomial distribution with increasing N, fixed p
;----------------------------------------------------


 ps=0
 psdirect,program,ps
 !p.multi=[0,1,2]
 !p.charsize=0.8

 for icase=1,6 do begin
    fac=2.^(icase-1.d0)
    p=0.5d0
    nn=10.d0*fac
    vtab=dindgen(nn+1)
    ptab=vtab*1.d0
    Q=1.D0-P
    nn_use=(n_elements(vtab)-1)<110d0
    print,nn_use,p

    for i=0l,nn_use  do begin

;make binomial using Gamma-function, and then going to
;Gaussian approximation if needed

       vv=vtab(i)
       logFact = LNGAMMA(nn+1)-LNGAMMA(vv+1)-LNGAMMA(nn-vv+1)
       IF(LOGFACT GT 700) THEN EFAC=1. - GAUSS_PDF((vV-0.5D0-NN*p)/SQRT(Nn*p*q))
       IF(LOGFACT lt 700) THEN EFAC=exp(logfact)
       ptab(i)=efac*p^vv*q^(nn-vv)
    endfor

    if(icase eq 1) then begin
       plot,vtab,ptab,xr=[0,100],xtitle='V',ytitle='prob',title='Binomial  p=0.5'
    endif
    oplot,vtab,ptab,col=icase,lines=icase-1

 endfor
 label_data,0.8,0.8,['10','20','40','80','160'],len=0.06,title='N='


;----------------------------------------------------
;binomial distribution with increasing N, keeping pn fixed
;----------------------------------------------------

 for icase=1,6 do begin

    fac=2.^(icase-1.d0)
    p=0.5d0/fac
    nn=100.d0*fac
    vtab=dindgen(nn+1)
    ptab=vtab*1.d0
    Q=1.D0-P
   nn_use=(n_elements(vtab)-1)<110d0
   print,nn_use,p

   for i=0l,nn_use  do begin
      vv=vtab(i)
```

```
      logFact = LNGAMMA(nn+1)-LNGAMMA(vv+1)-LNGAMMA(nn-vv+1)
      IF(LOGFACT GT 700) THEN EFAC=1. - GAUSS_PDF((vV-0.5D0-NN*p)/SQRT(Nn*p*q))
      IF(LOGFACT lt 700) THEN EFAC=exp(logfact)
      ptab(i)=efac*p^vv*q^(nn-vv)
   endfor

   if(icase eq 1) then begin
      plot,vtab,ptab,xr=[0,100],xtitle='V',ytitle='prob',$
           title='Binomial with Np=50'
   endif
   oplot,vtab,ptab,col=icase,lines=icase-1

endfor

 label_data,0.8,0.8,['100','200','400','800','1600','3200'],$
           len=0.06,y_i=0.1,title='N='

 label_data,0.01,0.9,['Poisson with mu=50'],len=0.1,psym=-6,shift=.001


;overplot Poisson

 mu=nn*p
 xtab=dindgen(110)
 poisson=xtab*0.d0
 for i=0.d0,n_elements(poisson)-1 do begin
    poisson(i)=mu^i/exp(lngamma(i+1))*exp(-mu)
 endfor
 oplot,xtab,poisson,psym=6,sym=.5
 psdirect,program,ps,/stop

 !p.charsize=1
end
```

- <u>ESIMERKKIOHJELMA: `exercise1_poisson1.pro`</u>

Esimerkkejä Poisson-jakaumista: Luennoilla ollut esimerkki (Bevington esim. 2.4).
-jakaumat on piirretty yhdistämällä varsinaiset diskreetit arvot jatkuvalla käyrällä (laskettu Gamma-funktion avulla)




```
;-------------------------------------------------
;STATISTICAL METHODS IN ASTRONOMY
; EXERCISES I
; 21.02.2008 Heikki Salo
; exercise1_bino1.pro
;-------------------------------------------------

 program='exercise1_poisson1'
 ps=0
 psdirect,program,ps

;------------------------------------------------
;Poisson distribution
;
;              mu^V
; prob(V) =    -----  exp(-mu)   V=0,1,2,3,...
;               V!
;------------------------------------------------
```

```
;Draw with in the case mu=1.69 and mu=11.48
;(corresponds to plots in page 31, except that the curves
;should be multiplied by the number of samples (100 and 60)

;make also a 10 finer grid


!p.multi=[0,2,1]

;-----------------------------------------------------------
mu=1.69
vtab=findgen(10)
ptab=vtab
for v=0,n_elements(vtab)-1 do begin
ptab(v)=mu^v/exp(lngamma(v+1))*exp(-mu)
endfor
nwin
plot,vtab,ptab,title='Poisson with mu='+string(mu),psym=6,syms=0.5


vtab=findgen(100)*.1
ptab=vtab
for i=0,n_elements(vtab)-1 do begin
v=vtab(i)
ptab(i)=mu^v/exp(lngamma(v+1))*exp(-mu)
endfor
oplot,vtab,ptab


;-----------------------------------------------------------
mu=11.48
vtab=findgen(23)
ptab=vtab
for v=0,n_elements(vtab)-1 do begin
ptab(v)=mu^v/exp(lngamma(v+1))*exp(-mu)
endfor

plot,vtab,ptab,title='Poisson with mu='+string(mu),psym=6,syms=0.5


vtab=findgen(230)*.1
ptab=vtab
for i=0,n_elements(vtab)-1 do begin
v=vtab(i)
ptab(i)=mu^v/exp(lngamma(v+1))*exp(-mu)
endfor

oplot,vtab,ptab
 psdirect,program,ps,/stop
end
```

Useful statistical procedures in IDL:

```
--------------------------------------------
PROBABILITY FUNCTIONS
--------------------------------------------

BINOMIAL-function
----------------------
Cumulative binomial probability distribution

Y=Binomial(V,N,P)

    V= minimum number of event occurring in
    N independent trials.
    P=probability in single event


GAUSS_PDF(V) function:
----------------------
  gives the probability Prob(x le V)
  for a Gaussian probability distribution
  with a mean of 0.0 and a variance of 1.0,

IDL> print,gauss_pdf(1)
     0.841345
IDL> print,gauss_pdf(2)
     0.977250
IDL> print,gauss_pdf(3)
     0.998650

GAUSS_CVF(P) function:
----------------------
  gives the cut-off value V so that Prob(x gt V)=P
  for a Gaussian probability  distribution
  with a mean of 0.0 and a variance of 1.0,

IDL> print,gauss_cvf(0.1)
      1.28155
IDL> print,gauss_cvf(0.01)
      2.32635
IDL> print,gauss_cvf(0.001)
      3.09025
```

```
CHISQR_PDF(V,df):
----------------------
  gives the probability Prob(x le V)
  for a Chi-square distribution with DF degrees of freedom.

CHISQR_CVF(P,df)
----------------------
  gives the cut-off value V so that Prob(x gt V)=P
  for a Chi-square distribution with DF degrees of freedom.


T_PDF(V,df):
----------------------
  gives the probability Prob(x le V)
  for a Student's t distribution with DF degrees of freedom.

T_CVF(P,df)
----------------------
  gives the cut-off value V so that Prob(x gt V)=P
  for a Student's t distribution with DF degrees of freedom.


F_PDF(V, DFn, DFd):
----------------------
  gives the probability Prob(x le V)
  for an F distribution with DFn,DFd degrees of freedom.

F_CVF(P, DFn, DFd)
----------------------
  gives the cut-off value V so that Prob(x gt V)=P
  for an F distribution with DFn,DFd degrees of freedom.
```

```
------------------------------------------
CALCULATION OF SAMPLE STATISTICS
------------------------------------------

  MEAN(X)             mean
  MEDIAN(X)           median
  MIN(X)              smallest element
  MAX(X)              largest  element
  STDDEV(X)           standard deviation (sqrt of sample variance)
  MEANABSDEV(X)       mean absolute deviation
  SKEWNESS(X)         skewness
  KURTOSIS(X)         kurtosis

MOMENT-function
----------------
RESULT = MOMENT(X, MDEV=variable, SDEV=variable)
  X=vector of sample values
  Result=[mean,sample variance, skewness, kurtosis]
  MDEV = mean absolute deviation
  SDEV = standard deviation = sqrt of sample variance

SORT-function
-------------
INDEX=SORT(X)
  returns indices so that X(index) in ascending order

-----------------------------------------
DISPLAYING DISTRIBUTIONS
-----------------------------------------

HISTOGRAM-function
-------------------

Y=HISTOGRAM(X)
  Y(i)=number of array elements X(j) with value i

Y=HISTOGRAM(X, BINSIZE=value,MIN=value,MAX=value)
  Y(I)=number of array elements in the specified
  (MAX-MIN)/BINSIZE intervals


HISTO_F-procedure (Auxillary procedure written in IDL)
```

```
-----------------------------------------------
pro histo_f,x,x1,x2,dx,xx,yy,gg
  x=input values
  histogram from x1 to x2 with step dx
  xx,yy return calculated values
  gg returns corresponding gaussian fit (if /gauss)
    /plot       -> plot data
    /oplot      -> oplot data
    /auto       -> automatic scaling with 50 part/bin
    auto=nbins -> automatic scaling with nbin bins
    /noscale    -> do not scale distribution (default: normalized area)
    color=col  -> plot with given color
    psym =sym  -> plot with given symbol type (try psym=10)
```

```
--------------------------------------------
GENERATTION OF RANDOM NUMBERS
--------------------------------------------

----------------------
* Uniform distribution
-----------------------

The RANDOMU function returns one or more uniformly-distributed,
floating-point, pseudo-random numbers in the range 0 < Y < 1.0.

  N=1000
  Y=RANDOMU(SEED,N)   ;N=1000 uniform random numbers

If the variable SEED in undefined before the call, IDL initializes
the random sequence by a value formed from time and date of call.
After the call SEED is defined, and is used at subsequent calls to
continue the random sequence.

If SEED is set to scalar value before the first call, yields a fixed
set of random numbers.


----------------------
* Normal distribution:
-----------------------

  N=1000
  Y=RANDOMN(seed, N)    ;N=1000 random numbers following Gaussian distribution
         ;with zero mean and unit variance

To create Gaussian numbers with mean MU and variance VARI, use
  Y=sqrt(VARI)*RANDOMN(seed,N)+MU

-----------------------
* Other distributions:
-----------------------

Either RANDOMU or RANDOMN accepts keywords:

BINOMIAL=[n,p] -->
  generate random numbers from a binomial distribution.
```

```
  If an event occurs with probability p, with n trials, then the
  number of times it occurs has a binomial distribution.

  N=1000
  Y=RANDOMN(seed,N,binomial=[10,.25])

POISSON=A  -->
  generate random numbers from a Poisson distribution with the mean A
  (mean number of rare events occurring during a unit time)

GAMMA=i    -->
  random number from gamma-distribution = waiting time to the i-th event
  in Poisson random process of unit mean.
  i=1 corresponds to exponential distribution
```