SURVO-KILPATEHTÄVÄ 4: ETÄISYYSJAKAUMA

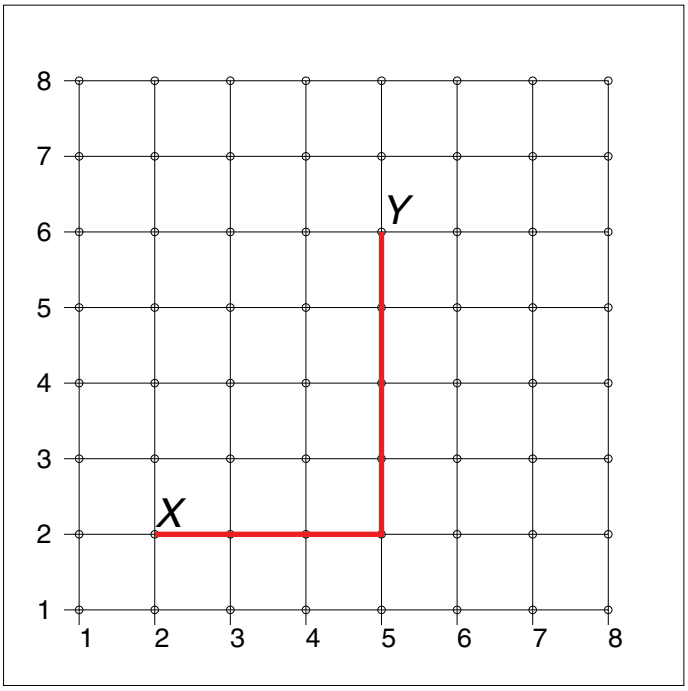
SEPPO MUSTONEN
12.3.2004

1. Tehtävän kuvaus

Käsittelen keväällä 2004 antamaani Survon käyttäjille tarkoitetun kilpatehtävän toista osaa. Kilpa oli tarkoitettu erityisesti Survosta kiinnostuneille opiskelijoille. Tehtävä esitettiin seuraavassa muodossa:

Ongelma liittyy "etäisesti" omaan, 40 vuotta sitten tekemääni, matematiikan väitöskirjatyöhöni "On distance distributions in networks", mutta sen tuntemisesta ei tehtävän ratkaisun kannalta ole mitään apua.

Tarkastellaan $n \times n$ pisteen muodostamaa neliönmuotoista "katuverkkoa" (kuvassa on tapaus $n = 8$).



Näistä n^2 pisteestä valitaan toisistaan riippumatta, umpimähkään kaksi pistettä (yllä X ja Y ovat valitut pisteet) ja olkoon D niiden välinen etäisyys verkkoa pitkin (eli tässä tapauksessa D on 3+4=7 yksikköä).

Kun $n = 8$, D voi saada arvoja $0, 1, 2, \ldots, 14$ ja yleisesti $n \times n$-verkossa mahdolliset D-arvot ovat $0, 1, 2, \ldots, 2n - 2$.

Jokaisella etäisyydellä D on tietty todennäköisyys ja nämä todennäköisyydet määräävät sen jakauman.

Yksinkertaisimmassa tapauksessa $n = 2$ mahdolliset X-Y-valinnat (16 kpl.) ja niitä vastaavat etäisyydet D ovat

X	Y	D	X	Y	D
(1,1)	(1,1)	0	(2,1)	(1,1)	1
(1,1)	(1,2)	1	(2,1)	(1,2)	2
(1,1)	(2,1)	1	(2,1)	(2,1)	0
(1,1)	(2,2)	2	(2,1)	(2,2)	1
(1,2)	(1,1)	1	(2,2)	(1,1)	2
(1,2)	(1,2)	0	(2,2)	(1,2)	1
(1,2)	(2,1)	2	(2,2)	(2,1)	1
(1,2)	(2,2)	1	(2,2)	(2,2)	0

eli D-jakauma on seuraava:

D	0	1	2
todennäköisyys	4/16	8/16	4/16
	1/4	1/2	1/4

Etäisyyden odotusarvo (keskiarvo) on siis $1/4 \times 0 + 1/2 \times 1 + 1/4 \times 2 = 1$ (jo jakauman symmetrisyyden vuoksi) ja keskihajonnaksi tulee
$\sqrt{1/4 \times 0 \times 0 + 1/2 \times 1 \times 1 + 1/4 \times 2 \times 2 - 1 \times 1} = \sqrt{1/2} = 0.7071...$

Tässä on suurin piirtein kaikki, mitä voidaan sanoa tapauksesta $n = 2$.

Varsinaisena tehtävänä on nyt tutkia tapauksia $n = 3, 4, 5, \ldots$ tukeutuen olennaisesti Survon tarjoamiin aineistonkäsittely- ja laskentakeinoihin. Käytettävissä ovat kaikki Survon operaatiot ja saahan sitä laatia sukrojakin.

Ensisijaisesti ei ole tarkoituksena lähteä johtamaan kynällä ja paperilla mitään yleisiä tuloksia, mutta sellaistakin saa tehdä, kun on ensin selvitellyt näitä erikoistapauksia ja vaikkapa yrittänyt arvata jotain jakauman yleisestä luonteesta. Yksi mielenkiintoinen kysymys on, lähestyykö jakauma luvun n kasvaessa normaalijakaumaa; sitä voi arvioida ihan intuitiivisestikin.

Survolle ominaiseen tyyliin ratkaisun tulisi olla "itsensä dokumentoiva" niin, että vastauksena lähetetystä toimituskentästä tai sen osasta ilmenevät kaikki ratkaisun vaiheet.

Tehtävää ei varmaankaan tarvitse selvittää yleisesti (tullakseen palkituksi) vaan riittää käsitellä (mieluiten näppärämmin kuin yllä tein tilanteessa $n = 2$) tapauksia $n = 3, 4, 5, \ldots$ ja laskeskella todennäköisyyksien lisäksi tunnuslukuja (keskiarvo, keskihajonta).

Korostan vielä erityisesti sitä, että arvostan ratkaisuja, joissa Survon käytöllä on merkittävä osuus.

Seuraava tarina ei ole mikään mallivastaus vaan siinä hahmotellaan erilaisia Survolle ominaisia ja hiukan teoreettispitoisiakin lähestymistapoja.

2. Ratkaisu valitulla arvolla n

Tarkastelen ongelmaa hiukan "kiertäen ja kaartaen". Päämääränäni on osoittaa, miten monenlaisin tavoin Survoa saattaa hyödyntää ongelmanratkaisussa.

Tutkitaan kysymystä ensin kiinteillä arvoilla $n = 3, 4, 5, \ldots$ Tarjolla on useita vaihtoehtoja. Esitän ensin menettelyn, joka mielestäni on kaikkein lyhin mitä tulee tarvittavien Survo-komentojen määrään, mutta jossa käytetään raakaa laskentavoimaa ja paljon muistitilaa. Nykyisillä koneilla sillä päästään tapauksiin, joissa n on muutamia kymmeniä. Käyttäen hieman enemmän ajatusvaivaa esitän toisen, huomattavasti tehokkaamman ratkaisun, jolla ylletään tilanteisiin, joissa n on vaikkapa kymmeniä tuhansia.

Raaka ratkaisu perustuu kaikkien mahdollisten pisteparien X, Y luettelointiin Survon COMB-operaation avulla. COMB on tarkoitettu erilaisiin kombinatoorisiin tehtäviin. Sillä voi listata mm. suoraan toimituskenttään permutaatioita, partitioita, osajoukkoja jne. COMB laskee myös pelkästään eri vaihtoehtojen lukumääriä. Pisteparien X, Y luettelointi, kun n on valittu, tapahtuu käyttäen joko optiota INTEGERS tai LATTICE (hila). Sovellan jälkimmäistä, jolloin luettelo syntyy tapauksessa $n = 2$ seuraavasti:

```
-------------------------------------------------------------------
COMB L,CUR+1 / L=LATTICE,4 MIN=1,1,1,1 MAX=2,2,2,2
Lattice points in 4 dimensions: N[L]=16
1 1 1 1
1 1 1 2
1 1 2 1
1 1 2 2
1 2 1 1
1 2 1 2
1 2 2 1
1 2 2 2
2 1 1 1
2 1 1 2
2 1 2 1
2 1 2 2
2 2 1 1
2 2 1 2
2 2 2 1
2 2 2 2
-------------------------------------------------------------------
```

Saatiin luettelo, jossa on 4 saraketta ja kussakin esiintyy vain arvoja 1,2 niin, että mukana on kaikki mahdolliset kombinaatiot. Kaksi ensimmäistä saraketta tulkitaan pisteen X koordinaateiksi verkossa ja kaksi jälkimmäistä pisteen Y koordinaateiksi.

Siirrytään hieman suurempaan arvoon $n = 8$, jolloin vastaava luettelo kasvaa $8^4 = 4096$ rivin mittaiseksi:

```
-------------------------------------------------------------------
COMB L,CUR+1 / L=LATTICE,4 MIN=1,1,1,1 MAX=8,8,8,8
Lattice points in 4 dimensions: N[L]=4096
1 1 1 1
1 1 1 2
```

```
1 1 1 3
1 1 1 4
1 1 1 5
1 1 1 6
1 1 1 7
1 1 1 8
1 1 2 1
1 1 2 2
...     (4096-20 tapausta tässä välissä)
8 8 7 7
8 8 7 8
8 8 8 1
8 8 8 2
8 8 8 3
8 8 8 4
8 8 8 5
8 8 8 6
8 8 8 7
8 8 8 8
```

Etäisyysjakauman määräämiseksi annetaan em. luettelolle DATA-määrite PISTEET (muuttujat olkoot X1,X2,Y1,Y2) ja talletetaan luettelo Survon uudeksi havaintotiedostoksi VERKKO8:

```
FILE COPY PISTEET TO NEW VERKKO8
COMB L,CUR+1 / L=LATTICE,4 MIN=1,1,1,1 MAX=8,8,8,8
Lattice points in 4 dimensions: N[L]=4096
DATA PISTEET
X1 X2 Y1 Y2
1 1 1 1
1 1 1 2
1 1 1 3
1 1 1 4
1 1 1 5
...
```

Pisteiden $X = (X_1, X_2)$ ja $Y = (Y_1, Y_2)$ etäisyys D on erotusten $X_1 - X_2$ ja $Y_1 - Y_2$ itseisarvojen summa. Etäisyydet lasketaan yksinkertaisesti uutena muuttujana D tiedostoon VERKKO8 komennolla

```
VAR D:2=abs(X1-X2)+abs(Y1-Y2) TO VERKKO8
```

ja etäisyyksien frekvenssijakauma (tunnuslukuineen) syntyy STAT- komennolla

```
-------------------------------------------------------------------
STAT VERKKO8,CUR+1 / VARS=D
Basic statistics: VERKKO8 N=4096
Variable: D        ~abs(X1-X2)+abs(Y1-Y2)
min=0        in obs.#1
max=14       in obs.#456
mean=5.25     stddev=2.687101 skewness=0.346775 kurtosis=-0.330837
autocorrelation=0.8152
lower_Q=3        median=5        upper_Q=7
D             f     %      *=16 obs.
       0     64   1.6 ****
       1    224   5.5 **************
       2    388   9.5 ************************
       3    496  12.1 ******************************
       4    552  13.5 **********************************
       5    560  13.7 **********************************
       6    524  12.8 ********************************
       7    448  10.9 ****************************
       8    336   8.2 ********************
       9    224   5.5 **************
      10    140   3.4 ********
      11     80   2.0 *****
      12     40   1.0 **
      13     16   0.4 *
      14      4   0.1 :
-------------------------------------------------------------------
```

Siten neljän Survo-komennon (COMB, FILE COPY, VAR, STAT) yhdistelmä on tuottanut kaikki tarpeelliset perustulokset (tapauksessa n=8).

Äskeisen menettelyn varjopuolena on se, että X-Y-yhdistelmien lukumäärän ollessa yleisesti n^4 suurilla n-arvoilla laskenta hidastuu ja tulee kapasiteettiongelmia.

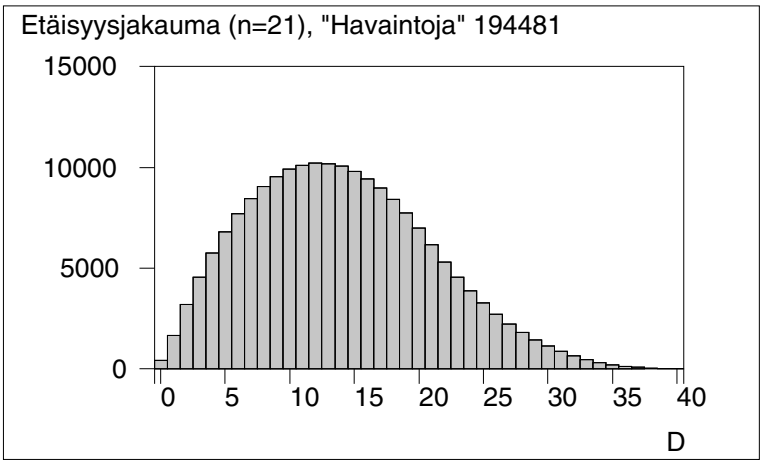
Jotta saataisiin jonkinlainen käsitys jakauman muodosta jo nyt, käsitellään tapaus $n = 21$ eli tehdään samat asiat kuin edellä tilanteessa $n = 8$. Nyt X-Y-yhdistelmiä on jo $21^4 = 194481$. Kun nämä muodostetaan (esim. 200000-riviseen toimituskenttään) komennolla

```
COMB L,CUR+1 / L=LATTICE,4 MIN=1,1,1,1 MAX=21,21,21,21
```

talletetaan tiedostoksi VERKKO21 ja lasketaan etäisyydet D, Survon kuvakaavio

```
-------------------------------------------------------------------
HEADER=Etäisyysjakauma_(n=21);_"Havaintoja"_194481
XSCALE=-0.5:,0(5)40,40.5:     D=-0.5(1)40.5
HOME=0,0 SIZE=1000,600 XDIV=2,7,1
DEVICE=PS,JAK21.PS
HISTO VERKKO21,D
-------------------------------------------------------------------
```

tuottaa seuraavan histogramman:



Havaitaan, että jakauma on oikealle vino. Se ei siis ilmeisesti lähesty normaalijakaumaa luvun n kasvaessa. Tämä on pääteltävissä jo siitäkin, että suuret etäisyydet (lähellä 40) ovat paljon harvinaisempia kuin pienet (lähellä 0).

Ennenkuin esittelen tehokkaamman ratkaisun, yritän näyttää, mitä voidaan sanoa jakauman odotusarvosta yleisesti erikoistapausten $n = 2, 3, 4, 5, 6, 7, 8$ pohjalta. Tämä on samalla pieni havainnollistus siitä, miten Survoa käytetään "laboratoriona" teoreettisemmissa ongelmissa.

Em. laskelmat on tehty noilla n-arvoilla valitsemalla paras tulostustarkkuus ensin komennolla

```
SYSTEM accuracy=16
```

jolloin taulukoituina odotusarvot ovat:

```
-------------------------------------------------------------------
 n E(n) (odotusarvo)                    n^2*E(n)     Erotuskaavio
 2 1.000000000000000(10:suhde)=1/1         4
 3 1.777777777777778(10:suhde)=16/9       16          12
 4 2.500000000000000(10:suhde)=5/2        40          24  12
 5 3.200000000000000(10:suhde)=16/5       80          40  16  4
 6 3.888888888888889(10:suhde)=35/9      140          60  20  4
 7 4.571428571428571(10:suhde)=32/7      224          84  24  4
 8 5.250000000000000(10:suhde)=21/4      336         112  28  4
-------------------------------------------------------------------
```

Nähdään, että odotusarvot vastaavat mitä ilmeisimmin yksinkertaisia rationaalilukuja, jotka on kaivettu esiin Survon suhde-muunnoksella. Kertomalla odotusarvot luvulla n^2 (nämähän ovat todennäköisyyksissä jakajina) saadaan puhtaita kokonaislukuja.
Kun muodostetaan peräkkäisten arvojen erotukset ja näiden erotukset jne., syntyy erotuskaavio, joka kolmannella askeleella päätyy vakiosarakkeeseen.

Tämä viittaa siihen, että luvun n neliöllä kerrottu odotusarvo $E(n)$ voidaan lausua yleisesti kolmannen asteen polynomina. Kyseinen polynomi pystytään johtamaan suoraan erotuskaaviosta, mutta jälleen raaka, survomainen tapa on käyttää polynomiregressiota.

Tehdään siis datataulukko POLY3, jossa muuttujina ovat $Y = n^2E(n)$, n, n^2 ja n^3 ja siitä lineaarinen regressioanalyysi:

```
-------------------------------------------------------------------
DATA POLY3
  Y  N  N2  N3
  4  2   4   8
 16  3   9  27
 40  4  16  64
 80  5  25 125
140  6  36 216
224  7  49 343
336  8  64 512

LINREG POLY3,CUR+1 / VARS=Y(Y),N(X),N2(X),N3(X) RESULTS=0
Linear regression analysis: Data KK, Regressand Y         N=7
Variable Regr.coeff.    Std.dev.    t     beta
N        -0.666667       0.000113  -5902 -0.012
N2       -0.000000       0.000024 -0.000 -0.000
N3        0.666667       0.000002
constant -0.000000       0.000157 -0.000
Variance of regressand Y=15024.00000 df=6
Residual variance=0.000000001 df=3
R=1.0000 R^2=1.0000
-------------------------------------------------------------------
```

"Selitys" on täydellinen ja regressiokertoimista päätellään, että

$$n^2E(n) = 2/3 \cdot (n^3 - n) \quad \text{eli} \quad E(n) = 2/3 \cdot (n-1)(n+1)/n.$$

Suurilla n-arvoilla $E(n) \approx 2n/3$ eli suhteessa verkon sivun pituuteen keskimääräinen etäisyys on asymptoottisesti 2/3.

Herää väistämättä ajatus siitä, että noin yksinkertainen tulos tulisi olla tajuttavissa suoraan. Yhdistämällä kaksi pientä oivallusta se onkin välittömästi todettavissa.

Ensinnäkin - ja tämä on tärkeää koko ongelman "parempien" ratkaisujen kannalta - tarkasteltu etäisyys $D = |X_1 - X_2| + |Y_1 - Y_2|$ on kahden toisistaan riippumattoman ja samaa jakaumaa noudattavan muuttujan $|X_1 - X_2|$ ja $|Y_1 - Y_2|$ summa, sillä ovathan X_1, X_2, Y_1, Y_2 toisistaan riippumattomia ja jopa samoin jakautuneita.

Kun n kasvaa ja tarkastellaan verkkoa kutistettuna niin, että koko sivun pituus on 1, tullaan rajalla jatkuvaan tilanteeseen, jossa yksikköneliöstä valitaan (tasaisen jakauman mukaisesti) kaksi pistettä ja tarkastellaan niiden välistä "city block" -etäisyyttä D. $E(D)$ on tällöin $2E(D_1)$, missä D_1 tarkoittaa etäisyyttä kahden yksikköjanalta, toisistaan riippumatta valitun pisteen X_1 ja X_2 välillä.

Nyt seuraa toinen pieni oivallus: Jos valitaan samalta janalta vielä kolmaskin piste X_3 umpimähkään, on todennäköisyys sille, että se löytyy pisteiden X_1 ja X_2

välistä - valintojen "symmetrian" nojalla - $1/3$ eli pisteiden X_1 ja X_2 välisen janan pituus on keskimäärin kolmasosa koko janan pituudesta. Siis $E(D) = 2/3$.

Koska viimeinen päättely on varsin heuristinen, esitän vielä toisen tavan, joka sekään ei edellytä mitään vankempaa tietoa kuten muuttujan $|X_1 - X_2|$ tiheysfunktion tuntemista ja odotusarvon laskemista integroimalla.

Käytän yleistä ehdollistamisperiaatetta. Olkoon kysytty odotusarvo x. Tarkastellaan pisteiden X_1,X_2 valintaa janan keskipisteen suhteen. Ne tulevat valituiksi eri puolilta keskipistettä todennäköisyydellä $1/2$, jolloin välijanan pituuden odotusarvo on $1/4 + 1/4 = 1/2$. Ne tulevat molemmat valituiksi samalta puolen keskipistettä samoin todennäköisyydellä $1/2$ ja tällöin välijanan pituuden odotusarvo on $x/2$. Odotusarvo x määräytyy siten ehdollisten odotusarvojen todennäköisyyksillä painotettuna summana yhtälöstä $x = 1/2 \cdot 1/2 + 1/2 \cdot x/2$, josta saadaan $x = 1/3$.

Edellä mainostettu tehokkaampi ratkaisu perustuu jo todettuun riippumattomuusominaisuuteen. Koska $D_1 = |X_1 - X_2|$ ja $D_2 = |Y_1 - Y_2|$ ovat riippumattomia ja samoin jakautuneita muuttujia, riittää tarkastella vain näistä toista ja määrätä kysytty etäisyysjakauma muuttujien summan jakaumana.

Jotta saataisiin käsitys etäisyyden $D_1 = |X_1 - X_2|$ käyttäytymisestä, lasketaan aluksi D_1:n jakauma Survolla vastaavalla tavalla kuin edellä etäisyydelle D eli nyt muodostetaan pelkästään mahdollisten X_1, X_2-yhdelmien listaus COMB-operaatiolla (tässä esimerkkinä arvolla $n = 8$):

```
-----------------------------------------------------------------
COMB L,CUR+1 / L=LATTICE,2 MIN=1,1 MAX=8,8
Lattice points in 2 dimensions: N[L]=64
1 1
1 2
1 3
1 4
1 5
1 6
1 7
1 8
2 1
2 2
...
7 7
7 8
8 1
8 2
8 3
8 4
8 5
8 6
8 7
8 8
-----------------------------------------------------------------
```

Annetaan listalle DATA-määrite, nimetään muuttujiksi X_1,X_2, kopioidaan arvot uuteen tiedostoon JANA8 ja lasketaan etäisyydet $D_1 = |X_1 - X_2|$:

```
-----------------------------------------------------------------
FILE COPY JANA TO NEW JANA8
VAR D1:2=abs(X1-X2) TO JANA8

DATA JANA
X1 X2
1 1
1 2
1 3
1 4
1 5
....
-----------------------------------------------------------------
```

Etäisyyksien D_1 frekvenssijakauma muodostetaan STAT-komennolla:

```
-----------------------------------------------------------------
STAT JANA8,CUR+1 / VARS=D1
Basic statistics: JANA8 N=64
Variable: D1       ~abs(X1-X2)
min=0        in obs.#1
max=7        in obs.#8
mean=2.625    stddev=1.914854 skewness=0.486626 kurtosis=-0.696927
autocorrelation=0.6250
lower_Q=1        median=2        upper_Q=4
D1            f     %
       0      8  12.5 ********
       1     14  21.9 **************
       2     12  18.8 ************
       3     10  15.6 **********
       4      8  12.5 ********
       5      6   9.4 ******
       6      4   6.3 ****
       7      2   3.1 **
-----------------------------------------------------------------
```

Frekvenssit laskevat lineaarisesti, poikkeuksena ensimmäinen arvo. Viimeistään kokeilemalla muillakin arvoilla n vakuututaan siitä, että yleisesti etäisyyksien frekvenssijakauma tulee olemaan muotoa

D_1	0	1	2	...	i	...	$n-2$	$n-1$
frekvenssi	n	$2(n-1)$	$2(n-2)$	...	$2(n-i)$	...	4	2

ja vastaavat pistetodennäköisyydet saadaan jakamalla luvulla n^2. Tämä tulos on pääteltävissä tarkasti esim. näin: Kun $D_1 = i > 0$, tämä etäisyys voi syntyä vain yhdistelmillä $(1, i+1)$,$(2, i+2)$, ...,$(n-i, n)$ ja $(i+1, 1)$,$(i+2, 2)$, ...,$(n, n-i)$, joita on yhteensä $2(n-i)$ kpl. Vastaavasti 0-etäisyys esiintyy vain n tavalla.

Kyseinen frekvenssijakauma on helpointa laskea ja tallettaa Survossa matriisitiedostoon:

```
-----------------------------------------------------------------
n=8
MAT F=ZER(n,1)                 / Aluksi F on nollavektori.
MAT RLABELS NUM(0) TO F        / riviotsikoiksi 0,1,...,n-1
MAT TRANSFORM F BY 2*(n+1-I#)  / yleiset frekvenssit
MAT F(1,1)=n                   / 1. frekvenssin "paikkaus"
MAT LOAD F
MATRIX F
T(F_by_2*(n+1-I#))&n
///             1
  0             8
  1            14
  2            12
  3            10
  4             8
  5             6
  6             4
  7             2
-----------------------------------------------------------------
```

D_1:n frekvenssijakauma syntyy siis tällä tyylillä huomattavasti helpommin. Koska D_1 ja D_2 ovat riippumattomia, lopullinen muuttujan $D = D_1 + D_2$ frekvenssijakauma saadaan nyt D_1:n jakauman konvoluutiona itsensä kanssa ja se lasketaan Survossa näin: (kts. myös sivua 15)

```
-----------------------------------------------------------------
MAT C=#CONVOLUTION(F,F)      / *C~CONVOLUTION(F,F) 15*1

MAT LOAD C
MATRIX C
CONVOLUTION(F,F)
///        Convol
C0             64
C1            224
C2            388
C3            496
C4            552
C5            560
C6            524
C7            448
C8            336
C9            224
C10           140
C11            80
C12            40
C13            16
C14             4
-----------------------------------------------------------------
```

Havaitaan luonnollisesti, että päädytään täsmälleen samaan jakaumaan kuin raa'alla tavalla tapauksessa $n = 8$. Laskenta on valtavan paljon nopeampaa ja

vähemmän tilaa vievää, kun ei ole tarvis muodostaa n^4 vaihtoehtoista X-Y-yhdistelmää vaan kaikki tieto mahtuu n alkion (F) ja $2n-1$ alkion (C) vektoreihin.

Esim. seuraavasta Survon laskentakaaviosta nähdään, että arvolla $n = 10000$ tulokset on saatu alle sekunnissa (1.6 GHz:in koneella).

```
-------------------------------------------------------------------
ACCURACY=16

n=10000
MAT F=ZER(n,1)                  / F aluksi nollavektori
MAT RLABELS NUM(0) TO F         / riviotsikoiksi 0,1,...,n-1
MAT TRANSFORM F BY 2*(n+1-I#)   / yleiset frekvenssit
MAT F(1,1)=n                    / 1. frekvenssin "paikkaus"

TIME COUNT START
MAT C=#CONVOLUTION(F,F)      / *C~CONVOLUTION(F,F) 19999*1
TIME COUNT END   0.741

MAT S=SUM(C)       / Summatarkistus
MAT_S(1,1)=10000000000000000
       n^4=10000000000000000
-------------------------------------------------------------------
```

Jos samaa yritettäisiin ensimmäisellä raa'alla tyylillä, X-Y-yhdelmien luettelointi veisi tilaa ainakin noin 80 miljoonaa gigatavua ja laskenta-aika olisi yli 12 vuotta jo pelkästään luetteloinnin osalta. Kun tehokkuudesta on kysymys, kannattaa siis aina hieman miettiä miten laskee. Tässä ongelmassa laskentatapojen eron merkitys ei ole käytännössä niin huomattava, koska asiallista mielenkiintoa käsitellä ongelmaa yksittäisillä, suurilla n-arvoilla ei liene.

3. Yleisiä tarkasteluja

Vaikka tehtävässä ei edellytetty välttämättä mitään yleisten tulosten johtamisia, on paikallaan lopuksi kuitenkin tarkastella D-jakaumaa ainakin odotusarvon, varianssin ja asymptoottisen käyttäytymisen osalta.

Koska $D = D_1 + D_2$ ja D_1 ja D_2 ovat riippumattomia ja samoin jakautuneita, riittää laskea perustunnusluvut muuttujalle $D_1 = |X_1 - X_2|$. D_1-jakauman pistetodennäköisyydet johdettiin jo edellä ja ne ovat

$$p_i = P(D_1 = i) = \begin{cases} 1/n, & i = 0, \\ 2(n-i)/n^2, & i = 1, 2, \ldots, n-1. \end{cases}$$

Tällöin

$$\begin{aligned} n^2 E(D_1) &= \sum_{i=1}^{n-1} i(n-i) = 2(\sum_{i=1}^{n-1} i - \sum_{i=1}^{n-1} i^2) \\ &= 2[n^2(n-1)/2 - (n-1)n(n+1)/6] \\ &= (n-1)n(n+1)/3. \end{aligned}$$

Koska $E(D) = 2E(D_1)$, saadaan

$$E(D) = 2/3\frac{n^2-1}{n}$$

mikä "arvattiin" jo aikaisemmin.

Etäisyyden D varianssi $D^2(D)$ voidaan laskea kaavasta $D^2(D) = E(D^2) - E(D)^2$, jolloin vastaavalla tavalla päädytään lausekkeeseen

$$D^2(D) = 2/3 \cdot (n^2-1)[1/2 - \frac{n^2-1}{3n^2}]$$

eli tästä todetaan, että etäisyyden D keskihajonta on asymptoottisesti $D(D) \approx n/3$.

D-jakauman pistetodennäköisyyksille (joita ei näissä tarkasteluissa ole tarvittu) voidaan johtaa suoraan konvoluutiosummien kautta lausekkeet, jotka sievenevät muotoon

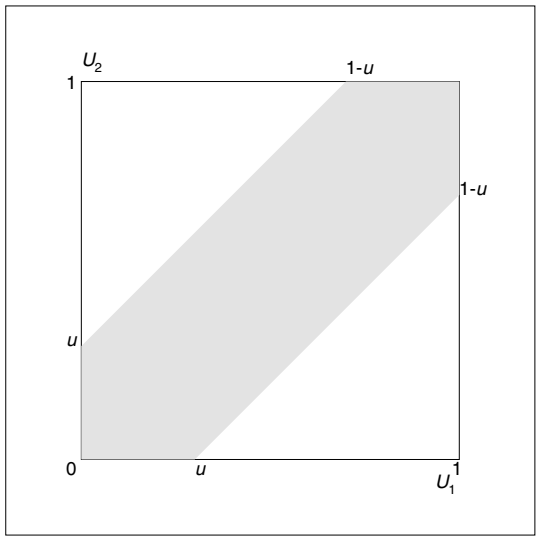
$$q_k = P(D=k) = \begin{cases} 1/n^2, & k = 0, \\ 4/n^4[nk(n-k) + \binom{k+1}{3}], & k = 1, 2, \ldots, n-1, \\ 4/n^4\binom{2n-k+1}{3}, & k = n, n+1, \ldots, 2n-2. \end{cases}$$

Varsinkin viimeinen tapaus (suuret etäisyydet) houkutelkoon suoriin kombinatoorisiin päätelmiin!

Katsotaan vielä, millainen on etäisyyden D rajajakauma parametrin n lähetessä ääretöntä. Tätä jakaumaa on helpointa tutkia käyttäen sitä edellä todettua seikkaa, että normeerattaessa verkon koko sivunpituus ykköseksi rajajakauma saadaan tarkastelemalla kahden yksikköneliöstä valitun satunnaispisteen etäisyyden Z jakaumaa "city-block"-metriikan mukaisesti.

Muuttuja Z on vuorostaan esitettävissä kahden riippumattoman muuttujan summana $Z = Z_1 + Z_2$, missä kumpikin vastaa yksikköjanalta satunnaisesti valittujen pisteiden etäisyyttä.

Koska muuttuja $U = Z1$ on muotoa $U = |U_1 - U_2|$, missä U_1 ja U_2 noudattavat toisistaan riippumatta tasaista jakaumaa välillä $(0, 1)$. Muuttujan U jakaumaan pääsee kiinni ehkä parhaiten kertymäfunktion kautta katsomalla tilannetta geometrisesti U_1, U_2-koordinaatistossa



U:n kertymäfunktiota arvolla u vastaa kuvassa harmaan alueen (rajoina suorat $u_1 - u_2 = u$ ja $u_1 - u_2 = -u$) pinta-ala. Se saadaan vähentämällä neliön pinta-alasta 1 alueen ulkopuolelle jääneiden kolmioiden yhteinen pinta-ala $(1-u)^2$ eli

$$P(U \leq u) = F_U(u) = 1 - (1-u^2), \quad 0 <= u <= 1.$$

Tiheysfunktio on tällöin

$$F_U^{'}(u) = f_U(u) = 2(1-u), \quad 0 \leq u \leq 1.$$

Summan $Z = Z_1 + Z_2$ tiheysfunktio syntyy konvoluutiona

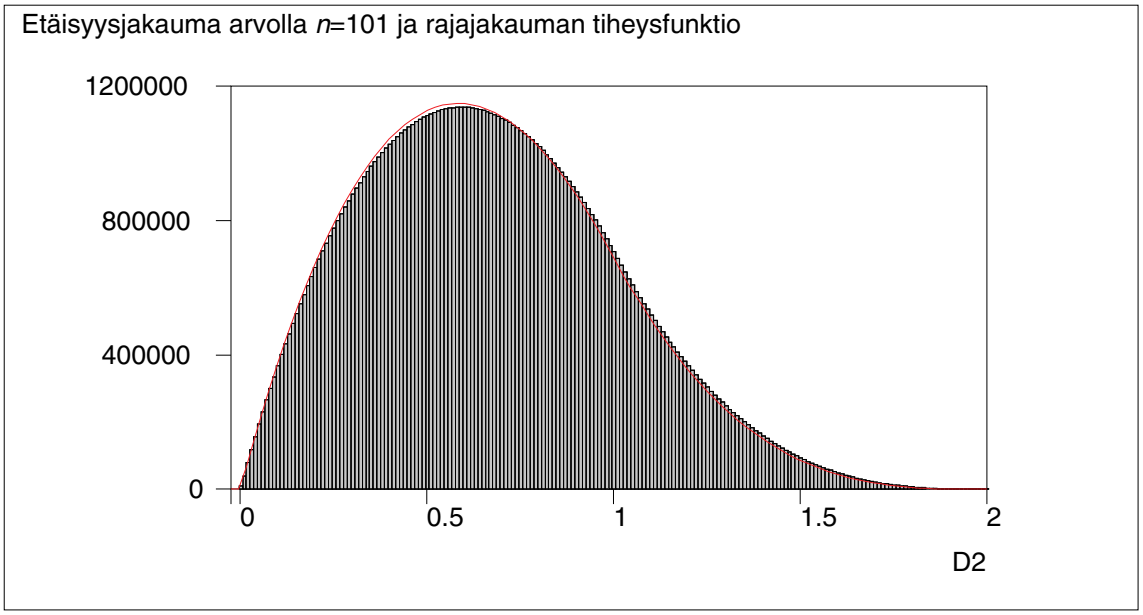
$$f_Z(z) = \int_{-\infty}^{\infty} f_{Z_2}(z-x) f_{Z_1}(x) dx, \quad 0 \leq z \leq 2.$$

Yksinkertaisella integroinnilla nähdään, että

$$f_Z(z) = \begin{cases} \int_0^z 4(1-z+x)(1-x)dx = z(z^2-6z+6), & 0 \leq z \leq 1, \\ \\ \int_{z-1}^1 4(1-z+x)(1-x)dx = (2-z)^3, & 1 \leq z \leq 2. \end{cases}$$

Analyyttinen esitys vaihtuu siististi pisteessä $z = 1$ niin, että molemmat funktiot samoinkuin niiden ensimmäisen, toisen ja kolmannen kertaluvun derivaatat yhtyvät. Kyseessä on siis "sileä" 3. asteen funktio koko vaihteluvälillä $(0, 2)$.

Kuvasta nähdään, miten lähellä D_2-jakauma on tätä rajajakaumaa, kun $n = 101$.



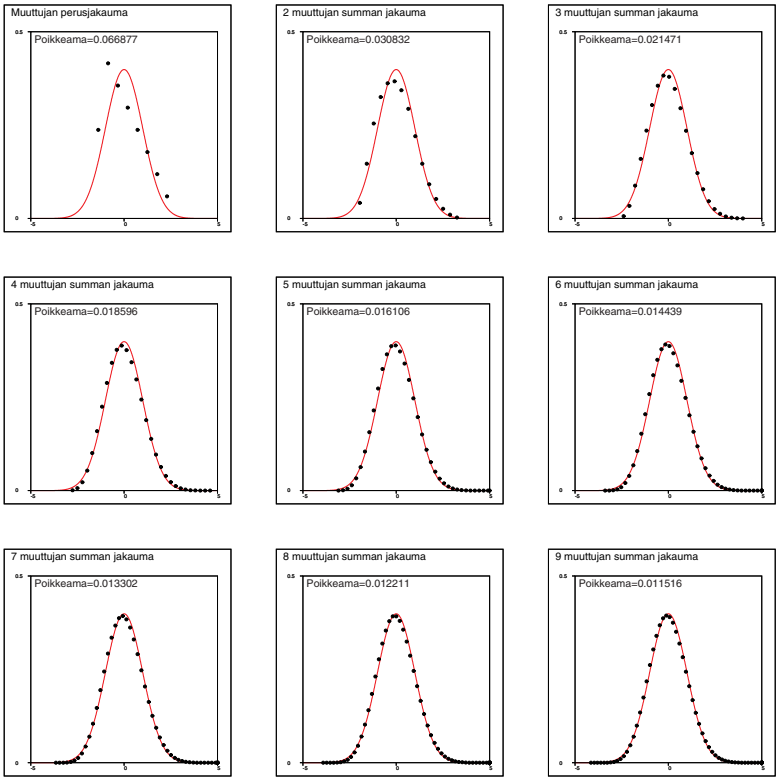
Rajajakauma ei siis ole lähelläkään normaalijakaumaa. Jos tehtävää yleistettäisiin niin, että siirryttäisiin tutkimaan vastaavaa etäisyyttä N-ulotteisessa hilassa, on selvää, että silloin dimension N kasvaessa (vaikka n olisi pienikin) jakauma lähestyy normaalijakaumaa todennäköisyyslaskennan keskeisen raja-arvolauseen mukaisesti.

Tätä saattaa helposti havainnollistaa Survossa käyttämällä sukroa /SUMMAJAK, joka muodostaa samaa diskreettiä todennäköisyysjakaumaa noudattavien, riippumattomien muuttujien summajakaumien pistetodennäköisyydet ja havainnollistaa summan lähentymistä kohti normaalijakaumaa sekä kuvallisesti että numeerisesti. Tutkittavan jakauman pistetodennäköisyydet annetaan matriisitiedostoksi talletettuna vektorina.

Jos siis esim. käynnistetään

```
/SUMMAJAK F  / NMIN=1 NMAX=9 PS=F8_
```

missä F on edellä laskettu 8 pisteen, yksiulotteisen hilan etäisyyksien frekvenssijakauma (matriisitiedostona F), /SUMMAJAK normeeraa sen automaattisesti todennäköisyyksiksi ja laskee konvoluutioina summajakamat arvoilla $N = 1, 2, \ldots, 9$ (eli 9-ulotteiseen "kuutioon" asti). Se näyttää kuvat peräkanaa ja täsmennyksen PS=F8_ ansiosta tallettaa nuo 9 kuvaa myös PostScript-tiedostoiksi F8_1.PS - F8_9.PS. Lisäksi pohjakuvana luodaan F8_0.PS, jossa on rajatilannetta vastaavan normaalijakauman tiheysfunktio. Näistä PostScript-tiedostoista on koostettu erikseen (EPS JOIN) seuraava kuvasarja



josta näkyy selvästi, miten keskeinen raja-arvolause tekee tehtävänsä. Kuvissa näkyvä poikkeama-arvo (viimeisessä 0.007914) on laskettu Kolmogorov-Smirnov-testisuureen kaltaisesti.

Edellä olevan kuvakoosteen tekotapa löytyy Survo-sivuilta osoitteesta
`http://www.survo.fi/galleria/072.html`

4. Pistetodennäköisyydet

Etäisyysjakauman yleiset pistetodennäköisyydet (sivulla 12) voidaan laskea konvoluutiosummina (kokonaistodennäköisyyksinä) eli muodossa

$$q_k = P(D = k) = \sum_{i=0}^{k} p_i p_{k-i}, \quad k = 0, 1, 2, \ldots, 2n-2,$$

missä (kuten todettiin sivulla 11)

$$p_i = \begin{cases} 1/n, & i = 0, \\ 2(n-i)/n^2, & i = 1, 2, \ldots, n-1. \end{cases}$$

Todennäköisyyksiä q_k vastaavat frekvenssit laskettiin edellä kullakin yksittäisellä n-arvolla Survon matriisikomennolla

```
MAT C=#CONVOLUTION(F,F)
```

Näillä tiedoin sivulla 12 mainitut q_k-todennäköisyydet voidaan johtaa yleisesti sijoittamalla p_i-todennäköisyyksien lausekkeet em. konvoluutiosummaan. Laskentaa hiukan häiritsee se, että p_0 on eri muotoa kuin muut p_i:t. Kun $k > n-1$,

p_0 ei kuitenkaan voi esiintyä konvoluutiosummassa. Tästä johtuu, että q_k:lle tulee erilainen lauseke riippuen siitä, onko $k = 0$ tai $k = 1, 2, \ldots, n-1$ tai $k = n, n+1, \ldots, 2n-2$. Vastaava kahtiajako tuli esiin jatkuvassa rajajakaumatarkastelussa sivulla 13.

Kun asiaa tutkitaan frekvenssien tasolla (p_i:t kerrottuna n^2:lla), havaitaan, että todennäköisyyksiä q_k vastaavat frekvenssit tulevat olemaan parametrien n ja k suhteen polynomeja, joilla korkein asteluku on 3.

Tällöin pääsee vähimmin pohdiskeluin lopputulokseen "mallintamalla" kysytyn frekvenssin $F(n,k)$ lausekkeen muodossa

$$F(n,k) = \sum_{i=0}^{3} \sum_{j=0}^{3} c_{ij} n^i k^j$$

ja estimoimalla parametrit c_{ij} erikoistapauksista saadun aineiston avulla käyttäen lineaarista regressioanalyysia.

Tarkastellaan esimerkkinä (ehkä hankalinta) tapausta $k = 1, 2, \ldots, n-1$, jolloin arvoilla $n = 4, 5, 6, 7, 8, 9$ on koottu seuraava NKF-aineisto:

```
-------------------------------------------------------------------
DATA NKF:(n,k,F) 4,1,48 4,2,68 4,3,64 5,1,80 5,2,124 5,3,136 5,4,120
6,1,120 6,2,196 6,3,232 6,4,232 6,5,200 7,1,168 7,2,284 7,3,352 7,4,376
7,5,360 7,6,308 8,1,224 8,2,388 8,3,496 8,4,552 8,5,560 8,6,524 8,7,448
9,1,288 9,2,508 9,3,664 9,4,760 9,5,800 9,6,788 9,7,728 9,8,624 END
-------------------------------------------------------------------
```

Aineisto on muodostettu sivulla 10 näkyvällä tekniikalla.

On kätevintä siirtää aineisto Survon datatiedostoksi (NKF2), varata siihen lisätilaa uusia muuttujia varten ja muodostaa ne (uudella) POWERS-komennolla:

```
-------------------------------------------------------------------
FILE COPY NKF TO NEW NKF2
FILE EXPAND NKF2,14,7
.....................................................
POWERS NKF2  / POW_VARS=n,k DEGREE=3 TYPE=2
Näin syntyneet lisämuuttujat näkyvät seuraavasti:
FILE STATUS NKF2
 Copy of data list NKF
FIELDS: (active)
   1 NA_   1 n        (#)
   2 NA_   1 k        (#)
   3 NA_   2 F        (###)
   4 NA-   2 n2       ~n^2
   5 NA-   2 n1k1     ~n*k
   6 NA-   2 k2       ~k^2
   7 NA-   2 n3       ~n^3
   8 NA-   2 n2k1     ~n^2*k
   9 NA-   2 n1k2     ~n*k^2
  10 NA-   2 k3       ~k^3
END
Survo data file NKF2: record=32 bytes, M1=21 L=64  M=10 N=33
-------------------------------------------------------------------
```

Lineaarinen regressioanalyysi antaa tuloksen:

```
-------------------------------------------------------------------------
MASK=XXYXXXXXXX / Muuttujien valinta malliin
LINREG NKF2,CUR+1 / RESULTS=0
Linear regression analysis: Data NKF2, Regressand F         N=33
Variable Regr.coeff.    Std.dev.    t     beta
n        -0.000000       0.000000
k        -0.666667       0.000000
n2        0.000000       0.000000
n1k1     -0.000000       0.000000
k2        0.000000       0.000000
n3       -0.000000       0.000000
n2k1      4.000000       0.000000
n1k2     -4.000000       0.000000
k3        0.666667       0.000000
constant  0.000000       0.000000
Variance of regressand F=53475.09091 df=32
Residual variance=0.000000000 df=23
R=1.0000 R^2=1.0000
-------------------------------------------------------------------------
```

Odotuksien mukaisesti syntyy "täydellinen" selitys ja regressiokertoimista on heti luettavissa, että

$$F(n,k) = 4(n^2k - nk^2) + 2/3(k^3 - k) = 4[nk(n-k) + \binom{k+1}{3}],$$

kun siis $k = 1, 2, \ldots, n-1$.

Totesin jo sivulla 12, että "suurten" etäisyyksien todennäköisyyden lausekkeen

$$q_k = 4/n^4 \binom{2n-k+1}{3}, \quad k = n, n+1, \ldots, 2n-2$$

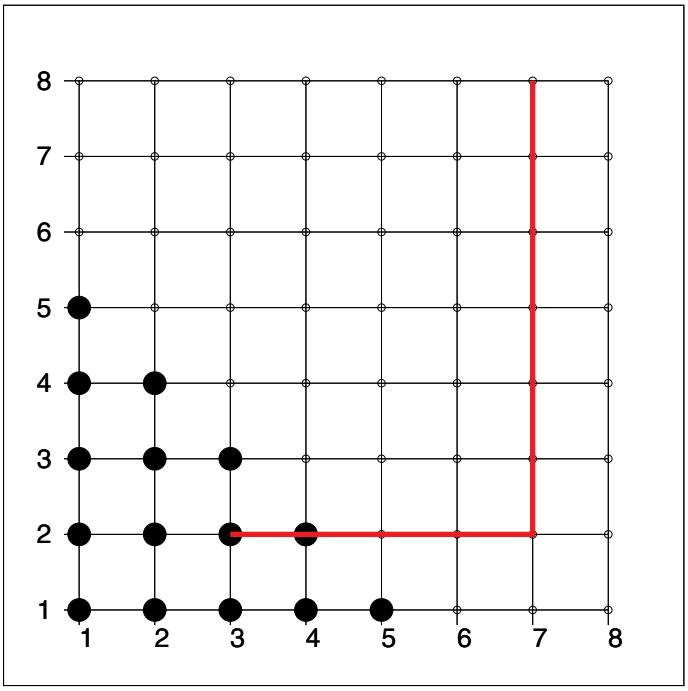
yksinkertaisuus houkuttelee kombinatoorisiin tulkintoihin.

Koska jokainen yli $n-1$-mittainen reitti kulkee ikäänkuin kulmasta vastakkaiseen kulmaan (ja näitä perusvaihtoehtoja on 4), sellaisten k-mittaisten reittien (tyyppiä ensin oikealle, sitten ylös), jotka lähtevät verkon vasemmasta alaosasta oikealle yläosaan, on edellisen perusteella

$$Q(n,k) = \binom{2n-k+1}{3}.$$

En tunne tälle lausekkeelle suoraa kombinatoorista selitystä. Seuraava tulkinta on kuitenkin sitä lähellä. Binomikertoimien tunnetusta palautuskaavasta $\binom{n}{m} = \binom{n-1}{m} + \binom{n-1}{m-1}$ seuraa tässä tapauksessa

$$Q(n,k) = Q(n,k+1) + \binom{2n-k}{2}.$$



Kuva esittää tilannetta $n = 8$, $k = 10$. Ylläoleva $Q(n,k)$:n lauseke kertoo, että mahdolliset reitit muodostuvat etäisyyttä $k+1$ vastaavista reiteistä (lukumäärä $Q(n,k+1)$), joita on lyhennetty lopusta yhdellä yksiköllä ja lisäksi reiteistä (joista mallina kuvaan punaisella piirretty), jotka törmäävät verkon yläreunaan ja joilla ei ole vastinetta edellä mainittujen reittien joukossa. Näitä törmäysreittejä on vain yksi kutakin kuvan vasemman alalaidan kolmiossa olevaa (pulleaa) pistettä kohti eli yhteensä kolmioluvun $\binom{2n-k}{2}$ (kuvassa $\binom{6}{2} = 15$) mukainen määrä.