Ohjeita SAS-ohjelmiston käyttöön

Kokoelma @CSC- ja SuperMenu-lehdissä
ilmestyneitä artikkeleita

Esa Lammi

E-mail: Esa.Lammi@csc.fi

CSC – Tieteellinen laskenta Oy

Versio 1.00 (23.2.2001)

Saatteeksi

Tämä kirjanen on tarkoitettu johdatukseksi SAS-ohjelmiston käyttöön. Opas ei sisällä kattavaa kuvausta mistään aihepiiristä, vaan sen tarkoitus on auttaa uutta käyttäjää alkuun ja ohjata hänet yksityiskohtaisemman tiedon lähteille. Tähän kirjaseen kootut artikkelit käsittelevät SAS:n käytön perusteita sekä mm. grafiikkaa, tilastollista analyysiä ja numeerisia menetelmiä.

Opas löytyy PDF-muodossa www-osoitteesta `http://www.csc.fi/oppaat/sas/`.

Sisältö

Lisätietoja

SAS-ohjelmiston käytöstä saa lisätietoja CSC:n koneiden komennolla `help sas` tai www-osoitteesta
`http://www.csc.fi/cschelp/sovellukset/stat/sas/`

Tekijänoikeudet

Tämän teoksen tekijänoikeudet kuuluvat CSC – Tieteellinen laskenta Oy:lle. Teoksen tai osia siitä voi kopioida ja tulostaa vapaasti henkilökohtaiseen käyttöön sekä Suomen yliopistojen ja korkeakoulujen kurssikäyttöön edellyttäen, että kopioon tai tulosteeseen liitetään tämä ilmoitus teoksen tekijästä ja tekijänoikeuksista. Teosta ei saa myydä tai sisällyttää osaksi muita teoksia ilman CSC:n lupaa.

SAS-ohjelmiston perusteita

SAS on tilastollinen ohjelmisto, joka käsittää useita eri analysointivälineitä. Perinteisesti tilastolliset proseduurit ovat sisältyneet STAT-pakettiin. Tämän lisäksi aikasarjoja käsittelevässä ETS-tuotteessa on tilastollisia proseduureja. Perus-SAS eli SAS/Base sisältää myös analysointivälineitä. Uudempina SASin analysointituotteina ovat Insight ja LAB. Myös laadunvalvontaan tarkoitettu tuote QC sisältää tilastollisen analyysin välineitä.

SAS-ohjelmiston kutsu

Ohjelmistoa kutsutaan X-ympäristössä komennolla

```
sas
```

Tällöin avautuu kolme ikkunaa:

- PROGRAM EDITOR -ikkuna komentojen antamista varten
- LOG-ikkuna lokitiedostoja varten
- OUTPUT-ikkuna tulostusta varten

Komennot kirjoitetaan PROGRAM EDITOR -ikkunan komentoriville. Jokainen komento päättyy aina puolipisteeseen. Suoritukseen komennot lähetetään ikkunan yläreunassa olevan Locals-painikkeen alta löytyvällä Submit-komennolla. Ohjelmistosta poistutaan valitsemalla hiirellä File-painikkeen alta kohta Exit.

LOG-ikkunaan tulostuu kaikki SAS-ajon suoritusta koskevat tiedot. Mikäli kaikki sujuu moitteettomasti, niin proseduurien käyttämät CPU-ajat näkyvät lokitiedostossa. Mahdolliset virheilmoitukset kirjataan myös lokitiedostoon.

OUTPUT-ikkunaan tulostuu ajon tuottamat tulokset. Mikäli SAS-ajossa käytetään grafiikkaa, niin se tulostuu erilliseen grafiikkaikkunaan. Tulostusikkunan sisältö on mahdollista tallentaa tiedostoon. Se tehdään valitsemalla sivun ylälaidan FILE-painikkeesta kohdan Save tai Save as.

SAS-ajovirta tiedostossa

SAS-ohjelmistoa on mahdollista käyttää myös siten, että koko SAS-ajovirta kirjoitetaan tiedostoon. Tiedoston loppu tulee olla muotoa `.sas`, esimerkiksi `tied.sas`. Tällöin SAS-ajo suoritetaan käyttöjärjestelmätasolla antamalla komento

sas tied

Lokitiedosto on tällöin tiedostossa `tied.log` ja tulostustiedosto tiedostossa `tied.lst`.

SAS-ohjelmiston käyttö ASSISTin avulla

SAS/ASSIST on valikkopohjainen toimintaympäristö, joka mahdollistaa SASin käytön ilman ohjelmiston tuntemista. Valikosta tehtyjen valintojen myötä SAS-ohjelmisto generoi valintoja vastaavan SAS-koodin, joka sitten suoritetaan.

ASSIST on nopea keino tutustua itse SAS-ohjelmistoon. Se sopii yhtä hyvin aloittelijalle kuin kokeneellekin SAS-käyttäjälle. ASSISTin generoima SAS-koodi on mahdollista tallentaa. Sitä voidaan editoida vastaamaan haluttua päämäärää. Tällöin syntynyttä SAS-tiedostoa voidaan ajaa ilman SAS/ASSIST-tuotetta juuri niin kuin edellä olevassa kohdassa `SAS-ajovirtatiedostossa` mainitaan.

SAS/ASSISTia käynnistettäessä on ensin kutsuttava itse SAS-ohjelmistoa aiemmin mainitulla `sas`-komennolla. PROGRAM EDITOR -ikkunan yläreunassa olevasta Globals-painikkeesta valitaan kohta SAS/ASSIST. Tällöin avautuu SAS/ASSISTin päävalikko, josta valintojen teko alkaa.

SAS-tiedoston muodostaminen

Tarkastellaan aluksi tilannetta, jossa SAS-tiedoston syötteet annetaan vasta itse ajossa. Tämä tapahtuu SASin Data-vaiheen avulla. Komento data aloittaa data-vaiheen.

```
libname lib '$HOME/alihak';
data lib.datat;
   input a b c;
cards;
1 2 3
2 3 4
3 4 5
4 5 6
5 6 7
;
```

Kaikki SASin komennot päättyvät puolipisteeseen. Rivi `lib.datat;` mudostaa nyt pysyvän sas-tiedoston nimeltään `datat` hakemistoviitteen lib osoittamaan paikkaan eli tässä tapauksessa kotihakemiston alla olevaan hakemistoon `alihak`. Käyttöjärjestelmätasolla tämä sas-tiedosto näkyy nimellä `datat.ssd01`. Mikäli tiedostonimi on yksiosainen, tässä tapauksessa siis pelkkä `datat`, niin silloin syntynyt sas-tiedosto on tilapäistiedosto, joka tuhoutuu ajon loputtua.

Komennolla input luetellaan muuttujat. Ne voivat olla vapaan formaatin mukaisia kuten edellä olevassa esimerkissä, pohjautua sarakesidonnaisuuteen tai olla format-määrittelyllä tehtyjä. Vapaan formaatin tapauksessa tyhjä toimii havaintojen erotinmerkkinä. Mikäli muuttujan perässä on $-merkki tyhjällä erotettuna, on kyseessä merkkimuotoinen muuttuja. Muutoin muuttuja on numeerinen.

Varsinaisten havaintoaineistojen syöttäminen aloitetaan `cards`-lauseella. Jokainen tietue kirjoitetaan omalle rivilleen tyhjällä erotettuna.

Mikäli havaintoaineisto on ASCII-tiedostona, on se aluksi saatettava SAS-tiedostoksi. Ennen SA-Sin datavaihetta on määriteltävä ASCII-tiedosto `filename`-lauseella

```
filename in '$HOME/alihak/tied.dat';
```

Tällöin äskeinen data-vaihe muuttuu seuraavaan muotoon:

```
libname lib '$HOME/alihak';
filename in '$HOME/alihak/tied.dat';
data lib.datat;
   infile in;
   input a b c;
```

ASCII-tiedosto ilmaistaan `filename`-lauseen tiedostoviitteellä, joka tässä tapauksessa on `in`. Tämän tiedostoviitteen nimen voi valita vapaasti. Valittu tiedostoviite tuodaan data-vaiheeseen `infile`-lauseella.

SAS-ohjelmisto luo automaattisesti jokaiselle käyttäjälle ensimmäisellä käyttökerralla kotihakemiston alle `sasuser`-nimisen alihakemiston. Tätä voi hyödyntää SAS-ajossa tallentamalla sinne pysyviä tiedostoja. Tällöin `libname`-lausetta ei tarvita hakemiston määrittelyyn, koska SAS tuntee sen oletuksena. Täten edellä mainittu data-vaihe muuttuu seuraavasti oletushakemistoa `sasuser` käytettäessä:

```
filename in '$HOME/alihak/tied.dat';
data sasuser.datat;
   infile in;
   input a b c;
```

Haluttaessa tarkistaa syntyneen sas-tiedoston oikeellisuus voidaan se tehdä `proc print` -proseduuria käyttäen. Tällöin alkuperäisen SAS-tiedoston sisältö saadaan tulostettua.

```
options ls=77 nodate nonumber;
libname lib '$HOME/alihak';
data lib.datat;
   input a b c;
cards;
1 2 3
2 3 4
3 4 5
4 5 6
5 6 7
;
proc print data=lib.datat;
title 'Havaintoaineiston tulostus';
run;
```

Alkuun lisätyllä `options`-lauseella säädellään tulostuksen rivinpituudeksi 77 merkkiä. Samalla halutaan, että päivämäärää eikä sivunumeroa tulosteta. Proseduurin `proc print` jälkeen on lisätty `title` määrittely, jolla saadaan tulostettua otsikko. Mikäli `title`-määrittely jätetään pois, tulee tulostuksen otsikoksi oletuksena `The Sas System`. Komento `run` käynnistää suorituksen välittömästi.

Tulostus näyttää nyt seuraavanlaiselta:

```
Havaintoaineiston tulostus

    OBS    A    B    C

     1     1    2    3
     2     2    3    4
     3     3    4    5
     4     4    5    6
     5     5    6    7
```

Tulostuksessa OBS numeroi havainnot. Mikäli OBS-sarake halutaan jättää pois, niin se onnistuu `print`-proseduurin alimääreellä `id`. Tällöin `id`:llä määritelty muuttuja tulee tulostuksen ensimmäiseksi sarakkeeksi korvaten OBS-sarakkeen.

SAS-tiedoston muokkaus

Syntynyttä sas-tiedostoa voi muokata data-vaiheessa hyvin monella tavalla. Muuttujien arvoja voi muuttaa. Uusia muuttujia voidaan ottaa mukaan. Havainnoista voidaan poimia mukaan haluttu osajoukko.

```
options ls=77 nodate nonumber;
libname lib '$HOME/alihak';
data aa;
   set lib.datat (firstobs=2 obs=3);
   d=sum(a,b,c);
proc print data=a;
title 'Havaintoaineiston tulostus';
run;
```

Tässä ajovirrassa luodaan uusi sas-tiedosto `aa`. Koska tämän tiedoston nimi on yksiosainen, se on tilapäistiedosto. Tätä tiedostoa hyödynnetään siis kertaluontoisesti vain tässä nimenomaisessa ajossa.

Alilauseella `set` kopioidaan sas-tiedoston `lib.datat` sisältö tiedostoon `aa` siten, että mukaan otetaan toisesta havainnosta alkaen kolme seuraavaa havaintoa. Samalla muodostetaan uusi muuttuja `d`, joka on muiden muuttujien summa. Tulokseksena saadaan seuraava listaus:

```
Havaintoaineiston tulostus

 OBS    A    B    C     D

  1     2    3    4     9
  2     3    4    5    12
```

Formaattien käyttö

Format-määrittelyllä voidaan antaa muuttujien arvoille havainnollisempia esityksiä. Tarkastellaan aiemmin luotua sas-tiedostoa `datat`, johon siis viitattiin nimellä `lib.datat`. Muuttujien arvot vaihtelivat yhden ja seitsemän välillä. Nyt halutaan tulostaa tämä sas-tiedosto siten, että muuttujien arvot edustavat viikonpäiviä: ykkönen on maanantai ja suurin luku seitsemän esittää sunnuntaita.

```
options ls=77 nodate nonumber;
libname lib '$HOME/alihak';
libname library '$HOME/alihak';
proc format library=library;
   value pv 1='Maanantai'
            2='Tiistai'
            3='Keskiviikko'
            4='Torstai'
            5='Perjantai'
            6='Lauantai'
            7='Sunnuntai';
proc print data=lib.datat;
   id a;
   format a pv. b pv. c pv.;
title 'Havaintoaineiston tulostus';
run;
```

Format-määrittelyt tehtiin tässä tapauksessa pysyviksi. Lauseen `libname` yhteydessä on käytettävä `library`-hakemistoviitettä. Kun formaatti liitetään muuttujaan, niin formaatin nimeä seuraa piste. Tämä menettely pätee numeerisille muuttujille kuten tässä. Jos muuttuja on merkkimuotoinen, niin siihen liittyvän formaatin nimen eteen tulee $-merkki.

Tuloksena saadaan viikonpäivien lista:

```
        Havaintoaineiston tulostus

A              B              C

Maanantai      Tiistai        Keskiviikko
Tiistai        Keskiviikko    Torstai
Keskiviikko    Torstai        Perjantai
Torstai        Perjantai      Lauantai
Perjantai      Lauantai       Sunnuntai
```

Perustunnusluvut

Perustunnusluvut antavat tutkittavasta aineistosta ennakkokäsityksen, jonka perusteella voidaan aloittaa tarkempi analysointi. Laajasti näitä tunnuslukuja antavat perus-SASin proseduurit *means* ja *univariate*. Ristiintaulukointia on mahdollista suorittaa *freq*-proseduuria hyväksikäyttäen.

Means-proseduurin perustunnusluvut

Tarkastellaan *means*-proseduurin tuottamia tunnuslukuja esimerkin avulla. Havaintoaineisto käsittää kolmenlaisia kolestroliarvojen mittaustuloksia: kokonaiskolestroli, ns. hyvä kolestroli ja ns. huono kolestroli.

Havaintoaineisto

TOTAL	LDL	HDL
11.18	9.71	1.02
7.51	5.94	1.29
7.62	6.38	0.97
7.98	6.70	0.97
10.20	8.69	1.26
8.36	6.58	1.25
7.54	6.18	1.17
4.59	2.55	1.82
7.10	4.97	1.60
8.13	6.88	0.83
5.72	4.02	1.41
5.93	4.44	1.18
8.08	6.65	1.14
8.25	6.89	0.91
6.78	5.09	1.16
7.85	6.16	1.15
9.42	7.62	1.08
6.32	4.30	1.67
10.00	8.24	1.10
11.96	10.48	0.89
12.20	10.44	0.64
8.37	6.63	0.88
8.52	6.97	1.14
7.61	5.83	1.10
11.80	9.74	1.41
13.25	11.65	0.98
10.20	8.37	1.41
8.97	7.64	0.88
7.83	5.83	1.21
8.20	6.78	1.18
8.23	5.50	1.77
11.29	9.74	1.05
8.90	7.33	1.25
7.82	6.46	0.85
8.08	6.49	1.21
8.46	6.94	1.10
13.60	12.20	0.79
7.87	5.39	1.03
7.22	4.87	2.08
8.50	6.34	1.26
13.20	10.94	1.54
9.26	6.64	1.04
8.00	6.20	0.97
11.00	8.57	1.82
9.82	8.37	0.89
10.48	8.11	1.49
12.00	9.98	1.48
12.60	10.24	1.41

Oletuksena *means*-proseduuri tulostaa jokaisesta muuttujasta havaintojen lukumäärän, keskiarvon, keskihajonnan, mimimin ja maksimin. Haluttaessa enemmän tunnuslukuja on kaikkiin tunnuslukuihin liittyvä avainsana mainittava proseduurin yhteydessä.

Tunnusluvut saadaan seuraavalla ajovirralla:

```
options linesize=77 nodate nonumber;
filename in  '$HOME/sas/test/testi.dat';
libname lib '$HOME/sas/test/';
data lib.kol;
   infile in;
   input total  ldl  hdl;
proc means data=lib.kol n mean std stderr
           sum var min max cv range nmiss
           skewness kurtosis;
title 'Means-proseduurin tulostus';
run;
```

Lauseella *options* määritellään rivin pituus ja estetään päivämäärän sekä rivinumeron tulostuminen. *Filename*-määrittely yksilöi alkuperäisen ASCII-tiedoston *testi.dat*. *Libname* määrittelee hakemiston, mihin muodostettava SAS-tiedosto halutaan sijoittaa. Rivi *data lib.kol* muodostaa pysyvän SAS-tiedoston nimeltään *kol.ssd01 libname*-lauseessa mainittuun hakemistoon. *Infile* kiinnittää aiemmin määritellyn ASCII-tiedoston lukemista varten. Muuttujat nimetään *input*-lauseella. *Proc means* -lauseessa luetellaan kaikki tunnusluvut, jotka halutaan tulostettaviksi. Ne ovat seuraavat:

- n = havaintojen lukumäärä
- mean = keskiarvo
- std = keskihajonta
- stderr = keskiarvon keskivirhe
- sum = havaintojen summa
- var = varianssi
- max = havaintojen maksimiarvo
- min = havaintojen minimiarvo
- cv = variaatiokerroin
- range = vaihteluväli
- nmiss = puuttuvien havaintojen lukumäärä
- skewness = vinous
- kurtosis = huipukkuus

Otsikko saadaan aikaiseksi *title*-lauseella. Mikäli tätä lauseketta ei anneta, tulostuu otsikoksi oletuksena *The SAS System*.

Tuloksena saadaan seuraava listaus:

```
                Means-proseduurin tulostus

Variable  N       Mean    Std Dev  Std Error          Sum
-----------------------------------------------------------
TOTAL    48  9.0375000  2.0962190  0.3025631  433.8000000
LDL      48  7.2637500  2.0983080  0.3028647  348.6600000
HDL      48  1.2027083  0.3031658  0.0437582   57.7300000
-----------------------------------------------------------

 Variable   Variance    Minimum     Maximum          CV
 ------------------------------------------------------
 TOTAL     4.3941340  4.5900000  13.6000000  23.1946776
 LDL       4.4028963  2.5500000  12.2000000  28.8873923
 HDL       0.0919095  0.6400000   2.0800000  25.2069299
 ------------------------------------------------------

   Variable      Range  Nmiss   Skewness    Kurtosis
   ---------------------------------------------------
   TOTAL     9.0100000      0  0.4839469  -0.2799828
   LDL       9.6500000      0  0.4061463  -0.1094778
   HDL       1.4400000      0  0.8301119   0.5743746
   ---------------------------------------------------
```

Univariate-proseduurin tulostus

Proc univariate -proseduuri tuottaa automaattisesti ilman lisämäärittelyjä huomattavan määrän tunnuslukuja. Tarkastellaan edellä luotua SAS-tiedostoa siten, että muuttujaksi valitaan ainoastaan *total*. Tällöin SAS-ajovirta on seuraava:

```
options linesize=77 nodate nonumber;
libname lib '$HOME/sas/test/';
proc univariate data=lib.kol freq;
   var total ;
title 'Univariate-proseduurin tulostus';
run;
```

Freq-määrittely *univariate*-proseduurin yhteydessä aikaansaa frekvenssitaulun tulostamisen. Tulostus näyttää kokonaisuudessaan seuraavalta:

```
        Univariate-proseduurin tulostus

             Univariate Procedure

Variable=TOTAL

                  Moments

N                  48  Sum Wgts          48
Mean           9.0375  Sum              433.8
Std Dev     2.096219  Variance      4.394134
Skewness    0.483947  Kurtosis      -0.27998
USS         4126.992  CSS           206.5243
CV          23.19468  Std Mean      0.302563
T:Mean=0     29.8698  Pr>|T|          0.0001
Num ^= 0          48  Num > 0            48
M(Sign)           24  Pr>=|M|         0.0001
Sgn Rank         588  Pr>=|S|         0.0001

               Quantiles(Def=5)

100% Max       13.6       99%       13.6
 75% Q3       10.34       95%       13.2
 50% Med      8.365       90%       12.2
 25% Q1       7.825       10%       6.78
 0% Min        4.59        5%       5.93
 1%            4.59
   Range          9.01
    Q3-Q1         2.515
   Mode           8.08

              Extremes

Lowest    Obs     Highest    Obs
4.59(       8)     12.2(      21)
5.72(      11)     12.6(      48)
5.93(      12)     13.2(      41)
6.32(      18)     13.25(     26)
6.78(      15)     13.6(      37)

        Univariate-proseduurin tulostus

             Univariate Procedure

Variable=TOTAL

               Frequency Table

                Percents                    Percents
Value Count  Cell   Cum  Value Count  Cell   Cum
4.59     1   2.1   2.1   8.37     1   2.1  52.1
5.72     1   2.1   4.2   8.46     1   2.1  54.2
5.93     1   2.1   6.3    8.5     1   2.1  56.3
6.32     1   2.1   8.3   8.52     1   2.1  58.3
6.78     1   2.1  10.4    8.9     1   2.1  60.4
 7.1     1   2.1  12.5   8.97     1   2.1  62.5
7.22     1   2.1  14.6   9.26     1   2.1  64.6
7.51     1   2.1  16.7   9.42     1   2.1  66.7
7.54     1   2.1  18.8   9.82     1   2.1  68.8
7.61     1   2.1  20.8     10     1   2.1  70.8
7.62     1   2.1  22.9   10.2     2   4.2  75.0
7.82     1   2.1  25.0  10.48     1   2.1  77.1
7.83     1   2.1  27.1     11     1   2.1  79.2
7.85     1   2.1  29.2  11.18     1   2.1  81.3
7.87     1   2.1  31.3  11.29     1   2.1  83.3
7.98     1   2.1  33.3   11.8     1   2.1  85.4
   8     1   2.1  35.4  11.96     1   2.1  87.5
8.08     2   4.2  39.6     12     1   2.1  89.6
8.13     1   2.1  41.7   12.2     1   2.1  91.7
 8.2     1   2.1  43.8   12.6     1   2.1  93.8
8.23     1   2.1  45.8   13.2     1   2.1  95.8
8.25     1   2.1  47.9  13.25     1   2.1  97.9
8.36     1   2.1  50.0   13.6     1   2.1 100.0
```

Univariate-proseduuri antaa *means*-proseduuriin verrattuna joitakin lisätunnuslukuja. Niitä ovat mm.:

- $T : mean = 0$ ilmoittaa t-testin testifunktion arvon testattaessa hypoteesia, että populaation keskiarvo = 0.
- $Num \neq 0$ kertoo nollasta eroavien havaintojen lukumäärän
- $M(Sign) = p - n/2$, missä p on nollaa suurempien arvojen lukumäärä ja n on nollasta poikkeavien arvojen lukumäärä
- Std Mean = keskiarvon keskivirhe
- Num > 0 kertoo positiivisten havaintojen lukumäärän
- huomattava joukko eri kvantiileja tulostetaan
- havaintojen ääripäistä tulostetaan oletuksena viisi pienintä ja viisi suurinta arvoa

Korrelaatioanalyysista

Kahden satunnaismuuttujan riippuvuutta kuvaavaa lukua kutsutaan korrelaatioksi. Se ilmaisee astetta, miten muutos yhdessä muuttujassa vaikuttaa vastaavaan muutokseen toisessa muuttujassa. Korrelaatiokertoimella voidaan verrata eri muuttujaparien suhdetta toisiinsa.

Korrelaatiokertoimen arvot vaihtelevat välillä $[-1,1]$. Kerroin saa arvon 1, jos y on x:n monikerta ja -1, jos kysymyksessä on negatiivinen monikerta. Lähellä nollaa oleva korrelaatiokertoimen arvo tarkoittaa, että muuttujat ovat korreloimattomia. Tämä pätee ainoastaan lineaariseen riippuvuuteen. Jos esimerkiksi x ja y ovat satunnaisesti valittuja pisteitä ympyrän kehältä, on niiden välinen lineaarinen korrelaatio pieni, mutta pisteiden välillä on voimakas riippuvuus.

Tutkittaessa korrelaatiokertointen merkitsevyyttä käytetään seuraavia merkitsevyystasoja α:

α	*Tulkinta*
< 0.05	melkein merkitsevä ero havaintojen ja nollahypoteesin välillä
< 0.01	merkitsevä ero
< 0.001	erittäin merkitsevä ero

Merkitsevyystasolla tarkoitetaan todennäköisyyttä, jolla nollahypoteesi virheellisesti hylätään.

Kontingenssikerroin

Kontingenssikerroin C kuvaa kahden nominaaliasteikon muuttujan välistä riippuvuutta. Sen arvot ovat välillä $[0,1]$ ja suuremmat arvot tarkoittavat voimakkaampaa riippuvuutta. Kontingenssikertoimen käyttökelpoisuus empiirisenä tunnuslukuna perustuu ensisijaisesti siihen, että tutkittavien muuttujien mittaustuloksilta ei vaadita kuin nominaaliasteikon taso. Tällöin tarkastelun pohjana on muuttujien arvoparien esiintymisfrekvenssit. Kontingenssikerroin määritellään seuraavalla lausekkeella:

$$C = \left(\frac{\chi^2}{n+\chi^2}\right)^{\frac{1}{2}},$$

missä

$$\chi^2 = \sum_{i=1}^{l}\sum_{j=1}^{m}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

n = kaikkien havaintojen lukumäärä

m,l = mahdollisten luokkien lukumäärä

e_{ij} = odotettu frekvenssi

= reunafrekvenssien tulo/havaintojen lukumäärä

o_{ij} = havaittu frekvenssi

Kontingenssikertoimen tilastollista merkitsevyyttä testataan χ^2-testisuureen avulla, joka on χ^2-jakautunut vapausastein $(l-1)(m-1)$.

Kontingenssikertoimella on myös heikkoutensa, joita käsitellään seuraavassa:

Jos havaitut ja odotetut frekvenssit ovat yhtä suuria eli $o_{ij} = e_{ij}$ kaikilla i:n ja j:n arvoilla, niin $\chi^2 = 0$, joten myös $C = 0$. Tältä osin kontingenssikerroin täyttää riippuvuudelle asetettavat vaatimukset.

Jos tarkasteltavien muuttujien välillä vallitsee voimakas riippuvuus, poikkeavat havaitut frekvenssit huomattavasti odotetuista. Tällöin χ^2:n arvo muodostuu suureksi ja myös kontingenssikertoimen arvo kasvaa. Kuitenkaan kontingenssikerroin ei koskaan saavuta arvoa 1.

Koska kontingenssikerroin ei voi koskaan saada negatiivisia arvoja, ei sen avulla voi päätellä muuttujien välisen riippuvuuden suuntaa.

Kontingenssikertoimien keskinäinen vertailu ei ole mielekästä, mikäli ne perustuvat eri kokoisiin kontingenssitaulukoihin. Tämä johtuu siitä, että kontingenssikertoimen arvo riippuu rivien ja sarakkeiden lukumääristä.

Kontingenssikertoimen käytön heikkoutena on myös se, että se ei ole vertailukelpoinen järjestyskorrelaatiokertoimien ja tavallisen korrelaatiokertoimen kanssa.

Esimerkki kontingenssikertoimen laskemisesta SAS-ohjelmistolla

Tarkastellaan seuraavassa erään työpaikan työntekijöiden sukupuolijakautuman ja siviilisäädyn välistä riippuvuutta 53 hengen otoksella. Tarvittava SAS-ohjelmiston ajovirta on seuraava:

```
* Tähdellä alkavat rivit ovat kommenttirivejä;
* Tulosteessa olevan rivin pituuden määrittäminen,;
* sivunumeroinnin sekä päivämäärän poistaminen ;
options linesize=77 nonumber nodate;
* Määritetään tilapäistiedosto havaintoja varten;
DATA A;
* Muuttujien nimeäminen;
   input tekija $ saaty $;
* Havaintoaineisto, jossa on kaikkiaan;
* 53 tietuetta eli havaintoa;
cards;
```

```
mies naimisissa
nainen naimaton
nainen naimisissa
nainen naimisissa
mies naimisissa
nainen naimaton
mies naimisissa
mies naimaton
mies naimisissa
mies naimisissa
mies naimisissa
nainen naimisissa
nainen naimisissa
mies naimisissa
mies naimisissa
mies naimaton
mies naimisissa
mies naimisissa
nainen naimaton
mies naimisissa
mies naimaton
mies naimisissa
nainen naimisissa
mies naimaton
mies naimaton
mies naimisissa
mies naimisissa
mies naimisissa
mies naimisissa
mies naimaton
nainen naimisissa
nainen naimaton
mies naimisissa
mies naimisissa
nainen naimisissa
mies naimisissa
mies naimaton
nainen naimaton
mies naimaton
mies naimisissa
mies naimisissa
mies naimisissa
mies naimaton
mies naimisissa
mies naimisissa
mies naimaton
nainen naimaton
mies naimaton
mies naimaton
mies naimaton
mies naimisissa
mies naimaton
mies naimisissa
;
RUN;
PROC FREQ DATA=A;
   TABLES tekija*saaty/all;
title 'Luokittelu';
RUN;
```

Saatu tulostus on seuraavanlainen:

```
            Luokittelu

     TABLE OF TEKIJA BY SAATY

  TEKIJA     SAATY

  Frequency|
  Percent  |
  Row Pct  |
  Col Pct  |naimaton|naimisis|  Total
  ---------+--------+--------+
  mies     |     14 |     26 |     40
           |  26.42 |  49.06 |  75.47
           |  35.00 |  65.00 |
           |  70.00 |  78.79 |
  ---------+--------+--------+
  nainen   |      6 |      7 |     13
           |  11.32 |  13.21 |  24.53
           |  46.15 |  53.85 |
           |  30.00 |  21.21 |
  ---------+--------+--------+
  Total          20       33       53
              37.74    62.26   100.00

 STATISTICS FOR TABLE OF TEKIJA BY SAATY

Statistic                     DF     Value        Prob
------------------------------------------------------
Chi-Square                     1     0.519       0.471
Likelihood Ratio Chi-Square    1     0.512       0.474
Continuity Adj. Chi-Square     1     0.153       0.695
Mantel-Haenszel Chi-Square     1     0.510       0.475
Fisher's Exact Test (Left)                       0.344
                    (Right)                      0.853
                    (2-Tail)                     0.522
Phi Coefficient                     -0.099
Contingency Coefficient              0.099
Cramer's V                          -0.099
```

Kun rivien ja sarakkeiden summafrekvenssit pidetään kiinnitettyinä, niin `Fisherin eksaktin testin` vasemman yksitahoisen testin p-arvo 0.344 on todennäköisyys sille, naimattomien miesten lukumäärä on $>= 14$.

Vastaavasti `Fisherin eksaktin testin` oikean yksitahoisen testin p-arvo 0.853 on todennäköisyys sille, naimattomien miesten lukumäärä on $<= 14$.

```
Statistic                                Value        ASE
-------------------------------------------------------
Gamma                                   -0.228      0.307
Kendall's Tau-b                         -0.099      0.140
Stuart's Tau-c                          -0.083      0.117

Somers' D C|R                           -0.112      0.157
Somers' D R|C                           -0.088      0.125

Pearson Correlation                     -0.099      0.140
Spearman Correlation                    -0.099      0.140

Lambda Asymmetric C|R                    0.000      0.000
Lambda Asymmetric R|C                    0.000      0.000
Lambda Symmetric                         0.000      0.000

Uncertainty Coefficient C|R              0.007      0.020
Uncertainty Coefficient R|C              0.009      0.024
Uncertainty Coefficient Symmetric        0.008      0.022

      Estimates of the Relative Risk (Row1/Row2)

                                              95%
Type of Study           Value      Confidence Bounds
-------------------------------------------------------
Case-Control            0.628       0.177      2.235
Cohort (Col1 Risk)      0.758       0.368      1.563
Cohort (Col2 Risk)      1.207       0.695      2.097

Sample Size = 53
WARNING:  25% of the cells have expected counts less
          than 5. Chi-Square may not be a valid test.

    SUMMARY STATISTICS FOR TEKIJA BY SAATY


Cochran-Mantel-Haenszel Statistics (Based on Table
                                        Scores)

Statistic Alternative Hypothesis DF   Value   Prob
-------------------------------------------------
   1      Nonzero Correlation     1   0.510  0.475
   2      Row Mean Scores Differ  1   0.510  0.475
   3      General Association     1   0.510  0.475

Estimates of the Common Relative Risk (Row1/Row2)
                                                95%
Type of Study Method           Value Confidence Bounds
-------------------------------------------------------
```

```
Case-Control  Mantel-Haenszel  0.628    0.175    2.251
  (Odds Ratio Logit            0.628    0.177    2.235

Cohort        Mantel-Haenszel  0.758    0.355    1.621
  (Col1 Risk) Logit            0.758    0.368    1.563

Cohort        Mantel-Haenszel  1.207    0.720    2.024
  (Col2 Risk) Logit            1.207    0.695    2.097

The confidence bounds for the M-H estimates are test-
                                                 based.

Total Sample Size = 53
```

Tulostuksesta nähdään, että kontingenssikertoimen arvo on 0.099, χ^2-testisuureen arvo on 0.519 ja todennäköisyys sille, että χ^2-jakautuneen muuttujan arvo on suurempi kuin tämä on 0.471. Täten voidaan todeta, että yllä olevassa havaintoaineistossa ei ole mitään riippuvuutta sivilisäädyn ja sukupuolen välillä.

Spearmannin järjestyskorrelaatiokerroin

Spearmannin järjestyskorrelaatiokerroin ρ kuvaa kahden vähintään ordinaaliasteikon muuttujan välistä riippuvuutta. Muuttujien havaintoarvot korvataan havaintojen suuruuden mukaisilla järjestysluvuilla $1, 2, \ldots, n$. Muuttujien välinen riippuvuus saadaan tarkastelemalla kyseisten muuttujien järjestyslukuja. Spearmannin järjestyskorrelaatiokerroin saadaan lausekkeesta

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n},$$

missä d_i on muuttujien järjestyslukujen erotus ja n on havaintojen lukumäärä. Testisuure ρ:n merkitsevyyden testaamiseksi on

$$\frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}},$$

joka on t-jakautunut vapausastein $n-2$.

Mikäli muuttujien havaintoarvojen suuruusjärjestykset ovat täsmälleen samat, vallitsee järjestyslukujen välillä täydellinen positiivinen riippuvuus. Tällöin $\sum_{i=1}^{n} d_i^2 = 0$, joten $\rho = 1$. ρ:n laskentakaavasta seuraa, että mitä suurempia ovat muuttujien järjestyslukujen erotukset, sitä pienempi on muuttujien välinen riippuvuus.

Spearmannin järjestyskorrelaatiokerroin ρ on järjestysluvuista laskettu tavallinen korrelaatiokerroin. Laskettaessa ρ:ta ylläolevan lausekkeen avulla edellytetään, että muuttujien järjestysluvuissa ei esiinny tasatuloksia eli sidoksia. Pieni sidosmäärä voidaan käsitellä käyttämällä tasatuloksista järjestyslukujen keskiarvoja. Sidosten määrän kasvaessa on ρ:n kaavaan otettava mukaan sidoksien vaikutuksen eliminoiva korjaustekijä.

Kendallin järjestyskorrelaatiokerroin

Kendallin järjestyskorrelaatiokerroin soveltuu samanlaisiin tilanteisiin kuin Spearmannin järjestyskorrelaatiokerroinkin. Sekin kuvaa kahden vähintään ordinaaliasteikon muuttujan välistä riippuvuutta. Kendallin τ saadaan lausekkeella

$$\tau = 1 - \frac{4p}{n(n-1)},$$

missä n on havaintojen lukumäärä ja p on niiden vaihtojen pienin lukumäärä, joilla epäjärjestyksessä oleva järjestyslukujono saadaan samaan järjestykseen kuin toinen jono. Merkitsevyyttä testataan testisuureella

$$S = \tau\frac{1}{2}n(n-1),$$

jonka jakauma löytyy taulukkokirjoista.

Kendallin järjestyskorrelaatiokertointen arvot ovat pienempiä kuin samasta aineistosta lasketut Spearmannin järjestyskorrelaatiokertoimet. Tutkittaessa molempien kertoimien tilastollista merkitsevyyttä saadaan huomata, että johtopäätökset riippuvuuden merkitsevyydestä ovat täsmälleen samat kummankin järjestyskorrelaatiokertoimen suhteen.

Spearmannin ρ ja Kendallin τ ovat parametrittömiä korrelaatioita. Parametrittömyys tarkoittaa, että mitään oletuksia muuttujan tapausten jakautumasta ei tehdä. Niiden tarkoituksena on määrittää, ovatko samojen tapausten kaksi järjestystä samanlaisia.

Pääero Spearmannin ja Kendallin kertoimien välillä on, että Kendallin τ on merkitsellisempi, kun havaintoaineisto sisältää suuren määrän jaettuja sijoja. Spearmannin ρ antaa tarkemman aproksimaation tulomomenttikertoimille, kun havaintoaineisto on jatkuva.

Spearmannin ja Kendallin järjestyskorrelaatiokertoimet SAS-ohjelmistolla

Tutkitaan seuraavassa esimerkissä, onko jääkiekon SM-liigassa kaudella 92/93 pelaavien joukkueiden budjetilla, sijoituksella 21 kierroksen jälkeen sekä kauden 91/92 loppusijoituksilla riippuvuutta keskenään. Tarvittava SAS-ohjelmiston ajovirta on seuraava:

```
options linesize=77;
TITLE '----------CORR SAMPLE------------';
DATA;
      INPUT budj 1-3 sija21 5-6 sija92 8-9;
CARDS;
  1  1  5      * TPS
```

```
  2  3  1      * Jokerit
3.5  4  4      * HIFK
3.5  6 11      * Tappara
  5  9  2      * JyPHT
  6  7  9      * Ilves
7.5 10  7      * Kalpa
7.5  2  3      * Ässät
  9 11 12      * Kiekko-Espoo
 10  5  8      * HPK
 11  8  6      * Lukko
 12 12 10      * Reipas
;
RUN;
proc corr kendall spearman;
RUN;
```

Tulostus on seuraavanlainen:

```
---------------CORR SAMPLE---------------
       Correlation Analysis

3 'VAR' Variables:  BUDJ     SIJA21   SIJA92

                      Simple Statistics

Variable N   Mean Std Dev Median Minimum Maximum

BUDJ    12 6.5000  3.5929 6.7500  1.0000 12.0000
SIJA21  12 6.5000  3.6056 6.5000  1.0000 12.0000
SIJA92  12 6.5000  3.6056 6.5000  1.0000 12.0000

 Spearman Correlation Coefficients /
  Prob > |R| under Ho: Rho=0 / N = 12

             BUDJ    SIJA21    SIJA92

BUDJ      1.00000   0.65264   0.45965
           0.0       0.0214    0.1327

SIJA21    0.65264   1.00000   0.55944
           0.0214    0.0       0.0586

SIJA92    0.45965   0.55944   1.00000
           0.1327    0.0586    0.0

Kendall Tau b Correlation Coefficients /
  Prob > |R| under Ho: Rho=0 / N = 12

            BUDJ     SIJA21    SIJA92

BUDJ     1.00000    0.55391   0.30773
          0.0        0.0131    0.1682

SIJA21   0.55391    1.00000   0.39394
          0.0131     0.0       0.0746

SIJA92   0.30773    0.39394   1.00000
          0.1682     0.0746    0.0
```

Sekä Spearmannin että Kendallin järjestyskorrelaatiot antavat saman merkitsevyyden tuloksille. Nollahypoteesin hylkäämisen virhetodennäköisyydet ovat 0.0214 ja 0.0131. Kummallakin korrelaatiokertoimella on muuttujien budjetin ja sijoituksella 21 kierroksen jälkeen välillä melkein merkitsevä riippuvuus. Muiden muuttujien välillä ei ole riippuvuutta keskenään.

Vastaava tulos saadaan Cedarilla olevalla Splus-ohjelmistolla. Ohjelmistoa kutsutaan komennolla:

```
Splus
```

Tällöin tullaan Splus-ohjelmistoon, jonka kehoite on >. Splus-ohjelmiston käskyt annetaan kehoitteen jälkeen.

```
cypress ~> Splus
S-PLUS : Copyright (c) 1988, 1999 MathSoft, Inc.
S : Copyright Lucent Technologies, Inc.
Version 5.1 Release 1 for Silicon Graphics...
Working data will be in /mnt/mds/csc/lammi/MySwork
> # tarkastellaan samaa esimerkkiä kuin edellä
> budj <- scan()
1: 1 2 3.5 3.5 5 6 7.5 7.5 9 10 11 12
13:
> # tyhjä rivi lopettaa havaintojen syötön
> # scan pelkillä tyhjillä sulkumerkeillä
> #ilmoittaa,
> # että aineisto annetaan päätteeltä
> sija21 <- scan()
1: 1 3 4 6 9 7 10 2 11 5 8 12
13:
> # lasketaan Spearmannin korrelaatio
> cor(budj,sija21)
[1] 0.6526356
# lopetetaan Splus-istunto
>q()
```

Tulos on vastaava kuin SAS-ohjelmistolla saatu Spearmannin järjestyskorrelaatio.

Pearsonin tulomomenttikorrelaatiokerroin

Pearsonin tulomomenttikorrelaatiokerroin poikkeaa oleellisesti edellä esitetyistä korrelaatiokertoimista. Se on perusjoukon korrelaatiokertoimen estimaattori, kun taas kontingenssikerroin, Spearmannin ja Kendallin järjestyskorrelaatiokertoimet eivät ole minkään perusjoukon parametrin estimaattoreita. Ne kuvaavat ainoastaan riippuvuutta havaintoaineistossa.

Pearsonin tulomomenttikorrelaatiokerroin kuvaa kahden vähintään intervalliasteikon muuttujan välistä riippuvuutta. Tällöin tarkastellaan muuttujien riippuvuutta havaintoarvoja hyväksi käyttäen. Se saadaan kaavalla

$$r_{xy} = \frac{s_{xy}}{s_x s_y},$$

missä

s_x = muuttujan x keskihajonta

s_y = muuttujan y keskihajonta

s_{xy} = muuttujien x ja y välinen kovarianssi.

Pearsonin tulomomenttikorrelaatiokerroin kuvaa lineaarisessa regressiossa käyrän sovituksen hyvyyttä. Se saa virheettömässä sovituksessa arvon $+1.0$ tai -1.0. Negatiivinen arvo ei tarkoita huonoa sovitusta vaan käänteistä riippuvuutta. Arvo nolla merkitsee,

että lineaarista riippuvuutta ei ole. Pearsonin tulomomenttikerroin on myös riippuvuuden mittari. Se ilmaisee kahden muuttujan välisen lineaarisen riippuvuuden voimakkuuden.

Esimerkki Pearsonin korrelaatiokertoimesta

Tarkastellaan vastaavaa aineistoa kuin yllä. Nyt muuttujiksi otetaan budjetin suuruus miljoonissa markoissa sekä tehdyt ja päästetyt maalit 21 kierroksen jälkeen. SAS-ajovirta on seuraava:

```
options linesize=77;
DATA;
      INPUT budj tehdyt paast;
CARDS;
16  86 52
14  89 64
10  72 71
10  80 70
8.5 63 68
8   63 66
7.5 69 88
7.5 75 54
6.8 46 84
6.5 66 59
5   63 60
4.5 61 97
;
RUN;
proc corr pearson;
RUN;
```

Tulostus on seuraava:

```
    Correlation Analysis
3 'VAR' Variables:  BUDJ     TEHDYT   PAAST

                  Simple Statistics

Variable  N    Mean Std Dev      Sum  Minimum  Maximum

BUDJ     12  8.6917  3.4105 104.3000   4.5000  16.0000
TEHDYT   12 69.4167 11.9199 833.0000  46.0000  89.0000
PAAST    12 69.4167 13.8266 833.0000  52.0000  97.0000

  Pearson Correlation Coefficients /
  Prob > |R| under Ho: Rho=0 / N = 12

            BUDJ    TEHDYT     PAAST

BUDJ     1.00000   0.78858  -0.46260
          0.0       0.0023    0.1299

TEHDYT   0.78858   1.00000  -0.51027
          0.0023    0.0       0.0901

PAAST   -0.46260  -0.51027   1.00000
          0.1299    0.0901    0.0
```

Havaitaan, että tehdyt maalit ja budjetin suuruus ovat merkitsevästi riippuvaisia keskenään.

Korrelaatiokertoimien väärinkäyttö

Korrelaatiokertoimia käytettäessä on pidettävä harkinta mukana tarkasteltavien muuttujien suhteen. Mitä hyvänsä kahta muuttujaa, jotka saavat sinällään järkeviä arvoja, ei voi välttämättä verrata toisiinsa.

Esimerkkinä virheellisestä soveltamisesta voidaan mainita tapaus, jossa henkilö niinä aamuina, kun heräsi sängystään kumisaappaat jalassa, tunsi päänsä olevan kipeä. Tästä hän päätteli, että kumisaappaat ovat epäterveelliset jalkineet.

Regressioanalyysi

Regressioanalyysi käsittelee tapauksia, joissa yhtä riippuvaa muuttujaa selitetään yhdellä tai useammalla riippumattomalla muuttujalla.

Kahden muuttujan regressioanalyysi

Kahden muuttujan lineaarinen yhteys edellyttää, että muuttujat voidaan ilmaista numerollisina suureina. Tällöin niiden yhteys voidaan kuvata yhtälöllä

$$y = bx + a$$

missä a ja b ovat vakioita sekä muuttujat (x, y) ovat havaintoarvoja. Pyrkimyksenä on löytää sellaiset arvot kulmakertoimelle b ja vakiotermille a, että eo. suora parhaiten kuvaa havaintopareja. Tätä kuvaajaa kutsutaan regressiosuoraksi.

Havaitut pisteet poikkeavat regressiosuoralta. Havaittujen pisteiden (x_i, y_i) ja suoralla olevien vastaavien pisteiden (x_i, Y_i) y-arvojen erotus pyritään minimoimaan. Jotta erimerkkiset erotukset eivät kumoisi toistensa vaikutusta, minimoidaan lauseketta $\sum_{i=1}^{n}(y_i - Y_i)^2$. Tätä kutsutaan pienimmän neliösumman menetelmäksi.

Regressiosuoran kulmakerroin b saadaan kaavasta

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

Sijoittamalla saatu b:n arvo regressiosuoran yhtälöön saadaan vakiotermi

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}$$

Havaintojen lukumäärä edellä on n.

Mikäli havaintoparien (x, y) riippuvuus on käyräviivaista, niin riippuvuutta kuvaava yhtälö määritellään seuraavaksi

$$y = a + b_1 x + b_2 x^2$$

Käytetty tekniikka parametrien b_1, b_2 ja a ratkaisemiseksi on vastaava kuin lineaarisen riippuvuuden tapauksessa. Ennustetun regressiokäyrän ja havaittujen pisteiden y-akselilla olevien arvojen erotusten neliöiden summa minimoidaan. Tuloksena saadaan regressiokäyrä.

Esimerkki kahden muuttujan regressioanalyysistä

Tutkitaan seuraavassa esimerkissä 25 suurehkon kunnan vähittäiskauppojen työntekijöiden kokonaislukumäärän ja liikevaihdon suhdetta vuodelta 1990. Selittäväksi muuttujaksi valitaan liikevaihto miljoonissa markoissa. Selitettävä muuttuja on työntekijöiden lukumäärä. Tarvittava SAS-ohjelmiston ajovirta on seuraava:

```
options linesize=77 nonumber nodate;
data kauppa;
   input kunta $ hlolkm vaihto;
cards;
Espoo 5317 4847
Turku 7087 5959
Oulu 3875 4052
Kuopio 3250 2769
Pori 3149 2932
Jyväskylä 3113 3101
Lappeenranta 2166 2027
Joensuu 2308 2358
Hyvinkää 1441 1388
Kokkola 1486 1427
Rovaniemi 1652 1754
Kouvola 1859 2017
Rauma 1278 1155
Savonlinna 1199 1206
Seinäjoki 1563 1800
Tuusula 703 814
Kemi 972 879
Riihimäki 1017 886
Iisalmi 921 1073
Kuusankoski 475 454
Salo 1292 1245
Tampere 7242 7032
Lahti 3954 3492
Kotka 2022 1829
Hämeenlinna 1851 1873
;
proc reg data=kauppa;
   model hlolkm = vaihto;
   plot hlolkm*vaihto/
        vplots=2 hplots=2;
title 'Työntekijät/liikevaihto';
run;
```

Selitettävän muuttujan nimi on `model`-lauseen yhtäsuuruusmerkin vasemmalla puolella. Selittävä muuttuja on puolestaan yhtäsuuruusmerkin oikealla puolella. Tuloksena saadaan seuraava listaus:

```
                          Työntekijät/liikevaihto

Model: MODEL1
Dependent Variable: HLOLKM

                          Analysis of Variance

                       Sum of          Mean
Source   DF      Squares       Square    F Value   Prob>F

Model     1 79412061.589 79412061.589   1400.276   0.0001
Error    23 1304369.8511 56711.732656
C Total  24  80716431.44

    Root MSE     238.14225     R-square       0.9838
```

```
    Dep Mean    2447.68000     Adj R-sq       0.9831
    C.V.           9.72931

                    Parameter Estimates

              Parameter    Standard   T for H0:
Variable DF    Estimate       Error  Parameter=0 Prob>|T|

INTERCEP  1 -124.011216 83.61540593       -1.483   0.1516
VAIHTO    1    1.101480  0.02943539       37.420   0.0001
```

```
                          Työntekijät/liikevaihto

       -+------+------+------+------+-
        |                             |
  8000 +                             +
        |                             |
        |                      1   1  |
        |                             |
        |                             |
  6000 +                             +
H       |                             |
L       |                  1          |
O       |                             |
L       |                             |
K 4000 +             1 1             +
M       |                             |
        |           21                |
        |                             |
        |         1                   |
  2000 +       13                     +
        |     22                      |
        |   13                        |
        |   21                        |
        |  1                          |
     0 +                             +
        |                             |
       -+------+------+------+------+-
        0     2000   4000   6000  8000

                    VAIHTO
```

Saatu regressioyhtälö on

$$\text{hlolkm} = 1.101 \times \text{vaihto} - 124.011.$$

Varsinainen tulostus jakaantuu kahteen osaan: varianssianalyysi ja parametrien estimaatit. Varianssianalyysin tulostusta tarkastellaan erikseen mallin selittämällä osalla (Model) ja selittämättä jääneellä osalla (Error). Kokonaisvapausasteet (C Total, DF) = havaintojen lukumäärä - 1. Mallin vapausaste = 1 ja selittämättä jääneen osan vapausasteet = 24 - 1.

Jakamalla havaintojen neliöiden summat vapausasteilla saadaan keskineliöpoikkeamat (Mean Square). F-arvo on mallin selittämän ja selittämättä jääneen osan keskineliöpoikkeamien suhde. Tähän liittyvä todennäköisyysarvo $(Prob > F) = 0.0001$ ilmoittaa, että malli selittää erittäin merkitsevästi havaintoaineiston vaihtelua.

Jäännöstermin hajonnan estimaatti on 'Root MSE'. Selitettävän muuttujan keskiarvon ilmaisee 'Dep Mean'. Variaatiokerroin (C.V.) saadaan kertomalla sadalla jäännöstermin hajonnan estimaatin ja selitettävän muuttujan keskiarvon osamäärä. Mallin selittävän ja selitettävän muuttujan välisen korrelaation neliö (R-square) ilmoittaa selitysasteen. Malli selittää työntekijöiden lukumäärän vaihtelua keskiarvonsa nähden 98,38%.

Tulostuksen jälkimmäisessä osassa saadaan parametrien estimaattien lisäksi niiden keskivirheet. Niiden avulla voidaan määritellä luotettavuusvälit parametrien estimaateille. Parametreille lasketaan t-arvo jakamalla estimaatin arvo keskivirheellä. Nollahypoteesi on, että parametrien arvot ovat nollia. Hylkäämisen virhetodennäköisyydet osoittavat, että erittäin varmasti (p-arvo =0.0001) muuttujan vaihto kerroin ei ole nolla. Sen sijaan vakiotermi on nolla merkitsevyystasolla 0.1516. Tulos on hyvin ymmärrettävä. Mikäli ei ole liikevaihtoa, niin ei ole myöskään työntekijöitä.

Graafisessa esityksessä numerot ilmoittavat havaintojen lukumäärän tarkastellussa kohdassa. Siitä näkee myös, että havainnot eivät sisällä ns. vieraita havaintoja.

Esimerkin ratkaisu Splus-ohjelmistolla

Tarkastellaan edellä käsiteltyä esimerkkiä. Ratkaistaan se nyt Cedarilla olevalla Splus-ohjelmistolla. Havaintoaineisto on tiedostossa `kauppa`. Kirjoitetaan tarvittava Splus-ohjelmiston ajovirta tiedostoon `ajo`. Tällöin tiedosto `ajo` on seuraava:

```
z <- scan("kauppa",list(name="",0,0))
ls.print(lsfit(z[[3]],z[[2]]))
```

Havaintojen syötössä ilmoitetaan ensimmäisen muuttujan olevan merkkimuotoisen ja kahden seuraavan numeerisia. Käsky `lsfit` suorittaa regressiosuoran määräämisen pienimmän neliösumman menetelmää käyttäen. Tulostuksen tiivistelmä saadaan `ls.print`-komennolla.

Esimerkin ajo suoritetaan tämän jälkeen komennolla

```
Splus < ajo > tied
```

Tulostiedosto `tied` sisältää vastaavat tiedot, kuin SAS-ohjelmiston tulostuksessa saatiin. Se on seuraavanlainen:

```
Working data will be in /mnt/mds/csc/lammi/MySwork
Residual Standard Error = 238.1423,
Multiple R-Square = 0.9838
N = 25,  F-statistic = 1400.276 on 1 and 23 df,
p-value = 0

               coef std.err   t.stat p.value
Intercept -124.0112 83.6154  -1.4831  0.1516
        X    1.1015  0.0294  37.4203  0.0000
```

Useamman muuttujan regressioanalyysi

Olkoon $x_1, x_2, \ldots, x_n$ selittäviä muuttujia ja y selitettävä muuttuja. Tällöin regressioyhtälö on

$$y = b_1 x_1 + b_2 x_2 + \ldots + b_n x_n + a$$

Kertoimille $b_1, b_2, \ldots, b_n$ ja a löydetään aina sellaiset arvot, jotka tuottavat regressiokäyrälle parhaan sovituksen.

Askeltava regressio

Askeltavan mallin valinnassa on pyrkimys jättää mallista pois tarpeettomia selittäjiä. Selittäviksi muuttujiksi valitaan silloin niitä, joiden avulla mallin selitysaste saadaan mahdollisimman suureksi. Pois jätetään sellaiset muuttujat, jotka eivät lisää oleellisesti selitysastetta. Askeltavan mallin valinnassa on useita eri vaihtoehtoja:

- eteenpäin (`forward`) -menetelmä alkaa ilman selittäviä muuttujia. Malliin lisätään muuttujia yksi kerrallaan testisuureen mukaisessa järjestyksessä. Lisättyä muuttujaa ei voida enää poistaa mallista. Uusia muuttujia ei oteta, kun muuttujaan liittyvä p-arvo ylittää 0,5.
- taaksepäin (`backward`) -menetelmän alussa kaikki selittävät muuttujat ovat mallissa. Mallista poistetaan muuttuja, jonka testisuure ei ole merkittävän suuri. Poistettavan muuttujan testisuure on pienin.
- askeltava (`stepwise`) -menetelmä on melkein samanlainen kuin eteenpäin-valinta. Ainoa ero on siinä, että askeltavassa mallissa mukaan valittu muuttuja voidaan myös poistaa sieltä.
- maksimoinnissa (`maxr`) muuttujat valitaan malliin siten, että joka kerta selitysaste R^2 maksimoidaan.
- minimointi (`minr`) on maksimoinnin vastakohta, joka valitsee malliin muuttujat minimoimalla joka vaiheessa R^2:n.

Esimerkki eteenpäin askeltavasta mallista

Tarkastellaan seuraavassa esimerkissä kuntien taloutta. Selitettäväksi muuttujaksi valitaan veroäyrit. Selittävinä muuttujina ovat kunnan väkiluku, pinta-ala, vähittäiskauppojen lukumäärä ja menot yhteensä. Veroäyrien ja menojen luvut ovat tuhansia. Kaikki tiedot ovat vuodelta 1990. Pinta-alojen yksikkö on km^2. Tarvittava SAS-ajovirta on seuraava:

```
options linesize=64 nonumber nodate;
data ayrit;
   input kunta $ ayrit vakiluku pintaala
         kaupat menot;
cards;
Espoo        15865942 172629 312  755 4221378
Turku        10821859 159180 243 1588 5464290
Oulu          6294649 101379 328  799 2826526
Kuopio        4738773  80613 779  626 2309382
Pori          4290482  76357 503  772 2179612
Jyväskylä     4217372  66526  97  645 2235552
Lappeenranta 3248469  54941 760  454 1465573
Joensuu      2683955  47554  82  490 1627738
Hyvinkää     2622471  40194 323  319  980674
Kokkola      1827460  34635 328  338  800710
Rovaniemi    2045501  33500  94  386  949659
Kouvola      2095537  31740  44  331  738415
Rauma        1846785  29755  51  335  811738
Savonlinna   1492233  28559 821  298  751348
Seinäjoki    1692645  27765 129  316  717265
Kemi         1541790  25374  91  229  801986
Riihimäki    1533298  25000 121  247  548036
Iisalmi      1168837  23979 763  219  557928
Kuusankoski  1472708  21788 114  125  435752
Salo         1344728  21660 144  305  511702
Tampere     11343169 172560 523 1614 5264481
Lahti        5688525  93551 135  840 2602631
Kotka        3298406  56634 268  487 1659990
Hämeenlinna  2760400  43417 167  414 1224405
;
proc reg data=ayrit;
   model ayrit = vakiluku pintaala kaupat menot
         /selection=forward;
title 'Kuntien talous';
run;
```

Tuloksena saadaan seuraava listaus:

```
                          Kuntien talous

  Forward Selection Procedure for Dependent Variable AYRIT

Step 1   Variable VAKILUKU Entered  R-square = 0.94735369
C(p) = 71.17161511

                 DF         Sum of Squares      Mean Square
F   Prob>F

Regression        1        301518158113294  301518158113294
 395.88   0.0001
Error            22         16755956589800  761634390445.43
Total            23        318274114703094

                 Parameter        Standard          Type II
Variable          Estimate           Error   Sum of Squares
F   Prob>F

INTERCEP  -694940.3111276 295551.61754182  4210899470073.7
   5.53   0.0281
VAKILUKU      76.64556450      3.85215332  301518158113294
 395.88   0.0001

Bounds on condition number:            1,            1
------------------------------------------------------------

Step 2   Variable KAUPAT Entered    R-square = 0.98501409
C(p) =  7.95223229

                 DF         Sum of Squares      Mean Square
F   Prob>F

Regression        2        313504489013602  156752244506801
 690.16   0.0001
Error            21        4769625689492.4  227125032832.97
Total            23        318274114703094

                 Parameter        Standard          Type II
Variable          Estimate           Error   Sum of Squares
F   Prob>F

INTERCEP  -299148.1568358 170343.71541218  700463378832.60
   3.08   0.0936
VAKILUKU     109.46865569      4.98393180  109572202613035
 482.43   0.0001
KAUPAT     -4463.78374221    614.45844322   11986330900307
  52.77   0.0001

Bounds on condition number:     5.613304,     22.45322
------------------------------------------------------------

Step 3   Variable PINTAALA Entered  R-square = 0.98887177
C(p) =  3.27160110
```

```
                DF         Sum of Squares      Mean Square
F   Prob>F

Regression       3        314732287447044  104910762482348
                        The SAS System
                                3 Thursday, April 15, 1993

 592.41   0.0001
Error           20        3541827256050.0  177091362802.50
Total           23        318274114703094

                Parameter        Standard          Type II
Variable         Estimate           Error   Sum of Squares
F   Prob>F

INTERCEP  -69626.68438261 173847.92668165  28405980068.640
   0.16   0.6930
VAKILUKU     110.84151775      4.43164532  110782795142394
 625.57   0.0001
PINTAALA    -922.82468997    350.47283209  1227798433442.4
   6.93   0.0159
KAUPAT     -4530.50610343    543.16506654   12320466673857
  69.57   0.0001

Bounds on condition number:     5.692094,     37.05708
------------------------------------------------------------

No other variable met the 0.5000 significance level for
entry into the model.

Summary of Forward Selection Procedure for Dependent
Variable AYRIT

        Variable   Number   Partial     Model
Step    Entered        In      R**2      R**2        C(p)
F   Prob>F

 1    VAKILUKU        1    0.9474    0.9474      71.1716
395.8831   0.0001
 2    KAUPAT          2    0.0377    0.9850       7.9522
 52.7742   0.0001
 3    PINTAALA        3    0.0039    0.9889       3.2716
  6.9331   0.0159
```

Estimaattien perusteella saadaan mallin yhtälöksi:

$$\text{ayrit} = 110.841 \times \text{vakiluku} - 922.825 \times \text{pintaala} - 4530.51 \times \text{kaupat} - 69626.7.$$

Tulostuksessa on eritelty mallin eri vaiheet otettaessa muuttujia mukaan. Vaiheesta 1 huomataan, että väkiluku selittää kuntien äyrimäärää 94.75%.

Toisessa vaiheessa mukaan otetaan lähinnä tärkein selittävä muuttuja. Se on vähittäiskauppojen lukumäärä. Huomataan, että selitysaste on tämän jälkeen jo 98.5%.

Kolmannessa vaiheessa mallia voidaan vielä hivenen parantaa ottamalla mukaan pinta-ala. Tällöin lopulliseksi selitysasteeksi tulee 98.89%. Kuntien menoja kuvaava muuttuja on mallin selittämisen kannalta tarpeeton. Loppuyhteenvedossa on esitetty myös jokaisen muuttujan tuoma selitysasteen lisäys (`Partial R**2`).

Otoskeskiarvoista

Tilastollisessa testauksessa oikean testin valinta on edelletyksenä havaintoaineiston sisältämän tiedon hyväksikäytölle. Havainnoista tai käytetystä teoriasta johdetaan väitteitä, joiden perusteella asetetaan hypoteesit H_0 ja H_1. Nollahypoteesi H_0 on tavallisesti yksinkertaisin olettamus havaintojen käyttäytymisestä.

Otoskeskiarvotestit luokitellaan otosten lukumäärän mukaan: yhden otoksen keskiarvotesti, kahden otoksen keskiarvotestit ja useamman otoksen keskiarvotestit.

Yhden otoksen keskiarvotesti, kun varianssi on tunnettu

Tutkitaan havaintoja $x_1, x_2, \ldots, x_n$, jotka ovat otos normaalijakaumasta $N(\mu, \sigma^2)$. Yhden otoksen keskiarvotestissä perusjoukon varianssi oletetaan tunnetuksi. Tällöin testataan, onko perusjoukon tuntematon odotusarvo μ yhtäsuuri kuin jokin ennalta annettu luku μ_0. Kaksisuuntaisessa testauksessa hypoteesit ovat:

$$H_0 : \mu = \mu_0,$$
$$H_1 : \mu \neq \mu_0.$$

Yksisuuntaisen testauksen hypoteesit ovat:

$$H_0 : \mu = \mu_0,$$
$$H_1 : \mu > \mu_0,$$

tai

$$H_0 : \mu = \mu_0,$$
$$H_1 : \mu < \mu_0.$$

Testisuure z noudattaa jakaumaa $N(0,1)$. Mikäli kaksisuuntaisessa testauksessa

$$| z | > \text{kriittinen arvo},$$

hylätään nollahypoteesi kriittiseen arvoon liittyvällä merkitsevyystasolla.Mikäli nollahypoteesi on voimassa, niin testisuure z saadaan kaavasta:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}},$$

missä

$\bar{x}$ = havaintojen aritmeettinen keskiarvo

n = havaintojen lukumäärä.

Mikäli otoksesta laskettu aritmeettinen keskiarvo poikkeaa merkitsevästi luvusta μ_0, niin tämän tulkitaan osoittavan, että myös perusjoukon odotusarvo poikkeaa hypoteettisesta luvusta μ_0 ja nollahypoteesi $\mu = \mu_0$ hylätään.

Yhden otoksen keskiarvotesti, kun varianssi on tuntematon

Olkoon nyt jakauman parametrit μ ja σ^2 tuntemattomia. Tutkittaessa tällöin onko perusjoukon keskiarvo μ yhtä suuri kuin ennalta annettu luku μ_0, käytetään asian selvittämiseksi t-testiä. Mahdolliset hypoteesit ovat täsmälleen samat kuin tapauksessa, jossa varianssi oli tunnettu. Mikäli nollahypoteesi $\mu = \mu_0$ on voimassa, niin testisuure t noudattaa t - jakaumaa vapausastein $f = n - 1$. Testisuure t saadaan kaavasta

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}},$$

jossa

$$s = \sqrt{\frac{1}{n-1} \times \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Mikäli otos $(x_1, x_2, \ldots, x_n)$ on lähtöisin jakaumasta $N(\mu_0, \sigma^2)$, niin testisuure t saa suurella todennäköisyydellä pieniä arvoja. Suuret testisuureen t arvot johtavat nollahypoteesin $\mu = \mu_0$ hylkäämiseen.

Esimerkki yhden otoksen t-testistä

Tutkitaan seuraavassa esimerkissä perusjoukon tunnettua keskiarvoa. Perusjoukoksi valitaan miesten elinikä, joka on Suomessa noin 72 vuotta. Suoritetaan miesten eliniän otos ottamalla HS:n kuolinilmoitusten miesten eliniät ajalla 11.1.–27.1. otokseksi. Kaikkiaan otokseen sisältyy 126 havaintoa.

Testaamiseen käytetään Cedarilla olevan SAS-ohjelmiston MEANS-proseduuria. Koska MEANS-proseduurilla otoskeskiarvoa testataan lukuun nolla, luodaan apumuuttuja erotus = otoskeskiarvo - 72. Tätä apumuuttujaa testataan nollaan. Käytettävä SAS-ohjelmiston ajovirta on seuraava:

```
options linesize=77 nodate;
libname lib '/csc/lammi/sas/base';
data lib.ikam; * erotellaan koko ;
               * aineistosta;
   set lib.ika; * miehet mukaanotettaviksi;
   keep sp ika erotus;
   if sp='m';
   erotus=ika-72;
proc means data=lib.ikam mean std n t prt;
   var erotus;
title 'Miesten elinikä';
run;
```

Tuloksena saadaan MEANS-proseduurin tuottamat arvot:

```
                    Miesten elinikä

Analysis Variable : EROTUS

      Mean       Std Dev    N             T  Prob>|T|
-----------------------------------------------------------
-3.3333333    14.9018120  126    -2.5108741    0.0133
-----------------------------------------------------------
```

Testisuureeseen liittyvä p-arvo 0.0133 osoittaa, että tarkastellulla ajanjaksolla miesten keski-iät poikkeavat melkein merkittävästi miesten keskimääräisestä eliniästä.

Kahden otoksen keskiarvotesti, kun varianssit ovat tunnettuja

Otosten oletetaan olevan peräisin normaalisesta perusjoukosta N(μ_1,σ_1^2) ja N(μ_2,σ_2^2), joiden varianssit ovat tunnettuja. Testataan otoskeskiarvojen yhtäsuuruutta, jolloin hypoteesit ovat kaksisuuntaisessa testauksessa:

$$H_0 : \mu_1 = \mu_2,$$
$$H_1 : \mu_1 \neq \mu_2.$$

Yksisuuntaisessa testauksessa ovat hypoteeesit vastaavasti:

$$H_0 : \mu_1 = \mu_2,$$
$$H_1 : \mu_1 > \mu_2,$$

tai

$$H_0 : \mu_1 = \mu_2,$$
$$H_1 : \mu_1 < \mu_2.$$

Testisuure z, jonka jakauma on N(0,1), saadaan kaavasta:

$$z = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

Kahden otoksen keskiarvotesti, kun perusjoukkojen varianssit ovat yhtä suuria, mutta tuntemattomia

Tarkasteltavat havaintoryhmät ovat riippumattomia otoksia jakaumista N(μ_1,σ^2) ja N(μ_2,σ^2), joissa yhteinen varianssi on tuntematon. Testisuure t saadaan kaavasta

$$t = \frac{\bar{x_1} - \bar{x_2}}{s \times \sqrt{1/n_1 + 1/n_2}},$$

jossa

$$s = \sqrt{\frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x_1})^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x_2})^2}{n_1 + n_2 - 2}}.$$

Testisuure t noudattaa t-jakaumaa vapausastein $f = n_1 + n_2 - 2$ ehdolla, että nollahypoteesi $H_0 : \mu_1 = \mu_2$ on tosi.

Hypoteesit ovat samat kuin tapauksessa, jossa perusjoukkojen varianssit ovat tunnettuja.

Kahden otoksen keskiarvotesti, kun perusjoukkojen varianssit ovat eri suuret ja tuntemattomat

Tutkittavien perusjoukkojen varianssien ollessa eri suuret ja tuntemattomat joudutaan testisuureeseen, jonka otosjakauma noudattaa vain likimäärin t-jakaumaa. Testisuure t saadaan kaavasta

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{s_i^2/n_1 + s_2^2/n_2}}.$$

Vapausasteluku saadaan ratkaistua yhtälöstä

$$\frac{1}{f} = \frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1},$$

jossa

$$c = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2}.$$

Näin määrätty vapausasteluku täyttää ehdon

$$min(n_1 - 1, n_2 - 2) \leq f \leq n_1 + n_2 - 2.$$

Hypoteesit ovat samat kuin edellä.

Esimerkki kahden otoksen keskiarvojen testauksesta

Tutkitaan edellisessä esimerkissä olleen aineiston perusteella, onko miesten ja naisten eliniällä eroa. Tarvittava testaus tehdään SAS-ohjelmiston TTEST-proseduurilla. Luokittelevana muuttujana on sukupuoli = sp.

```
options linesize=74 nodate nonumber;
libname lib '/csc/lammi/sas/base';
proc ttest data=lib.ika;
   class sp;
   var ika;
title 'Miesten ja naisten elinikien vertailu';
run;
```

Tuloksena saadaan listaus 1.

Saatu p-arvo 0.0077 osoittaa, että miesten ja naisten eliniät poikkeavat merkittävästi toisistaan.

```
                    Miesten ja naisten elinikien vertailu

                                TTEST PROCEDURE

Variable: IKA

SP       N          Mean       Std Dev    Std Error      Minimum      Maximum
-------------------------------------------------------------------------------
m      126  68.66666667  14.90181197   1.32755892  24.00000000  95.00000000
n      129  73.93023256  16.32443818   1.43728664  11.00000000  96.00000000

Variances        T       DF    Prob>|T|
---------------------------------------
Unequal    -2.6902    251.9      0.0076
Equal      -2.6873    253.0      0.0077

For HO: Variances are equal, F' = 1.20    DF = (128,125)    Prob>F' = 0.3069
```

Listaus 1: *Kahden otoksen keskiarvojen testaus.*

Varianssianalyysi

Varianssianalyysi on väline analysoida kokeellista dataa. Siinä riippuvaa muuttujaa pyritään selittämään itsenäisten muuttujien avulla. Itsenäisten luokittelumuuttujien kombinaatiot muodostavat kokeellisen mallin.

Yksisuuntainen varianssianalyysi

Yksisuuntaisessa varianssianalyysissä itsenäiset muuttujat jaetaan luokkiin. Luokkia käsitellään jollain tavoin. Pyritään selvittämään, onko tällä käsittelyllä vaikutusta riippuvaan muuttujaan.

Tarkastellaan mallia, missä Y_{ij} on itsenäinen muuttuja ja on normaalisti jakautunut

$$Y_{ij} \sim N(\mu_i, \sigma^2),$$

missä $j = 1,\ldots,n_i$, $i = 1,\ldots,k$. Kullekin i:lle $Y_{i1},\ldots,Y_{in_i}$ ovat otoksia normaailijakumasta, jonka keskiarvo on μ_i. Lisäksi k otosta ovat kaikki riippumattomia.

Nollahypoteesina on tässä tarkastelussa

$$\mu_1 = \mu_2 = \ldots = \mu_k$$

Vaihtoehtoinen hypoteesi on, että jollekin i:lle ja j:lle $\mu_i \neq \mu_j$. Olkoon nyt

$$N = \sum n_i$$

missä n_i in i:nnen otoksen koko ja N kaikkien havaintojen määrä. Otoksen i havaintojen keskiarvoa merkitään $\bar{Y_i}$:llä

$$\bar{Y_i} = \frac{\sum_j Y_{ij}}{n_i}$$

Kaikkien havaintojen keskiarvoa merkitään $\bar{Y}$:llä

$$\bar{Y} = \frac{\sum_i \sum j Y_{ij}}{N} = \frac{\sum_i n_i \bar{Y_i}}{N}$$

Olkoon

$$\begin{aligned} T^2 &= \sum_i n_i (\bar{Y_i} - \bar{Y})^2 \\ &= \sum_i N_i \bar{Y_i}^2 - N\bar{Y}^2 \end{aligned}$$

$$\begin{aligned} S^2 &= \sum_i \sum_j (Y_{ij} - \bar{Y_i})^2 \\ &= \sum_i \sum_j Y_{ij}^2 - \sum_i n_i \bar{Y_i}^2 \end{aligned}$$

Tällöin testisuure F saadaan T^2:n ja S^2:n avulla, jolloin

$$F = \frac{T^2/(k-1)}{S^2/(N-k)}$$

Mikäli nollahypoteesi on tosi, niin testisuure F noudattaa jakaumaa

$$F \sim F_{k-1,N-k}$$

Nollahypoteesi joudutaan hylkäämään, mikäli

$$F > F_{k-1,N-k}$$

Esimerkki varianssianalyysistä

Tarkastellaan seuraavassa esimerkissä jääkiekon SM-liigan runkosarjan lopputaulukoita kolmelta kaudelta: 1996-1997, 1997-1998 ja 1998-1999. Tutkitaan joukkueiden saamia pistemääriä näiltä kausilta.

Tarvittava `sas`-ajovirta näyttää seuraavanlaiselta:

```
options ls=77 nodate;
libname lib '$HOME/sas/stat';
data lib.sarjat;
   input seura $ vuosi pisteet;
cards;
jok 1997 74
tps 1997 71
hpk 1997 63
ilv 1997 59
jyp 1997 56
kes 1997 51
ass 1997 48
tap 1997 44
hif 1997 42
luk 1997 37
sai 1997 34
kal 1997 21
tps 1998 66
hif 1998 63
ilv 1998 60
jok 1998 53
tap 1998 51
sai 1998 49
ass 1998 48
kes 1998 46
luk 1998 46
hpk 1998 41
jyp 1998 37
kal 1998 16
tps 1999 81
hif 1999 74
jok 1999 65
hpk 1999 60
ilv 1999 60
sai 1999 54
kes 1999 49
jyp 1999 48
tap 1999 48
ass 1999 46
luk 1999 39
kal 1999 24
;
```

```
proc format;
   value $nimi 'ass'='Ässät'
               'hif'='HIFK'
               'hpk'='HPK'
               'ilv'='Ilves'
               'jok'='Jokerit'
               'jyp'='JYP'
               'kal'='KalPa'
               'kes'='Blues'
               'luk'='Lukko'
               'sai'='SaiPa'
               'tap'='Tappara'
               'tps'='TPS';
proc anova data=lib.sarjat;
class seura ;
model pisteet=seura ;
means seura / duncan waller;
means seura / lsd tukey cldiff;
format seura $nimi.;
title 'SM-liiga kausilla 1996-1999';
run;
```

Tässä sas-ajossa muodostetaan ensin hakemistoon

```
$HOME/sas/stat
```

sas-tiedosto sarjat, johon havainnot tallennetaan. Format-proseduurilla saadaan seurojen nimet tulostettua kokonaisuudessaan.

Anova-proseduurin class määrittelyllä ilmoitetaan luokitteleva muuttuja seura. Means-alilauseella määritellään tarvittavat testit. duncan-määreellä suoritetaan Duncanin moniulotteinen vaihteluvälitesti. waller-määre mahdollistaa Waller-Duncanin testin suorittamisen.

Määre lsd suorittaa parittaiset t-testit. Tukey mahdollistaa Tukeyn vaihteluvälitestin ja cldiff antaa luottamusvälit parittaisille keskiarvojen erotuksille.

Tuloksena saadaan listaus 1. P-arvo 0.0001 osoittaa, että seurat-muuttuja pystyy erittäin hyvin selittämään saatuja pisteitä. Selitysaste R-Square on 77.6%.

Waller-Duncanin testi (listaus 2) luokittelee joukkueet kuuteen eri luokkaa. Joukkueet voivat esiintyä useammassa eri luokassa. Ainoastaan TPS ja KalPa kuuluvat yhteen luokkaan. TPS:n kanssa samaan luokkaan kuuluvat Jokerit, Ilves ja HIFK. KalPa on sen sijaan ainoana omassa luokassaan.

Duncanin testi (listaus 3) ryhmittelee joukkueet viiteen eri luokkaan. Yhteisenä piirteenä Waller-Duncanin testille on, että tässäkin joukkueet TPS ja KalPa kuuluvat ainoina ainoastaan yhteen luokkaan. Kolmeen eri luokkaan kuuluvat Ilves, HIFK, HPK ja Blues.

LSD-testin (listaukset 4–7) merkittäviä poikkeamia merkitään merkkijonolla '***'. Parittaisilla keskiarvotesteillä on määritelty myös 95%:n luottamusrajat. Tulos on hyvin samansuuntainen kuin Waller-Duncanin ja Duncanin testeillä saadut.

Tukeyn testillä (listaukset 8–11) keskiarvojen erotus pitää olla vähintään 24.548, jotta se olisi merkittävä 0.05:n tasolla. LSD-testissä pienin merkitsevyysero oli 14.051. Tästä on se seuraus, että Tukeyn testillä TPS:n pistemäärä entisten kolmen Jokereiden, Ilveksen ja HIFK:n lisäksi ei poikkea 0.05:n tasolla merkitsevästi myöskään HPK:sta ja Bluesista. Lisäksi Tukeyn testin mukaan KalPan ja Lukon pistemäärien ero ei ole merkittävä tarkastellulla tasolla.

```
                        SM-liiga kausilla 1996-1999                         1

                      Analysis of Variance Procedure
                         Class Level Information

Class    Levels    Values

SEURA        12    Blues HIFK HPK Ilves JYP Jokerit KalPa Lukko SaiPa TPS
                   Tappara Ässät


                   Number of observations in data set = 36


                        SM-liiga kausilla 1996-1999                         2

                      Analysis of Variance Procedure

Dependent Variable: PISTEET
                                    Sum of           Mean
Source                   DF        Squares         Square  F Value    Pr > F

Model                    11     5767.333333     524.303030     7.54    0.0001

Error                    24     1668.666667      69.527778

Corrected Total          35     7436.000000

                  R-Square            C.V.       Root MSE       PISTEET Mean

                  0.775596        16.45723       8.338332            50.66667


Source                   DF       Anova SS    Mean Square  F Value    Pr > F

SEURA                    11     5767.333333     524.303030     7.54    0.0001
```

Listaus 1: *Saatujen pisteiden selittäminen.*

```
                    SM-liiga kausilla 1996-1999                        3

                Analysis of Variance Procedure

    Waller-Duncan K-ratio T test for variable: PISTEET

NOTE: This test minimizes the Bayes risk under additive loss and
      certain other assumptions.

      Kratio= 100  df= 24  MSE= 69.52778  F= 7.540915
               Critical Value of T= 2.00389
           Minimum Significant Difference= 13.643

Means with the same letter are not significantly different.

      Waller Grouping              Mean      N  SEURA

                  A             72.667      3  TPS
                  A
            B     A             64.000      3  Jokerit
            B     A
            B     A   C         59.667      3  Ilves
            B     A   C
            B     A   C         59.667      3  HIFK
            B         C
            B     D   C         54.667      3  HPK
                  D   C
            E     D   C         48.667      3  Blues
            E     D   C
            E     D   C         47.667      3  Tappara
            E     D   C
            E     D   C         47.333      3  Ässät
            E     D   C
            E     D   C         47.000      3  JYP
            E     D
            E     D             45.667      3  SaiPa
            E
            E                   40.667      3  Lukko

                  F             20.333      3  KalPa
```

Listaus 2: *Waller-Duncanin testi.*

```
                    SM-liiga kausilla 1996-1999                        4

                  Analysis of Variance Procedure

       Duncan's Multiple Range Test for variable: PISTEET

     NOTE: This test controls the type I comparisonwise error rate, not
           the experimentwise error rate

                Alpha= 0.05  df= 24  MSE= 69.52778

       Number of Means     2     3     4     5     6     7
       Critical Range  14.05 14.76 15.21 15.53 15.77 15.96

       Number of Means     8     9    10    11    12
       Critical Range  16.10 16.22 16.32 16.40 16.46

     Means with the same letter are not significantly different.

           Duncan Grouping              Mean      N  SEURA

                      A            72.667      3  TPS
                      A
                B     A            64.000      3  Jokerit
                B     A
                B     A   C        59.667      3  Ilves
                B     A   C
                B     A   C        59.667      3  HIFK
                B         C
                B     D   C        54.667      3  HPK
                B     D   C
                B     D   C        48.667      3  Blues
                      D   C
                      D   C        47.667      3  Tappara
                      D   C
                      D   C        47.333      3  Ässät
                      D   C
                      D   C        47.000      3  JYP
                      D   C
                      D   C        45.667      3  SaiPa
                      D
                      D            40.667      3  Lukko

                      E            20.333      3  KalPa
```

Listaus 3: *Duncanin testi.*

```
                    SM-liiga kausilla 1996-1999                        8

                  Analysis of Variance Procedure

                              Lower    Difference     Upper
          SEURA           Confidence    Between   Confidence
        Comparison           Limit       Means       Limit

     KalPa    - Lukko        -34.385     -20.333      -6.282   ***
```

Listaus 4: *LSD-testi.*

```
                    SM-liiga kausilla 1996-1999                    5

                  Analysis of Variance Procedure

               T tests (LSD) for variable: PISTEET

   NOTE: This test controls the type I comparisonwise error rate not
         the experimentwise error rate.

      Alpha= 0.05  Confidence= 0.95  df= 24  MSE= 69.52778
                  Critical Value of T= 2.06390
              Least Significant Difference= 14.051

Comparisons significant at the 0.05 level are indicated by '***'.

                          Lower    Difference     Upper
       SEURA           Confidence    Between   Confidence
    Comparison            Limit       Means       Limit

  TPS     - Jokerit       -5.385       8.667      22.718
  TPS     - Ilves         -1.051      13.000      27.051
  TPS     - HIFK          -1.051      13.000      27.051
  TPS     - HPK            3.949      18.000      32.051   ***
  TPS     - Blues          9.949      24.000      38.051   ***
  TPS     - Tappara       10.949      25.000      39.051   ***
  TPS     - Ässät         11.282      25.333      39.385   ***
  TPS     - JYP           11.615      25.667      39.718   ***
  TPS     - SaiPa         12.949      27.000      41.051   ***
  TPS     - Lukko         17.949      32.000      46.051   ***
  TPS     - KalPa         38.282      52.333      66.385   ***

  Jokerit - TPS          -22.718      -8.667       5.385
  Jokerit - Ilves         -9.718       4.333      18.385
  Jokerit - HIFK          -9.718       4.333      18.385
  Jokerit - HPK           -4.718       9.333      23.385
  Jokerit - Blues          1.282      15.333      29.385   ***
  Jokerit - Tappara        2.282      16.333      30.385   ***
  Jokerit - Ässät          2.615      16.667      30.718   ***
  Jokerit - JYP            2.949      17.000      31.051   ***
  Jokerit - SaiPa          4.282      18.333      32.385   ***
  Jokerit - Lukko          9.282      23.333      37.385   ***
  Jokerit - KalPa         29.615      43.667      57.718   ***

  Ilves   - TPS          -27.051     -13.000       1.051
  Ilves   - Jokerit      -18.385      -4.333       9.718
  Ilves   - HIFK         -14.051       0.000      14.051
  Ilves   - HPK           -9.051       5.000      19.051
  Ilves   - Blues         -3.051      11.000      25.051
  Ilves   - Tappara       -2.051      12.000      26.051
  Ilves   - Ässät         -1.718      12.333      26.385
  Ilves   - JYP           -1.385      12.667      26.718
  Ilves   - SaiPa         -0.051      14.000      28.051
  Ilves   - Lukko          4.949      19.000      33.051   ***
  Ilves   - KalPa         25.282      39.333      53.385   ***

  HIFK    - TPS          -27.051     -13.000       1.051
  HIFK    - Jokerit      -18.385      -4.333       9.718
  HIFK    - Ilves        -14.051       0.000      14.051
  HIFK    - HPK           -9.051       5.000      19.051
```

Listaus 5: *LSD-testi.*

```
                    SM-liiga kausilla 1996-1999                          6

                    Analysis of Variance Procedure

                                    Lower     Difference     Upper
          SEURA                  Confidence    Between    Confidence
       Comparison                   Limit       Means       Limit

   HIFK     - Blues               -3.051       11.000      25.051
   HIFK     - Tappara             -2.051       12.000      26.051
   HIFK     - Ässät               -1.718       12.333      26.385
   HIFK     - JYP                 -1.385       12.667      26.718
   HIFK     - SaiPa               -0.051       14.000      28.051
   HIFK     - Lukko                4.949       19.000      33.051   ***
   HIFK     - KalPa               25.282       39.333      53.385   ***

   HPK      - TPS                -32.051      -18.000      -3.949   ***
   HPK      - Jokerit            -23.385       -9.333       4.718
   HPK      - Ilves              -19.051       -5.000       9.051
   HPK      - HIFK               -19.051       -5.000       9.051
   HPK      - Blues               -8.051        6.000      20.051
   HPK      - Tappara             -7.051        7.000      21.051
   HPK      - Ässät               -6.718        7.333      21.385
   HPK      - JYP                 -6.385        7.667      21.718
   HPK      - SaiPa               -5.051        9.000      23.051
   HPK      - Lukko               -0.051       14.000      28.051
   HPK      - KalPa               20.282       34.333      48.385   ***

   Blues    - TPS                -38.051      -24.000      -9.949   ***
   Blues    - Jokerit            -29.385      -15.333      -1.282   ***
   Blues    - Ilves              -25.051      -11.000       3.051
   Blues    - HIFK               -25.051      -11.000       3.051
   Blues    - HPK                -20.051       -6.000       8.051
   Blues    - Tappara            -13.051        1.000      15.051
   Blues    - Ässät              -12.718        1.333      15.385
   Blues    - JYP                -12.385        1.667      15.718
   Blues    - SaiPa              -11.051        3.000      17.051
   Blues    - Lukko               -6.051        8.000      22.051
   Blues    - KalPa               14.282       28.333      42.385   ***

   Tappara  - TPS                -39.051      -25.000     -10.949   ***
   Tappara  - Jokerit            -30.385      -16.333      -2.282   ***
   Tappara  - Ilves              -26.051      -12.000       2.051
   Tappara  - HIFK               -26.051      -12.000       2.051
   Tappara  - HPK                -21.051       -7.000       7.051
   Tappara  - Blues              -15.051       -1.000      13.051
   Tappara  - Ässät              -13.718        0.333      14.385
   Tappara  - JYP                -13.385        0.667      14.718
   Tappara  - SaiPa              -12.051        2.000      16.051
   Tappara  - Lukko               -7.051        7.000      21.051
   Tappara  - KalPa               13.282       27.333      41.385   ***

   Ässät    - TPS                -39.385      -25.333     -11.282   ***
   Ässät    - Jokerit            -30.718      -16.667      -2.615   ***
   Ässät    - Ilves              -26.385      -12.333       1.718
   Ässät    - HIFK               -26.385      -12.333       1.718
   Ässät    - HPK                -21.385       -7.333       6.718
   Ässät    - Blues              -15.385       -1.333      12.718
   Ässät    - Tappara            -14.385       -0.333      13.718
```

Listaus 6: *LSD-testi.*

```
                    SM-liiga kausilla 1996-1999                          7

                    Analysis of Variance Procedure

                                    Lower     Difference     Upper
                SEURA           Confidence     Between    Confidence
              Comparison           Limit        Means       Limit

Ässät     - JYP              -13.718        0.333       14.385
Ässät     - SaiPa            -12.385        1.667       15.718
Ässät     - Lukko             -7.385        6.667       20.718
Ässät     - KalPa             12.949       27.000       41.051   ***

JYP       - TPS              -39.718      -25.667      -11.615   ***
JYP       - Jokerit          -31.051      -17.000       -2.949   ***
JYP       - Ilves            -26.718      -12.667        1.385
JYP       - HIFK             -26.718      -12.667        1.385
JYP       - HPK              -21.718       -7.667        6.385
JYP       - Blues            -15.718       -1.667       12.385
JYP       - Tappara          -14.718       -0.667       13.385
JYP       - Ässät            -14.385       -0.333       13.718
JYP       - SaiPa            -12.718        1.333       15.385
JYP       - Lukko             -7.718        6.333       20.385
JYP       - KalPa             12.615       26.667       40.718   ***

SaiPa     - TPS              -41.051      -27.000      -12.949   ***
SaiPa     - Jokerit          -32.385      -18.333       -4.282   ***
SaiPa     - Ilves            -28.051      -14.000        0.051
SaiPa     - HIFK             -28.051      -14.000        0.051
SaiPa     - HPK              -23.051       -9.000        5.051
SaiPa     - Blues            -17.051       -3.000       11.051
SaiPa     - Tappara          -16.051       -2.000       12.051
SaiPa     - Ässät            -15.718       -1.667       12.385
SaiPa     - JYP              -15.385       -1.333       12.718
SaiPa     - Lukko             -9.051        5.000       19.051
SaiPa     - KalPa             11.282       25.333       39.385   ***

Lukko     - TPS              -46.051      -32.000      -17.949   ***
Lukko     - Jokerit          -37.385      -23.333       -9.282   ***
Lukko     - Ilves            -33.051      -19.000       -4.949   ***
Lukko     - HIFK             -33.051      -19.000       -4.949   ***
Lukko     - HPK              -28.051      -14.000        0.051
Lukko     - Blues            -22.051       -8.000        6.051
Lukko     - Tappara          -21.051       -7.000        7.051
Lukko     - Ässät            -20.718       -6.667        7.385
Lukko     - JYP              -20.385       -6.333        7.718
Lukko     - SaiPa            -19.051       -5.000        9.051
Lukko     - KalPa              6.282       20.333       34.385   ***

KalPa     - TPS              -66.385      -52.333      -38.282   ***
KalPa     - Jokerit          -57.718      -43.667      -29.615   ***
KalPa     - Ilves            -53.385      -39.333      -25.282   ***
KalPa     - HIFK             -53.385      -39.333      -25.282   ***
KalPa     - HPK              -48.385      -34.333      -20.282   ***
KalPa     - Blues            -42.385      -28.333      -14.282   ***
KalPa     - Tappara          -41.385      -27.333      -13.282   ***
KalPa     - Ässät            -41.051      -27.000      -12.949   ***
KalPa     - JYP              -40.718      -26.667      -12.615   ***
KalPa     - SaiPa            -39.385      -25.333      -11.282   ***
```

Listaus 7: *LSD-testi.*

```
                    SM-liiga kausilla 1996-1999                    9

                  Analysis of Variance Procedure

   Tukey's Studentized Range (HSD) Test for variable: PISTEET

   NOTE: This test controls the type I experimentwise error rate.

      Alpha= 0.05  Confidence= 0.95  df= 24  MSE= 69.52778
           Critical Value of Studentized Range= 5.099
             Minimum Significant Difference= 24.548

Comparisons significant at the 0.05 level are indicated by '***'.

                      Simultaneous            Simultaneous
                          Lower    Difference     Upper
        SEURA          Confidence    Between   Confidence
     Comparison           Limit       Means       Limit

  TPS     - Jokerit      -15.881       8.667      33.215
  TPS     - Ilves        -11.548      13.000      37.548
  TPS     - HIFK         -11.548      13.000      37.548
  TPS     - HPK           -6.548      18.000      42.548
  TPS     - Blues         -0.548      24.000      48.548
  TPS     - Tappara        0.452      25.000      49.548   ***
  TPS     - Ässät          0.785      25.333      49.881   ***
  TPS     - JYP            1.119      25.667      50.215   ***
  TPS     - SaiPa          2.452      27.000      51.548   ***
  TPS     - Lukko          7.452      32.000      56.548   ***
  TPS     - KalPa         27.785      52.333      76.881   ***

  Jokerit - TPS          -33.215      -8.667      15.881
  Jokerit - Ilves        -20.215       4.333      28.881
  Jokerit - HIFK         -20.215       4.333      28.881
  Jokerit - HPK          -15.215       9.333      33.881
  Jokerit - Blues         -9.215      15.333      39.881
  Jokerit - Tappara       -8.215      16.333      40.881
  Jokerit - Ässät         -7.881      16.667      41.215
  Jokerit - JYP           -7.548      17.000      41.548
  Jokerit - SaiPa         -6.215      18.333      42.881
  Jokerit - Lukko         -1.215      23.333      47.881
  Jokerit - KalPa         19.119      43.667      68.215   ***

  Ilves   - TPS          -37.548     -13.000      11.548
  Ilves   - Jokerit      -28.881      -4.333      20.215
  Ilves   - HIFK         -24.548       0.000      24.548
  Ilves   - HPK          -19.548       5.000      29.548
  Ilves   - Blues         -13.548      11.000      35.548
  Ilves   - Tappara       -12.548      12.000      36.548
  Ilves   - Ässät         -12.215      12.333      36.881
  Ilves   - JYP           -11.881      12.667      37.215
  Ilves   - SaiPa         -10.548      14.000      38.548
  Ilves   - Lukko          -5.548      19.000      43.548
  Ilves   - KalPa          14.785      39.333      63.881   ***

  HIFK    - TPS          -37.548     -13.000      11.548
  HIFK    - Jokerit      -28.881      -4.333      20.215
  HIFK    - Ilves        -24.548       0.000      24.548
  HIFK    - HPK          -19.548       5.000      29.548
```

Listaus 8: *Tukeyn testi.*

SM-liiga kausilla 1996-1999 10

Analysis of Variance Procedure

SEURA Comparison	Simultaneous Lower Confidence Limit	Difference Between Means	Simultaneous Upper Confidence Limit	
HIFK - Blues	-13.548	11.000	35.548	
HIFK - Tappara	-12.548	12.000	36.548	
HIFK - Ässät	-12.215	12.333	36.881	
HIFK - JYP	-11.881	12.667	37.215	
HIFK - SaiPa	-10.548	14.000	38.548	
HIFK - Lukko	-5.548	19.000	43.548	
HIFK - KalPa	14.785	39.333	63.881	***
HPK - TPS	-42.548	-18.000	6.548	
HPK - Jokerit	-33.881	-9.333	15.215	
HPK - Ilves	-29.548	-5.000	19.548	
HPK - HIFK	-29.548	-5.000	19.548	
HPK - Blues	-18.548	6.000	30.548	
HPK - Tappara	-17.548	7.000	31.548	
HPK - Ässät	-17.215	7.333	31.881	
HPK - JYP	-16.881	7.667	32.215	
HPK - SaiPa	-15.548	9.000	33.548	
HPK - Lukko	-10.548	14.000	38.548	
HPK - KalPa	9.785	34.333	58.881	***
Blues - TPS	-48.548	-24.000	0.548	
Blues - Jokerit	-39.881	-15.333	9.215	
Blues - Ilves	-35.548	-11.000	13.548	
Blues - HIFK	-35.548	-11.000	13.548	
Blues - HPK	-30.548	-6.000	18.548	
Blues - Tappara	-23.548	1.000	25.548	
Blues - Ässät	-23.215	1.333	25.881	
Blues - JYP	-22.881	1.667	26.215	
Blues - SaiPa	-21.548	3.000	27.548	
Blues - Lukko	-16.548	8.000	32.548	
Blues - KalPa	3.785	28.333	52.881	***
Tappara - TPS	-49.548	-25.000	-0.452	***
Tappara - Jokerit	-40.881	-16.333	8.215	
Tappara - Ilves	-36.548	-12.000	12.548	
Tappara - HIFK	-36.548	-12.000	12.548	
Tappara - HPK	-31.548	-7.000	17.548	
Tappara - Blues	-25.548	-1.000	23.548	
Tappara - Ässät	-24.215	0.333	24.881	
Tappara - JYP	-23.881	0.667	25.215	
Tappara - SaiPa	-22.548	2.000	26.548	
Tappara - Lukko	-17.548	7.000	31.548	
Tappara - KalPa	2.785	27.333	51.881	***
Ässät - TPS	-49.881	-25.333	-0.785	***
Ässät - Jokerit	-41.215	-16.667	7.881	
Ässät - Ilves	-36.881	-12.333	12.215	
Ässät - HIFK	-36.881	-12.333	12.215	
Ässät - HPK	-31.881	-7.333	17.215	
Ässät - Blues	-25.881	-1.333	23.215	

Listaus 9: *Tukeyn testi.*

SM-liiga kausilla 1996-1999 11

Analysis of Variance Procedure

SEURA Comparison		Simultaneous Lower Confidence Limit	Difference Between Means	Simultaneous Upper Confidence Limit	
Ässät	- Tappara	-24.881	-0.333	24.215	
Ässät	- JYP	-24.215	0.333	24.881	
Ässät	- SaiPa	-22.881	1.667	26.215	
Ässät	- Lukko	-17.881	6.667	31.215	
Ässät	- KalPa	2.452	27.000	51.548	***
JYP	- TPS	-50.215	-25.667	-1.119	***
JYP	- Jokerit	-41.548	-17.000	7.548	
JYP	- Ilves	-37.215	-12.667	11.881	
JYP	- HIFK	-37.215	-12.667	11.881	
JYP	- HPK	-32.215	-7.667	16.881	
JYP	- Blues	-26.215	-1.667	22.881	
JYP	- Tappara	-25.215	-0.667	23.881	
JYP	- Ässät	-24.881	-0.333	24.215	
JYP	- SaiPa	-23.215	1.333	25.881	
JYP	- Lukko	-18.215	6.333	30.881	
JYP	- KalPa	2.119	26.667	51.215	***
SaiPa	- TPS	-51.548	-27.000	-2.452	***
SaiPa	- Jokerit	-42.881	-18.333	6.215	
SaiPa	- Ilves	-38.548	-14.000	10.548	
SaiPa	- HIFK	-38.548	-14.000	10.548	
SaiPa	- HPK	-33.548	-9.000	15.548	
SaiPa	- Blues	-27.548	-3.000	21.548	
SaiPa	- Tappara	-26.548	-2.000	22.548	
SaiPa	- Ässät	-26.215	-1.667	22.881	
SaiPa	- JYP	-25.881	-1.333	23.215	
SaiPa	- Lukko	-19.548	5.000	29.548	
SaiPa	- KalPa	0.785	25.333	49.881	***
Lukko	- TPS	-56.548	-32.000	-7.452	***
Lukko	- Jokerit	-47.881	-23.333	1.215	
Lukko	- Ilves	-43.548	-19.000	5.548	
Lukko	- HIFK	-43.548	-19.000	5.548	
Lukko	- HPK	-38.548	-14.000	10.548	
Lukko	- Blues	-32.548	-8.000	16.548	
Lukko	- Tappara	-31.548	-7.000	17.548	
Lukko	- Ässät	-31.215	-6.667	17.881	
Lukko	- JYP	-30.881	-6.333	18.215	
Lukko	- SaiPa	-29.548	-5.000	19.548	
Lukko	- KalPa	-4.215	20.333	44.881	
KalPa	- TPS	-76.881	-52.333	-27.785	***
KalPa	- Jokerit	-68.215	-43.667	-19.119	***
KalPa	- Ilves	-63.881	-39.333	-14.785	***
KalPa	- HIFK	-63.881	-39.333	-14.785	***
KalPa	- HPK	-58.881	-34.333	-9.785	***
KalPa	- Blues	-52.881	-28.333	-3.785	***
KalPa	- Tappara	-51.881	-27.333	-2.785	***
KalPa	- Ässät	-51.548	-27.000	-2.452	***

Listaus 10: *Tukeyn testi.*

```
                    SM-liiga kausilla 1996-1999                    12

                  Analysis of Variance Procedure

                         Simultaneous            Simultaneous
                             Lower    Difference     Upper
          SEURA          Confidence    Between   Confidence
       Comparison           Limit       Means       Limit

KalPa    - JYP           -51.215     -26.667      -2.119   ***
KalPa    - SaiPa         -49.881     -25.333      -0.785   ***
KalPa    - Lukko         -44.881     -20.333       4.215
```

Listaus 11: *Tukeyn testi.*

Otoskoosta

Suunniteltaessa otantaa jostakin populaatiosta tulee välittömästi eteen ongelma, kuinka suuri otannan on vähintään oltava. Kysymykseen ei löydy yksikäsitteistä vastausta. Tarvittava otoskoko riippuu sekä käytetystä otantamenetelmästä että halutusta tarkkuudesta. Käytettävä analyysimenetelmä saattaa edellyttää tiettyä havaintojen vähimmäismäärää. Monimuuttujaisissa tilastomenetelmissä pidetään eräänä otoskoon suosituksena, että otosalkioita on oltava vähintään se määrä, kuin on mitattavia tulosmuuttujia.

Otoskoon määrittäminen keskiarvon avulla

Tarkastellaaan otoskoon ratkaisemista, kun koko populaatio on normaalijakautunut ja sen keskiarvo on $\bar{Y}$. Onnistuneen otoksen tapauksessa halutaan, että otoskeskiarvo $\bar{y}$ poikkeaa mahdollisimman vähän populaation keskiarvosta $\bar{Y}$. Sille asetetaankin ehto

$$|\bar{y} - \bar{Y}| \leq d,$$

missä $\bar{y}$ on otoskeskiarvo ja d on ennalta valittu vakio. Todennäköisyys sille, että otoskeskiarvo poikkeaa populaation keskiarvosta pitäsi olla pieni luku α, esim. 1%.

$$P\{|\bar{y} - \bar{Y}| > d\} \leq \alpha$$

Tällöin valittu luottamustaso on $(1-\alpha)$.

Jos tunnetaan populaation koko N ja sen varianssi S^2, niin saadaan keskiarvopoikkeaman d, todennäköisyyden α ja otoskoon n välille yhtälö

$$n = \left(\frac{z_\alpha}{d}\right)^2 \left(\frac{N-n}{N}\right) S^2. \qquad (1)$$

Kaavassa (1) z_α on taulukoista saatava normaalijakautunut kiinteä arvo sille, että syntynyt virhe on pienempi kuin α. Jos otantasuhde n/N on pieni, voidaan kaavasta (1) saatava otoskoon kaava kirjoittaa muodossa

$$n = \left(\frac{z_\alpha}{d}\right)^2 S^2. \qquad (2)$$

Populaation keskihajontaa ei aina tunneta. Sille voi kuitenkin tehdä arvauksia ja laskea otoskoko sen jälkeen.

Otoskoon laskemisesta

Tutkitaan hajonnan S sekä otos- ja populaatiokeskiarvojen välin d vaikutusta otoskokoon eri luottamustasoilla.

Lasketaan otoskoko Cedarilla olevalla SAS-ohjelmistolla käyttäen hyväksi kaavaa (2). Tutkitaan, mikä merkitys tarvittavalle otoskoolle on otoskeskiarvojen vaihtelulla eli otos- ja populaatiokeskiarvojen poikkeamien muutoksilla. Lisäksi selvitetään otoskeskiarvon hajonnan merkitystä otoskokoon.

Annetaan keskiarvopoikkeaman vaihdella 5:stä 20:een. Vastaavasti estimoidun populaation keskihajonta saa arvoja 5:stä 35:een. Tarvittava SAS-ohjelmiston ajovirta on seuraava:

```
options ls=64 nodate nonumber;
data a (drop=z95 z99);
   z95=1.96;
   z99=2.58;
   do vali=5 to 20 by 5;
      do hajont=5 to 35 by 5;
      n95=ceil((z95*hajont/vali)**2);
      n99=ceil((z99*hajont/vali)**2);
      output;
    end;
   end;
proc print data=a;
title 'Otoskoot 95%:n ja 99%:n luottamustasolla';
run;
```

Ajovirrassa lasketaan otoskoot 95%:n ja 99%:n luottamusväleille. Print-proseduuri tulostaa otoskoot $n95$ ja $n99$ eri parametrien arvoilla.

Otoskoot 95%:n ja 99%:n luottamustasolla

OBS	VALI	HAJONT	N95	N99
1	5	5	4	7
2	5	10	16	27
3	5	15	35	60
4	5	20	62	107
5	5	25	97	167
6	5	30	139	240
7	5	35	189	327
8	10	5	1	2
9	10	10	4	7
10	10	15	9	15
11	10	20	16	27
12	10	25	25	42
13	10	30	35	60
14	10	35	48	82
15	15	5	1	1
16	15	10	2	3
17	15	15	4	7
18	15	20	7	12
19	15	25	11	19
20	15	30	16	27
21	15	35	21	37
22	20	5	1	1
23	20	10	1	2
24	20	15	3	4
25	20	20	4	7
26	20	25	7	11
27	20	30	9	15
28	20	35	12	21

Suurimmat otoskoot annetuilla luottamusväleillä tarvitaan silloin, kun otos- ja populaation keskiarvot poikkeavat hyvin vähän toisistaan, mutta otoskeskiarvon hajonta on kuitenkin suuri.

Esimerkki otoskoon laskemisen soveltamisesta

Tarkastellaan seuraavassa esimerkissä EU-äänestystä Uudenmaan vaalipiirissä. Pyritään selvittämään tarvittavaa otoskokoa kyllä-äänten vaihtelusta tietyllä välillä 95%:n luotettavuustasolla. Aineisto voidaan katsoa normaalisti jakautuneeksi, kun jätetään suuret kyllä-kaupungit Espoo ja Vantaa tarkastelun ulkopuolelle. Tarvittavat tunnusluvut saadaan seuraavalla SAS-koodilla.

```
options ls=64 nodate nonumber;
data a;
   input kunta $13. kylla ei;
cards;
artjärvi      357 614
askola        1137 1158
hanko         4036 2076
hyvinkää      14639 8575
inkoo         1563 1308
järvenpää     12440 6023
karjaa        3004 1906
karjalohja    390 422
karkkila      2523 2486
kauniainen    4704 653
kerava        11200 5008
kirkkonummi   10316 4558
lapinjärvi    860 1026
liljendahl    373 510
lohja         5416 3069
lohjan_kunta  6230 4302
loviisa       3331 1227
myrskylä      479 676
mäntsälä      4116 3868
nummi-pusula  1378 1870
nurmijärvi    10132 6310
orimattila    4030 3538
pernaja       1207 966
pohja         1653 1236
pornainen     946 997
porvoo        8651 3165
porvoon_mlk   7683 4401
pukkila       360 637
ruotsinpyhtää 1030 778
sammatti      403 346
sipoo         5646 3123
siuntio       1474 1036
tammisaari    5307 2987
tuusula       9725 5641
vihti         7442 4933
;
data sasuser.aa;
   set a;
   keep kunta kylla ei hyv k_pros;
   hyv=kylla+ei;
   k_pros=kylla/hyv;
proc plot data=sasuser.aa;
   plot hyv*k_pros;
       vpos=20 hpos=20;
title 'EU-äänestys Uudellamaalla';
proc print data=sasuser.aa;
   id kunta;
proc means data=sasuser.aa;
run;
```

Ohjelmassa lasketaan annettujen kyllä- ja ei-äänien perusteella `means`-proseduurilla muuttujille keskiarvo ja keskihajonta. Muuttuja `hyv` on hyväksyttyjen äänien lukumäärä ja `k_pros` on kyllä-äänten prosentuaalinen osuus. Kyllä-äänten määrä esitetään graafisesti kyllä-äänten prosentuaalisen osuuden suhteen.

```
      EU-äänestys Uudellamaalla

Plot of HYV*K_PROS.

Legend: A = 1 obs, B = 2 obs,
etc.

    HYV |
  30000 +
        |
        |
        |
        |         A
        |
  20000 +
        |          A
        |         A A
        |         A A
        |
        |        AA  A
  10000 +        A
        |      AA C
        |         A
        |      A  A  A  A
        |    A  AB
        |   BBABAA
      0 +      AA
        -+-----+-----+-----+-
       0.25  0.50  0.75  1.00

                K_PROS
```

```
                The SAS System

KUNTA           KYLLA      EI       HYV       K_PROS

artjärvi          357     614       971     0.36766
askola           1137    1158      2295     0.49542
hanko            4036    2076      6112     0.66034
hyvinkää        14639    8575     23214     0.63061
inkoo            1563    1308      2871     0.54441
järvenpää       12440    6023     18463     0.67378
karjaa           3004    1906      4910     0.61181
karjalohja        390     422       812     0.48030
karkkila         2523    2486      5009     0.50369
kauniainen       4704     653      5357     0.87810
kerava          11200    5008     16208     0.69102
kirkkonummi     10316    4558     14874     0.69356
lapinjärvi        860    1026      1886     0.45599
liljendahl        373     510       883     0.42242
lohja            5416    3069      8485     0.63830
lohjan_kunta     6230    4302     10532     0.59153
loviisa          3331    1227      4558     0.73080
myrskylä          479     676      1155     0.41472
mäntsälä         4116    3868      7984     0.51553
nummi-pusula     1378    1870      3248     0.42426
nurmijärvi      10132    6310     16442     0.61623
orimattila       4030    3538      7568     0.53251
pernaja          1207     966      2173     0.55545
pohja            1653    1236      2889     0.57217
pornainen         946     997      1943     0.48688
porvoo           8651    3165     11816     0.73214
porvoon_mlk      7683    4401     12084     0.63580
pukkila           360     637       997     0.36108
ruotsinpyhtää    1030     778      1808     0.56969
sammatti          403     346       749     0.53805
sipoo            5646    3123      8769     0.64386
siuntio          1474    1036      2510     0.58725
tammisaari       5307    2987      8294     0.63986
tuusula          9725    5641     15366     0.63289
vihti            7442    4933     12375     0.60137
```

```
                EU-äänestys Uudellamaalla

Variable  N       Mean       Std Dev      Minimum
-----------------------------------------------------
KYLLA    35    4405.17       4022.40   357.0000000
EI       35    2612.26       2078.26   346.0000000
HYV      35    7017.43       6015.52   749.0000000
K_PROS   35  0.5751287     0.1116038     0.3610832
-----------------------------------------------------

                Variable       Maximum
                ----------------------
                KYLLA         14639.00
```

```
EI              8575.00
HYV            23214.00
K_PROS        0.8781034
----------------------
```

Havaitaan, että Uudenmaan läänin 35:n kunnan kyllä-äänten keskiarvo on 4405.17 ja keskihajonta 4022.4.

Määritetään nyt tarvittava otoskoko 95%:n luotettavuusvälillä. Halutaan tarkastella kuntia, joiden kyllä-äänten keskiarvopoikkeama on korkeintaan 4000. Populaation keskihajontana käytetään muuttujalle kyllä saatua arvoa. Otoskoko saadaan käyttäen hyväksi aiempaa otoskoon määrittämistä.

```
options ls=77 nodate nonumber;
data otos (drop =  z95);
   z95=1.96;
     n95=ceil((z95*4022.4/4000)**2);
proc print data=otos;
title 'Otoskoko 95%:n luottamustasolla';
run;
```

Saatu tulos on hyvin yksiselitteinen.

```
     Otoskoko 95%:n luottamustasolla

               OBS    N95

                1      4
```

Neljän kunnan mukaanotto riittää siihen että, otokseen mukaan tulevien kuntien kyllä-äänten keskiarvo poikkeaa kaikista 35:n kunnan kyllä-äänten keskiarvosta korkeintaan 4000. Virhetodennäköisyys on korkeintaan 5%.

Tehdään tämän tiedon perusteella satunnaisotos tutkittavasta aineistosta. Sitä varten annetaan satunnaislukuja tuottavalle funktiolle `ranuni` jokin positiivinen luku siemenluvuksi. Se palauttaa arvon nollan ja ykkösen väliltä. Tämän arvon on oltava pienempi tai yhtäsuuri kuin otoksen koon suhde koko populaatioon eli 4/35. Ratkaisu ei tietystikään ole yksikäsitteinen. Otannan koko on noin neljä. Kyseessä ei ole tarkka arvo otoskoolle. Vaihtamalla siemenlukua saadaan erilainen otanta.

```
options ls=77 nodate nonumber;
data tulos;
   set sasuser.aa;
   if ranuni(7599)<=(4/35);
proc print data=tulos;
   id kunta;
proc means data=tulos;
run;
```

Tällä siemenluvulla saadaan otannaksi seuraavat neljä kuntaa.

```
                Kuntien satunnaisotos

KUNTA             KYLLA      EI       HYV       K_PROS

karjaa             3004     1906      4910     0.61181
kerava            11200     5008     16208     0.69102
porvoo             8651     3165     11816     0.73214
ruotsinpyhtää      1030      778      1808     0.56969

              Kuntien satunnaisotos

Variable  N Mean       Std Dev       Minimum      Maximum
-----------------------------------------------------------
KYLLA     4 5971.25    4751.89        1030.00     11200.00
EI        4 2714.25    1813.54    778.0000000      5008.00
HYV       4 8685.50    6530.50        1808.00     16208.00
K_PROS    4 0.6511656 0.0737842     0.5696903    0.7321429
-----------------------------------------------------------
```

Saavutettu otos täyttää annetut vaatimukset. Otoskeskiarvo 5971.25 sisältyy vaaditulle luottamusvälille.

Otoskoon määrittäminen suhteen avulla

Mikäli tiedetään populaation tietyn kiinnostavuuden omaavien alkioiden määrän suhteen koko populaatioon vaihtelevan tietyissä rajoissa, voidaan tätä suhteen vaihtelua käyttää hyväksi otoskokoa määrättäessä. Olkoon tämä suhteen arvo p. Sen keskihajonta σ_p voidaan silloin ilmaista p:n ja otoskoon n avulla

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}.$$

Kun p on likimain normaalijakautunut, voidaan keskiarvopoikkeama d ilmaista p:n ja luottamustasoon liittyvän testiarvon z_c:n avulla, jolloin

$$d = z_c\sigma_p = \frac{z_c\sqrt{p(1-p)}}{\sqrt{n}}.$$

Tästä saadaan otoskoko n ratkaistua:

$$n = \frac{z_c^2(p(1-p))}{d^2}$$

Huomataan eo. kaavasta, että otoskoko on suurimmillaan, kun $p = 0.5$.

Käyttämällä hyväksi eo. kaavaa lasketaan seuraavassa otoskokoja, kun keskiarvopoikkeama vaihtelee 0.01:stä 0.2:een 0.01 välein. Tarkasteltava suhde olkoon 0.05. Otoskoot määritellään 95%:n ja 99%:n luottamusväleille.

```
options ls=64 nodate ;
data a (drop=z95 z99);
   z95=1.96;
   z99=2.58;
   do vali=.01 to .2  by .01;
      n95=ceil(0.0475*(z95**2)/(vali)**2);
      n99=ceil(0.0475*(z99**2)/(vali)**2);
      output;
    end;
proc print data=a;
  title 'Otoskoot 95%:n ja 99%:n luottamustasolla';
run;
```

Tulostuksessa huomataan tarkkuuden merkitys. Luottamusvälin muutos vaikuttaa merkittävästi otoskokoon.

```
Otoskoot 95%:n ja 99%:n luottamustasolla

      OBS    VALI     N95     N99

        1    0.01    1825    3162
        2    0.02     457     791
        3    0.03     203     352
```

```
 4    0.04     115     198
 5    0.05      73     127
 6    0.06      51      88
 7    0.07      38      65
 8    0.08      29      50
 9    0.09      23      40
10    0.10      19      32
11    0.11      16      27
12    0.12      13      22
13    0.13      11      19
14    0.14      10      17
15    0.15       9      15
16    0.16       8      13
17    0.17       7      11
18    0.18       6      10
19    0.19       6       9
20    0.20       5       8
```

Tarkka otoskoko

Aiemmassa satunnaisotantaa käsittelevässä esimerkissä otannan koko saattoi vaihdella siemenlukua vaihtelemalla. Seuraavassa esimerkissä käsitellään tarkalleen halutun suuruista otantaa. Siemenluvun vaihtaminen muuttaa ainoastaan mukaan tulevia kuntia, mutta ei niiden määrää. Tarkka otoskoko neljä saadaan saadaan seuraavilla SAS-ohjelmiston lauseilla:

```
options ls=64 nodate;
data tarkka (drop=k n);
   retain k 4 n;
   if _n_=1 then n=total;
   set sasuser.aa nobs=total;
   if ranuni(224890)<=k/n then
      do;
          output;
          k=k-1;
      end;
   n=n-1;
   if k=0 then stop;
run;

proc print data=tarkka;
   title 'Tarkka otoskoko';
proc means data=tarkka;
run;
```

Otoksen kokoa edustaa arvo k. Sille annetaan alkuarvoksi 4 `retain`-lauseessa. Havaintojen kokonaismäärä on n. Sitä ei tarvitse tuntea. Data-vaihe keskeytyy, kun tarvittava määrä kuntia on löytynyt.

Tuloksena saadaan seuraava otos:

```
                Tarkka otoskoko

KUNTA          KYLLA      EI       HYV       K_PROS

kauniainen      4704      653      5357     0.87810
lohja           5416     3069      8485     0.63830
porvoo          8651     3165     11816     0.73214
pukkila          360      637       997     0.36108

                Tarkka otoskoko

Variable  N           Mean        Std Dev        Minimum
-----------------------------------------------------------
KYLLA     4        4782.75        3412.30    360.0000000
EI        4        1881.00        1427.76    637.0000000
HYV       4        6663.75        4607.32    997.0000000
K_PROS    4      0.6524081      0.2178418      0.3610832
 -----------------------------------------------------------

                    Variable       Maximum
                    ----------------------
                    KYLLA          8651.00
                    EI             3165.00
                    HYV           11816.00
                    K_PROS        0.8781034
                    ----------------------
```

Tässäkin otannassa otoskeskiarvo 4782.75 sisältyy saadulle luottamusvälille.

Faktorianalyysista

Monimuuttujamenetelmiin kuuluva faktorianalyysi on menetelmä, jonka avulla selvitetään tutkittavassa muuttujajoukossa vallitsevat suuret linjat. Havaintoaineistosta muodostetaan uusia muuttujia, joita kutsutaan faktoreiksi. Niillä pyritään tiiviimmin ja paremmin havainnollistamaan tutkittavaa aineistoa.

Faktorianalyysin mallitus

Tarkastellaan havaintoja muuttujilla $X_1, X_2, \ldots, X_p$, joiden keskiarvovektori on μ ja kovarianssimatriisi σ. Oletetaan jatkossa, että $\mu = 0$.

Faktorianalyysissä oletetaan, että on olemassa m tutkittavaa faktoria (missä $m < p$) $f_1, f_2, \ldots, f_m$. Kukin muuttuja X_j on näiden faktoreiden f_j ja satunnaismuuttujan e_j lineaarinen funktio. Tällöin

$$X_j = \lambda_{j1} f_1 + \ldots + \lambda_{jm} f_m + e_j, \qquad (1)$$

missä $j = 1, \ldots, p$. Yhtälössä painoja λ_{jk} kutsutaan faktorin latauksiksi niin, että λ_{jk} on j:nnen muuttujan lataus k:nnelle faktorille. Termi e_j kuvaa j:nteen muuttujaan liittyvää satunnaisvaihtelua. Faktoreita f_i kutsutaan yhteisfaktoreiksi ja satunnaismuuttujia e_j erityisfaktoreiksi.

Yhtälö (1) kirjoitetaan usein matriisimuodossa, jolloin se on

$$X = \Lambda f + e,$$

missä $X^t = [X_1, X_2, \ldots, X_p]$, $f^t = [f_1, f_2, \ldots, f_m]$, $e^t = [e_1, e_2, \ldots, e_p]$ ja

$$\Lambda = \begin{pmatrix} \lambda_{11} & \ldots & \lambda_{1m} \\ \lambda_{21} & \ldots & \lambda_{2m} \\ \vdots & \ldots & \vdots \\ \lambda_{p1} & \ldots & \lambda_{pm} \end{pmatrix}.$$

Faktorianalyysin heikkoutena on suuri määrä oletuksia, joita joudutaan tekemään muuttujista ja faktoreista. Erityisfaktoreiden oletetaan olevan riippumattomia sekä toisistaan että yhteisfaktoreista. Yhteisfaktorit ovat riippumattomia toisistaan. Koska muuttujien X_i keskiarvo oletettiin nollaksi, niin myös faktoreilla tulee olla keskiarvo nolla. Yhteisfaktorit on mahdollista valita niin, että niiden varianssi on yksi. Tämä on mahdollista yhtälön (1) mukaan, koska jokaiseen yhteisfaktoriin liittyy skaalaustekijä.

Yksittäinen havainto voidaan kirjoittaa muodossa

$$x_{rj} = \sum_{k=1}^{m} \lambda_{jk} f_{rk} + e_{rj},$$

missä x_{rj} on r:s havainto muuttujalla j, f_{rk} on k:nnen yhteisfaktorin faktoripistemäärä r:nnelle havainnolle ja e_{rj} on j:nnen erikoisfaktorin arvo r:nnelle havainnolle. Pyritään siis löytämään sellaiset faktorit, että niihin liittyvät lataukset ovat mahdollisimman suuria.

Koska eri faktorit ovat keskenään riippumattomia, saadaan yhtälöstä (1) muuttujan X_j varianssi

$$\begin{aligned} Var(X_j) &= \lambda_{j1}^2 + \lambda_{j2}^2 + \ldots + \lambda_{jm}^2 + Var(e_j) \\ &= \sum_{k=1}^{m} \lambda_{jk}^2 + \psi_j, \end{aligned}$$

kun e_j:n varianssi on ψ_j.

Muuttujien X_i ja X_j väliseksi kovarianssiksi saadaan yhtälöstä (1)

$$Cov(X_i, X_j) = \sum_{k=1}^{m} \lambda_{ik} \lambda_{jk}$$

Siten matriisin X kovarianssimatriisi σ on

$$\sigma = \Lambda \Lambda^t + \Psi \qquad (2)$$

missä

$$\Psi = \begin{pmatrix} \psi_1 & 0 & \ldots & 0 \\ 0 & \psi_2 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & \psi_p \end{pmatrix}$$

Yhtälö (2) on tärkeä faktorianalyysissä. Se ilmentää, että faktorit selittävät tarkasti kovarianssimatriisin ei-lävistäjätermit. Faktorien latauksien löytäminen tarkoittaa siis X:n kovarianssimatriisin faktorisointia.

Latausten merkitys

Faktorointi voidaan suorittaa monimuuttujadataa, kovarianssimatriisia tai korrelaatiomatriisia hyväksi käyttäen. Useimmiten faktoroinnin perustana on korrelaatiomatriisi. Saatu ratkaisu ei ole yksikäsitteinen. Mikä hyvänsä faktoreiden ortogonaalinen rotaatio antaa uuden faktoreiden joukon, joka toteuttaa X:n kovarianssimatriisin yhtälön.

Rotaation tarkoituksena on löytää sellaiset lataukset, että ratkaisun selitys tulee mahdollisimman yksinkertaiseksi. Pyritään saamaan mahdollisimman paljon sekä suuria että pieniä latauksia. Suorakulmaisissa rotaatioratkaisuissa faktorit ovat toisistaan riippumattomia. Niistä lasketut faktoripistemäärät soveltuvat hyvin monien erilaisten jatkoanalyysien pohjaksi. Vinokulmaisissa rotaatioratkaisuissa tulkinta vaikeutuu erityisesti tapauksissa, joissa faktorit korreloituvat voimakkaasti.

Tunnetuimpia rotaatiomenetelmiä ovat

- quartimaxin pääperiaatteena on rotatoida faktoreiden latausten alkuarvot siten, etta yhdellä faktorilla on mahdollisimman suuri lataus ja muut hyvin lähellä nollaa.
- varimax on yleisesti käytetty rotaatio. Se keskittyy faktorin latausten muodostaman matriisin sarakkeiden yksinkertaistamiseen, kun quartimax pyrkii yksinkertaistamaan latausten muodostaman matriisin rivit.
- equimax on sekoitus kahdesta edellisesta rotaatiosta.
- oblique ei ole ortogonaalirotaatio, kuten edella olevat. Alkufaktoriakseleiden sallitaan rotatoituvan vapaasti parhaan tuloksen saavuttamiseksi.

Esimerkki faktorianalyysistä

Tarkastellaan 24 suurehkoa kaupunkia. Pyritään löytämään aineistosta kaupunkeja kuvaavia piirteitä. Alkuperäisen aineiston muuttujat ovat kaupungin nimi, veroäyrit, väkiluku, pinta-ala, vähittäiskauppojen lukumäärä ja menot yhteensä. Veroäyrien ja menojen luvut ovat tuhansia. Pinta-alojen yksikkö on km^2. Tiedot ovat vuodelta 1990. Aineisto esiintyi myös SuperMenu 2/93:ssa regressioanalyysiä käsitelleessä artikkelissa.

Esimerkki suoritetaan Cedarilla olevalla SAS-ohjelmistolla. Tarvittava ajovirta on seuraava:

```
options ls=64 nodate nonumber;
   data faktori;
      title 'Viisi kuntamuuttujaa';
input kunta $ ayrit vakiluku pintaala kaupat menot;
cards;
Espoo       15865942 172629 312  755 4221378
Turku       10821859 159180 243 1588 5464290
Oulu         6294649 101379 328  799 2826526
Kuopio       4738773  80613 779  626 2309382
Pori         4290482  76357 503  772 2179612
Jyväskylä    4217372  66526  97  645 2235552
Lappeenranta 3248469  54941 760  454 1465573
Joensuu      2683955  47554  82  490 1627738
Hyvinkää     2622471  40194 323  319  980674
Kokkola      1827460  34635 328  338  800710
Rovaniemi    2045501  33500  94  386  949659
Kouvola      2095537  31740  44  331  738415
Rauma        1846785  29755  51  335  811738
Savonlinna   1492233  28559 821  298  751348
Seinäjoki    1692645  27765 129  316  717265
Kemi         1541790  25374  91  229  801986
Riihimäki    1533298  25000 121  247  548036
Iisalmi      1168837  23979 763  219  557928
Kuusankoski  1472708  21788 114  125  435752
Salo         1344728  21660 144  305  511702
Tampere     11343169 172560 523 1614 5264481
Lahti        5688525  93551 135  840 2602631
Kotka        3298406  56634 268  487 1659990
Hämeenlinna  2760400  43417 167  414 1224405
;
   proc factor data=faktori simple corr;
      title2'Pääkomponenttianalyysi';


   proc factor data=faktori priors=smc msa
               rotate=promax reorder;
               title2 'PROMAX-rotaation käyttö';
run;
```

Ajovirrassa suoritetaan kaksi faktorointia. Ensimmäisessä vaiheessa halutaan tulostettavaksi keskiarvot, keskihajonnat ja korrelaatiot. Ne saadaan optioilla `simple` ja `corr`. Rotaatiota ei suoriteta.

Toisessa vaiheessa halutaan selvittää rotaation vaikutus. Sitä varten käytetään `promax`-rotaatiomenetelmää. Optiolla `priors=smc` selvitetään, kuinka suuren osan muuttujan vaihtelusta selittää kyseessä oleva faktorirakenne. Kunkin muuttujaparin väliset osittaiskorrelaatiokertoimet saadaan tulostettua määreellä `msa`. Määreellä `reorder` saadaan fakrorimatriisin rivit järjestettyä suuruusjärjestykseen. Ensimmäiseksi tulee muuttujat, joilla on korkein lataus ensimmäisellä faktorilla.

Tulostus on seuraava:

```
                Viisi kuntamuuttujaa
              Pääkomponenttianalyysi

Means and Standard Deviations from 24 observations

                  AYRIT    VAKILUKU    PINTAALA
   Mean      3997333.08  61220.4167  300.833333
   Std Dev   3719946.91  47239.5542  254.678086

                 KAUPAT       MENOT
   Mean      538.833333  1736948.79
   Std Dev   383.164588  1450286.78

                Correlations

             AYRIT  VAKILUKU  PINTAALA    KAUPAT     MENOT

AYRIT      1.00000   0.97332   0.12045   0.80046   0.92609
VAKILUKU   0.97332   1.00000   0.17733   0.90656   0.98193
PINTAALA   0.12045   0.17733   1.00000   0.14138   0.15127
KAUPAT     0.80046   0.90656   0.14138   1.00000   0.95930
MENOT      0.92609   0.98193   0.15127   0.95930   1.00000
                    Viisi kuntamuuttujaa
                  Pääkomponenttianalyysi

Initial Factor Method: Principal Components

              Prior Communality Estimates: ONE

 Eigenvalues of the Correlation Matrix:  Total = 5
                                      Average = 1

                              1           2           3
        Eigenvalue       3.8072      0.9702      0.2129
        Difference       2.8371      0.7572      0.2065
        Proportion       0.7614      0.1940      0.0426
        Cumulative       0.7614      0.9555      0.9981

                              4           5
        Eigenvalue       0.0064      0.0033
        Difference       0.0031
        Proportion       0.0013      0.0007
        Cumulative       0.9993      1.0000
```

Ensimmäisen faktoroinnin tulostuksesta nähdään, että suurin ominaisarvo selittää 76.1% standardoidusta varianssista. Kaksi suurinta selittävät yhdessä 95.6% vastaavasta varianssista. Kolmanneksi suurimman ominaisarvon mukaanotto lisää selitysasteen 99.8%. Nämä kolme komponenttia ovat riittäviä mihin sovellukseen hyvänsä.

```
1 factors will be retained by the MINEIGEN
                                 criterion.
```

```
           Factor Pattern

                     FACTOR1

          AYRIT      0.94988
          VAKILUKU   0.99407
          PINTAALA   0.20454
          KAUPAT     0.94183
          MENOT      0.99394

  Variance explained by each factor

                FACTOR1
               3.807247


Final Communality Estimates: Total = 3.807247

   AYRIT  VAKILUKU  PINTAALA    KAUPAT     MENOT
0.902281  0.988182  0.041837  0.887039  0.987908
```

Ainoassa mukaan kelpuutetussa faktorissa korkeimmat lataukset ovat väkiluvulla (99.41) ja menoilla (99.39). Lopulliset kommunaliteettiestimaatit osoittavat, että tällä yhdellä faktorilla tulevat neljä muuttujaa hyvin estimoiduiksi. Pinta-alaa ei sen sijaan saada lasketuksi riittävän hyvin yhdellä faktorilla.

Väkiluvun ja menojen hallitsemaa faktoria voidaan kutsua elinvoimaksi.

```
                    Viisi kuntamuuttujaa
                 PROMAX-rotaation käyttö

Initial Factor Method: Principal Factors

      Partial Correlations Controlling all other Variables

              AYRIT  VAKILUKU  PINTAALA    KAUPAT     MENOT

AYRIT       1.00000   0.86900  -0.49206  -0.72058   0.11872
VAKILUKU    0.86900   1.00000   0.51103   0.39684   0.36031
PINTAALA   -0.49206   0.51103   1.00000  -0.28231  -0.08377
KAUPAT     -0.72058   0.39684  -0.28231   1.00000   0.68254
MENOT       0.11872   0.36031  -0.08377   0.68254   1.00000
```

Jos data on sopiva yhteisfaktorimallille, niin osittaiskorrelaatiot ovat pienempiä kuin vastaavat tavalliset korrelaatiokeroimet. Toisen faktoroinnin tulostuksesta nähdään, että esimerkiksi väkiluvun ja menojen osittaiskorrelaatiokerroin on 0.36. Vastaava tavallinen korrelaatiokerroin näiden muuttujien välillä on 0.98.

```
Kaiser's Measure of Sampling Adequacy:
            Over-all MSA = 0.66571621

   AYRIT  VAKILUKU  PINTAALA    KAUPAT     MENOT
0.616470  0.679578  0.130852  0.662832  0.817611
```

Kaiserin `msa` on yhteenveto kullekin muuttujalle ja kaikille yhdessä, kuinka paljon pienempiä ovat osittaiskorrelaatiokertoimet kuin tavalliset korrelaatiokertoimet. Kaiserin `msa`:n arvot 0.8 tai suuremmat ovat hyviä ja alle 0.5 olevat kertoimet huonoja. Täten nähdään, että menojen arvo on hyvä. Pinta-ala arvollaan 0.13 on heikko. Kaikille muuttujille yhteinen `msa`-arvo 0.67 osoittaa, että lisämuuttujien mukaanottaminen saattaisi parantaa analyysin luotettavuutta.

```
        Prior Communality Estimates: SMC

   AYRIT  VAKILUKU  PINTAALA    KAUPAT     MENOT
0.989029  0.995264  0.287373  0.978799  0.991371
```

Muuttujien selitysastetta kuvaava `smc`-arvo osoittaa, että muuttujien äyrit, väkiluku, kauppojen liikevaihto ja kunnan menot tulevat hyvin selitetyiksi saadulla faktorirakenteella. Selitysasteet ovat kaikilla yli 97%. Sen sijaan muuttujan pinta-ala vaihtelusta faktorirakenne pystyy selittämään vain 28.7%.

```
        Eigenvalues of the Reduced Correlation Matrix:
            Total = 4.24183608  Average = 0.84836722

                         1          2          3
        Eigenvalue    3.7899     0.2680     0.1959
        Difference    3.5219     0.0721     0.1996
        Proportion    0.8935     0.0632     0.0462
        Cumulative    0.8935     0.9566     1.0028

                         4          5
        Eigenvalue   -0.0037    -0.0083
        Difference    0.0047
        Proportion   -0.0009    -0.0020
        Cumulative    1.0020     1.0000
```

Ominaisarvoissa on nähtävissä, että yksi on muihin verrattuna huomattavan suuri. Se korostaa yhden faktorin merkitystä.

```
3 factors will be retained by the PROPORTION criterion.

           Factor Pattern

          FACTOR1   FACTOR2   FACTOR3

VAKILUKU  0.99519   0.01588   0.09432
MENOT     0.99426  -0.01465  -0.07590
AYRIT     0.94997  -0.10066   0.28127
KAUPAT    0.93898   0.01198  -0.31430
PINTAALA  0.16390   0.50720   0.05810

     Variance explained by each factor

          FACTOR1   FACTOR2   FACTOR3
         3.789933  0.267994  0.195930
            Viisi kuntamuuttujaa
             PROMAX-rotaation käyttö

Initial Factor Method: Principal Factors

  Final Communality Estimates: Total = 4.253857

   AYRIT  VAKILUKU  PINTAALA    KAUPAT     MENOT
0.991682  0.999547  0.287488  0.980604  0.994536
```

Toisessa faktoroinnissa, jossa on käytetty rotaatiota, on korkeimmat lataukset ensimmäisellä faktorilla väkiluvulla (0.995) ja menoilla (0.994). Pinta-alalla on tässäkin tapauksessa pienin lataus (0.16). Toisella faktorilla sen sijaan suurin lataus on muuttujalla pinta-ala (0.507). Kolmannella faktorilla suurimmat lataukset itseisarvoltaan ovat muuttujilla kauppojen liikevaihto (-0.31) ja äyrit (0.28). Muuttujien selitysasteet ovat samaa luokkaa kuin edelläkin eli neljän muuttujan vaihtelut pystytään selittämään saaduilla yhteisfaktoreilla yli 98 prosenttisesti. Pinta-alan vaihtelut selittyvät näillä faktoreilla vain 28.7%.

Elinvoimaa kuvaava faktori on merkittävin tässäkin tapauksessa. Toista faktoria voidaan kutsua pinta-alan mukaan luontoystävälliseksi. Kolmas faktori kuvaa kulutusjuhlaa.

```
      Viisi kuntamuuttujaa
     PROMAX-rotaation käyttö

Prerotation Method: Varimax

   Orthogonal Transformation Matrix

          1         2         3

1      0.77077   0.61641   0.16108
2     -0.20230  -0.00296   0.97932
3      0.60414  -0.78742   0.12241

        Rotated Factor Pattern

           FACTOR1   FACTOR2   FACTOR3

AYRIT      0.92250   0.36439   0.08888
VAKILUKU   0.82084   0.53912   0.18741
MENOT      0.72346   0.67268   0.13652
KAUPAT     0.53143   0.82624   0.12451
PINTAALA   0.05882   0.05377   0.53022
```

Varimax-rotaation tärkein faktori on elinvoima. Kulutusjuhla edustaa tässä toista faktoria. Luontoystävällisyys on nyt kolmantena faktorina.

```
        Variance explained by each factor

          FACTOR1   FACTOR2   FACTOR3
         2.334054  1.561504  0.358299


Final Communality Estimates: Total = 4.253857

   AYRIT  VAKILUKU  PINTAALA    KAUPAT     MENOT
0.991682  0.999547  0.287488  0.980604  0.994536

             Viisi kuntamuuttujaa
            PROMAX-rotaation käyttö

Rotation Method: Promax

Target Matrix for Procrustean Transformation

           FACTOR1   FACTOR2   FACTOR3

AYRIT      1.00000   0.08434   0.00074
VAKILUKU   0.69619   0.26994   0.00681
MENOT      0.48026   0.52834   0.00265
KAUPAT     0.19443   1.00000   0.00206
PINTAALA   0.00166   0.00174   1.00000

       Procrustean Transformation Matrix

                1         2         3

        1      1.30817  -0.57272  -0.16794
        2     -0.62696   1.52115  -0.21050
        3     -0.08779  -0.11019   1.89729

    Normalized Oblique Transformation Matrix

                1         2         3

        1      0.62229   0.43949   0.02544
        2     -0.35712   0.00316   1.03736
        3      1.30381  -1.43044   0.16253

              Inter-factor Correlations

           FACTOR1   FACTOR2   FACTOR3

FACTOR1    1.00000   0.73802   0.27523
FACTOR2    0.73802   1.00000   0.29312
FACTOR3    0.27523   0.29312   1.00000

   Rotated Factor Pattern (Std Reg Coefs)

            FACTOR1   FACTOR2   FACTOR3

 AYRIT      0.99382   0.01485  -0.03453
 VAKILUKU   0.73660   0.30250   0.05713
 KAUPAT     0.17024   0.86230  -0.01476
 MENOT      0.52499   0.54550  -0.00223
 PINTAALA  -0.00339  -0.00947   0.53976
```

Tässäkin tapauksessa faktorit noudattavat samaa järjestystä kuin edellä: elinvoima, kulutusjuhla ja luontoystävällisyys.

```
            Reference Axis Correlations

                  FACTOR1   FACTOR2   FACTOR3

        FACTOR1    1.00000  -0.71517  -0.09130
        FACTOR2   -0.71517   1.00000  -0.13872
        FACTOR3   -0.09130  -0.13872   1.00000
                 Viisi kuntamuuttujaa
                PROMAX-rotaation käyttö

Rotation Method: Promax

   Reference Structure (Semipartial Correlations)

                  FACTOR1   FACTOR2   FACTOR3

         AYRIT      0.66781   0.00992  -0.03288
         VAKILUKU   0.49496   0.20215   0.05439
         KAUPAT     0.11440   0.57624  -0.01406
         MENOT      0.35277   0.36453  -0.00213
         PINTAALA  -0.00228  -0.00633   0.51390

Variance explained by each factor eliminating
other factors

           FACTOR1   FACTOR2   FACTOR3
         0.828495  0.505932  0.268335

      Factor Structure (Correlations)

                  FACTOR1   FACTOR2   FACTOR3

         AYRIT      0.99527   0.73819   0.24335
         VAKILUKU   0.97558   0.86288   0.34853
         KAUPAT     0.80258   0.98362   0.28485
         MENOT      0.92696   0.93230   0.30215
         PINTAALA   0.13818   0.14624   0.53605

Variance explained by each factor ignoring
other factors

            FACTOR1   FACTOR2   FACTOR3
          3.464811  3.147541  0.640484


   Final Communality Estimates: Total = 4.253857

   AYRIT  VAKILUKU  PINTAALA    KAUPAT     MENOT
0.991682  0.999547  0.287488  0.980604  0.994536
```

Faktorianalyysin tulosta ei yleensä voi pitää lopullisena. Sen sijaan se antaa hyvän pohjan jatkoanalyyseille.

Kanonisesta analyysista

Kanonisella analyysillä voidaan tutkia kahden muuttujajoukon välistä yhteyttä. Kumpikin joukko voi sisältää useita muuttujia. Kanonista analyysiä voidaan pitää regressioanalyysin erikoistapauksena, vaikkakin se poikkeaa sekä matemaattisessa mielessä että tulkinnan suhteen selvästi regressioanalyysistä.

Kanoniset muuttujat

Kanonisella analyysillä lasketaan kanonisia korrelaatioita kahden muuttujaryhmän välillä. Kummankin muuttujaryhmän sisällä muodostetaan sellaiset lineaarikombinaatiot, että niiden mukaan laskettujen kanonisten muuttujien välinen korrelaatio on mahdollisimman suuri.

Olkoon aineistossa selittävät muuttujat $x_1, x_2, \ldots, x_p$ ja selitettävät muuttujat $y_1, y_2, \ldots, y_q$. Kanoniset muuttujat voidaan silloin laskea seuraavasti:

$$
\begin{aligned}
\hat{x}_1 &= a_{11}x_1 + a_{21}x_2 + \ldots + a_{p1}x_p \\
\hat{x}_2 &= a_{12}x_1 + a_{22}x_2 + \ldots + a_{p2}x_p \\
&\vdots \\
\hat{x}_s &= a_{1s}x_1 + a_{2s}x_2 + \ldots + a_{ps}x_p \\
\hat{y}_1 &= b_{11}y_1 + b_{21}y_2 + \ldots + b_{q1}y_q \\
\hat{y}_2 &= b_{12}y_1 + b_{22}y_2 + \ldots + b_{q2}y_q \\
&\vdots \\
\hat{y}_s &= b_{1s}y_1 + b_{2s}y_2 + \ldots + b_{qs}y_q
\end{aligned}
$$

missä s on $s = min(p,q)$.

Tehtävänä on löytää sellaiset kanoniset kertoimet a_{ij} ja b_{ij}, että korrelaatio $\hat{x}_i$:n ja $\hat{y}_j$:n välillä olisi mahdollisimman suuri. Suurin löydetty korrelaatio on ensimmäinen kanonien korrelaatio. Usein on hyödyllistä normalisoida kanoniset kertoimet siten, että kunkin kanonisen muuttujan varianssi on 1.

Vastaavalla tavalla etsitään muita kanonisia korrelaatiokertoimia. Vaatimuksena uusille kanonisille korrelaatiokertoimille on se, että ne eivät saa korreloida edellisten kanonisten korrelaatiokertoimien kanssa. Niitä löydetään s kappaletta. Toisesta eteenpäin olevat kanoniset korrelaatiokertoimet saattavat kuitenkin olla hyvin pieniä.

Esimerkki kanonisesta korrelaatiosta

Tarkastellaan seuraavassa esimerkissä kolmen presidenttiehdokkaan kannatusta vuoden 1994 presidentinvaalien ensimmäisellä kierroksella sekä kolmea kuntatietoja kuvaavaa muuttujaa vuodelta 1990. Pyritään selvittämään, löytyykö näiden eri joukkojen muuttujien väliltä riippuvuutta.

Tarvittava SAS-ajovirta on seuraava:

```
options ls=64 nodate nonumber;
data a;
   input kunta $ vakiluku pintaala kaupat rehn
         vayrynen ahtisaar;
   label vakiluku='Väkiluku'
         pintaala='Pinta-ala'
         kaupat = 'Kaupat'
         rehn   = 'Rehn'
         vayrynen='Väyrynen'
         ahtisaar='Ahtisaari';
cards;
Espoo       172629 312  755 38.2  6.5 22.2
Turku       159180 243 1588 22.1  8.7 29.6
Oulu        101379 328  799 17.1 21.9 28.6
Kuopio       80613 779  626 23.6 19.2 26.0
Pori         76357 503  772 16.1 12.2 32.5
Jyväskylä    66526  97  645 17.6 15.8 31.7
Lappeenranta 54941 760  454 20.5 17.8 31.7
Joensuu      47554  82  490 18.5 13.6 31.4
Hyvinkää     40194 323  319 24.9 10.3 31.5
Kokkola      34635 328  338 27.8 22.3 22.6
Rovaniemi    33500  94  386 17.5 29.3 23.3
Kouvola      31740  44  331 24.2 11.5 27.8
Rauma        29755  51  335 18.4 10.1 37.2
Savonlinna   28559 821  298 16.3 19.2 34.4
Seinäjoki    27765 129  316 14.9 28.5 22.8
Kemi         25374  91  229  7.0 35.0 30.0
Riihimäki    25000 121  247 20.8  9.0 36.5
Iisalmi      23979 763  219 15.0 27.0 21.9
Kuusankoski  21788 114  125 20.5 10.3 40.6
Salo         21660 144  305 19.0 13.6 33.7
Tampere     172560 523 1614 21.9  8.3 28.4
Lahti        93551 135  840 22.6  9.3 31.1
Kotka        56634 268  487 21.8  9.2 39.6
Hämeenlinna  43417 167  414 20.0  8.3 32.6
;
proc cancorr data=a all
   vprefix=ehdokas vname='Candidates'
   wprefix=kunnat wname='Communal aspects';
 var rehn vayrynen ahtisaar;
 with vakiluku pintaala kaupat;
title 'Rehnin, Väyrysen ja Ahtisaaren kannatus v.
       1994';
title2 'tarkasteltuna v. 1990 kuntatietojen
        valossa';
run;
```

Tarkasteluun on otettu mukaan 24 kaupunkia. Pinta-alan yksikkö on km^2. Muuttuja kaupat kuvaa vähittäiskauppojen lukumäärää. Määrellä `vprefix` nimetään juoksevalla numerolla `var-` lauseesta saatavat kanoniset muuttujat. Vastaavasti `wprefix` nimeää `with`-lauseesta saatavat kanoniset muuttujat.

Tulostuksena saadaan seuraava listaus:

```
Rehnin, Väyrysen ja Ahtisaaren kannatus v. 1994
    tarkasteltuna v. 1990 kuntatietojen valossa
```

Means and Standard Deviations

3 Candidates
3 Communal aspects
24 Observations

Variable	Mean	Std Dev	Label
REHN	20.262500	5.669661	Rehn
VAYRYNEN	15.704167	7.966096	Väyrynen
AHTISAAR	30.320833	5.344073	Ahtisaari
VAKILUKU	61220	47240	Väkiluku
PINTAALA	300.833333	254.678086	Pinta-ala
KAUPAT	538.833333	383.164588	Kaupat

Rehnin, Väyrysen ja Ahtisaaren kannatus v. 1994
tarkasteltuna v. 1990 kuntatietojen valossa

Correlations Among the Original Variables

Correlations Among the Candidates

	REHN	VAYRYNEN	AHTISAAR
REHN	1.0000	-0.5984	-0.2006
VAYRYNEN	-0.5984	1.0000	-0.4993
AHTISAAR	-0.2006	-0.4993	1.0000

Suurin korrelaatio presidenttiehdokkaiden välillä on Rehnillä ja Väyrysellä -0.5984.

Correlations Among the Communal aspect

	VAKILUKU	PINTAALA	KAUPAT
VAKILUKU	1.0000	0.1773	0.9066
PINTAALA	0.1773	1.0000	0.1414
KAUPAT	0.9066	0.1414	1.0000

Kunnallisten muuttujien joukossa suurin korrelaatio 0.9066 on kauppojen lukumäärän ja väkiluvun välillä.

Correlations Between the Candidates and the Communal aspect

	VAKILUKU	PINTAALA	KAUPAT
REHN	0.4934	0.0410	0.2389
VAYRYNEN	-0.4108	0.1398	-0.3665
AHTISAAR	-0.2531	-0.2091	-0.1555

Kunnallisten muuttujien ja presidenttiehdokkaiden välillä suurin korrelaatio 0.4934 löytyy Rehnin ja väkiluvun väliltä.

Rehnin, Väyrysen ja Ahtisaaren kannatus v. 1994
tarkasteltuna v. 1990 kuntatietojen valossa

Canonical Correlation Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approx Standard Error	Squared Canonical Correlation
1	0.718639	0.596172	0.100829	0.516442
2	0.655755	.	0.118850	0.430014
3	0.194983	.	0.200587	0.038018

Eigenvalues of INV(E)*H
= CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative
1	1.0680	0.3136	0.5736	0.5736
2	0.7544	0.7149	0.4052	0.9788
3	0.0395	.	0.0212	1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approx F	Num DF	Den DF	Pr > F
1	0.26514240	3.5430	9	43.95788	0.0022
2	0.54831597	3.3295	4	38	0.0197
3	0.96198174	0.7904	1	20	0.3845

Multivariate Statistics and F Approximations

S=3 M=-0.5 N=8

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.265142	3.543	9	43.958	0.0022
Pillai's Trace	0.984475	3.2563	9	60	0.0028
Hotelling-Lawley Trace	1.861956	3.4481	9	50	0.0023
Roy's Greatest Root	1.068006	7.12	3	20	0.0019

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

Ensimmäinen kanoninen korrelaatio on 0.718639. Se on selvästi suurempi kuin korrelaatio Rehnin ja väkiluvun välillä. Todennäköisyys, että kaikki kanoniset korrelaatiot ovat nollia, on vain 0.0022. Ensimmäisen kanonisen korrelaatiokertoimen olemassaolo on tilastollisesti merkitsevä.

Muut kanoniset korrelaatiokertoimet voidaan jättää tarkastelun ulkopuolelle, koska niiden sovitettuja kanonisia korrelaatiokertoimia (`Adjusted Canonical Correlations`) ei ole saatu.

Rehnin, Väyrysen ja Ahtisaaren kannatus v. 1994
tarkasteltuna v. 1990 kuntatietojen valossa

Canonical Correlation Analysis

Raw Canonical Coefficients for the Candidates

	EHDOKAS1	EHDOKAS2	EHDOKAS3	
REHN	0.1826817135	0.2501070806	-0.067161458	Rehn
VAYRYNEN	0.0186515454	0.2542642095	0.0071563005	Väyrynen
AHTISAAR	-0.024126307	0.2529168149	-0.179244551	Ahtisaari

Raw Canonical Coefficients for the Communal aspect

	KUNNAT1	KUNNAT2	KUNNAT3	
VAKILUKU	0.000048046	0.0000137234	-7.345343E-6	Väkiluku
PINTAALA	-0.000112061	0.0007607871	0.0039193785	Pinta-ala
KAUPAT	-0.004585904	-0.004047994	0.0009489232	Kaupat

Standardized Canonical Coefficients for the Candidates

	EHDOKAS1	EHDOKAS2	EHDOKAS3	
REHN	1.0357	1.4180	-0.3808	Rehn
VAYRYNEN	0.1486	2.0255	0.0570	Väyrynen
AHTISAAR	-0.1289	1.3516	-0.9579	Ahtisaari

Standardized Canonical Coefficients for the Communal aspect

	KUNNAT1	KUNNAT2	KUNNAT3	
VAKILUKU	2.2697	0.6483	-0.3470	Väkiluku
PINTAALA	-0.0285	0.1938	0.9982	Pinta-ala
KAUPAT	-1.7572	-1.5510	0.3636	Kaupat

Koska muuttujat on mitattu eri yksiköillä, on parempi tulkita standardisoituja kertoimia kuin raakaker-

toimia.

Presidenttiehdokkaiden ensimmäinen kanonien muuttuja painottuu selvästi Rehniin. Se on

$$\text{ehdokas1} = 1.0357 * \text{Rehn} + 0.1486 * \text{Väyrynen} - 0.1289 * \text{Ahtisaari}.$$

Vastaavasti kuntatietojen ensimmäinen kanoninen muuttuja on

$$\text{kunnat1} = 2.2697 * \text{Väkiluku} - 0.0285 * \text{Pinta-ala} - 1.7572 * \text{Kaupat}.$$

Rehnin, Väyrysen ja Ahtisaaren kannatus v. 1994
tarkasteltuna v. 1990 kuntatietojen valossa

Canonical Structure

Correlations Between the Candidates
and Their Canonical Variables

	EHDOKAS1	EHDOKAS2	EHDOKAS3	
REHN	0.9727	-0.0651	-0.2228	Rehn
VAYRYNEN	-0.4068	0.5021	0.7632	Väyrynen
AHTISAAR	-0.4109	0.0558	-0.9100	Ahtisaari

Rehnin korrelaatio presidenttiehdokkaiden joukosta muodostetun ensimmäisen kanonisen muuttujan kanssa on 0.9727. Vastaavat korrelaatiot Väyrysellä ja Ahtisaarella ovat -0.4068 ja -0.4109. Väyrynen toimii vaimentavana muuttujana, koska Väyrysen kanoninen kerroin (0.1486) ja korrelaatio kanonisen muuttujan kanssa (-0.4068) ovat eri merkkiset.

Correlations Between the Communal aspect
and Their Canonical Variables

	KUNNAT1	KUNNAT2	KUNNAT3	
VAKILUKU	0.6716	-0.7235	0.1596	Väkiluku
PINTAALA	0.1255	0.0894	0.9881	Pinta-ala
KAUPAT	0.2964	-0.9359	0.1901	Kaupat

Kuntatietojen joukossa on väkiluvun korrelaatio ensimmäisen kanonisen muuttujan kanssa 0.6716.

Kauppojen lukumäärä toimii kuntatietojen vaimentavana tekijänä, koska sen korrelaation kunnallisten kanonisten muuttujien kanssa on 0.2934 ja vastaava kanoninen kerroin on -1.7572. Pinta-alakin on vaimentava tekijä, mutta sen merkitys on vähäinen.

Correlations Between the Candidates and the
Canonical Variables of the Communal aspect

	KUNNAT1	KUNNAT2	KUNNAT3	
REHN	0.6990	-0.0427	-0.0434	Rehn
VAYRYNEN	-0.2923	0.3292	0.1488	Väyrynen
AHTISAAR	-0.2953	0.0366	-0.1774	Ahtisaari

Correlations Between the Communal aspect and
the Canonical Variables of the Candidates

	EHDOKAS1	EHDOKAS2	EHDOKAS3	
VAKILUKU	0.4827	-0.4744	0.0311	Väkiluku
PINTAALA	0.0902	0.0586	0.1927	Pinta-ala
KAUPAT	0.2130	-0.6137	0.0371	Kaupat

Rehnin, Väyrysen ja Ahtisaaren kannatus v. 1994
tarkasteltuna v. 1990 kuntatietojen valossa

Canonical Redundancy Analysis

Raw Variance of the Candidates
Explained by

	Their Own Canonical Variables			The Opposite Canonical Variables	
	Proportion	Cumulative Proportion	Canonical R-Square	Proportion	Cumulative Proportion
1	0.3684	0.3684	0.5164	0.1902	0.1902
2	0.1306	0.4990	0.4300	0.0562	0.2464
3	0.5010	1.0000	0.0380	0.0190	0.2655

Raw Variance of the Communal aspect
Explained by

	Their Own Canonical Variables			The Opposite Canonical Variables	
	Proportion	Cumulative Proportion	Canonical R-Squared	Proportion	Cumulative Proportion
1	0.4511	0.4511	0.5164	0.2330	0.2330
2	0.5234	0.9745	0.4300	0.2251	0.4580
3	0.0255	1.0000	0.0380	0.0010	0.4590

Standardized Variance of the Candidates
Explained by

	Their Own Canonical Variables			The Opposite Canonical Variables	
	Proportion	Cumulative Proportion	Canonical R-Squared	Proportion	Cumulative Proportion
1	0.4268	0.4268	0.5164	0.2204	0.2204
2	0.0865	0.5133	0.4300	0.0372	0.2576
3	0.4867	1.0000	0.0380	0.0185	0.2761

Presidenttiehdokkaiden ensimmäinen kanoninen muuttuja selittää kuntatietoja vain 22.04%. Toinen ja kolmas kanoninen muuttuja eivät tuo oleellista lisää. Kaikki kolme kanonista muuttujaa selittävät kuntatietoja vain 27.61%.

Rehnin, Väyrysen ja Ahtisaaren kannatus v. 1994
tarkasteltuna v. 1990 kuntatietojen valossa

Canonical Redundancy Analysis

Standardized Variance of the Communal aspect
Explained by

	Their Own Canonical Variables			The Opposite Canonical Variables	
	Proportion	Cumulative Proportion	Canonical R-Squared	Proportion	Cumulative Proportion
1	0.1849	0.1849	0.5164	0.0955	0.0955
2	0.4691	0.6540	0.4300	0.2017	0.2972
3	0.3460	1.0000	0.0380	0.0132	0.3104

Kuntatietojen kanoniset muuttujat selittävät myös huonosti vastapuolen muuttujia. Ensimmäinen kuntatietojen kanoninen muuttuja selittää presidenttiehdokkaita 9.55%. Toinen kanoninen muuttuja tuo se-

littämiseen lisäarvoa. Kuitenkin kaikki kolme kuntatietojen kanonista muuttujaa selittävät yhdessä vain 31.04% presidenttiehdokkaista.

```
Squared Multiple Correlations Between the Candidates and
the First 'M' Canonical Variables of the Communal aspect

M           1            2            3

REHN      0.4886       0.4904       0.4923 Rehn
VAYRYNEN 0.0855       0.1939       0.2160 Väyrynen
AHTISAAR 0.0872       0.0885       0.1200 Ahtisaari
```

Kuntatietojen 1. kanoninen muuttuja selittää hyvin Rehniä (48.86%). Sen sijaan Väyrystä ja Ahtisaarta se ei selitä ollenkaan.

```
Squared Multiple Correlations Between the Communal aspect
and the First 'M' Canonical Variables of the Candidates

M           1            2            3

VAKILUKU 0.2330       0.4580       0.4590 Väkiluku
PINTAALA 0.0081       0.0116       0.0487 Pinta-ala
KAUPAT   0.0454       0.4221       0.4234 Kaupat
```

Presidenttiehdokkaiden ensimmäinen kanonien muuttuja selittää heikosti väkilukua (23.3%). Pinta-alaa ja kauppojen lukumäärää se ei selitä ollenkaan.

Yhteenvetona voidaan todeta, että Väyrynen ja kauppojen lukumäärä toimivat vaimentavina tekijöinä. Ne vahvistavat yhdessä Rehnin ja väkiluvun korrelaatiota.

Erotteluanalyysista

Erotteluanalyysissä tutkitaan, onko ryhmien välillä merkittäviä eroavaisuuksia joidenkin muuttujien suhteen. Lähtökohtana on, että on olemassa vähintään kaksi ryhmää, joissa on mitattu ainakin kahden muuttujan arvoja. Jos ryhmät ovat erillisiä, halutaan tietää:

- a) ne muuttujat, joiden osalta ryhmät eroavat
- b) eroavatko ryhmät joidenkin muuttujien avulla paremmin.

Erottelufunktion muodostaminen

Erotteluanalyysissä pyritään löytämään jokin erottelufunktio, jonka avulla yksilö sijoitetaan oikeaan ryhmään. Oletetaan, että on olemassa kaksi ryhmää ja kummassakin ryhmässä p-dimensioinen muuttujavektori $\bar{x}$ saa arvoja eri tiheysfunktioiden $f_i(\bar{x}), i = 1,2$ mukaisesti.

Silloin yksilön ryhmä voidaan määrittää seuraavalla päättelysäännöllä:

$$\frac{f_1(\bar{x})}{f_2(\bar{x})} \geq \frac{\pi_2 C(1|2)}{\pi_1 C(2|1)} \qquad (1)$$

Jos epäyhtälö (1) on voimassa, niin yksilö kuuluu populaatioon 1. Muuten se kuuluu populaatioon 2. Kaavassa (1) π_i on prioritodennäköisyys sille, että yksilö kuuluu ryhmään i ja $C(i|j)$ on kustannus sille, että populaation j yksilö luokitellaan väärään populaatioon i. Kun

$$\pi_2 C(1|2) = \pi_1 C(2|1),$$

niin päätössääntö sijoittaa yksilön siihen populaatioon, mille todennäköisyys on suurempi.

Kun populaatiot ovat normaalijakautuneita, saadaan i:nnelle populaatiolle

$$X \sim N_p(\bar{\mu}_i, \Sigma),$$

missä $i = 1,2$. Kovarianssit oletetaan yhtäsuuriksi. Tällöin tiheysfunktioiksi saadaan

$$f_i(\bar{x}) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp(-\frac{1}{2}(\bar{x}-\bar{\mu}_i)^T \Sigma^{-1}(\bar{x}-\bar{\mu}_i))$$

Kirjoittamalla

$$\begin{aligned} L &= \Sigma^{-1}(\bar{\mu}_1 - \bar{\mu}_2) \text{ ja} \\ k &= \ln((\pi_2 C(1|2))/(\pi_1 C(2|1))). \end{aligned}$$

saadaan epäyhtälön (1) mukainen päättelysääntö muotoon

$$L^T x - \frac{1}{2} L^T(\bar{\mu}_1 + \bar{\mu}_2) \geq k \Rightarrow \text{kuuluu populaatioon 1}$$

$$L^T x - \frac{1}{2} L^T(\bar{\mu}_1 + \bar{\mu}_2) < k \Rightarrow \text{kuuluu populaatioon 2}$$

Päätössääntö perustuu Fisherin lineaariseen erottelufunktioon. Huom! Eo. epäyhtälöissä $k = 0$, kun $\pi_2 C(1|2) = \pi_1 C(2|1)$.

Määritellään

$$U = L^T X - \frac{1}{2} L^T(\bar{\mu}_1 + \bar{\mu}_2)$$

ja oletetaan, että $L^T \bar{\mu}_1 > L^T \bar{\mu}_2$. Jos X kuuluu populaatioon 1, niin

$$U = U_1 \sim N_1(\frac{1}{2}\alpha, \alpha)$$

Vastaavasti X:n ollessa peräisin populaatiosta 2 saadaan

$$U = U_2 \sim N_1(-\frac{1}{2}\alpha, \alpha)$$

missä

$$\begin{aligned} \alpha &= (\bar{\mu}_1 - \bar{\mu}_2)^T \Sigma^{-1}(\bar{\mu}_1 - \bar{\mu}_2) \\ &= L^T(\bar{\mu}_1 - \bar{\mu}_2) \end{aligned}$$

Edellä olevissa kaavoissa esiintyvä α on kahden populaation keskiarvojen eäisyyden mitta. Sitä kutsutaan `Mahalanobiksen etäisyydeksi`.

Yksilön ryhmä määritetään laskemalla erottelupistemäärät. Oikea ryhmä on se, johon liittyvä erottelupistemäärä on suurempi. Se saadaan lausekkeesta

$$w_i = -\Sigma \pi_j (C(i|j) f_j(\bar{x}),$$

missä $j = 1,2$. Vastaavasti useamman ryhmän tapauksessa oikea ryhmä on se, jolle w_i on suurin.

$$w_i = L_i^T \bar{x} - \frac{1}{2} L_i^T \bar{x}_i + \ln(\pi_i)$$

missä

$$\begin{aligned} m &= \text{populaatioiden lukumäärä} \\ i &= \text{1,2,...,m} \\ \bar{x}_i &= \text{populaation i keskiarvovektori} \\ L_i &= S^{-1}\bar{x}_i \\ S &= \text{ryhmien sisäisen hajonnan estimaatti} \end{aligned}$$

Esimerkki erotteluanalyysistä

Tarkastellaan presidentinvaalien 1. kierroksen prosenttuaalista ääniosuutta kolmen ehdokkaan kesken. Ehdokkaiksi on valittu eniten äänia saaneet ehdokkaat: Ahtisaari, Rehn ja Väyrynen. Heidän prosenttuaalista ääniosuutta tutkitaan kaikissa Kymen, Uudenmaan ja Lapin vaalipiirien kunnissa.

Jokaisella ehdokkaalla on vaalipiiri, missä he ovat keskimääräistä suositumpia. Tässä valinnassa Ahtisaaren suosikkialuetta on Kymen vaalipiiri, Rehnin Uusimaa ja Väyrysen Lappi.

Tarkastellaan nyt kaikkia valittujen vaalipiirien kuntia. Pyritään määrittämään, miten hyvin eri kunnat sopivat tähän jaotteluun. Yleisestä profiilista poikkeava kunta sijoitetaan siihen vaalipiiriin, mihin se äänestyskäyttäytymisen perusteella paremmin kuuluisi.

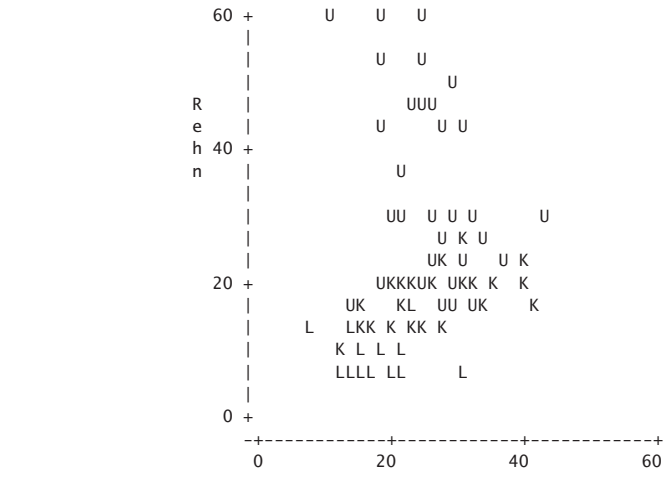
Analyysi suoritetaan Cedarilla olevalla sas-ohjelmistolla. Tarvittava ajovirta on seuraava:

```
options ls=64 nodate nonumber;
proc format;
   value lnimi
     1='Uusimaa'
     2='Kymi'
     3='Lappi';
   value lsym
     1='U'
     2='K'
     3='L';
run;

data sasuser.vaalit;
   input rehn vayrynen ahtisaar laani  @@;
   format laani lnimi.;
   label rehn='Rehn'
         vayrynen='Väyrynen'
         ahtisaar='Ahtisaari';
cards;
18.0 36.1 14.3 1 21.6 25.2 23.9 1
38.2 6.5 22.2 1 44.2 5.5 30.7 1
24.9 10.3 31.5 1 59.2 7.9 17.7 1
28.5 10.4 26.4 1 49.8 6.2 29.3 1
17.5 19.6 29.3 1 15.3 14.2 32.4 1
58.5 4.8 10.8 1 28.8 9.3 28.9 1
46.0 6.0 23.4 1 41.8 21.5 18.6 1
67.4 8.3 15.5 1 23.3 7.5 36.7 1
26.4 9.8 33.1 1 44.9 5.2 28.1 1
28.9 26.5 19.6 1 29.4 18.2 21.6 1
16.3 23.9 27.1 1 26.4 13.2 27.2 1
22.4 19.7 25.7 1 54.4 8.1 25.1 1
31.4 6.6 42.7 1 22.5 22.7 25.7 1
45.3 5.6 26.6 1 45.4 9.0 25.2 1
20.1 31.3 18.4 1 30.7 16.2 32.0 1
19.0 16.8 28.8 1 54.3 6.9 18.3 1
45.7 9.8 23.8 1 59.9 4.6 25.3 1
27.5 12.5 27.2 1 29.6 7.9 29.7 1
25.5 13.8 28.0 1 16.5 20.4 33.6 2
19.0 29.4 20.7 2 20.4 10.3 36.0 2
19.1 23.0 26.4 2 16.8 10.7 41.3 2
18.4 30.9 18.2 2 18.0 22.2 32.4 2
21.8 9.2 39.6 2 24.2 11.5 27.8 2
20.5 10.3 40.6 2 20.5 17.8 31.7 2
16.6 43.0 15.4 2 14.7 37.6 17.6 2
11.4 51.1 11.8 2 11.7 41.7 20.7 2
26.7 15.5 30.4 2 13.7 32.1 27.8 2
14.6 31.0 24.9 2 8.7 57.7 12.6 2
14.9 39.9 16.4 2 12.9 41.7 22.8 2
21.1 26.5 22.9 2 11.6 49.0 12.9 2
21.5 25.8 22.2 2 19.8 24.6 31.1 2
16.1 39.0 21.3 2 13.4 54.6 14.9 2
11.2 48.9 15.2 3 16.0 30.0 21.1 3
7.0 35.0 30.3 3 9.0 43.7 21.2 3
5.9 58.6 17.5 3 7.9 44.9 17.4 3
6.5 44.6 19.3 3 11.2 42.0 18.0 3
7.7 47.0 12.9 3 6.3 49.1 21.6 3
6.0 62.3 12.7 3 6.3 61.0 13.7 3
13.3 43.1 20.3 3 17.5 29.3 23.3 3
7.8 46.7 16.8 3 13.4 44.2 13.8 3
5.9 59.7 15.4 3 11.3 45.5 18.3 3
5.1 57.6 16.7 3 7.6 51.1 20.9 3
13.3 53.9 7.2 3 5.5 51.7 21.4 3
;
proc plot data=sasuser.vaalit;
   plot rehn*vayrynen=laani
        /vpos=25 hpos=40;
   plot     rehn*ahtisaar=laani
        / vpos=25 hpos=40;
   plot     ahtisaar*vayrynen=laani
        / vpos=25 hpos=40;
   title 'Kannatukset Uudellamaalla,
           Kymeessä ja Lapissa';
     format laani lsym.;
run;
proc discrim data=sasuser.vaalit;
     class laani;
     var rehn vayrynen ahtisaar;
run;
```

Tulostuksessa nähdään ensin kaikkien ehdokkaiden kahdenkeskinen vertailu graafisesti. Symboli osoittaa vaalipiiriä siten, että 'U' tarkoittaa Uuttamaata, 'K' Kymiä ja 'L' Lappia. Yksi merkki kuvaa yhtä kuntaa ko. vaalipiirissä.

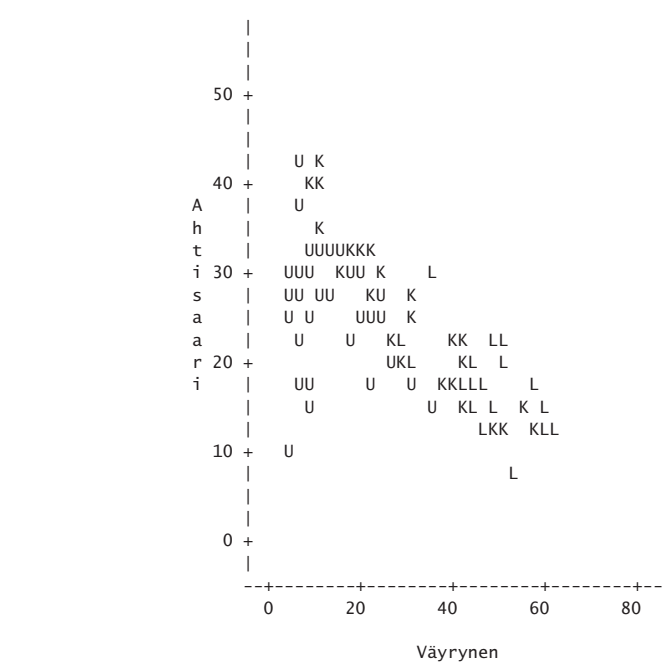
Tulostus on seuraavanlainen:

```
Kannatukset Uudellamaalla, Kymeessä ja Lapissa

Plot of REHN*VAYRYNEN.  Symbol is value of LAANI.

  80 +
     |
     |
     |
     |     U
     |
  60 +   U U
     |
     |    UU
     |    U
R    |    UU
e    |   U       U
h 40 +
n    |    U
     |
     |    UUU UU   U
     |     U UK
     |    UKU   UU
  20 +      K  U KUKKU
     |      KU  UKU LL U KK
     |               K  KKKL   LK
     |                    LL KK  K
     |                  L   LLLL  LLL
     |
   0 +
     --+--------+--------+--------+--------+--
       0        20       40       60       80

                    Väyrynen

NOTE: 24 obs hidden.
Kannatukset Uudellamaalla, Kymeessä ja Lapissa

Plot of REHN*AHTISAAR.  Symbol is value of LAANI.

  80 +
     |
     |
     |
     |          U
     |
```



NOTE: 19 obs hidden.
Kannatukset Uudellamaalla, Kymeessä ja Lapissa

Plot of AHTISAAR*VAYRYNEN. Symbol is value of LAANI.



NOTE: 13 obs hidden.
Kannatukset Uudellamaalla, Kymeessä ja Lapissa

Discriminant Analysis

86 Observations	85 DF Total
3 Variables	83 DF Within Classes
3 Classes	2 DF Between Classes

Class Level Information

LAANI	Frequency	Weight	Proportion	Prior Probability
Kymi	27	27.0000	0.313953	0.333333
Lappi	22	22.0000	0.255814	0.333333
Uusimaa	37	37.0000	0.430233	0.333333

Kannatukset Uudellamaalla, Kymeessä ja Lapissa

Discriminant Analysis Pooled Covariance Matrix Information

Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
3	11.5653865

Kannatukset Uudellamaalla, Kymeessä ja Lapissa

Discriminant Analysis

Pairwise Generalized Squared Distances Between Groups

$$D^2(i|j) = (\bar{X}_i - \bar{X}_j)' \, COV^{-1} \, (\bar{X}_i - \bar{X}_j)$$

Generalized Squared Distance to LAANI

From LAANI	Kymi	Lappi	Uusimaa
Kymi	0	2.84032	4.09801
Lappi	2.84032	0	12.32258
Uusimaa	4.09801	12.32258	0

Kannatukset Uudellamaalla, Kymeessä ja Lapissa

Discriminant Analysis Linear Discriminant Function

$$Constant = -.5 \, \bar{X}_j' \, COV^{-1} \, \bar{X}_j \qquad Coefficient\ Vector = COV^{-1} \, \bar{X}_j$$

LAANI

	Kymi	Lappi	Uusimaa	Label
CONSTANT	-116.28543	-125.82482	-109.12211	
REHN	2.21076	2.24398	2.26647	Rehn
VAYRYNEN	3.06826	3.26745	2.87390	Väyrynen
AHTISAAR	4.12128	4.18529	3.94641	Ahtisaari

Kannatukset Uudellamaalla, Kymeessä ja Lapissa

Discriminant Analysis

Classification Summary for Calibration Data: SASUSER.VAALIT

Resubstitution Summary using Linear Discriminant Function

Generalized Squared Distance Function:

$$D_j^2(X) = (X-\bar{X}_j)' \, COV^{-1} \, (X-\bar{X}_j)$$

Posterior Probability of Membership in each LAANI:

$$Pr(j|X) = \exp(-.5 \, D_j^2(X)) \, / \, SUM_k \exp(-.5 \, D_k^2(X))$$

Number of Observations and Percent Classified into LAANI:

From LAANI	Kymi	Lappi	Uusimaa	Total
Kymi	16	8	3	27
	59.26	29.63	11.11	100.00
Lappi	3	19	0	22
	13.64	86.36	0.00	100.00
Uusimaa	8	0	29	37
	21.62	0.00	78.38	100.00
Total	27	27	32	86
Percent	31.40	31.40	37.21	100.00
Priors	0.3333	0.3333	0.3333	

Error Count Estimates for LAANI:

	Kymi	Lappi	Uusimaa	Total
Rate	0.4074	0.1364	0.2162	0.2533

```
Priors 0.3333        0.3333        0.3333
```

Tuloksena saadaan, että Kymen vaalipiirin 27:stä kunnasta 16 kuntaa on tyypillisiä tämän vaalipiirin kuntia. Kahdeksan niistä sopisi äänestyskäyttäytymisen perusteella Lapin läänin kunniksi ja kolme Uudellemaalle. Lappi on homogeenisin näistä vaalipiireistä. Peräti 19 Lapin 22:sta kunnasta kuvaa Lapin vaalipiirin äänestyskäyttäytymistä. Kolme poikkeavaa Lapin vaalipiirin kuntaa sopisivat Kymen vaalipiiriin. Uudenmaan 37:stä kunnasta 29 edustaa tyypillistä uusmaalaista kuntaa. Loput 8 sopisivat paremmin Kymen vaalipiiriin. Yksikään Uudenmaan kunnista ei ole äänestyskäyttäytymiseltään tyypillinen Lapin vaalipiirin kunta.

Klusterianalyysista

Klusterianalyysin perustarkoituksena on löytää luonnollisia ryhmittelyjä. Ryhmä voi muodostaa koko populaation tai olla otos jostakin suuremmasta populaatiosta. Klusterianalyysi pyrkii sijoittamaan yksilöiden joukon keskenään poissulkeviin tyhjentäviin ryhmien joukoksi siten, että yksilöt ryhmän sisällä ovat samankaltaisia keskenään ja erilaisia toisten ryhmien yksilöiden kesken. Tätä kutsutaan osituksekseksi.

Klusterianalyysin rakenne

Klusterianalyysissä tarkasteltava havaintoaineisto voidaan esittää matriisimuodossa seuraavasti

$$X = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \dots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix}$$

$$X = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix},$$

missä X:n rivit edustavat n eri tapausta ja sarakkeet p eri muuttujaa.

Havaintovektoreiden $X_1, X_2, \dots, X_n$ ollessa itsenäisiä ne muodostavat klusterit $T_1, T_2, \dots, T_g$. Tällöin uskottavuusfunktio maksimoidaan jokaisessa ryhmittelyssä. Suurimman uskottavuusfunktion tuottava ryhmittely on optimaalinen.

Etäisyyden mitat

Klusterointimenetelmät poikkeavat toisistaan huomattavasti. Eri menetelmän valinta saattaa tuottaa täysin toisenlaisen tuloksen. Siksi menetelmän valinta onkin klusteroinnissa tärkeätä. Tarkasteltaessa etäisyyteen perustuvia klusterointimenetelmiä on perustana metriikka. Matriisin X pisteparien muodostaman numeerisen etäisyysfunktion $d(X_i, X_j)$ sanotaan olevan metrinen, jos se tyydyttää seuraavat ehdot:

$$d(X_i, X_j) \geq 0; d(X_i, X_j) = 0, \text{ jos } X_i = X_j; \text{ (i)}$$

$$d(X_i, X_j) = d(X_j, X_i); \text{ (symmetrisyys) (ii)}$$

$$d(X_i, X_k) + d(X_j, X_k) \geq d(X_i, X_j) \text{ (iii)}$$

Tässä (iii) on kolmion epäyhtälö. Merkitään etäisyyttä $d(X_i, X_j)$ d_{ij}:llä. Silloin kaikkien pisteparien etäisyydet muodostavat symmetrisen matriisin

$$D = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \vdots & \vdots & \dots & \vdots \\ d_{n1} & d_{n2} & \dots & d_{nn} \end{pmatrix}$$

Yleisin etäisyyden mitta on Minkowskin metriikka

$$d_{ij} = \left\{ \sum_{l=1}^{p} |X_{il} - X_{jl}|^k \right\}^{1/k}.$$

Euklidisen metriikan yleistys on painotettu Euklidinen metriikka

$$d_{ij} = \sqrt{\sum_{l=1}^{p} \omega_l (X_{il} - X_{jl})^2}.$$

Kun yo. kaavassa annetaan painon arvoksi $w_l = 1/s_l^2$, missä s_l on muuttujan X_l laskettu keskihajonta. Tämä tunnetaan Pearsonin tai χ^2:n etäisyysmittana. Tällöin

$$d_{ij} = \sqrt{\sum_{l=1}^{p} (X_{il} - X_{jl})^2 / s_l^2}.$$

Mahalonobiksen etäisyysmitta on myös standardoitu metriikka. Se saadaan kaavalla

$$d_{ij} = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)},$$

missä S on otosten $X_1, X_2, \dots, X_n$ muodostaman populaation estimoitu kovarianssimatriisi.

Similariteettisyys

Similariteettimatriisin C elementeille tehdään oletus

$$c_{ij} \geq 0, c_{ij} = 1, \text{ kun } i = j; \text{ (i)}$$

$$c_{ij} = c_{ji} \text{ (symmetrisyys). (ii)}$$

Similariteettikertoimet saavat arvoja välillä $[0,1]$. Kertoimien määritys perustuu binäärimatriisin summaustauluun. Eri vaihtoehtojen lukumäärät muodostavat taulun alkiot. Tapauksille i ja j tämä taulu on seuraava:

		Tapaus i: 1	Tapaus i: 0	
Tapaus j	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	p

Yleisimmin käytetyt similariteettikertoimet saadaan eo. taulukosta seuraavasti:

- Russel ja Rao

$$c_{ij} = a/p$$

- Sokal ja Michener (yksinkertainen sovitus)

$$c_{ij} = (a+d)/p$$

- Jaccard

$$c_{ij} = a/(a+b+c)$$

Etäisyyksien ja similariteettien suhde

Suuri etäisyys tapausten välillä merkitsee pientä similariteettiä. Sama pätee myös kääntäen eli pieniä etäisyyksiä vastaa suuret similariteetit. Similariteetin ja etäisyyden välinen yhteys voidaan ilmaista kaavalla

$$c_{ij} = (1+d_{ij})^{-1}$$

Käänteinen yhteys saadaan vastaavasti, kunhan similariteettimatriisi $C = \{c_{ij}\}$ on positiivinen semidefiniitti. Tällöin

$$d_{ij} = \sqrt{c_{ij} - 2c_{ij} + c_{jj}}.$$

Muuttujien klusterointi

Edellä on tarkasteltu klusterointia havaintojen suhteen. Monesti on tarkoituksenmukaista tarkastella muuttujia ja jakaa ne samankaltaisiin ryhmiin. Jaon perustana käytetään Pearsonin korrelaatiokertoimien muodostamaa matriisia.

Pearsonin korrelaatiomatriisiin R voidaan soveltaa edellä olleita etäisyys- ja similariteettimenetelmiä. Tällöin similariteettimatriisi on $C = |r_{ij}|$. Etäisyysmenetelmällä tarkastellaan silloin matriisia $D = \{1 - r_{ij}^2\}$. Klustereiden muodostamisen raja-arvona on $d_0^2 = (1 - r_0^2)$, missä r_0 on taulukoista saatava havainnoista riippuva luku.

Esimerkki muuttujien klusteroinnista

Eräässä kyselyssä tehtiin ihmisille 96 erilaista kysymystä. Kysymykset olivat osittain päällekkäisiä. Klusterianalyysillä pyritään selvittämään, kuinka kysymysten määrää voitaisiin yhdistelemällä pienentää. Kysely tehtiin 103:lle ihmiselle. Tarvittava SAS-ohjelmiston ajovirta on seuraava:

```
options linesize=64 nodate nonumber;
libname lib '\$HOME/lammi/sas/stat';
proc varclus data=lib.a short ;
   var k1-k96;
title 'Kysely';
run;
```

Proseduurin `varclus` optio `short` lyhentää tulostusta. Tulostuksesta jää tätä optiota käyttämällä pois standardisoidut pistekertoimet, korrelaatiot kunkin muuttujan ja klusterikomponentin välillä sekä klusterikomponenttien väliset korrelaatiot.

Tulostus on seuraava:

```
                          Kysely

      Oblique Principal Component Cluster Analysis

     103 Observations       PROPORTION  =         0
      96 Variables          MAXEIGEN    =         1

                          Kysely

      Oblique Principal Component Cluster Analysis

              Cluster summary for 12 cluster(s)

                    Cluster   Variation Proportion
Cluster   Members  Variation  Explained  Explained
-----------------------------------------------------
      1         9    9.00000    5.78594    0.6429
      2        15   15.00000    8.65957    0.5773
      3        10   10.00000    6.05746    0.6057
      4        13   13.00000    7.30593    0.5620
      5         7    7.00000    3.71680    0.5310
      6         6    6.00000    3.19880    0.5331
      7        10   10.00000    6.11571    0.6116
      8         5    5.00000    2.92295    0.5846
      9         9    9.00000    5.83638    0.6485
     10         3    3.00000    1.92632    0.6421
     11         6    6.00000    3.49587    0.5826
     12         3    3.00000    2.31434    0.7714

   Total variation explained = 57.33606
   Proportion = 0.5973
```

Tulostuksesta nähdään, että kysymykset voidaan ryhmitellä kahteentoista eri ryhmään annettujen vastausten perusteella. Tällä klusterointijaolla pystytään selittämään 59.7% kokonaisvaihtelusta.

```
                          R-squared with
                         ------------------
                          Own       Next    1-R**2
            Variable   Cluster   Closest    Ratio
Cluster   1----------------------------------------
            K29         0.6233    0.4586   0.6959
            K31         0.3819    0.244    0.8185
            K68         0.6684    0.4756   0.6322
            K69         0.7182    0.448    0.5109
            K77         0.6744    0.5122   0.6674
            K91         0.7452    0.683    0.8042
            K92         0.7647    0.6467   0.6661
            K93         0.7662    0.5506   0.5202
            K95         0.4436    0.2571   0.7489
Cluster   2----------------------------------------
            K3          0.3196    0.2284   0.8818
            K8          0.6162    0.4910   0.7541
            K30         0.2354    0.1077   0.8569
            K34         0.6721    0.4342   0.5795
            K35         0.7802    0.5019   0.4413

            K38         0.6252    0.2742   0.5164
            K40         0.5796    0.391    0.6913
            K48         0.3711    0.1918   0.7781
            K51         0.6074    0.3315   0.5873
            K52         0.6685    0.3620   0.5195
            K56         0.4621    0.4130   0.9163
            K84         0.7231    0.4646   0.5172
            K88         0.6736    0.4440   0.5872
            K94         0.7006    0.3972   0.4967
            K96         0.6250    0.4034   0.6285
```

Cluster	3			
	K11	0.6460	0.4302	0.6212
	K18	0.5442	0.3295	0.6798
	K24	0.7086	0.5070	0.5911
	K33	0.5204	0.3931	0.7903

Kysely

Oblique Principal Component Cluster Analysis

	R-squared with		
Variable	Own Cluster	Next Closest	1-R**2 Ratio
K42	0.7205	0.4599	0.5174
K47	0.3682	0.2874	0.8867
K58	0.6646	0.3966	0.5558
K60	0.5509	0.4706	0.8484
K66	0.6545	0.4711	0.6533
K90	0.6797	0.5945	0.7901
Cluster 4			
K2	0.4551	0.4418	0.9762
K4	0.6209	0.3739	0.6056
K5	0.5387	0.3400	0.6990
K6	0.7002	0.4684	0.5639
K7	0.6671	0.3077	0.4809
K10	0.6461	0.4255	0.6160
K12	0.7310	0.4165	0.4611
K16	0.3730	0.2108	0.7944
K39	0.5956	0.5013	0.8110
K46	0.4682	0.3419	0.8081
K53	0.5957	0.4523	0.7382
K59	0.4909	0.2749	0.7021
K62	0.4236	0.2308	0.7494
Cluster 5			
K26	0.5918	0.3266	0.6062
K28	0.5996	0.2374	0.5251
K70	0.4936	0.4229	0.8775
K73	0.6664	0.3452	0.5094
K74	0.4481	0.2247	0.7119
K76	0.3936	0.1983	0.7563
K81	0.5238	0.3621	0.7466
Cluster 6			
K15	0.6550	0.4012	0.5761
K21	0.3858	0.3005	0.8780
K23	0.6581	0.2481	0.4548
K55	0.4907	0.2525	0.6813
K63	0.4701	0.1384	0.6151
K71	0.5391	0.2828	0.6426
Cluster 7			
K1	0.5934	0.4265	0.7090
K9	0.7375	0.4112	0.4459
K13	0.4367	0.3085	0.8146
K17	0.6940	0.4358	0.5423
K22	0.3521	0.2080	0.8181
K41	0.6132	0.4785	0.7417
K49	0.5908	0.3352	0.6156
K57	0.7049	0.5880	0.7163
K65	0.7639	0.4742	0.4491
K89	0.6293	0.3253	0.5495
Cluster 8			
K36	0.5260	0.2994	0.6765
K37	0.6649	0.4146	0.5725
K64	0.5839	0.2283	0.5392
K78	0.4454	0.1810	0.6772

Kysely

Oblique Principal Component Cluster Analysis

	R-squared with		
Variable	Own Cluster	Next Closest	1-R**2 Ratio
K83	0.7028	0.3466	0.4548
Cluster 9			
K19	0.7193	0.6725	0.8570
K20	0.7201	0.6737	0.8577
K25	0.4428	0.2523	0.7453
K43	0.5837	0.3766	0.6678
K44	0.6947	0.4232	0.5293
K45	0.7383	0.5037	0.5273
K67	0.6362	0.6070	0.9258
K72	0.6688	0.5008	0.6634
K80	0.6324	0.3806	0.5934
Cluster 10			
K27	0.6604	0.1686	0.4085
K75	0.6863	0.2717	0.4307
K79	0.5796	0.1784	0.5116
Cluster 11			
K14	0.5388	0.3116	0.6699
K32	0.6265	0.5157	0.7711
K54	0.5104	0.3682	0.7748
K85	0.6728	0.4167	0.5610
K86	0.4799	0.2051	0.6543
K87	0.6673	0.4449	0.5993
Cluster 12			
K50	0.7381	0.2808	0.3641
K61	0.7323	0.5207	0.5584
K82	0.8439	0.4594	0.2888

No cluster meets the criterion for splitting.

Jokainen muuttuja on lueteltu omassa klusterissaan. Tulostuksessa on mukana muuttujan korrelaatiokertoimen neliö omaan klusterikomponenttiin. `Next Closest` ilmoittaa muuttujan seuraavaksi korkeimman korrelaatiokertoimen neliön klusterikomponentteihin nähden. Suhdeluku $1 - R^2$ ilmoittaa suhteen

$$1 - R^2 = \frac{1 - r_0^2}{1 - r_n^2},$$

missä r_0 on korrelaatiokerroin oman klusterikomponentin kanssa ja r_n vastaavasti korrelaatiokerroin lähinnä seuraavan klusterikomponentin kanssa. Mitä pienempi suhdeluku $1 - R^2$ on, sitä parempi on klusterointi.

Kysely

Oblique Principal Component Cluster Analysis

Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster
1	36.500384	0.3802	0.3802
2	43.168284	0.4497	0.4464
3	45.578202	0.4748	0.4564
4	47.653541	0.4964	0.4672
5	49.435718	0.5150	0.4528
6	50.843861	0.5296	0.4528
7	52.295898	0.5447	0.4528
8	53.510175	0.5574	0.4528
9	54.432846	0.5670	0.4528
10	55.547572	0.5786	0.5310
11	56.486118	0.5884	0.5310
12	57.336064	0.5973	0.5310

Number of Clusters	Maximum Second Eigenvalue in a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	8.389564	0.0811	.
2	2.980692	0.0978	0.9832
3	2.389088	0.1284	0.9947
4	2.095219	0.1335	0.9773
5	1.763640	0.1569	0.9737
6	1.575470	0.2205	0.9667
7	1.352037	0.2254	0.9667
8	1.250220	0.2354	0.9762
9	1.230466	0.2354	0.9762
10	1.139379	0.2354	0.9762

11	1.040362	0.2354	0.9762
12	0.959590	0.2354	0.9762

Suuri maksimisuhde $1 - R^2$:lle osoittaa, että muuttujien klusterointi sisältää virhemahdollisuuksia. $1 - R^2$ voi saada arvoja väliltä $[0,1]$. Tämän pohjalta voidaan kuitenkin suuntaa-antavasti yhdistellä alkuperäisiä kysymyksiä.

Aikasarjoista

Aikasarjat kuvaavat yhden tai useamman suureen muuttumista ajan suhteen. Havaintojen sisältämän tiedon perusteella kehitetään aikasarjan pohjalla olevan tapahtuman mallia ja estimoidaan mallin parametrejä. Tapahtuman tulevaa kehitystä voidaan ennustaa. Eri aikasarjojen välille voidaan kehittää niiden riippuvuutta kuvaavaa mallia. Tällöin voidaan ennustaa yhdessä aikasarjassa tehtyjen muutosten vaikutusta toisen aikasarjan tulevaan käyttäytymiseen.

Aikasarjojen tyypit

Aikasarjoja voidaan aineistonsa perusteella jakaa useampaan erilaiseen luokkaan. Tunnetuimmat ovat seuraavat:

- Taloudellinen aikasarja (*economic time series*) käsittelee taloustieteellisiä aineistoja
- Fysikaalinen aikasarja (*physical time series*) käsittelee fysikaalisia ilmiöitä, esimerkiksi ilman lämpötilan mittaamista peräkkäisinä tunteina, päivinä tai kuukausina.
- Kauppatavara-aikasarja (*marketing time series*) käsittelee tuotteiden myyntiä pyrkien ennustamaan tulevaa kehitystä. Tärkeää on eri aikasarjojen vertailu. Myynnin tulosta ja mainonnan kustannusta voidaan tarkastella yhdessä.
- Väestöaikasarja (*demographic time series*) kuvaa vuosittaista väestömäärää. Sen avulla pyritään ennustamaan väestömäärän tulevia muutoksia.
- Laadunvalvonnassa (*process control*) on pyrkimyksenä havaita muutokset tuotteen valmistuksessa mittaamalla prosessin laatua kuvaavaa muuttujaa.
- Binääriprosessi (*binary process*) on aikasarja, joka voi saada vain kaksi eri arvoa, tavallisesti 0 ja 1.
- Pisteprosessi (*point process*) on aikasarja, joka esiintyy satunnaisesti ajan suhteen. Se sisältää tietoa tapahtumien esiintymisistä tietyllä aikavälillä ilmoittaen sekä tapahtumien lukumäärän että niiden välisen ajan.

Aikasarja-aineiston tunnusmerkit

Aikasarjan taustalla oleva prosessi voi olla jatkuva-aikainen tai diskreettiaikainen. Aikasarja voi sisältää minkä tahansa tai kaikki seuraavista osista: trendi, kausivaihtelu, jaksottainen vaihtelu tai satunnaisvaihtelu.

- trendi ilmentää jatkuvasti kasvavaa tai vähenevää komponenttia.
- kausivaihtelussa havaintoarvot muuttuvat säännöllisessä (esim. vuosi) ajanjaksossa
- jaksottaisessa vaihtelussa havaintoarvojen muutokset noudattavat yleisempää periodista vaihtelua.
- satunnaisvaihtelu jää jäljelle, kun edellä mainitut ominaisuudet on poistettu aineistosta. Osa satunnaisvaihtelusta voidaan selittää todennäköisyysmalleilla, kuten autoregressiivisillä malleilla tai liukuvien keskiarvojen malleilla.

Aikasarjan havaintoaineistoon voidaan tehdä analyysiä helpottavia muunnoksia, jolloin pyritään poistamaan trendi. Tavoitteena on tällöin, että muunnetun aikasarjan satunnaiskomponentti on stationaarinen. Ei-stationaarista trendin sisältävää aikasarjaa on myös mahdollista tutkia.

Esimerkki aikasarjasta

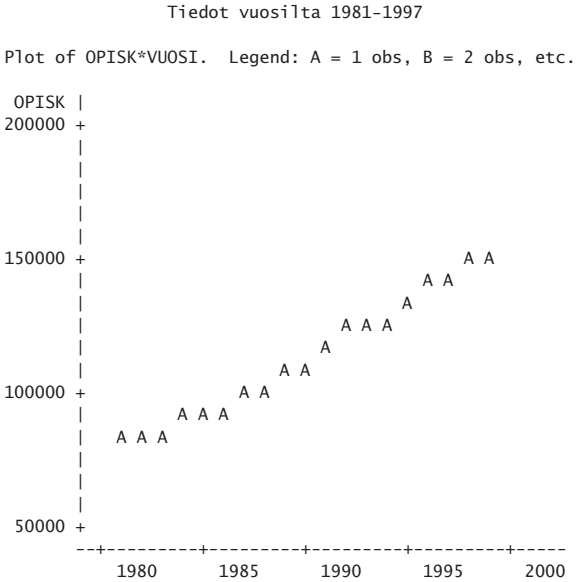
Tutkitaan korkeakoulujen opiskelijoiden, opettajien, muun henkilökunnan ja määrärahojen kehitystä ja pyritään ennustamaan tulevaisuuden määriä. Tiedot ovat KOTA-tietokannasta. Määrärahat ovat tuhansia markkoja.

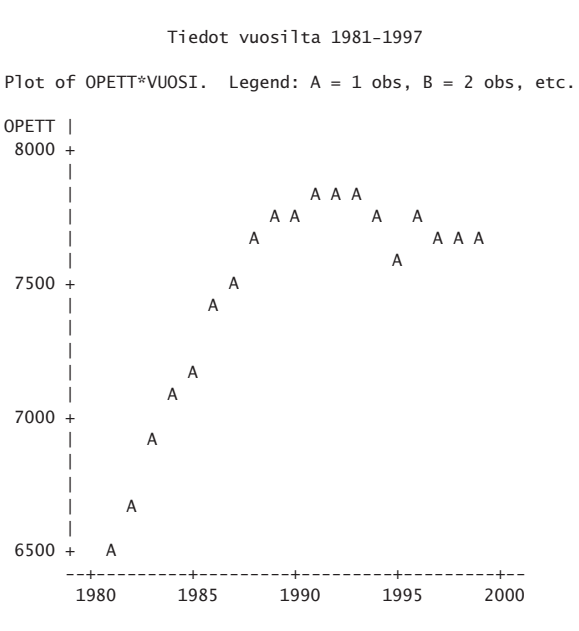
Ennustaminen suoritetaan tutkimalla ensin graafisesti olemassaolevia tietoja. Tähän tarkoitukseen käytetään Cedarilla olevaa SAS-ohjelmistoa. Tehdään tiedostoon `ajo.sas` tarvittava ajovirta, jolloin suoritus tapahtuu käskyllä `sas ajo`. Tarvittava ajovirta on seuraava:

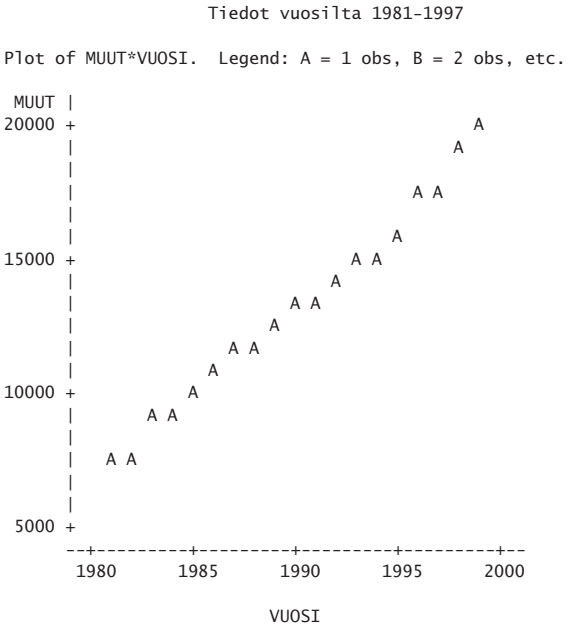
```
options ls=77 nodate nonumber;
data sasuser.kota;
   input vuosi opisk opett muut rahat;
cards;
1981     84187     6471     7419     1138082
1982     86389     6625     7742     1289106
1983     86287     6938     8785     1517560
1984     88564     7109     9493     1662292
1985     90720     7169    10191     1821997
1986     94311     7436    10445     1949854
1987     98137     7512    11475     2159918
```

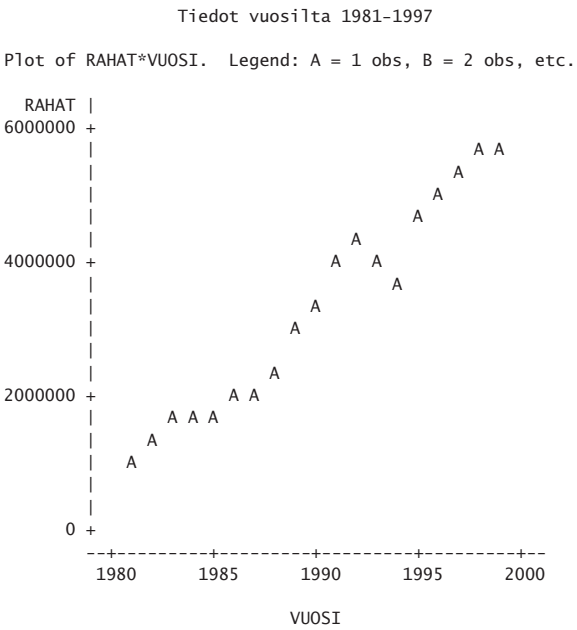
```
1988    102193     7625    11800    2477875
1989    109061     7731    12619    2835728
1990    110680     7788    13174    3232108
1991    115573     7812    13595    4073661
1992    122227     7828    13770    4206667
1993    126123     7814    14650    3885154
1994    128267     7722    14675    3829787
1995    135107     7550    15791    4547123
1996    138173     7737    17284    5116044
1997    142818     7706    17514    5331106
1998    147263     7637    19043    5606424
1999    151910     7668    19800    5815160
;
proc plot data=sasuser.kota;
   plot opisk*vuosi/vpos=20 hpos=40;
   plot opett*vuosi/vpos=20 hpos=40;
   plot muut*vuosi/vpos=20 hpos=40;
   plot rahat*vuosi/vpos=20 hpos=40;
run;
```

Tulostus on tällöin tiedostossa `ajo.lst`.









Graafisesta esityksestä huomataan, että kaikki tutkittavat muuttujat sisältävät trendin. Valitaan ennustamiseen SAS-ohjelmiston FORECAST-proseduuri. Se soveltuu trendien käsittelemiseen. Tarvittava ajovirta on seuraava:

```
options ls=77 nodate;
libname lib '$HOME/sas/ets';
filename gsasfile '$HOME/sas/ets/opisk.gsf';
goptions nodisplay device=ps gsfmode=replace
         gaccess=gsasfile;
goptions gsfname=gsasfile;
goptions noprompt handshake=xonxoff gsflen=72
         autofeed;
proc forecast data=sasuser.kota lead=10
         out=a outfull;
   id vuosi;
   var opisk opett muut rahat;
data b;
   set a;
```

```
   if _type_='FORECAST' and vuosi < 2000
       then delete;
   if _type_='FORECAST' and vuosi > 1999
       then _type_='LKM';
   if _type_='ACTUAL' then _type_='LKM';
proc print data=b;
   id vuosi;
   title 'Ennuste vuosille 2000-2009';
   format opisk f6.0 opett f5.0  muut f5.0
          rahat f8.0;
proc gplot data=b;
   plot opisk*vuosi = _type_/
   haxis=1981 to 2009;
symbol1 i=join v=star h=2;
run;
```

Ajovirrassa käytetään hyväksi edellä luotua sastiedostoa `sasuser.kota`. Hakemistoviite `sasuser` on SAS-ohjelmiston automaattisesti jokaiselle käyttäjälle varaama hakemistoviite. Ennustuksessa käytetään askeltavaa autoregressiivistä menetelmää (`stepar`), joka on oletuksena. Se soveltuu käsiteltäessä lineaarista trendimallia.

Menetelmä `stepar` käyttää pienimmän neliösumman menetelmää. Tämä pienentää trendin vaikutusta.

Tulostus menee tilapäistiedostoon `a`, joka tulostetaan modifioituna samassa ajovirrassa. Luottamusvälit saadaan parametrillä `outfull`. Oletuksena on 95%:n luottamusvälit. Ennustus halutaan suorittaa kymmenelle seuraavalle vuodelle.

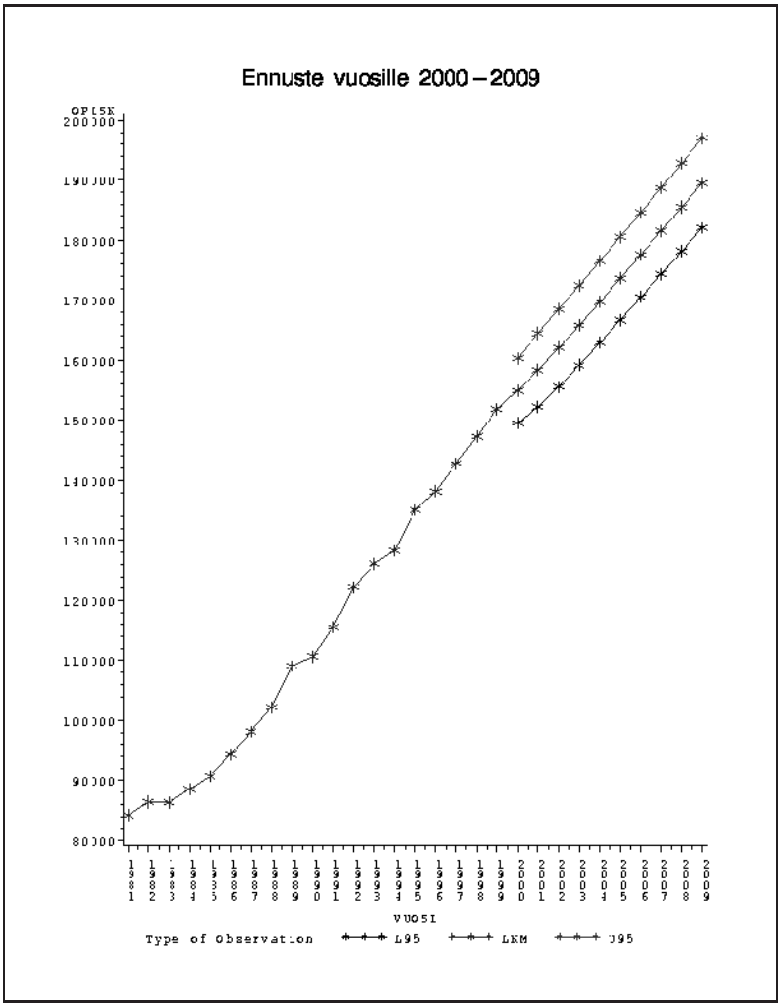
Tulostus on seuraava:

Ennuste vuosille 2000-2009

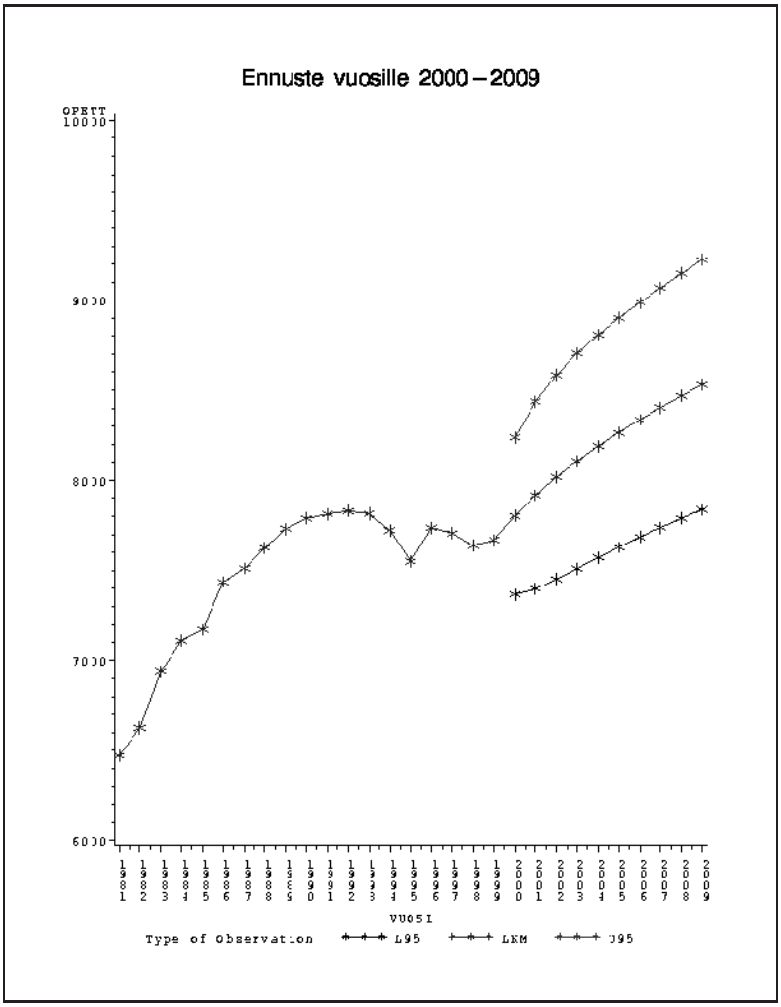
VUOSI	_TYPE_	_LEAD_	OPISK	OPETT	MUUT	RAHAT
1981	LKM	0	84187	6471	7419	1138082
1982	LKM	0	86389	6625	7742	1289106
1983	LKM	0	86287	6938	8785	1517560
1984	LKM	0	88564	7109	9493	1662292
1985	LKM	0	90720	7169	10191	1821997
1986	LKM	0	94311	7436	10445	1949854
1987	LKM	0	98137	7512	11475	2159918
1988	LKM	0	102193	7625	11800	2477875
1989	LKM	0	109061	7731	12619	2835728
1990	LKM	0	110680	7788	13174	3232108
1991	LKM	0	115573	7812	13595	4073661
1992	LKM	0	122227	7828	13770	4206667
1993	LKM	0	126123	7814	14650	3885154
1994	LKM	0	128267	7722	14675	3829787
1995	LKM	0	135107	7550	15791	4547123
1996	LKM	0	138173	7737	17284	5116044
1997	LKM	0	142818	7706	17514	5331106
1998	LKM	0	147263	7637	19043	5606424
1999	LKM	0	151910	7668	19800	5815160
2000	LKM	1	154961	7802	19985	5983635
2000	L95	1	149594	7367	19027	5529537
2000	U95	1	160329	8238	20944	6437734
2001	LKM	2	158391	7918	20421	6212770
2001	L95	2	152255	7397	19371	5655826
2001	U95	2	164528	8438	21471	6769714
2002	LKM	3	162048	8018	20971	6515193
2002	L95	3	155588	7451	19889	5949418
2002	U95	3	168509	8586	22053	7080967
2003	LKM	4	165842	8108	21574	6838105
2003	L95	4	159194	7511	20471	6232643
2003	U95	4	172489	8706	22677	7443567
2004	LKM	5	169716	8190	22201	7131672
2004	L95	5	162926	7571	21078	6498066
2004	U95	5	176507	8809	23324	7765278
2005	LKM	6	173640	8265	22839	7389154
2005	L95	6	166717	7629	21696	6745762
2005	U95	6	180563	8902	23982	8032546
2006	LKM	7	177593	8336	23482	7636222
2006	L95	7	170538	7685	22318	6976080
2006	U95	7	184648	8988	24647	8296364
2007	LKM	8	181563	8404	24128	7897481
2007	L95	8	174372	7738	22941	7222437
2007	U95	8	188754	9069	25314	8572524
2008	LKM	9	185544	8468	24774	8176504
2008	L95	9	178213	7789	23564	7490386
2008	U95	9	192876	9148	25984	8862622
2009	LKM	10	189531	8531	25421	8460813
2009	L95	10	182054	7837	24187	7761741
2009	U95	10	197009	9225	26655	9159886

Tulostuksessa on jokaisen kymmenen ennustevuoden kohdalla mainittuna ennustettu arvo sekä sille 95%:n ala- ja yläluottamusväli. Muuttujan _type_ arvo *lkm* edustaa vuosina 1981 - 1999 todellista arvoa ja vuodesta 2000 lähtien ennustettua arvoa.

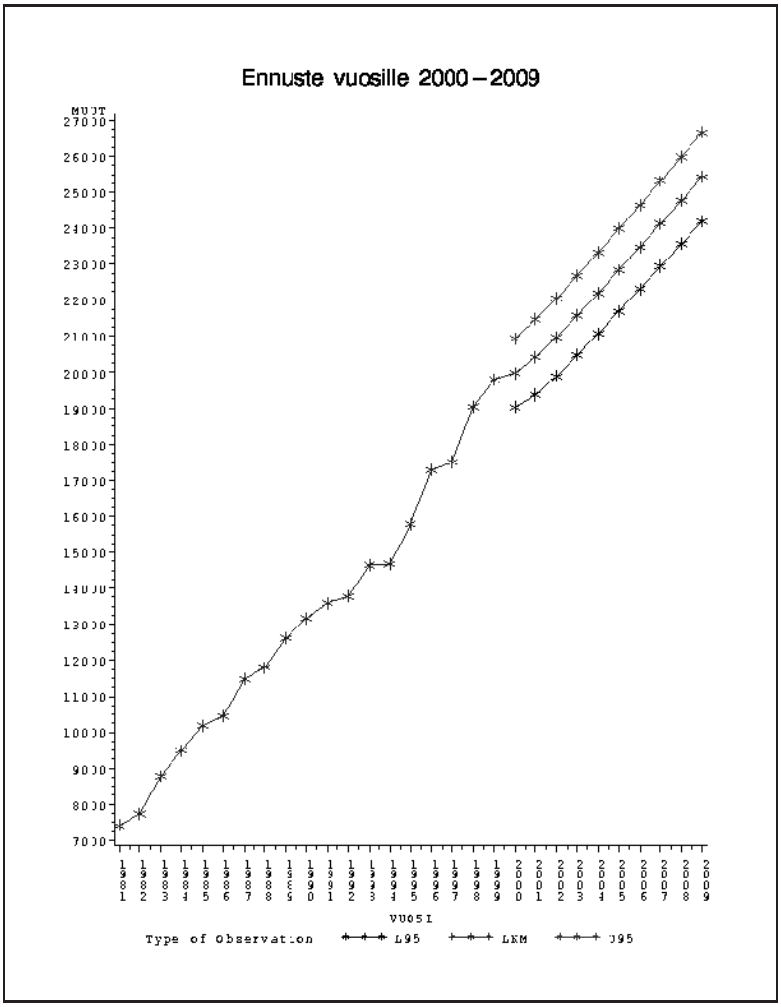
Ohjelmiston tuottama kuvaaja on muuttujalle *opisk*. Vastaavalla tavalla saadaan kuvaajat muillekin muuttujille. Kaikkien näiden muuttujien kuvaajat esitetään seuraavassa.



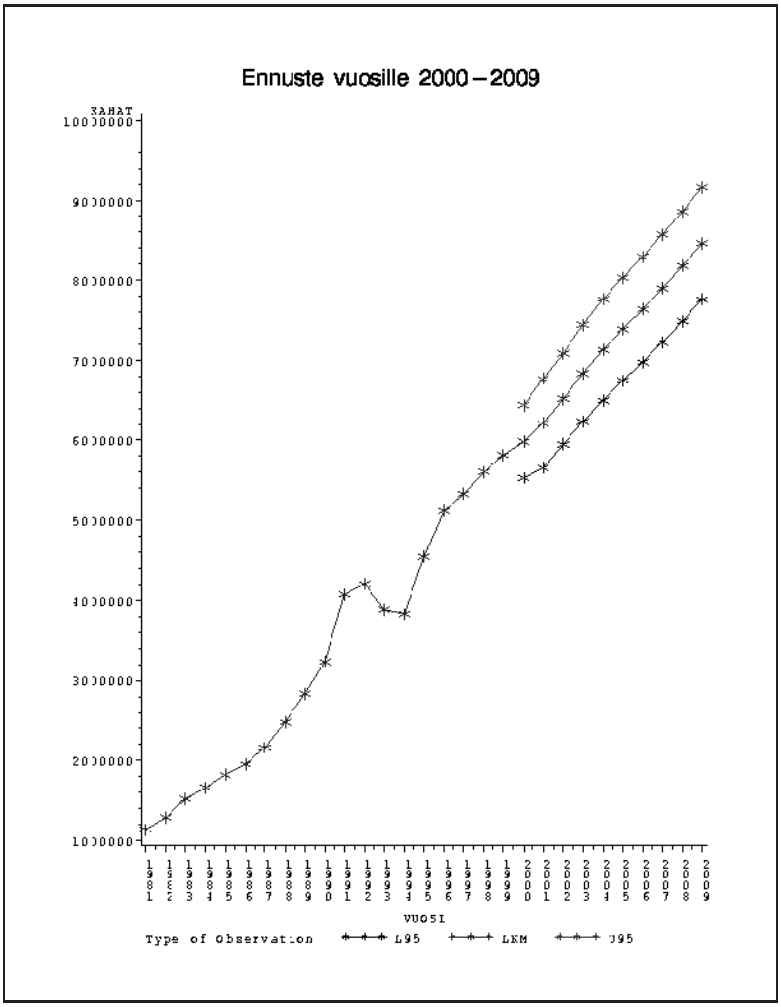
Kuva 1: *Opiskelijoiden määrä vuosina 1981-1999 ja ennustettu määrä vuosina 2000-2009*



Kuva 2: *Opettajien määrä vuosina 1981-1999 ja ennustettu määrä vuosina 2000-2009*



Kuva 3: *Muun henkilökunnan määrä vuosina 1981-1999 ja ennustettu määrä vuosina 2000-2009*



Kuva 4: *Määrärahojen määrä vuosina 1981-1999 ja ennustettu määrä vuosina 2000-2009*