

SAS-ohjelmiston perusteet

Luentorunko/päiväkirja

Ari Virtanen 11.12.08
(päivitetään luentojen edetessä)

<http://mtl.uta.fi/~maarvi/SASesite.html>

<http://mtl.uta.fi/tilasto/sas/syksy08/>

Harjoitukset

<http://mtl.uta.fi/tilasto/sas/syksy08/EGharj1.pdf>

<http://mtl.uta.fi/tilasto/sas/syksy08/EGharj2.pdf>

<http://mtl.uta.fi/tilasto/sas/syksy08/EGharj3.pdf>

<http://mtl.uta.fi/tilasto/sas/syksy08/EGharj4.pdf>

<http://mtl.uta.fi/tilasto/sas/syksy08/EGharj5x6.pdf>

HUOM. 6. Harjoitusten palautukselle annettiin lisää aikaa lauantaihin 20.12. klo 10 saakka.

Ilmoitusasioita:

Harjoituksia on kuusi kertaa: 13.11., 20.11., 27.11., 4.12., 10.12. ja 11.12.

Jokaisessa harjoituksessa on tarjolla 6 pistettä ja lisäksi bonuspisteitä.

Harjoitustyön osuus on 10 pistettä, vierailuluennon kuunteleminen (tai korvaava suoritus) 4 pistettä.

Läpikäytyraja on 50% pisteistä, 5/5 100% eli 50 pistettä (mutta bonuspisteiden avulla on mahdollista saada enemmänkin kuin 100%).

Vierailuluento **to 27.11. klo 16-17 ls A3107 Pinni A**. Jatketaan tavallisissa luennoissa 17-18 ML42. Keskiviikon 26.11. opetus on peruutettu ja harjoitukset palautetaan sähköisesti.

Tiedostojen siirtoalue mm. harjoitusten ratkaisuja varten:

<http://mtl.uta.fi/lister/>

Tunnus on "kurssilainen" ja salasana "ArinKurssinTiedostot"

Tiedoston nimen alkuun aina oma käyttäjätunnus.

Luennot ke 29.10.

Esimerkkejä SAS-ohjelmiston käytöstä

Tätä (eli Opiskelutilanne [Matematiikka/tilastotiede](#))

voi käydä kauhistelemassa omin päin:

https://intra.uta.fi/yksikot/tkk/hr/tilastoja/koht_aineet/v2008/

Vierailuluento tästä aihepiiristä:

<http://vos.uta.fi/rap/>

Esimerkki 1. Enterprise Guide eli EG

Lasketaan kokonaismerkinnän arvosana kokonaisuuden opintojaksojen ollessa seuraavat:

Lasketaan arvosana opintokokonaisuudelle

OPINTOJAKSO	OPINTOPISTEET	ARVOSANA
kurssiA	9	2
kurssiB	8	3
kurssiC	7	4
kurssiD	1	5

Ratkaisu:

<http://mtl.uta.fi/tilasto/sas/syksy08/EGstart.pdf>

<http://mtl.uta.fi/tilasto/sas/syksy08/esimerkit1x7/painotettu.egp>

E:\SAS\arvosana

Esimerkki 2. Tehdään sama varsinaisella SAS-ohjelmistolla.

```
data sasuser.arvosana; *tehdään pysyvä tiedosto arvosana;
  input kurssi $ pisteet arvosana;

datalines;
kurssiA 9 2
kurssiB 8 3
kurssiC 7 4
kurssiD 1 5
;
run;
```

Tarkistetaan, että aineiston luonti onnistui.

Ensin rakenne:

```
proc contents data=sasuser.arvosana;
run;
```

Sitten varsinainen aineisto:

```
proc print data=sasuser.arvosana;
run;
```

Lopuksi painotetun keskiarvon laskeminen:

```
proc means data=sasuser.arvosana n mean;
  var arvosana;
  weight pisteet;
run;
```

Esimerkki 3. EG

Aineistona on erään kesän opiskelijavalinnan hakijoiden sukupuoli, pitkän matematiikan arvosana yo-kirjoituksissa sekä valintakoepistemäärä. Hakijat, jotka eivät ole kirjoittaneet pitkää matematiikkaa, on jätetty pois. Tietosuojasyistä mukana on vain tietyn minipistemäärän valintakokeessa saaneet ja heidän kohdalla koepisteet on standardisoitu. Aineisto on Excel-muodossa, ja ensimmäiseksi se luetaan EG:hen ja tallennetaan osaksi projektia.

Tehtävänä on tutkia t-testillä, onko sukupuolella merkitystä valintakoemenestykseen. Muodostetaan myös laatikko-jana -kuvio.

E:\SAS\valinta (oma muistitikku) Alla olevaa linkkiä ei kannata avata IE:llä (ilmeisesti mikä muu selain tahansa käy. Koska datatiedostot puuttuvat, prosessia ei voi sellaisenaan ajaa uudelleen.)
<http://mtl.uta.fi/tilasto/sas/syksy08/esimerkit1x7/valinta.egp>

Esimerkki 4. EG

Luultavasti sukupuolta merkittävämpi selittäjä valintakoemenestykselle on ylioppilaskirjoitusten pitkän matematiikan arvosana. Tutkitaan tätä 1-suuntaisella varianssianalyysillä.

Tulosten tulkinnan helpottamiseksi koodataan ensin matematiikan yo-arvosana uudelleen niin että 1 on paras, 2 toiseksi paras jne.

E:\SAS\valinta

Esimerkki 5. EG

Tutkitaan 2-suuntaisella varianssianalyysillä, onko sukupuolella ja pitkän matematiikan arvosanalla yhteisvaikutusta.

1-suuntainen varianssianalyysi löytyy helposti nimensä perusteella. Muistamalla, että varianssianalyysissä on kyse lineaarisista malleista, osaa etsiä 2- (tai k-)suuntaista varianssianalyysia valikosta Analyze - ANOVA – Linear Models

Konkreettisemmän käsityksen saamiseksi muodostetaan vielä taulukko, johon lasketaan keskiarvot (ja havaintojen lukumäärät) eri osaryhmissä.

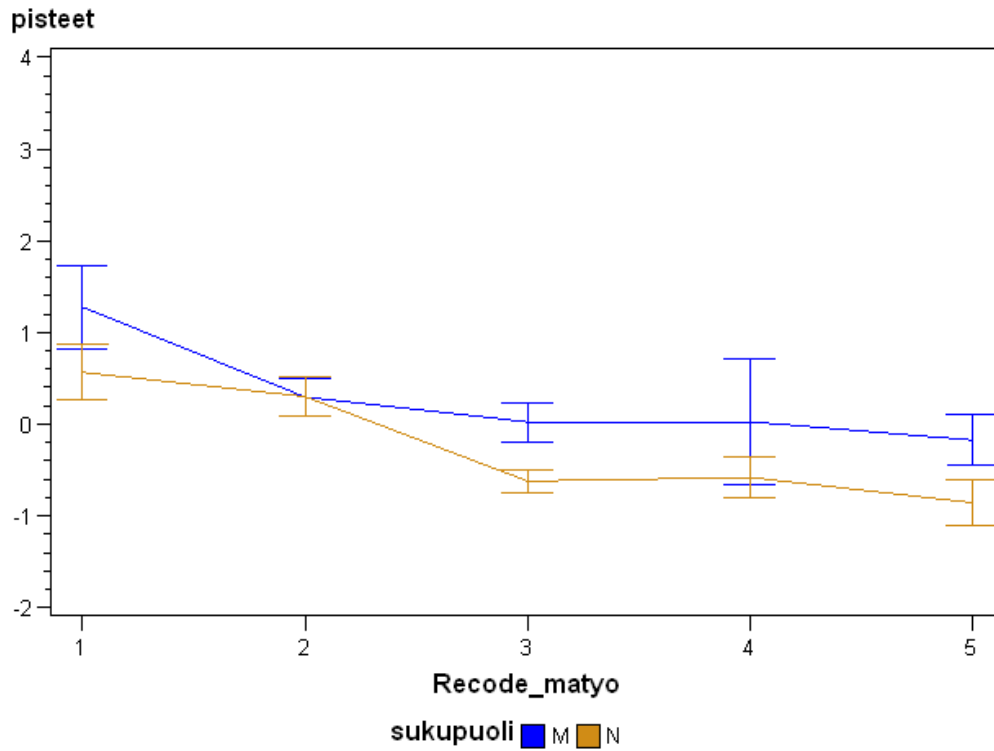
E:\SAS\valinta

Luennoilla 29.10. päästiin tähän saakka.

Luennot keskiviikko 5.11.

Tarkastellaan vielä hieman esimerkkiä 5. Vaikka p-arvon mukaan tilastollisesti merkitsevää yhteisvaikutusta ei olekaan, kuvio kertoo jotain muuta.

Means plot of pisteet by Recode_matyo and sukupuoli



Esimerkki 6. EG

Aineistona (seurantaVK) on kahtena eri vuonna opintojensa aloittaneiden matematiikan ja tilastotieteen opiskelijoiden standardisoidut valintakoe pisteet ja tähän mennessä suoritettut matematiikan ja tilastotieteen keskimääräiset opintopisteet vuotta kohden.

Aineisto on tiedostossa seurantaVK.xls. Luetaan aineisto EG:hen, muodostetaan pisteparvi ja lasketaan korrelaatiokerroin. Tulokset ovat hieman yllättäviä, joten jatketaan analyysia tekemällä se erikseen eri vuosina aloittaneille.

<http://mtl.uta.fi/tilasto/sas/syky08/esimerkit1x7/seuranta.egp> (kopioi ja avaa muulla selaimella kuin IE, muistitikulla E:\SAS\seuranta), data: [seuranta.sas7bdat](#)

komennot Graph → Scatter Plot,
Analyze → Multivariate → Correlations

Tarkoitus oli tämän jälkeen jatkaa regressioanalyysillä, mutta unohdetaan se. Miksi?

Esimerkki 7. EG

Tiedostoon normaali on generoitu muuttujien $x_1 - x_5$ satunnaislukuja tasaisesta jakaumasta väliltä $(0,1)$. Jotta saataisiin esimerkki suodatuksesta, muutamaan kohtaan on laitettu kuitenkin väärä arvo (liian suuri).

Suodatetaan väärät arvot pois ja muodostetaan näin saadussa aineistossa muuttujien $x_1 - x_5$ summamuuttuja X ja testataan, noudattaako se teorian mukaista normaalijakaumaa.

Suodatus: Data → Filter and Query

Jakaumaoletuksen testaus: Describe → Distribution Analysis

Vertailun vuoksi generoidaan R:llä otos normaalijakaumasta

```
> #teoreettinen keskiarvo
> 5*(1-0)/2
[1] 2.5
> #teoreettinen hajonta
> sqrt(5*(1-0)^2/12)
[1] 0.6454972
>
> x <- rnorm(1000,mean=2.5,sd=sqrt(5*(1-0)^2/12))
> write(t(t(x)), "E:\\SAS\\summamuuttuja\\generoitu", sep="\n")
>
```

ja tehdään vastaava testi tälle aineistolle.

E:\SAS\summamuuttuja

<http://mtl.uta.fi/tilasto/sas/syksy08/esimerkit1x7/summamuuttuja/>

(kansio sisältää datatiedostot)

Luentojen päälähde:

MTT:n selvityksiä 140



Opas SAS Enterprise Guiden käyttäjälle

EG versio 4.1

Timo Hurme



<http://www.mtt.fi/mtts/pdf/mtts140.pdf>

Data-aineistot:

2 Oppaassa käytettävät datat

Kaikki kolme oppaassa käytettävää aineistoa löytyvät internetistä (<http://www.mtt.fi/mtts/pdf/egdata/EGdatat.zip>). Aineistoja tarvitaan oppaan harjoitustehtäviä tehtäessä, mutta harjoitusmateriaalina voi vaihtoehtoisesti käyttää myös muita vastaavan tyyppisiä aineistoja. Käytettävät aineistot kannattaa tallentaa omalle tietokoneelle esimerkiksi kansioon C:\SAS\EGKURS, jota käytetään oppaan esimerkkien tallennuspaikkana.

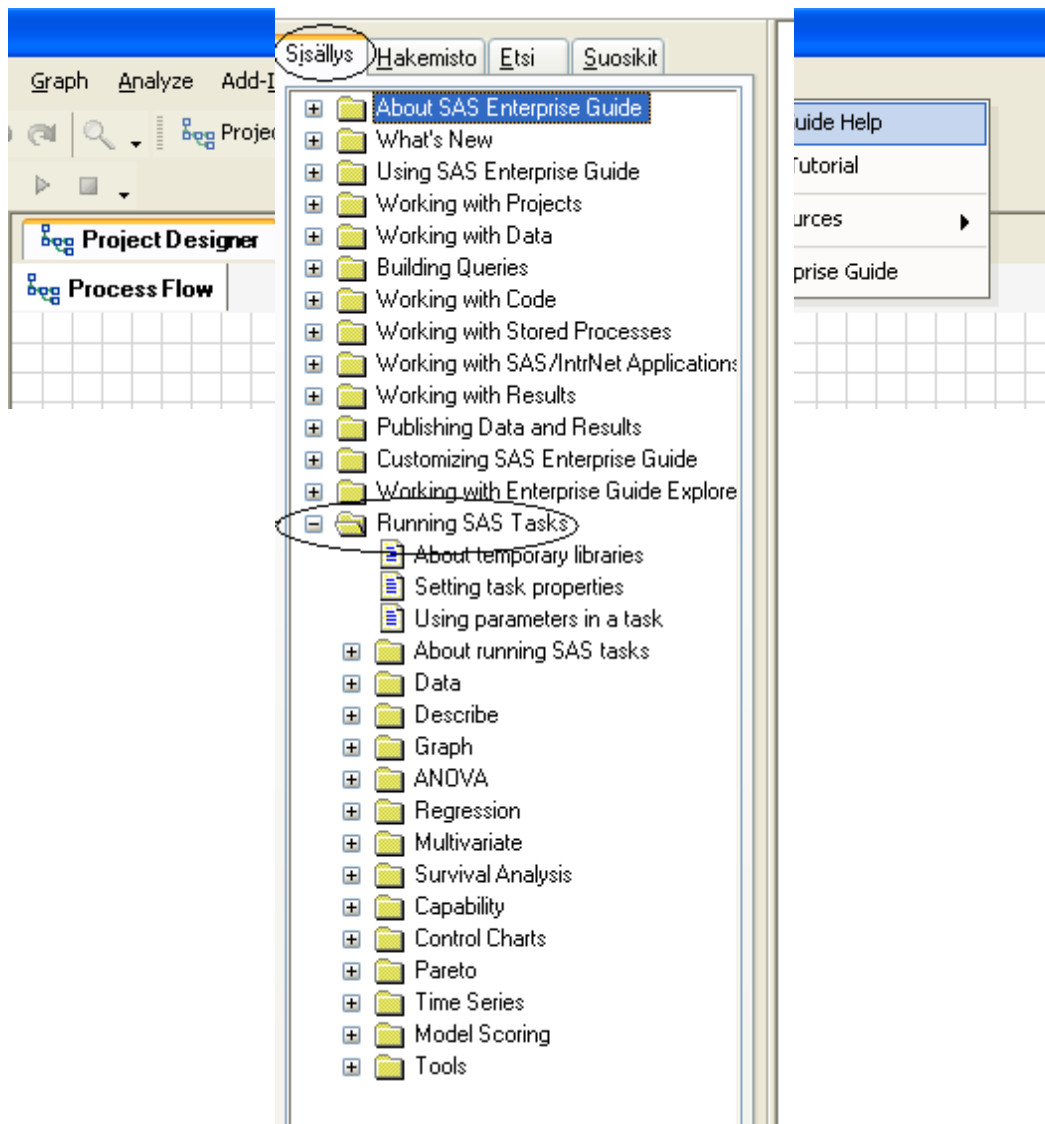
E:

www.mtt.fi/mtts/pdf/egdata/EGdatat.zip

Lue jossain vaiheessa s. 7-11.

Enterprise Guiden ohje

s.11-13



6 Aineiston käyttöönotto s. 13-

EGteht 1.

Tilapäisen kirjastoviitteen luominen `Tools` → `Assign Library` s.14-15

Kirjastoviitteen asettaminen oletukseksi s. 16

Tämä esimerkki käsitellään heti kun joku esittää perustelun kirjastoviittausten hyödyllisyydelle (bonus piste tarjolla perusteluista).

□

Esimerkki 8.

Tee pieni havaintotiedosto, sille jokin analyysi, tallenna data muistitikulle, avaa data, lisää pari havaintoa, tee uusiksi äskeinen analyysi, sulje ja talleta projekti. Avaa projekti uudelleen ja tarkista, mikä data on käytössä.

EGteht 2.

Datatiedostojen kopiointi (tehtävä s. 15)

Excel-data `GomexData.xls` → SAS-data `Gomex_Data.sas7bdat` (tehtävä s. 22)

`E:\SAS\EGKURS\EGmoniste.egp`

Tämän jälkeen vaihdettiin etenemisjärjestystä.

Datan muokkaus Query Builder työkalulla

Lue s. 30-34 ja kokeile `Data` → `Filter and Query`

EGteht 5.

Dataan `Gomex_Data.sas7bdat` uudet muuttujat `tsato` ja `sqrt_sato` (tehtävä s. 34). Tulosdatalle

nimeksi gomez_muok

Kysymys: mitenkä uusi muuttuja voidaan lisätä suoraan alkuperäiseen datamatriisiin?

Luennot torstai 6.11.2008

Luennoitsija paikalla jo klo 16.00 vastaamassa kysymyksiin jo käsitellyistä asioista. Uutta asiaa aletaan käsittelemään varttia yli. Menetellään toistaiseksi luennoilla näin.

Esimerkki 9.

Lasketaan EG:llä viiden kertotaulu suoraan datamatriisiin, jossa alunperin vain kerrottavat.

Ks. vaiheet:

<http://mtl.uta.fi/tilasto/sas/syky08/sarakkeenlis.pdf>

EGteht 3.

Suomenkieliset muuttujien nimet ja selitystekstit dataan
Gomex_Data.sas7bdat (tehtävä s. 27)

Avaa data, poista kirjoitussuojaus.

E:\SAS\EGKURS\EGmoniste.egp

Luennoitsijan kone kieltäytyi yhteistyöstä formaattien luonnissa. Asiaan palataan seuraavilla luennoilla.

Muuttujien luokittelu, muuttujan luominen aritmeettisella operaatiolla, muuttujan poistaminen ja muuttujien järjestyksen vaihtaminen sekä datan havaintorivien järjestäminen jätetään harjoitustehtäviin.

EGteht 6.

Aineisto **Viljelijä.xls**, s. 42 tehtävät:

Luo SAS-data, poista Taavetti, tallenna **vilj2006_muok**.

Kyselyllä uusi data **vilj_tiedot**, jossa muuttujina *id, nimi, sukup, paino, pituus* ja *ika*.

vilj2006_muok à **tila_tiedot**, jossa muuttujina *id, peltoa, nautoja, metsaa, tuotsuun, tulot, ajanmuk* ja *tulevais*.

Lopuksi **vilj_tiedot** à **vilj_nainen** (sukup= 2) ja **vilj_mies** (sukup= 1).

Tallenna muistikulle, näitä käytetään myöhemmin.

Data → Filter and Query

Jätettiin opiskelijoille tauon ajaksi tehtäväksi.

Esimerkki 10.

Regressioanalyysi, sama esimerkki

Girth Height and Volume for Black Cherry Trees

kuin kurssilla *Introduction to R programming* (Junior Professor Uwe Ligges)

Girth = ympärysmitta, mutta koska kyseessä halkaisija, otetaan selittäväksi muuttujaksi sen neliö. Mallinnettavasta tilanteesta johtuen kannattaa kokeilla vakiotermin poisjättämistä mallista.

Description

This data set provides measurements of the girth, height and volume of timber in 31 felled black cherry trees. Note that girth is the diameter of the tree (in inches) measured at 4 ft 6 in above the ground.

Girth Height Volume

1 8.3 70 10.3

2 8.6 65 10.3

...

<http://mtl.uta.fi/tilasto/sas/syksy08/puut/>

Datan oletuskansio E:\SAS\luennot\esimerkkejä\RA

Analyze → Regression → Linear

EGteht 7.

Yhdistäminen (s. 43) **Vilj_nainen + vilj_mies** à **vilj_molemmat**

Data → Append Table

EGteht 8.

Yhdistäminen rinnakkain (s. 45) **Vilj_molemmat + tila_tiedot** à **vilj06_takaisin**

Käsiteltiin vastaavat esimerkit luennoilla. Tehtävät Egteht7-8 voi tehdä ominpäin.

Datan kääntäminen

Jätetään harjoitustehtäväksi

Luennot ke 12.11.

Tarvittavat datat löytyvät mm. kansiossa <http://mtl.uta.fi/tilasto/sas/syksy08/dataEG/>

EGteht 4.

Formaatin luonti ja yhdistäminen muuttujaan datassa Gomex_Data.sas7bdat (tehtävä s. 30)

Formaatin säilyttäminen on ongelmallista käytettäessä useassa koneessa EG:tä, edes kirjastoviittaus ei tunnu auttavan. Ongelma on kuitenkin ratkaistavissa.

Komento Data → Create Format

Ks. tilannekuvat monisteen s. 28-30.

Esimerkki 11.

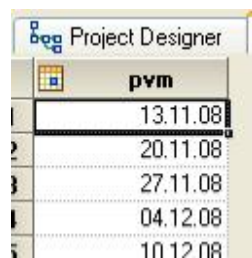
Luetaan harjoitusten päivämäärät sisältävä tekstitiedosto harjpvm.txt (kansiossa dataEG) EG:hen. ja kokeillaan erilaisia formaatteja; tilannekuvat ja kommentit

<http://mtl.uta.fi/tilasto/sas/syksy08/pvmformaatti.pdf>

Lähtökohtana siis

pvm
2008-11-13
2008-11-20
2008-11-27
2008-12-04
2008-12-10
2008-12-11

Tavoiteltava lopputulos voisi olla jotain tällaista:



The screenshot shows a window titled 'Project Designer' with a table named 'pvm'. The table contains five rows of dates. The first row is highlighted with a black border.

	pvm
1	13.11.08
2	20.11.08
3	27.11.08
4	04.12.08
5	10.12.08

Sivunumeroviittausket alla ovat Hurmeen monisteen lukua 9.

EGteht 9.

Datan listaus, vilj2006_muok, s. 72

Describe → List Data

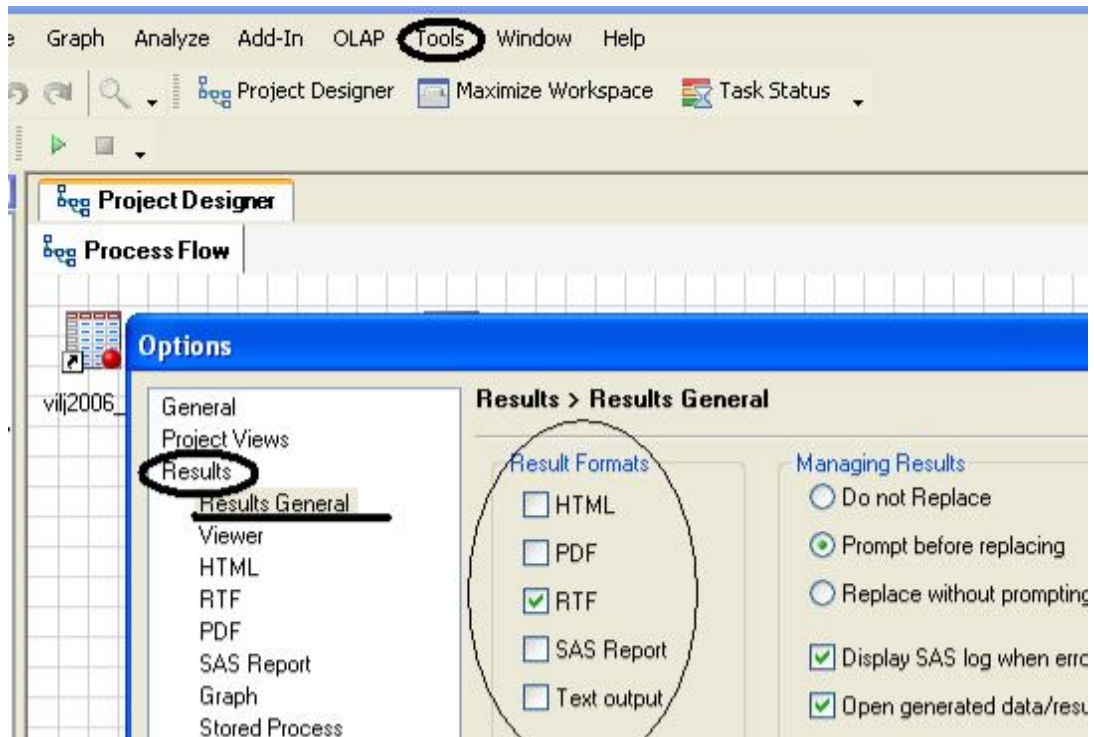
Kokeillaan myös järjestämistä: Data → Sort Data

EGteht 10.

Ikämuuttujan tunnuslukuja sukupuolen ja tuontantosuunnan eri luokkakombinaatioissa, vienti wordiin ja muokkaaminen, vilj2006_muok, s. 79

Describe → Summary Tables

Harjoitellaan erityisesti menettelyä välilehdellä Summary Tables



Tulostusmuotoa muutetaan valinnalla Tools → Options

Describe → Summary Tables sovellus aineistoon SAIDIT:

ANALYSIS MANAGER

Summary Tables

				N	paino
					Mean
sex	laani	siskm	sosryhma		

	11	1	6	1	3750.00
1	0	0	4	1	3460.00
			5	1	3210.00

Luennot ke 19.11.

Esimerkki 12.

Tutustuminen aineistoon SAIDIT Datan luonnehdinta -työkalulla:
Describe → Characterize Data

EGteht 11.

Ikämuuttujan ika jakauma, vilj2006_muok, s. 75
Describe → Distribution Analysis
Eryteisesti poikkeavat havainnot (välilehti Tables – Extreme values)

EGteht 12.

Pituusmuuttujan jakauma sukupuolen ja tuontantosuunnan eri luokkakombinaatioissa,
vilj2006_muok, s. 77 Describe → Summary statistics
Perehtyminen välilehden Results valintaan Specify ways

Itseopiskeluksi perehtyminen **wizard-työkaluihin**

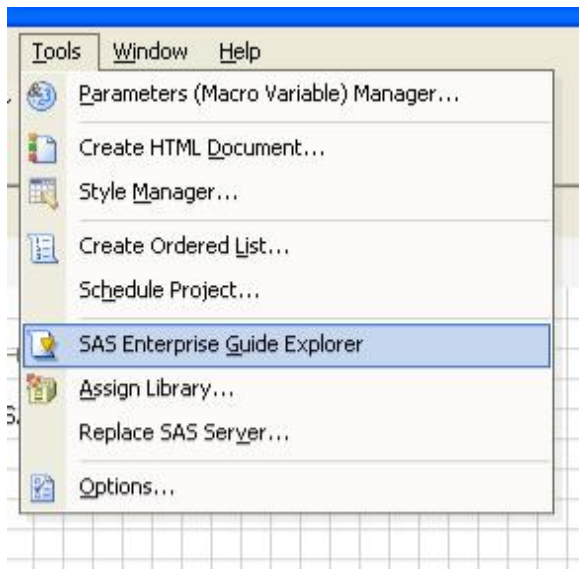
Describe → Wizards → Summary Statistics

Describe → Wizards → Summary Tables

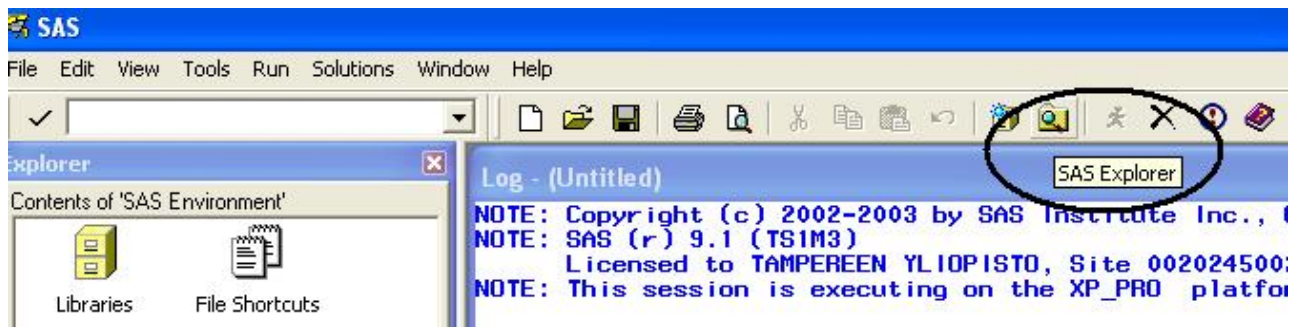
Monisteen s. 79-84

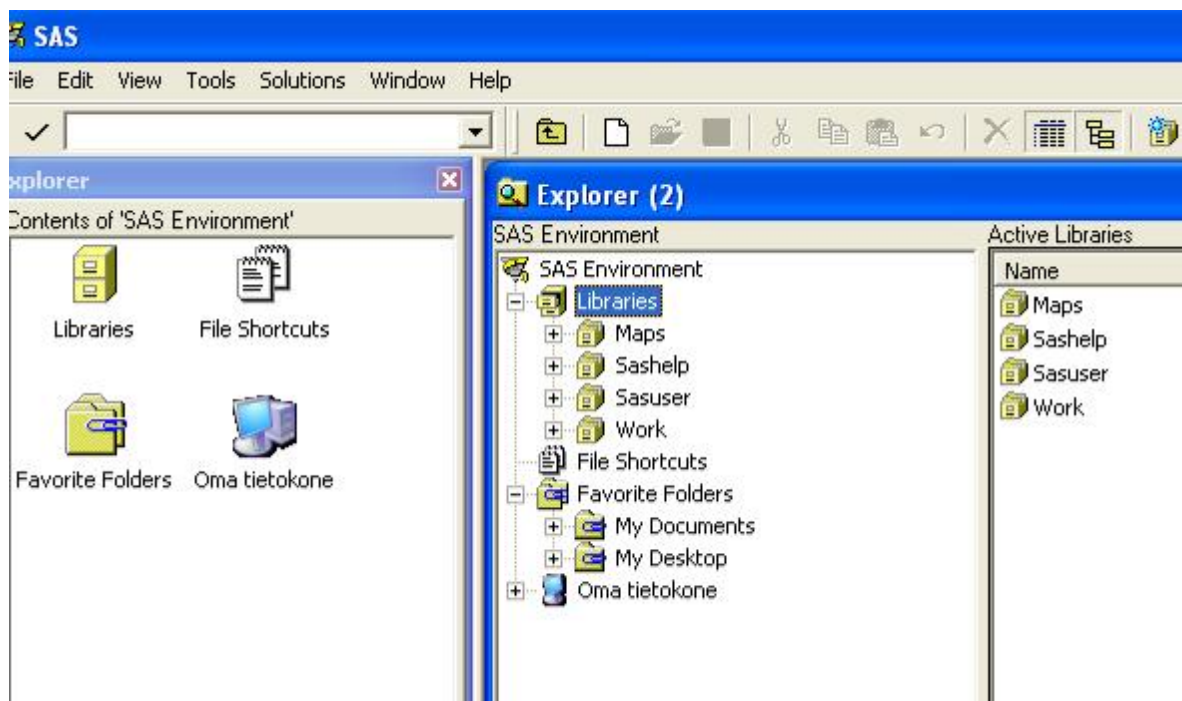
Kirjastoviittaukset SASUSER ja WORK.

EG:



SAS:

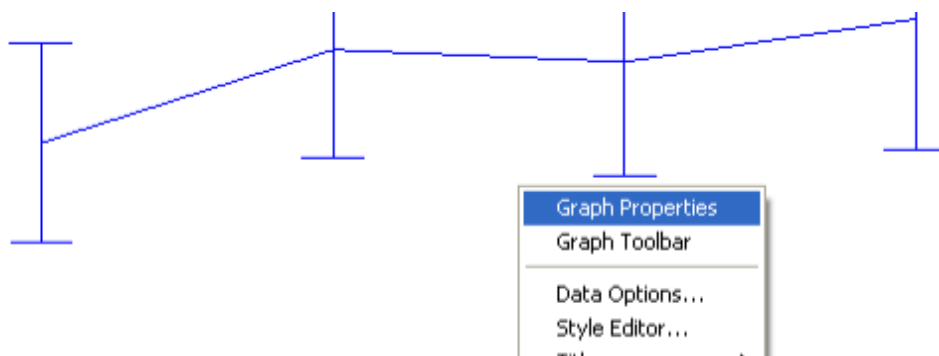




Esimerkki 13.

Kuvion graafisten ominaisuuksien muokkaus

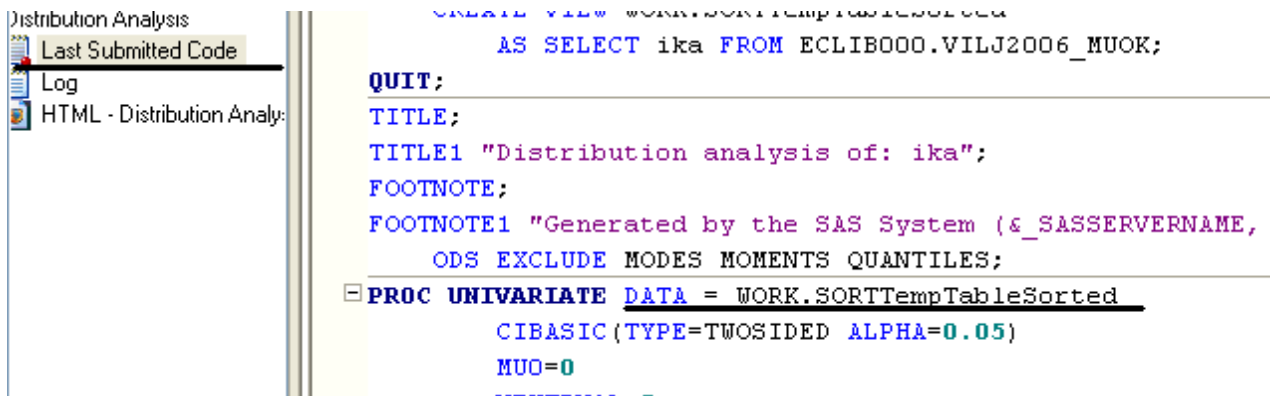
Klikkaa hiiren oikealla näppäimellä tulostetta



Jollet pääse luennotte, ks. videoleike <http://mtl.uta.fi/tilasto/sas/syksy08/grafiikka.avi>
(hidas ladattavaksi)

Esimerkki 14.

SAS-koodin etsintä



```
CREATE VIEW WORK.SORTTempTableSorted
AS SELECT ika FROM ECLIB000.VILJ2006_MUOK;

QUIT;
TITLE;
TITLE1 "Distribution analysis of: ika";
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (&_SASSERVERNAME,
ODS EXCLUDE MODES MOMENTS QUANTILES;
PROC UNIVARIATE DATA = WORK.SORTTempTableSorted
CIBASIC (TYPE=TWO SIDED ALPHA=0.05)
MUO=0
```

Esimerkki 15.

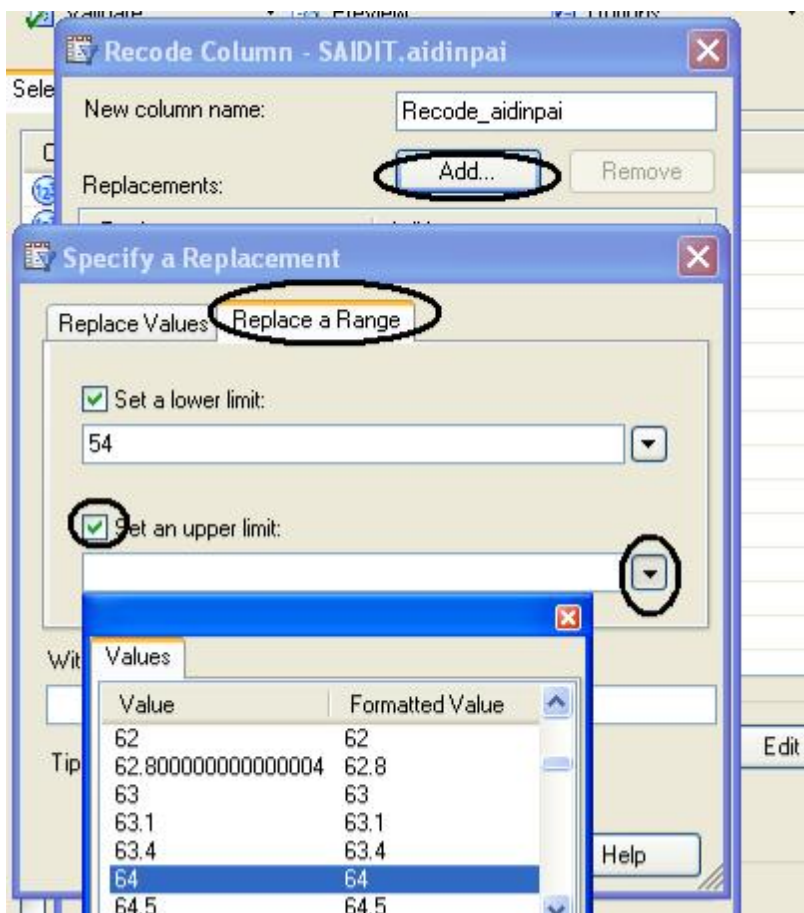
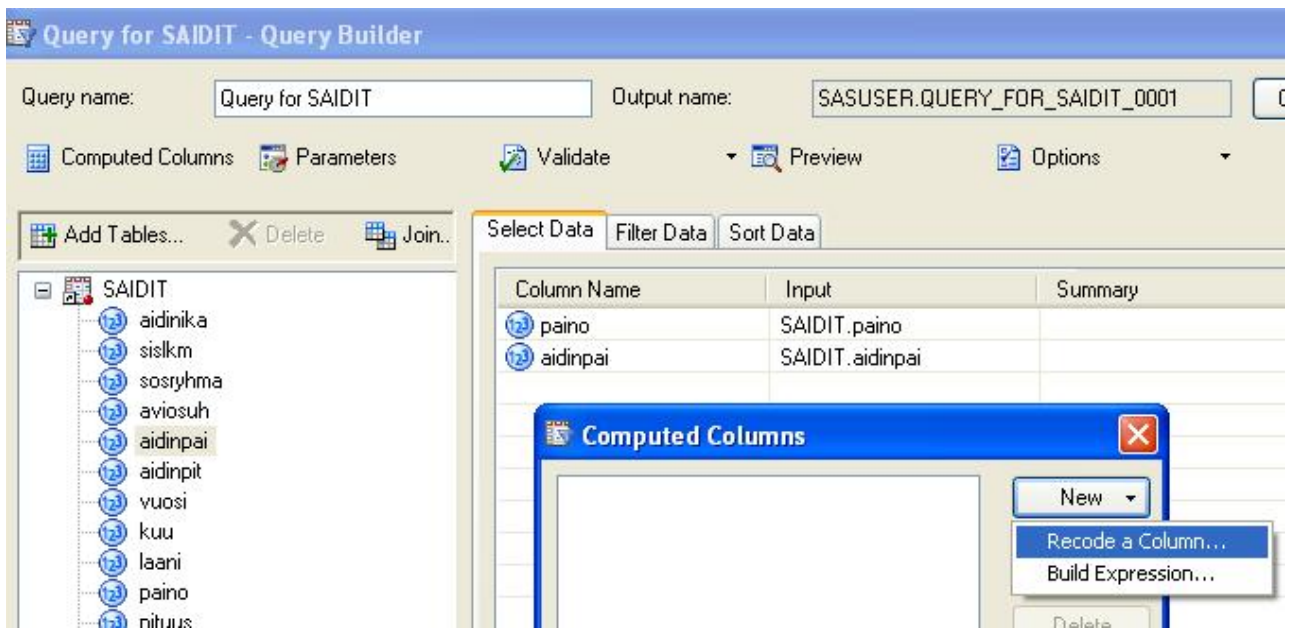
Äidin painon -muuttujan luokittelu datamatriisissa SAIDIT neljään yhtä suureen luokkaan.

Tehdään luokitus ensin formaattien avulla

ks. <http://mtl.uta.fi/tilasto/sas/syksy08/luokitus1911.pdf>

E:\SAS\luennot\esimerkkejä\luokitus

Seuraavaksi sama: Data → Filter and Query



Luennot to 27.11.

Vierailu

Apulaisjohtaja Jarkko Jännes tietokonekeskuksen Valos-tiimistä:

SAS:n käyttö opetustoimen valtionosuuksien laskennassa

Harjoitustöistä sopiminen

Jos muuta harjoitustyötä ei keksi, niin voi tehdä monisteen

http://www.sascommunity.org/wiki/Statistics_Using_SAS_Enterprise_Guide

tehtäviä niin että kustakin luvusta **5, 6 ja 7** tekee kaksi tehtävää niin että kohtia tulee yhteensä vähintään 10 (luennoilla oli virheellisesti puhetta luvuista 6,7 ja 8).

Harjoitusten 3 tehtävän 2 ja 3 käsittely

Seuraavat (viimeiset) luennot

Esimerkki 16.

Otsikkotietojen muuttaminen



Esimerkki 17.

SAS Raportti, monisteen s. 102-103

File → New → Report

Harjoitustyön saa raportoida tällä. Tämän käyttö ei kuitenkaan onnistu mikroluoka 42 koneilla:

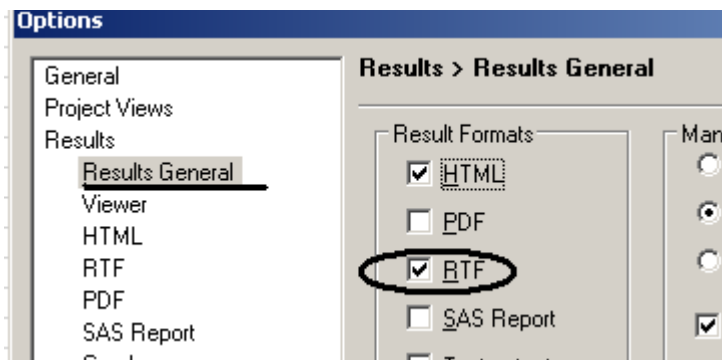
WARNING: SAS Report XML requires 9.1.3 service pack 4 or better and is not supported on this version of SAS.

Pitäisi ajaa korjauspaketit osoitteesta

http://ftp.sas.com/techsup/download/hotfix/ent_guide41.html#41eg06

Tämä onnistuu kotikoneeseen, mutta teknisten ongelmien vuoksi korjauspaketteja ei ole ajettu mikroluokan koneissa.

Harjoitustyön raportointi onnistuu kuitenkin Wordilla. Rastitetaan vain valikosta Tools → Options välilehdestä Result General vaihtoehto RTF.



Tällöin tulosteina olevia kuvia, tekstejä yms. pystyy kopioimaan siirtämään Wordiin Copy-Paste –toiminnolla.

Esimerkki 18.

Avaa ensin kansioista <http://mtl.uta.fi/tilasto/sas/syksy08/dataEG/> **uusipuut**

Määrittele siihen liittyen makromuuttuja muuttuja

Tools → Parameters (Macro Variable) Manager

Kokeile sitten makromuuttujan käyttöä komennossa

Graph → Bubble Plot

Esimerkki 19.

Datamatriisi **uusipuut**

Data → Rank

Graph → Bubble Plot

Korotetaan sitten vielä järjestysnumero (rank) toiseen

Data → Filter and Query

Esimerkki 20.

Avaa ensin kansiota <http://mtl.uta.fi/tilasto/sas/syksy08/dataEG/> **koepisteetA** ja **koepisteetB**
Standardisoi (Data → Standardize Data) datamatriisin **koepisteetA** muuttujat *koe1* ja *koe2*. Anna saadulle datamatriisille nimi **koepisteetC**.

Esimerkki 21.

Vertaile (Data → Compare Data), ovatko tiedostot **koepisteetB** ja **koepisteetC** samat (kirjastoviite ECLIB000).

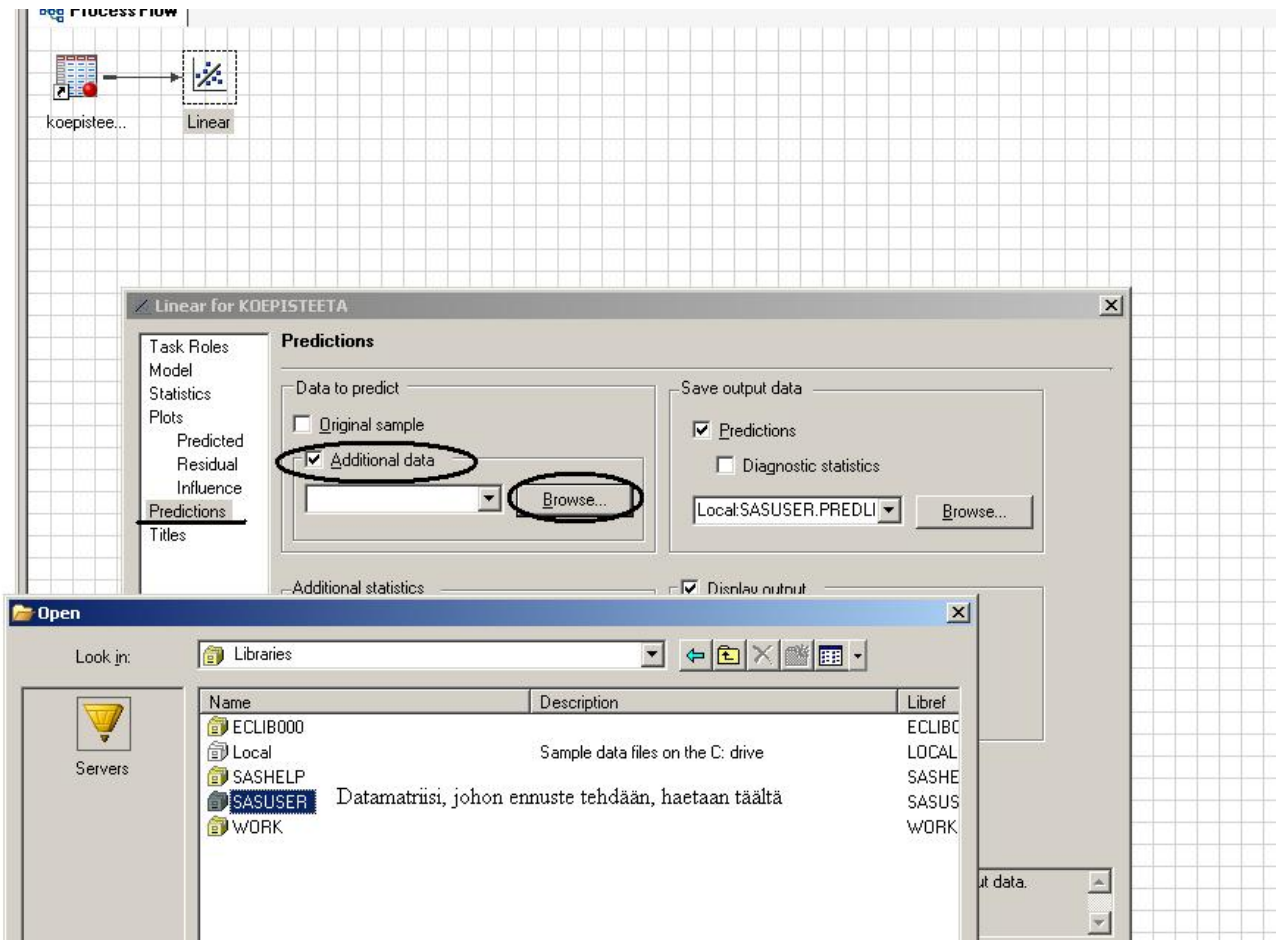
Esimerkki 22.

Edustakoon datamatriisi kahden välikokeen tuloksia. Opiskelija saa ensimmäisestä välikokeesta (*koe1*) 13 pistettä ja on hyväksyttävästä syystä pois toisesta välikokeesta. Mikä olisi oikeudenmukainen arvio toisen välikokeen (*koe2*) pistemääräksi?

Tehdään tämä ensin standardisoiduilla pisteillä datamatriisin **koepisteetB** perusteella. (Tätä ei konkreettisesti laskettu). Kokeillaan sitten ennustamista regressioanalyysillä Analyze → Regression → Linear datamatriisiin **koepisteetA**. Kumpikohan olisi oikeudenmukaisempi menettely?

Esimerkki 23.

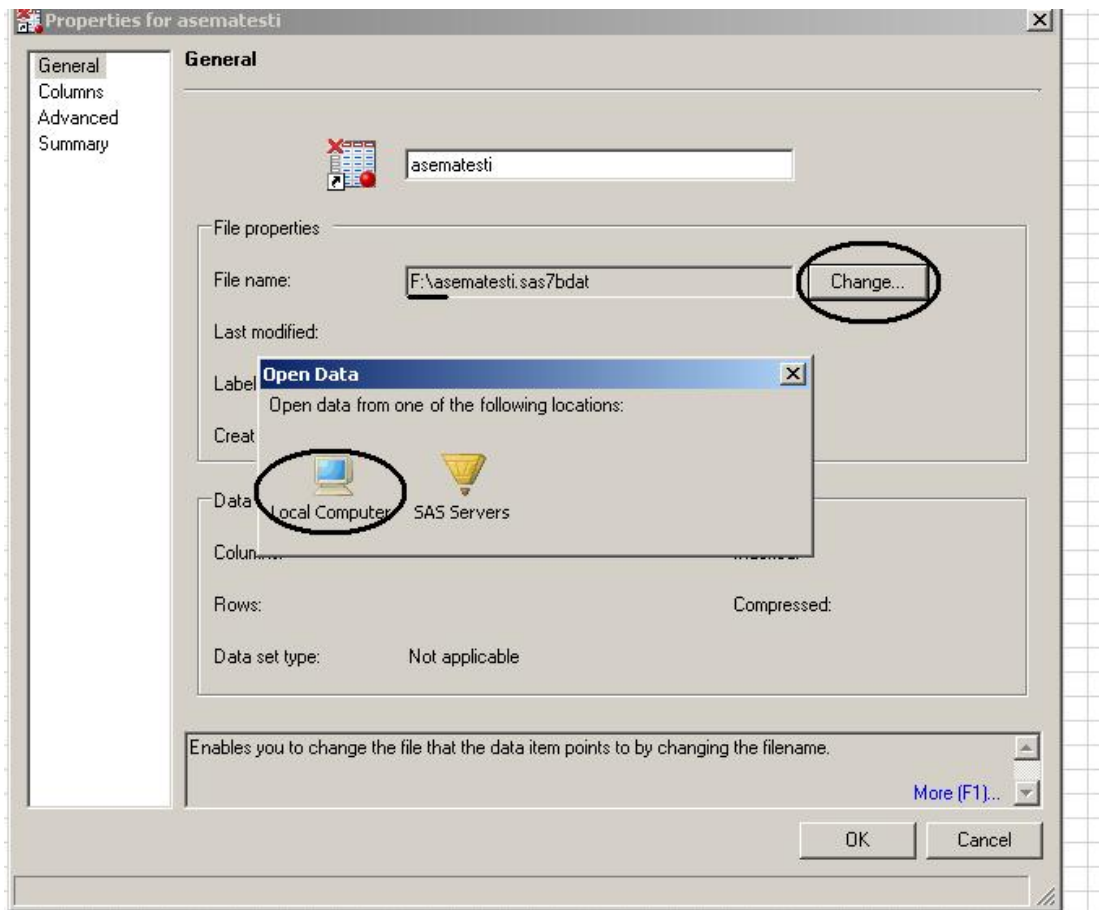
Jatkoa edelliseen. Tiedostossa **ekakoe** on muitakin sellaisia ensimmäisessä kokeessa olleita, jotka olivat poissa toisesta kokeesta. Lasketaan heilläkin pisteet regressioanalyysillä. Tämä sivuutettiin luennoilla



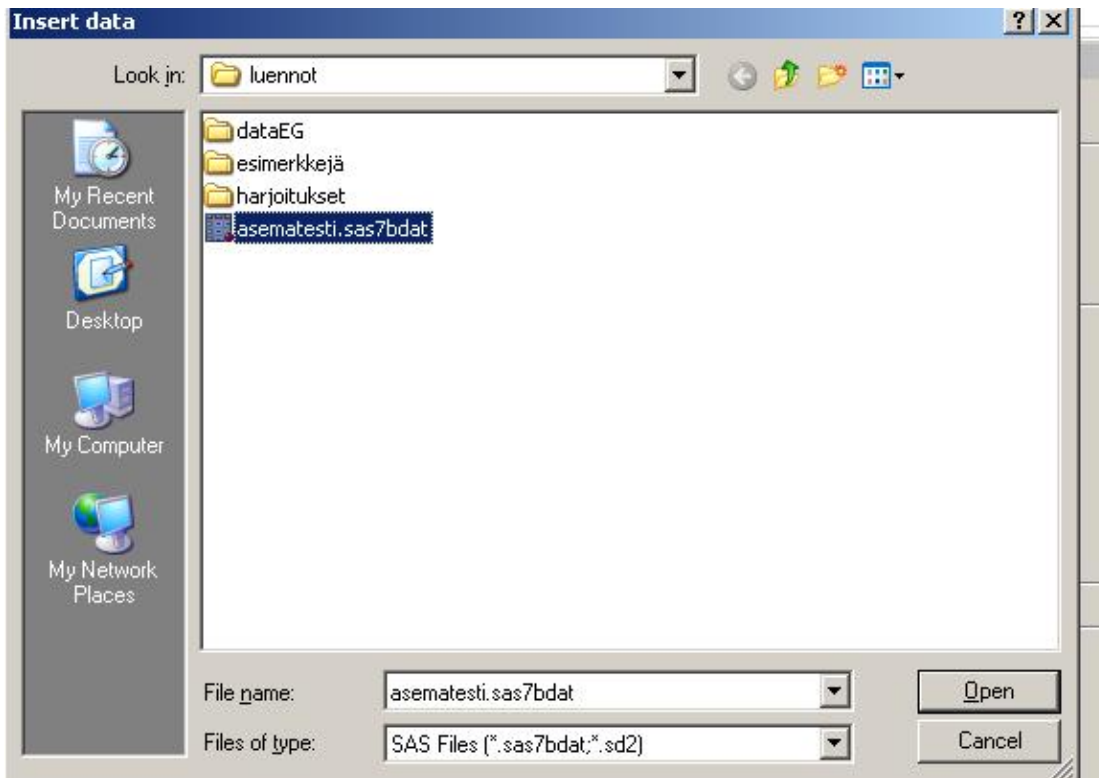
DATAMATRIISIN OSOITE VÄÄRÄ

Jos datiedostosi on tehty esim. Linnan mikroluokassa ja asematunnus ei toimi mikroluokassa 42, niin asia on korjattavissa seuraavasti:

Klikkaa hiiren oiekalla näppäimellä datamatriisia, valitse Properties. Jatka sitten valinnoilla Change – Local Computer



Ja valitse datamatriisi uudelleen (esimerkiksi muistitikulta) Open-komennolla



Tämän jälkeen datamatriisi on taas käytettävissä projektissasi.