HELSINGIN YLIOPISTON
TILASTOTIETEEN LAITOS

Juha Puranen – Mikko Virtanen – Mika Lahdenkari – Heikki Hyhkö – Kimmo Vehkalahti

Data-analyysi I

Survo-kurssi

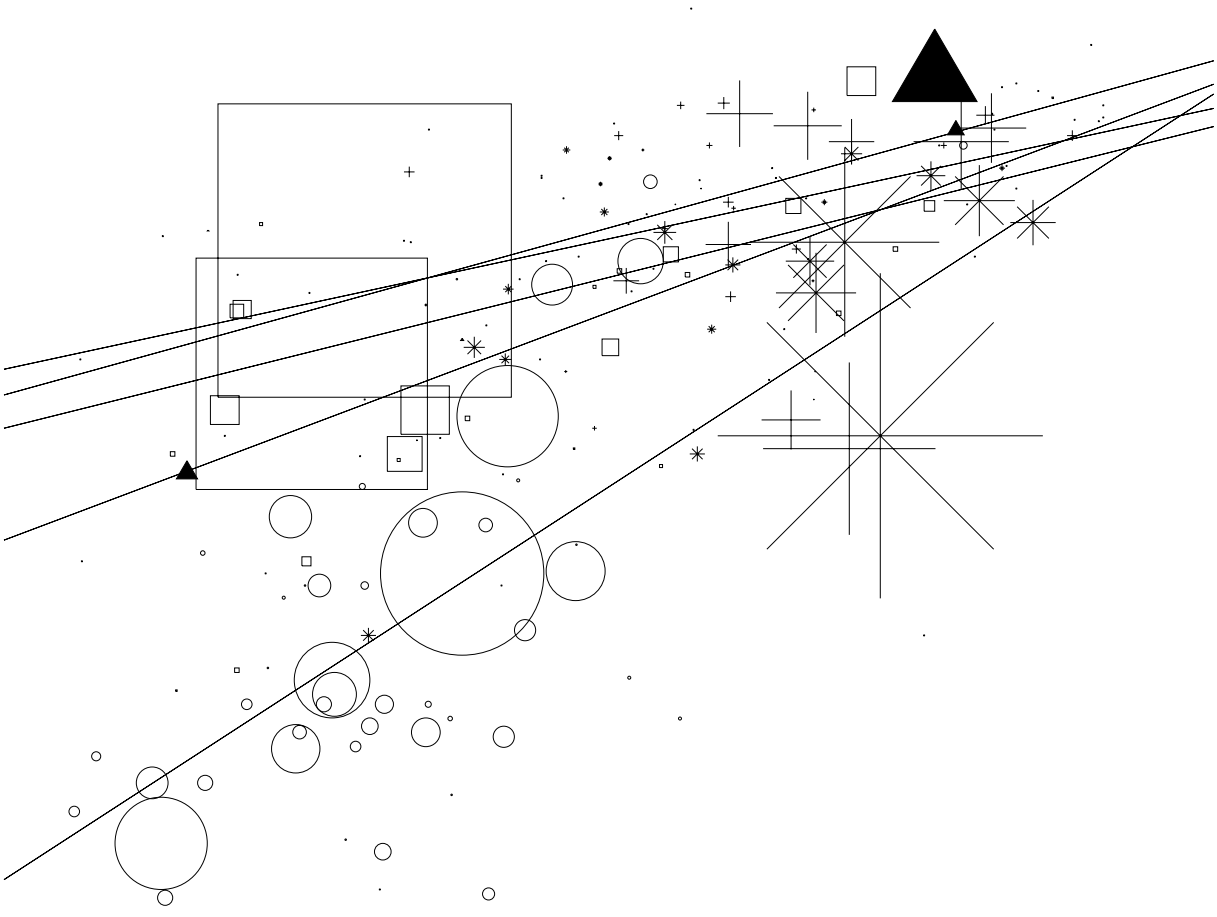
lukuvuonna 2003-2004

Moniste sisältää Helsingin yliopiston tilastotieteen laitoksen (**www.valt.helsinki.fi/til**) opintojaksoon **Data-analyysi I** kuuluvan **Survo-kurssin** harjoitustehtävät. Kurssiin kuuluu luentojen lisäksi 30 tuntia ohjattua harjoittelua atk-luokassa. Kaikkia tehtäviä ei ehditä yleensä käsitellä harjoitusten aikana. Loput on tarkoitettu tehtäväksi itsenäisesti omalla ajalla.

Survo on professori *Seppo Mustosen* kehittämä ohjelmisto *tekstin ja numeerisen tiedon luovaan käsittelyyn*. Kurssilla käytetään Survon Windows-versiota **SURVO MM**. Verkkosivuilta **www.survo.fi** löytyy runsaasti esimerkkejä Survon käytöstä ja tietoa mm. sen historiasta. Kirja **Survo ja minä** (Mustonen 1996) soveltuu hyvin kurssin oheislukemistoksi.

Monisteen lopussa on lyhyet ohjeet kurssin harjoitustyön laatimiseksi. Tarkempia neuvoja löytyy lehtori *Juha Purasen* ylläpitämältä sivulta **noppa5.pc.helsinki.fi/uudet/da1htm** .

SURVO MM on saatavana erittäin edullisesti yliopistojen ja korkeakoulujen opetus- ja tutkimuskäyttöön, ks. **www.survo.fi/hinnasto** (oppilaitoslisenssit). Helsingin yliopiston osalta tietoja löytyy atk-osaston sivulta **www.helsinki.fi/atk/tilasto/SURVohjmyy.html** .



Sisällys

Tehtävä 1: Tutustuminen Survoon

Survon verkkosivuilla **www.survo.fi** todetaan, että Survo tarjoaa monipuolisen työskentely-ympäristön mm. seuraavilla tehtäväalueilla:

- tekstinkäsittely ja taulukkolaskenta
- esitys- ja tilastografiikka
- painettujen julkaisujen ja verkkosivujen teko
- toistuvien raporttien automatisointi
- laajojen tietokantojen hallinta
- tilastollinen laskenta ja analyysi
- numeerinen laskenta, matriisitulkki
- asiantuntijasovellusten ohjelmointi
- opetusohjelmien laatiminen

Tietojenkäsittelyjärjestelmänä Survolla on poikkeuksellisen pitkä historia, sillä ensimmäisillä Survoilla tehtiin töitä jo 1960-luvulla. Viimeisen 20 vuoden ajan Survolle on ollut ominaista sen *editoriaalinen käyttötapa*, joka tarkoittaa kaikkien tehtävien hoitamista järjestelmän omalla teksturilla. Tekstin sekaan voi kirjoittaa mm. komentoja ja laskutehtäviä, jotka Survo automaattisesti tunnistaa kun ne aktivoidaan. Tulokset ilmestyvät välittömästi samaan tekstinkäsittelytilaan ja käyttäjä voi jatkaa niiden pohjalta työskentelyään.

Survon uudessa Windows-versiossa (**SURVO MM**) komennot ja operaatiot aktivoidaan joko hiirellä (kaksoisklikkaus) tai `Esc`-napilla. Valikoissa valinnat tapahtuvat joko hiirellä (tavallinen klikkaus) tai `Enter`iä painamalla. Survossa voi asiat tehdä monella vaihtoehtoisella tavalla. Kannattaa noudattaa seuraavia suosituksia (näitä harjoitellaan myöhemmin):

Kirjoita **komennot isoilla kirjaimilla** (esim. `LINREG`, `CLASSIFY`, `PRINT`).

Kirjoita **täsmennykset isoilla ja yhteen** (esim. `RESULTS=CSUMS,RSUMS`).

Sijoita **talletuskomento** (SAVE) toimituskentän **ensimmäiselle riville**.

Talleta työsi (toimituskenttäsi) ennen kuin lähdet seikkailemaan Survon valikoissa tai kokeilemaan uusia toimintoja. Muista myös tallettaa työsi ennen kuin poistut Survosta.

Älä ahda kaikkia töitäsi samaan toimituskenttään. Talleta mieluummin jokainen tehtävä omaan kenttäänsä. **Nimeä lyhyesti ja ytimekkäästi**, esim. `DA1T1`, `DA1T2` jne.

Erota komentokaaviot toisistaan **rajariveillä**, etteivät niiden täsmennykset sekaannu.

Eräitä komentoja ja näppäinyhdistelmiä:

`SCRATCH`	Toimituskentän tyhjennys	`Alt-F9`	Rivin lisäys
`DD`	Työhakemiston sisällön selailu	`Alt-F10`	Rivin poisto
`SAVE <nimi>`	Toimituskentän talletus levylle	`Alt-F3`	Rivin kopiointi
`LOAD <nimi>`	Talletetun kentän käyttöönotto	`Alt-F4`	Alueen kopiointi
`SHOW <nimi>`	Toisen toimituskentän selailu	`Ctrl-End`	Rivin tyhjennys

Survo käynnistetään kuten Windows-ohjelmat yleensä, klikkaamalla sen kuvaketta.

Survosta poistutaan joko `F8`-napilla tai kaksoisklikkaamalla alarivin `EXIT`-painiketta.

Opetusohjelmat ja neuvontajärjestelmä

Survon laajaan suomenkieliseen opetusohjelmasarjaan pääsee esimerkiksi Survon alkuvalikosta (START) tai `DEMO`-painikkeen kautta (`OPETUS`). Käy alkajaisiksi läpi opetusohjelmasarjan kohdat 1. Yleisiä tietoja, 2. Opetusohjelmien käyttö ja 5. Neuvonta käytön aikana.

Survon neuvontajärjestelmä eli kyselysysteemi on jatkuvasti ajantasainen hypertekstimuotoinen tietolähde. Se toimii omassa ikkunassaan, josta voi tehdä valintoja ja jatkokyselyjä sekä hiiren että näppäimistön avulla. Kyselysysteemiin päästään mm. aktivoimalla avainsana, jonka perään on kirjoitettu kysymysmerkki (esim. `STAT?`). Kyselyn voi aktivoida myös hiiren oikealla napilla (kaksoisklikkaus) tai näppäinyhdistelmällä `F2-F1` (ts. ensin painetaan kerran `F2` ja sitten kerran `F1`), jolloin saadaan tietoa kohdistimen osoittamasta sanasta.

Kyselysysteemin ohjeissa on komentojen parametrit usein esitetty merkkien < ja > välissä. Esimerkiksi `CORR <data>,L` tarkoittaa, että parametriksi tarvitaan tilastoaineisto. Merkintä `L` viittaa yleensä tulostuksen aloitusriviin. Komento voisi olla kokonaisuudessaan esimerkiksi `CORR KULUTUS,END+2` tai `CORR SUOMI,CUR+1` (`END` tarkoittaa kentän viimeistä käytössä olevaa riviä ja `CUR` sitä riviä jolla kohdistin on kun komento aktivoidaan).

Tehtävä 2: Työskentely tilastotieteen laitoksen atk-luokassa

Atk-luokan koneiden käyttöön tarvitaan Helsingin yliopiston atk-osaston myöntämä mikroverkon käyttäjätunnus. Sivulta **www.mv.helsinki.fi/mikroverkko** saa lisätietoja.

Talleta työsi omalle verkkolevyllesi (yleensä `R:\`). Luo esimerkiksi hakemisto (eli kansio) `R:\DA1`, jossa toimit aina harjoitusten aikana. Voit tehdä sen Survossa seuraavasti:

```
MD R:\DA1
```

Siirry näin luomaasi työhakemistoon (jatkossa tee tämä aina Survo-työskentelyn aluksi!):

```
CD R:\DA1
```

Nyt hakemiston nimi näkyy Survo-ikkunan oikeassa yläkulmassa. Se tarkoittaa, että kaikki työskentelysi jäljet tallettuvat sinne, ja ovat seuraavalla kerrallakin käytettävissäsi. Varmuuskopioita on silti hyvä tehdä esimerkiksi levykkeelle tai vaikka USB-porttiin liitettävälle taskukokoiselle muistipalikalle.

Töiden hallinta käy kätevästi, kun opettelet käyttämään Survon työpöytäohjelmia (ks. lisää kyselyllä `DESKTOP?`). Esim. komennoista

```
DD
```

tai

```
DM R:\DA1 A:\
```

saattaa olla hyötyä, kun tiedostoja kertyy enemmän. Myös eräät käyttöjärjestelmän komennot (esim. `DIR`, `COPY`, `DEL`) ovat hyödyllisiä. Survossa näitä komentoja voidaan aktivoida kuten Survon omia komentoja laittamalla niiden eteen merkki >. Esimerkiksi näin:

```
>DIR A:
```

Lisää ohjeita töiden hallinnasta löydät Survon alkuvalikon (`START`) kohdasta 8.

Tehtävä 3: Survon käytön perusteet

Erittäin tärkeän osan tutkimusprosessista muodostaa **oman työn dokumentointi**. Se tarkoittaa sitä että pystyy myöhemmin ottamaan esille aiemmin tallettamansa työn ja palauttamaan mieliin *mitä tuolloin ajatteli*. Tutkimustyössä joudutaan usein palaamaan muutama askel taaksepäin, esimerkiksi erilaisten tarkistusten vuoksi. Jos työstä on jäljellä pelkät tulokset, voi olla mahdotonta tai ainakin kovin työlästä johtaa niitä uudelleen.

Survon käyttö perustuu **tekstinkäsittelyyn**. Työskentely tapahtuu **toimituskentässä**, johon kirjoitetaan tekstiä, komentoja, piirroskaavioita jne. Survolla työskennellessä **työ dokumentoi itsensä**, koska työohjeet ja vastaavat tulokset tallentuvat **samaan tilaan**. Dokumentaatiota on helppo kehittää niin yksityiskohtaiseksi kuin haluaa, koska sekaan voi vapaasti kirjoittaa omia kommenttejaan ja huomioitaan ja korostaa niitä mm. **erilaisilla väreillä**.

Jotta dokumentteja työskentelystä jäisi jäljelle, on toimituskenttä aika ajoin talletettava levylle. **Talletus on sinun vastuullasi**, Survo ei huolehdi siitä puolestasi. Tässä tehtävässä tutustutaankin Survon käytön keskeisiin tehtäviin: toimituskentän tallettamiseen sekä tekstin kirjoittamiseen ja muokkaamiseen. Mitä paremmin hallitsee nämä perusasiat, sitä helpompaa Survon käyttö on.

Seuraavilla sivuilla tarkastellaan vaiheittain, miten uusi toimituskenttä perustetaan, nimetään ja talletetaan sekä miten tekstiä muokataan, kopioidaan ja muotoillaan. Sen jälkeen voidaankin siirtyä tilastoaineiston tallentamiseen ja tutkimiseen.

Uusi toimituskenttä perustetaan vaiheittain seuraavasti:

Nykyisen kentän tyhjentäminen:

Siirry kentän ensimmäiselle riville painamalla riittävän monta kertaa HOME-nappia.

Jos rivillä on ennestään tekstiä, poista se painamalla Ctrl-End. Mikäli rivi ei tyhjentynyt, paina uudelleen (mahdollisista värikkäistä teksteistä poistuvat ensin vain värit).

Kirjoita rivin alkuun komento SCRATCH ja aktivoi se. (Oliko hankala sana kirjoittaa? Se on tarkoituskin, jottei vahingossa tuhoaisi työtään. Komennon saa kyllä helpomminkin näppäinyhdistelmällä F2 ja 1. Kokeile!)

Uuden kentän nimeäminen:

Kirjoita kentän talletuskomento ja kommentti ensimmäiselle riville. Tämä on suositus, joka kannattaa ottaa rutiiniksi. Hyvien nimien keksiminen on usein atk:ssa vaikeinta. Tiedostojen nimissä kannattaa pysytellä korkeintaan kahdeksan merkin pituudessa, ja säästää pidemmät selitykset kommentteihin. Toimituskentät saavat automaattisesti tiedostopäätteen .EDT joten sitä ei tarvitse kirjoittaa. Olisi hyvä varata tästä eteenpäin joka tehtävälle oma toimituskenttänsä. Nimet on syytä valita systemaattisesti, esim. DA1T3, TEHT3 tms.

Kommentti on vapaamuotoinen teksti komennon jäljessä. Erottimena on kauttaviiva, jonka molemmin puolin on oltava ainakin yksi tyhjä (välilyönti). Ensimmäinen rivi voisi olla kokonaisuudessaan esimerkiksi seuraavanlainen:

```
SAVE DA1T3 / tekstinkäsittelyn harjoittelua
```

Kentän tallettaminen:

Aktivoi SAVE-komento. Riittää että kohdistin on samalla rivillä. Tällä tavalla voit tallettaa kentän juuri niin usein kuin haluat. Ensimmäiselle riville pääset aina helposti HOME-napilla. Paina se vaikka pohjaan hetkeksi, minkä jälkeen voit aktivoida talletuskomennon ja palata sitten jatkamaan työskentelyä.

Uusi kenttä on nyt perustettu. Jätä aina muutama tyhjä rivi tulostusta, töiden hallintaa ym. varten. Saadaksesi materiaalia tekstin muokkausta varten kirjoita tämän tehtävän kolmen ensimmäisen kappaleen tekstit toimituskenttään. Älä välitä tavutuksista, rivijaosta äläkä korostuksista. Talleta kenttä välillä.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003            R:\DA1\    1000  100 0
  1 *SAVE DA1T3 / tekstinkäsittelyn harjoittelua
  2 *
  3 *
  4 *DM R:\DA1 A:\
  5 *
  6 *>dir a:\*.edt
  7 *
  8 *Erittäin tärkeän osan tutkimusprosessista muodostaa oman työn
  9 *dokumentointi. Se tarkoittaa sitä että pystyy myöhemmin ottamaan
 10 *esille aiemmin tallettamansa työn ja palauttamaan mieliin mitä
 11 *tuolloin ajatteli.
 12 *
```

Toimituskentän yläreunan palkissa näkyy kuluvan ajan lisäksi nykyisen työhakemiston nimi. Kaikki työskentelyn kuluessa syntyvät tiedostot tallettuvat tähän hakemistoon. Hakemistoa voidaan vaihtaa ja kokonaan uusia perustaa, mutta se ei ole näissä harjoituksissa tarpeen. Laajemmissa töissä on järkevää perustaa jokaiselle työkokonaisuudelle, projektille tms. oma hakemistonsa.

Tutki työhakemistosi sisältöä DD-komennolla. Takaisin toimituskenttään pääset F8-napilla. Voit tehdä hakemistosi sisällöstä myös luettelon ja tallettaa sen omaksi toimituskentäkseen, josta pääset helposti käsiksi varsinaisiin työkenttiisi. Vakiintunut nimi tällaiselle kentälle on INDEX, mutta voit myös nimetä sen miten haluat. Luettelon saat aikaan samannimisellä komennolla (INDEX). Huomaa, että edellä antamasi kommentit tulevat myös automaattisesti mukaan luetteloon helpottaen sen selailua.

Tekstin muokkaus, kopiointi ja muotoilu

Tekstinkäsittely Survolla muistuttaa millä tahansa teksturilla työskentelyä. Eroavaisuuksia on, sillä Survon toimiala on huomattavasti tekstureita laajempi. Tulee myös muistaa että Survon editorin pelisäännöt on laadittu jo kauan ennen nykyisiä tekstureita, eikä pelisääntöjä ole syytä mennä muuttelemaan. Survossa yhteensopivuus aiempiin versioihin on asia, josta pidetään tiukasti kiinni.

Tekstin muokkaaminen tapahtuu näppäinten ja komentojen avulla. Ellei siirrytä lisäystilaan `Ins`-napilla, kirjoitetaan entisen tekstin päälle. Kappaleita ei automaattisesti tasata vaan se tehdään tarvittaessa erikseen `TRIM`-komennoilla.

Uusi rivi lisätään nykyisen alapuolelle painamalla `Alt-F9`. Nykyinen rivi poistetaan vastaavasti `Alt-F10`:llä. `Ctrl-End` tyhjentää rivin kohdistimesta oikealle. Yksi rivi kopioidaan kohdistimen osoittamaan paikkaan napilla `Alt-F3` (kirjoita kopioitavan rivin numero ja paina `Enter`). Yksityiskohtaista ohjausta näppäimien käytössä löydät suomenkielisestä opetussarjasta (`DEMO` → `OPETUS` → 4. Tekstinkäsittelyn alkeita).

Kokeile kaikkia edellä mainittuja näppäinyhdistelmiä. Jos saat tekstisi aivan sekaisin, ota se uudelleen esiin `LOAD`-komennolla. Tämä on Survon yleinen *"undo"*- eli *"eiku"*-tekniikka, jota edistyneemmätkin käyttäjät soveltavat tuon tuosta: jos et ole aivan varma mitä jokin komento tekee, kokeile, mutta talleta sitä ennen kenttäsi. Sen jälkeen voit helposti lavastaa saman tilanteen ja yrittää uudelleen kunnes onnistut.

Kokeile myös alueen kopiointia. Ideana on määritellä kentässä suorakulmio merkkaamalla sen vasen yläkulma ja oikea alakulma. Kaikki tapahtuu napilla `Alt-F4`, jota painetaan useammassa vaiheessa. Ensimmäinen painallus antaa vielä mahdollisuuden siirtyä oikeaan paikkaan (alueen vasempaan yläkulmaan). Toinen painallus aloittaa varsinaisen merkkauksen. Kolmas painallus annetaan oikeassa alakulmassa, minkä jälkeen siirrytään sinne mihin alue halutaan kopioida. Alue näkyy nyt kentässä korostettuna. Neljäs painallus kopioi alueen (tätä vaihetta voi toistaa rajattomasti). Alkuperäisestä paikasta alue poistuu jos painetaan `Ctrl-End`. Pelkkä `Del` lopettaa ja jättää alkuperäisen alueen paikalleen.

Muotoile lopuksi tekstikappaleet `TRIM`-komennolla. Katso lisätietoja kyselysysteemistä.

Tehtävä 4: Tilastoaineiston tallentaminen

Tarkastellaan esimerkkinä tietoja 111 henkilön koepisteistä. Perusta uusi toimituskenttä ja kirjoita aineisto siihen oheisen mallin mukaan. Aineiston nimi on siis `EXAMS` ja sen ainoan muuttujan nimi `Scores`. Ole huolellinen, aineistoa tarvitaan lukuisissa tehtävissä.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003           R:\DA1\    1000  100 0
  1 *SAVE DA1T4 / tilastoaineiston tallentaminen
  2 *
  3 *PVM / Torstai, 23. tammikuuta 2003 klo 11.24 (vko 4, pvä 23/365)
  4 *
  5 *DATA EXAMS:(Scores)
  6 * 77 79 62 70 87 75 94 89 71 82 75 65 93 63 66 62 76 75 75 75 82 70 86
  7 * 84 86 74 81 64 91 82 87 87 75 92 78 72 69 60 72 92 69 61 81 80 85 69
  8 * 89 72 52 73 81 78 83 81 44 70 56 82 54 73 76 79 75 71 83 49 77 87 88
  9 * 90 70 66 67 69 90 84 83 72 68 77 78 90 77 80 72 79 80 75 51 48 72 37
 10 * 91 68 79 78 64 31 73 86 68 72 85 60 89 73 88 52 90 97 93 END
 11 *
 12 *Lasketaan perustunnusluvut:
 13 *
 14 *STAT EXAMS,CUR+1 / tulokset seuraavalta riviltä alkaen
 15 *Basic statistics: EXAMS N=111
 16 *Variable: Scores
 17 *min=31       in obs.#98
 18 *max=97       in obs.#110
 19 *mean=75       stddev=12.36417 skewness=-0.928991 kurtosis=1.158136
 20 *lower_Q=68.76471 median=75.73333 upper_Q=83.61538
```

Tallennusvirheiden varalta on syytä vielä tarkistaa tallennetut luvut. STAT-operaatiolla voidaan lisäksi laskea perustunnusluvut, jotka antavat nopeasti yleiskäsityksen aineistosta. Vertaa omaa aineistoasi oheiseen valmiiksi laskettuun esimerkkiin. Onko havaintoja (`N`) yhtä paljon? Ovatko `Scores`-muuttujan pienin (`min`) ja suurin (`max`) arvo samat kuin esimerkissä? Entä keskiarvo (`mean`) ja keskihajonta (`stddev`)? Tai alakvartiili (`lower_Q`), mediaani (`median`) ja yläkvartiili (`upper_Q`)? STAT tulostaa näiden tietojen lisäksi myös luokitetun frekvenssijakauman, josta saa käsityksen muuttujan jakauman muodosta.

Tehtävä 5: Aineiston talletus havaintotiedostoon

Myös useamman muuttujan havaintomatriisi voidaan tallettaa toimituskenttään DATA-taulukoksi (ks. DATA?). Tässä tehtävässä harjoitellaan kuitenkin aineiston talletusta suoraan havaintotiedostoon. Useimmiten on järkevintä pitää tilastoaineisto erillään toimituskentästä.

Esimerkkinä tarkastellaan pientä palkka-aineistoa. Eräs helppo tapa perustaa havaintotiedosto on seuraava. Kirjoitetaan ensin DATA-määrittely, aineiston muuttujanimet ja yksi tyypillinen havainto. Sen jälkeen varsinainen havaintotiedosto perustetaan /DATACOPY-sukrolla. Mahdolliset lisätiedot aineistosta (kuten esim. alkuperä ja muuttujien kuvaukset) kannattaa liittää mukaan havaintotiedostoon LINES-täsmennyksen avulla.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003            R:\DA1\    1000  100 0
  1 *SAVE DA1T5 / aineiston talletus havaintotiedostoon
  2 *
  3 *Tietoja palkka-aineistosta:
  4 *
  5 *Dataset  : SALARY.DAT
  6 *Topic    : Faculty salaries at a liberal arts college (1980)
  7 *Source   : Survey    Variables: 4      Cases    : 85
  8 *  1. RANK  : 1 - Instructor or other      3 - Associate professor
  9 *             2 - Assistant professor      4 - Full professor
 10 *  2. YRS   : Years in rank
 11 *  3. MF    : 0 - Male   1 - Female
 12 *  4. SALARY: in thousands
 13 *
 14 *Määritellään pieni data ja kirjoitetaan sen ensimmäinen havainto:
 15 *
 16 *DATA P
 17 *RANK YRS MF SALARY
 18 *   1   1  1   16.5
 19 *
 20 *Perustetaan havaintotiedosto tämän mallin perusteella:
 21 *
 22 */DATACOPY P TO PALKKA
 23 *LINES=5,12 (tiedoston dokumentti mukaan)
 24 *
```

Palkka-aineiston havainnot 2-85:

	RANK	YRS	MF	SALARY		RANK	YRS	MF	SALARY		RANK	YRS	MF	SALARY
2	1	2	1	14.5	30	2	2	1	18.0	58	3	6	1	26.0
3	1	2	1	14.5	31	2	2	1	19.5	59	4	0	0	25.5
4	1	2	1	15.0	32	2	3	1	18.0	60	4	0	0	25.5
5	1	2	1	15.5	33	2	3	1	18.0	61	4	0	0	26.0
6	1	2	1	15.5	34	2	3	1	19.0	62	4	0	0	28.0
7	1	8	1	14.5	35	2	4	0	18.5	63	4	0	1	25.5
8	2	0	0	16.5	36	2	4	1	18.5	64	4	0	1	27.0
9	2	0	0	18.5	37	2	4	1	20.0	65	4	1	0	27.5
10	2	1	0	16.0	38	2	5	0	18.5	66	4	1	1	27.5
11	2	1	0	17.0	39	2	5	1	16.0	67	4	2	0	28.0
12	2	1	0	17.0	40	2	6	1	18.0	68	4	3	0	28.0
13	2	1	0	17.0	41	2	7	1	20.0	69	4	3	0	28.0
14	2	1	0	17.0	42	2	9	0	18.0	70	4	3	0	28.0
15	2	1	0	17.5	43	3	0	0	20.5	71	4	4	0	28.5
16	2	1	1	16.5	44	3	0	0	20.5	72	4	4	1	27.0
17	2	1	1	16.5	45	3	0	0	22.0	73	4	5	0	29.0
18	2	1	1	17.0	46	3	0	1	20.0	74	4	6	0	29.0
19	2	1	1	17.0	47	3	1	1	22.5	75	4	6	0	29.5
20	2	1	1	17.0	48	3	2	0	23.5	76	4	7	0	30.0
21	2	1	1	17.0	49	3	2	0	23.5	77	4	7	1	30.0
22	2	1	1	18.0	50	3	3	0	23.5	78	4	7	1	30.0
23	2	1	1	18.0	51	3	3	1	23.5	79	4	9	1	30.0
24	2	1	1	19.0	52	3	4	0	23.5	80	4	9	1	32.0
25	2	1	1	19.0	53	3	4	0	23.5	81	4	10	0	34.5
26	2	1	1	20.0	54	3	4	0	23.5	82	4	10	1	31.0
27	2	2	0	17.0	55	3	4	0	24.0	83	4	10	1	34.0
28	2	2	1	16.5	56	3	4	1	24.0	84	4	15	0	40.5
29	2	2	1	17.0	57	3	5	0	25.0	85	4	15	1	40.5

Havaintotiedostoja hallitaan FILE-alkuisilla komennoilla (ks. FILE?). /DATACOPY-sukro tulostaa kenttään valmiin FILE SHOW -komennon, jonka aktivoimalla voit selata aineistoa ja lisätä siihen havaintoja.

Uusia havaintoja talletettaessa on FILE SHOW -toiminnolle annettava tiedoston muuttamislupa F3-napilla tai käyttämällä OPTIONS-täsmennystä. Tämä varotoimenpide on tarpeellinen, jottei aineistoja selaillessa tulisi vahingossa muuttaneeksi niiden sisältöä.

Syötä palkka-aineisto havaintotiedostoon. Kokeile kumpaakin seuraavista tavoista:

Havainto kerrallaan (riveittäin): Siirry seuraavan muuttujan kohdalle Tab-napilla. Havainnon viimeisen muuttujan jälkeen pääset seuraavan alkuun Enterillä.

Muuttuja kerrallaan (sarakkeittain): Aloita ensimmäisestä muuttujasta (tässä RANK). Paina Enter muuttujanarvojen välissä. Kun ensimmäinen muuttuja on valmis, siirry seuraavaan (YRS). Kiinnitä kohdistin ko. sarakkeelle napeilla F2 ja Enter. Tällöin Enterin painallus siirtää kohdistimen saman muuttujan seuraavaan havaintoarvoon.

Havaintotiedosto tallettuu levylle sitä mukaa kun tallennat tietoja. **Huomaa, että Survossa desimaalierottimena käytetään pistettä.** Kun olet syöttänyt koko aineiston, paina F8, niin pääset takaisin toimituskenttään. Tee tarkistukset laskemalla muuttujien keskiarvot, hajonnat ja korrelaatiot ja vertaamalla oheiseen esimerkkiin.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003            R:\DA1\    1000  100 0
 25 *
 26 *CORR PALKKA CUR+2 / tulostus alkaa 2 riviä komennon alapuolelta
 27 *
 28 *Means, std.devs and correlations of PALKKA  N=85
 29 *Variable  Mean        Std.dev.
 30 *RANK      2.741176    1.001819
 31 *YRS       3.270588    3.260106
 32 *MF        0.541176    0.501259
 33 *SALARY    22.31765    5.988504
 34 *Correlations:
 35 *              RANK    YRS     MF      SALARY
 36 * RANK          1.0000  0.3315 -0.3341  0.9068
 37 * YRS           0.3315  1.0000  0.0623  0.6206
 38 * MF           -0.3341  0.0623  1.0000 -0.2245
 39 * SALARY        0.9068  0.6206 -0.2245  1.0000
 40 *
```

Talleta toimituskenttäsi ja käy päivittämässä työhakemiston sisällysluettelo (ks. tehtävä 3). Tutki tehtävän lopuksi DD-ohjelmalla työhakemistosi sisältöä.

Muista nämä loppukuviot myös jatkossa. Harjoitustyönkin tekeminen on huomattavasti helpompaa, kun sinulla on kunnon dokumentit tehtävistä.

Tehtävä 6: Koepisteiden talletus havaintotiedostoon

Poimi 111 henkilön koepisteet sisältävä EXAMS-aineisto toimituskenttään. Tämä tapahtuu parhaiten selaamalla aikaisemmin (tehtävässä 4) talletettua toimituskenttää SHOW-komennolla ja poimimalla tarvittavat rivit nykyiseen kenttään L-napilla. Poistu lopuksi SHOW:sta napilla F8.

Kirjoita oheisessa esimerkissä oleva EXAMS-datan tiedostodokumentti toimituskenttään ja talleta aineisto havaintotiedostoon /DATACOPY-sukrolla.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003            R:\DA1\    1000  100 0
  1 *SAVE DA1T6 / koepisteiden talletus havaintotiedostoon
  2 *
  3 *SHOW DA1T4 / haetaan tähän rivit 5-10 aiemmasta tehtävästä
  4 *DATA EXAMS:(Scores)
  5 * 77 79 62 70 87 75 94 89 71 82 75 65 93 63 66 62 76 75 75 75 82 70 86
  6 * 84 86 74 81 64 91 82 87 87 75 92 78 72 69 60 72 92 69 61 81 80 85 69
  7 * 89 72 52 73 81 78 83 81 44 70 56 82 54 73 76 79 75 71 83 49 77 87 88
  8 * 90 70 66 67 69 90 84 83 72 68 77 78 90 77 80 72 79 80 75 51 48 72 37
  9 * 91 68 79 78 64 31 73 86 68 72 85 60 89 73 88 52 90 97 93 END
 10 *
 11 *Aineiston taustatiedot:
 12 *Dataset  : EXAMS.DAT            Source   : Calculus II class
 13 *Topic    : Final exam scores    Variables: 1     Cases : 111
 14 *
 15 */DATACOPY EXAMS TO PISTEET
 16 *LINES=12,13
 17 *
```

Tarkastele `FILE SHOW` -komennolla millaisen havaintotiedoston sait aikaan. Perusta siihen uusi muuttuja Z komennolla

```
VAR Z=#STD(Scores) TO PISTEET
```

Selaile aineistoasi ja selvitä kyselysysteemistä millaisen muuttujan teit (ks. VAR?).

Tehtävä 7: Havaintotiedoston rakenne

Tutkitaan aiemmin tallennetun PALKKA-aineiston rakennetta `FILE STATUS` -operaatiolla. Jos sinulla ei tässä vaiheessa ole PALKKA-aineistoa, niin kopioi se käyttöösi komennolla

```
>COPY H:\DA1\PALKKA.SVO
```

Aineiston rakenteen kuvaus koostuu tiedostodokumentista, (aktiivisten) muuttujien tiedoista sekä eräistä teknisistä tiedoista. Oheisen esimerkin riveillä 5-17 on tiedostodokumentti, jonka /DATACOPY-sukro on sinne tallettanut. Riveillä 19-22 näkyvät muuttujien tiedot avainsanojen `FIELDS:` ja `END` välissä. Kustakin muuttujasta on näkyvissä sen järjestysnumero aineistossa, tyyppi (N=numeerinen, S=sanallinen), aktivointitiedot, numeerisen muuttujan tyyppi (1,2,4,8) tai sanallisen muuttujan pituus sekä muuttujan nimi ja selitys.

Täydennä PALKKA-aineiston kuvausta lisäämällä muuttujille pidemmät selitykset (ks. oheinen esimerkki). Korvaa `FILE STATUS` komennolla `FILE UPDATE` ja aktivoi se, jolloin havaintotiedoston rakenne päivittyy.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003            R:\DA1\    1000  100 0
  1 *SAVE DA1T7 / havaintotiedoston rakenne
  2 *
  3 *
  4 *FILE UPDATE PALKKA / kannattaa vaihtaa STATUS samantien UPDATE:ksi!
  5 * Copy of data matrix P
  6 *
  7 *Dataset  : SALARY.DAT
  8 *Topic    : Faculty salaries at a liberal arts college (1980)
  9 *Source   : Survey    Variables: 4      Cases    : 85
 10 *  1. RANK  : 1 - Instructor or other      3 - Associate professor
 11 *             2 - Assistant professor     4 - Full professor
 12 *  2. YRS   : Years in rank
 13 *  3. MF    : 0 - Male   1 - Female
 14 *  4. SALARY: in thousands
 15 *
 16 * Created by /DATACOPY
 17 * on Monday July 16 2001 21:48:38 Week=29 Day=197
 18 *FIELDS: (active)
 19 *   1 NA_   1 RANK     Arvo            (####)
 20 *   2 NA_   1 YRS      Palvelusvuodet (###)
 21 *   3 NA_   1 MF       Sukupuoli (0=mies, 1=nainen)
 22 *   4 NA_   4 SALARY   Palkka (tuhat) (####.#)
 23 *END
 24 *Survo data file PALKKA: record=28 bytes, M1=9 L=64  M=4 N=85
```

Tutki aineistoasi `FILE SHOW` -komennolla (kiinnitä huomiota ikkunan alimpaan riviin).

Survossa on siis neljä numeerisen tiedon tyyppiä (1,2,4,8) eri käyttötarkoituksiin:

1: kokonaisluku väliltä [0,254]
2: kokonaisluku väliltä [-32768,32766]
4: reaaliluku, 7 merkitsevää numeroa
8: reaaliluku, 15 merkitsevää numeroa

Tilaa säästyy yleensä merkittävästi, kun käyttää kokonaisluvuille tyyppejä 1 ja 2. Näin vältetään myös helpoimmin turhilta desimaaleilta. Desimaaliluvuille riittää useimmiten tyyppi 4 kuten SALARY-muuttujalle edellä. (Yleensä tilasto-ohjelmissa on käytettävissä vain tyyppiä 8 vastaava vaihtoehto kaikkia numeerisia muuttujia varten.)

Havaintotiedoston rakenteeseen kuuluu vielä yksi rivi (tässä 24), joka kertoo tiivistetysti eräät tekniset tiedot. Niistä ensimmäinen on tietueen pituus (tässä 28 tavua) eli paljonko tilaa on varattu kutakin havaintoa kohden. Survon havaintotiedostot ovat ns. suorasaantitiedostoja, joiden tietueet ovat kiinteänmittaisia. Niiden käsittely on siitä syystä erittäin nopeaa. Tietueen pituutta voi muuttaa vain kopioimalla datan toiseksi uudella nimellä. Toisaalta datan voi alunperin perustaa niin että tilaa on "riittävästi".

Seuraavaksi rivillä 24 `M1` kertoo kuinka monelle muuttujalle tiedostossa kaikkiaan on tilaa (tässä `M1=9`). `L=64` ilmoittaa muuttujanimitietojen maksimipituuden tavuina. `M` ilmaisee kuinka monta muuttujaa tiedostossa tällä hetkellä on (`M=4`). `N` on havaintojen lukumäärä (`N=85`) kuten yleensä. PALKKA-datan muuttujat varaavat tällä hetkellä tilaa yhteensä 1+1+1+4=7 tavua. Näin ollen PALKKA-dataan voi luoda 9-4=5 uutta muuttujaa, joiden yhteinen pituus on korkeintaan 28-7=21 tavua.

Tehtävä 8: Täsmennykset, rajarivit ja osa-aineistot

Ennen kuin harjoitellaan osa-aineistojen poimintaa, on syytä selventää mitä täsmennyksillä yleisesti tarkoitetaan ja minkälaisia sääntöjä niiden käyttöön liittyy.

Täsmennykset ovat toimituskenttään hyvin vapaasti sijoitettavia tiiviitä ohjeita, joilla vaikutetaan Survon operaatioiden toimintaan. Täsmennyksen tunnistaa yhtäsuuruusmerkistä, jonka vasemmalla puolella on itse täsmennyssana (isoilla kirjaimilla) ja oikealla puolella sen kulloinenkin sisältö. Täsmennyksiä ei yleensä aktivoida.

Täsmennyksiä saa siis sijoitella kenttään vapaasti, mutta toisaalta monet täsmennykset ovat Survon operaatioille yhteisiä. Sekaannusten välttämiseksi on hyvä tietää ne yksinkertaiset säännöt joilla tätä vapautta sopivasti rajoitetaan.

Operaatiot etsivät täsmennyksiä järjestyksessä kentän alusta loppuun. Täsmennysten kesken vallitsee tiukka **hierarkia**: ensimmäisellä sijalla on itse komentorivi, toisella nykyinen **osakenttä** ja kolmantena globaali osakenttä, jos sellainen on määritelty. Osakentät rajataan pystyttämällä toimituskenttään **rajarivejä**.

Rajarivi kannattaa tehdä näin:

- Kirjoita tyhjän rivin alkuun yksi piste.
- Kopioi riviä itsensä perään koko rivin täydeltä painamalla `Alt-F3` ja `Enter`.

Tällöin Survo piirtää rajarivin selkeästi erottuvana kaksoisviivana. (Monisteen esimerkeissä rajarivi näkyy pisteviivana.) Jos et ole varma, pitäisikö rajarivi laittaa, laita varmuuden vuoksi. Ylimääräiset ovat harvemmin haitaksi.

Osa-aineistot

Muuttujien suhteen tarkastelua voidaan rajoittaa osa-aineistoihin täsmennyksillä VARS ja MASK tai asettamalla aktivointitiedot suoraan datatiedostoon `FILE ACTIVATE` -operaatiolla (joka käynnistyy myös napilla `Alt-F6`). **Havaintojen suhteen** on käytettävissä kolme täsmennystä: `IND`, CASES ja `SELECT`.

Poimitaan toimituskenttään erilaisia osa-aineistoja PALKKA-datasta. Muista että täsmennyksistä saa yksityiskohtaisia tietoja kyselysysteemistä.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003          R:\DA1\    1000  100 0
  1 *SAVE DA1T8 / täsmennykset, rajarivit ja osa-aineistot
  2 *
  3 *Poimitaan osa-aineisto valinnoilla IND=MF,1 ja VARS=RANK,SALARY
  4 *FILE LOAD PALKKA,CUR+2
```

Tarkastele annettujen täsmennysten vaikutusta saadun osa-aineiston perusteella. Siirry sitten kentän loppuun. **Tee rajarivi**. Siirry tyhjälle riville. Kopioi rivin 4 komento sinne ja varusta se seuraavilla täsmennyksillä:

```
IND=RANK,2,3  SELECT=A*B   A=YRS,0,2   B=SALARY,25,45
```

Aktivoi FILE LOAD. Mitä tapahtui ja miksi? Entä jos vaihdat SELECT-täsmennyksessä kertomerkin (*) plusmerkiksi (+)?

Tehtävä 9: Osa-aineistojen tunnuslukuja

Laske STAT-operaatiolla erilaisia perustunnuslukuja PALKKA-aineiston osa-aineistoista:

- korkeintaan neljä vuotta virassa toimineiden mediaanipalkka (miehet ja naiset erikseen sekä kaikki yhdessä)
- alle vuoden toimineiden professoreiden (Associate & Full) lukumäärä
- miesten osuus henkilökunnasta
- palvelusvuosien vaihteluväli 20000-25000 ansaitsevien osalta

Tehtävä 10: Satunnaisotoksen poimiminen aineistosta

Tehtävänä on nyt poimia useita 20 havainnon suuruisia satunnaisotoksia PALKKA-aineistosta. Poiminta tapahtuu kahdessa vaiheessa:

Vaihe 1: Perustetaan PALKKA-aineistoon muuttuja SL, johon generoidaan (pseudo)satunnaislukuja väliltä (0,1) RND-funktiolla.

Vaihe 2: PALKKA-aineisto järjestetään SL-muuttujan perusteella suuruusjärjestykseen, minkä jälkeen 20 ensimmäistä havaintoa talletetaan toiseen tiedostoon nimeltä OTOS.

Tee nämä vaiheet omalla PALKKA-aineistollasi. Laske sen jälkeen keskipalkka **otoksen perusteella** (katso oheista esimerkkiä).

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003            R:\DA1\    1000  100 0
  1 *SAVE DA1T10 / satunnaisotoksen poimiminen aineistosta
  2 *
  3 *
  4 *VAR SL=RND(0) TO PALKKA
  5 *FILE SORT PALKKA BY SL TO OTOS / NSORT=20 (vain 20 ensimmäistä talteen)
  6 *CORR OTOS END+2 / VARS=SALARY
  7 *
  8 *OTOS  N=20
  9 *Variable  Mean        Std.dev.
 10 *SALARY    21.60000    7.557847
 11 *
```

Entä jos poimitaan uusi otos? Aktivoi kaikki komennot uudelleen ja vertaa tuloksia. Mitä sanoisit 20 havainnon otoksen perusteella lasketusta keskipalkasta?

Tehokkain tapa tehdä uusintapoimintoja on ns. **jatkuva aktivointi**. Sen käyttö edellyttää että aktivoitavat komennot ovat peräkkäisillä riveillä (kuten oheisessa esimerkissä). Vie kohdistin joko hiirellä tai nuolinapeilla ensimmäisen aktivoitavan rivin alkuun. Paina kerran nappia F2 ja sen jälkeen Esc. Toista muutaman kerran ja vertaile tuloksia.

Tehtävä 11: Harjoitustyöaineiston poimiminen

Tämän tehtävän tekemiseen tarvitset KDATA-aineistoa, jossa on tietoja suomalaisten kulutuskäyttäytymisestä (N=3052). Aineistosta poimitaan 300 havainnon suuruinen yksinkertainen satunnaisotos, joka on samalla tämän kurssin harjoitustyöaineisto. KDATA-aineisto on käytettävissä atk-luokan koneissa. Otoksen poiminta tapahtuu vaiheittain.

Vaihe 1: Kopioi KDATA-aineisto itsellesi otoksen poimintaa varten:

```
>COPY H:\DA1\KDATA.SVO
```

Vaihe 2: Generoi muuttujaan RN (pseudo)satunnaislukuja väliltä (0,1). Anna RND-funktion siemenluvuksi (sulkujen sisään) **oma syntymäaikasi** muodossa ppkkvvvv:

```
VAR RN=RND(         ) TO KDATA
```

Vaihe 3: Järjestä aineisto muuttujan RN suhteen ja talleta sen 300 ensimmäistä havaintoa harjoitustyöaineistoksi. Saat keksiä aineistollesi paremmankin nimen kuin mallissa:

```
FILE SORT KDATA BY RN TO AINEISTO / NSORT=300
```

Tarkista että aineistossasi on 300 havaintoa:

```
FILE SHOW AINEISTO
```

Lopuksi voit poistaa poiminnassa käytetyn KDATA-aineiston:

```
FILE DEL KDATA
```

Tehtävä 12: Suhteellisen osuuden laskeminen

Laske edellisessä tehtävässä muodostettuun aineistoon uusi muuttuja E%, joka kertoo ruokiin ja juomiin käytettyjen varojen prosentuaalisen osuuden kaikista kulutusmenoista:

$$E\% = 100 \times \frac{\text{Ruokiin ja juomiin käytetyt rahat}}{\text{Kulutusmenot yhteensä}}$$

Käytännössä nimittäjä voi olla joidenkin havaintojen osalta nolla, joten on varmistettava ettei nollalla jakoa pääse syntymään. Oheisessa esimerkissä käytetään tähän if-then-else -rakennetta. Nollatapauksiin talletetaan puuttuvan tiedon merkki (MISSING).

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003            R:\DA1\    1000  100 0
  1 *SAVE DA1T12 / suhteellisen osuuden laskeminen
  2 *
  3 *
  4 *FILE SHOW AINEISTO     / aineiston sisällön selailu
  5 *FILE ACTIVATE AINEISTO / muuttujien kuvauksien selailu
  6 *Poimitaan kenttään (FILE STATUS) ja muokataan, saadaan muistilappu:
  7 *
  8 *    K01U     Elintarvikkeet ja alkoholittomat juomat
  9 *    K012U    Alkoholittomat juomat (sisältyy edelliseen)
 10 *    K021U    Alkoholijuomat
 11 *    K01_12U  Kulutusmenot yhteensä
 12 *
 13 *VAR E%=if(K01_12U=0)then(MISSING)else(osuus) TO AINEISTO
 14 *                                        osuus=100*(K01U+K021U)/K01_12U
 15 *STAT AINEISTO CUR+1 / VARS=E%
 16 *
```

Tarkastele tekemääsi muuttujaa selailemalla aineistoa ja laskemalla siitä perustunnusluvut.

Tehtävä 13: Luokittelumuuttujan muodostaminen

Tutki STAT-komennolla maakuntajakoa kuvaavan muuttujan NUTS3 frekvenssijakaumaa aineistossasi. Muuttujan koodaus on seuraava (kurssin kotisivulta):

```
 1 Uusimaa          6 Pirkanmaa      11 Pohjois-Savo     16 Keski-Pohjanmaa
20 Itä-Uusimaa      7 Päijät-Häme    12 Pohjois-Karjala  17 Pohjois-Pohjanmaa
 2 Varsinais-Suomi  8 Kymenlaakso    13 Keski-Suomi      18 Kainuu
 4 Satakunta        9 Etelä-Karjala  14 Etelä-Pohjanmaa  19 Lappi
 5 Kanta-Häme      10 Etelä-Savo     15 Pohjanmaa        21 Ahvenanmaa
```

Muodosta maakuntamuuttujan jakaumaa hyväksi käyttäen uusi luokittelumuuttuja joka jakaa maan sopiviin alueisiin. Luokkien nimet ja lukumäärän saat päättää itse. Tehtävä hoituu kätevimmin CLASSIFY-operaatiolla. Tutustu siihen kyselysysteemin avulla.

Tehtävä 14: Ristiintaulukointi

Ristiintaulukoimalla voidaan tutkia kahden tai useamman luokitellun muuttujan välisiä riippuvuuksia. Käytännössä kahden muuttujan taulukot ovat usein tärkeimpiä. Harjoitellaan niiden muodostamista Survon TAB-operaatiolla. Useampiulotteisia taulukoita tehdään aivan vastaavalla tavalla.

Kopioi aluksi käyttöösi Survon historiallinen esimerkkiaineisto KUNNAT:

```
>COPY <Survo>\U\D\KUNNAT.SVO
```

(Tässä <Survo> on lyhennysmerkintä, joka viittaa automaattisesti siihen hakemistoon, johon Survo on koneessa asennettu. Kirjoita se täsmälleen mallin mukaan.)

Tutustu KUNNAT-tiedoston rakenteeseen (FILE STATUS) ja sisältöön (FILE SHOW).

Tehtävänä on tutkia veroäyrin jakautumista Etelä-, Keski- ja Pohjois-Suomessa. Oheisessa esimerkissä alueluokitus on jo tehty valmiiksi Lääni-muuttujan perusteella. Täydennä komentokaavio Äyri-muuttujan osalta.

Katso ensin STAT:illa, miten Äyri on jakautunut koko aineistossa, jotta löydät sille järkevän luokituksen. Kyselysysteemistä (mm. avainsanalla TABLE?) näet miten numeerinen muuttuja luokitellaan. Kokeile sekä tasavälistä että vaihtelevanpituista luokitusta. Sopiva luokkien lukumäärä on n. 4-5.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003            R:\DA1\    1000  100 0
  1 *SAVE DA1T14 / ristiintaulukointi
  2 *
  3 *
  4 *>COPY <Survo>\U\D\KUNNAT.SVO
  5 *
  6 *FILE SHOW KUNNAT
  7 *
  8 *TAB KUNNAT END+2 / VARIABLES=Äyri,Lääni (sarakemuuttuja,rivimuuttuja)
  9 *Muuttujien luokitukset annetaan täsmennyksinä:
 10 *Lääni=/UUS/TUR/AHV/HÄM/KYM(Etelä),&
 11 *      /MIK/KAR/KUO/KES/VAA(Keskiosa),/OUL/LAP(Pohjola)
 12 *Äyri=
 13 *
```

Pohdi näin saamasi frekvenssitaulukon sisältöä. Lisää sen jälkeen kenttään uusi täsmennys RESULTS=C%,R%,CSUMS,RSUMS ja tutki miten se vaikuttaa TAB-komennon tulostukseen. Miten prosenttitaulukoita tulkitaan? Voit jättää minkä tahansa RESULTS-avainsanan pois. Kokeile myös vaihtaa muuttujien järjestys VARIABLES-täsmennyksessä.

Tehtävä 15: Khi-neliötesti

Kahden muuttujan frekvenssitaulukossa muuttujien riippuvuuksia voidaan tutkia tarkemmin khi-neliötestin avulla. Oletuksena TAB-operaatio tulostaakin testin lyhyesti taulukon alapuolelle. Jos p-arvo on "pieni", voidaan hyvällä syyllä epäillä että taulukon rivi- ja sarakemuuttujien välillä vallitsee jokin riippuvuus.

Khi-neliötesti perustuu havaittujen frekvenssien (taulukon solujen arvojen) ja reunasummien (marginaalien) avulla laskettavien, ns. odotettujen frekvenssien erotuksiin. Mitä suurempi ero havaitun ja odotetun frekvenssin välillä on, sitä enemmän testisuure kasvaa ja p-arvo vastaavasti pienenee.

Pelkän p-arvon sijasta on usein hyödyllistä tutkia testisuureen solukohtaisia kontribuutioita riippuvuuden luonteen selvittämiseksi. Tähän on Survossa käytettävissä mm. sukro /X2. Se tulostaa kenttään havaitut ja odotetut frekvenssit sekä khi-neliön kontribuutiot summasarakkeilla varustettuina.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003          R:\DA1\    1000  100 0
  1 *SAVE DA1T15 / khi-neliötesti
  2 *
  3 *
  4 *Poimitaan tulostaulukko edellisestä tehtävästä: (Huom. Ilman summia!)
  5 *SHOW DA1T14
  6 *
  7 *TABLE KUNNAT1 A,B,F   N=464
  8 A        Lääni    Etelä Keskiosa  Pohjola
  9 *Äyri    *****
 10 *alle_15             77        1         0
 11 *15-16              114       34         8
 12 *16-17               36       93        38
 13 Byli_17               2       33        28
 14 *Chi_square=230.0 df=6 P=0.0000
 15 *
 16 */X2 KUNNAT1
 17 *
```

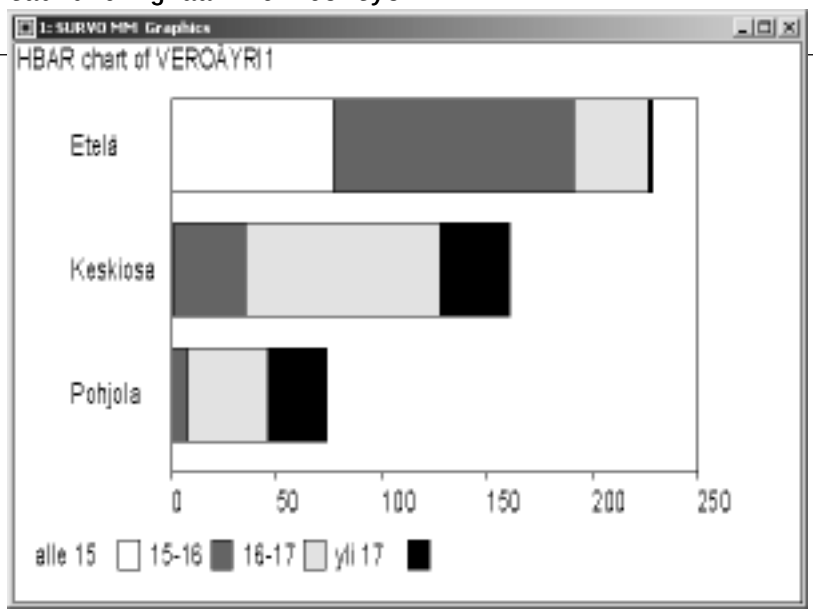
Tee nämä jatkotarkastelut jostakin edellä laatimastasi ristiintaulukosta. Pohdi erityisesti khi-neliötestisuureen muodostumista alueittain tulosmatriisien perusteella.

Tehtävä 16: Pylväsdiagrammit

Tehtävänä on piirtää tehtävässä 14 muodostetuista frekvenssitaulukoista erilaisia pylväsdiagrammeja. Se tapahtuu varsin suoraviivaisesti. TAB-operaatiolla muodostettu taulukko on vain kuvanpiirtoa varten nimettävä DATA-taulukoksi (ks. DATA? - kohta 3).

Oheisessa esimerkissä on poimittu esille aiemmin tehty taulukko, jossa `Äyri` on sarake- ja `Lääni` rivimuuttujana. Riville 15 on kirjoitettu DATA-määritelmä, jossa kerrotaan että havainnot alkavat riviltä G+2 (10) ja loppuvat riville H (12). Muuttujien otsikot ovat rivillä G (8). Kannattaa suosia symbolisia rivitunnuksia, sillä ne eivät muutu jos rivejä lisäillään ja poistetaan yläpuolelta. Nämä tunnukset ovat TABin tekemiä, ja ne voivat olla sinulla eri, riippuen siitä missä järjestyksessä olet mitäkin tehnyt tehtävässä 14.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003          R:\DA1\    1000  100 0
  1 *SAVE DA1T16 / pylväsdiagrammit
  2 *
  3 *
  4 *Poimitaan tulostaulukko edellisestä tehtävästä: (Huom. Ilman summia!)
  5 *SHOW DA1T14
  6 *
  7 *TABLE KUNNAT1 G,H,F   N=464
  8 G         Äyri alle_15   15-16   16-17  yli_17
  9 *Lääni    ****
 10 *Etelä               77     114      36       2
 11 *Keskiosa             1      34      93      33
 12 HPohjola              0       8      38      28
 13 *Chi_square=230.0 df=6 P=0.0000
 14 *
 15 *DATA VEROÄYRI1,G+2,H,G
 16 *
 17 *GPLOT VEROÄYRI1 / taulukon graafinen esitys
 18 *
```



Kuva ilmestyy oheisen kaltaiseen kuvaikkunaan (monivärisenä).

Oletuksena Survo piirtää pylväsdiagrammin, jonka pylväät ovat vaakasuorassa.

Kuvatyyppiä muutetaan TYPE-täsmennyksellä. Katso kyselysysteemistä vaihtoehdot ja kokeile.

Pylväskuvien lisäksi frekvenssiaineistosta voi piirtää sektori- eli piirakkadiagrammeja.

Kuvat ilmestyvät omiin graafisiin ikkunoihinsa, joiden oletusasetus on kolme vuorottelevaa ikkunaa ruudun oikeassa laidassa. Muita asetuksia on tarjolla alareunan pehmonapiston kautta (SYSTEM → GRAPH). Kuvia voi sulkea yksittäin kuten Windows-ikkunoita yleensä. Kaikki kuvaikkunat sulkeutuvat alareunan pehmonapilla C ja myös automaattisesti kun poistutaan Survosta F8:lla tai EXIT-painikkeella.

Kuvia tarkastellessasi mieti:

- Mikä on kuvioiden tulkinta?
- Mitkä kuvioista ovat helposti tulkittavissa sanallisesti?
- Mitä puutteita kuvissa esiintyy?

Tehtävä 17: Kuvan yksityiskohdat

Survo valitsee kuvan yksityiskohdat kuten otsikon, asteikot ja kuvan mittasuhteet automaattisesti. Tarvittaessa mitä tahansa niistä voi muuttaa erilaisilla täsmennyksillä.

Hae jokin edellä laatimasi piirroskaavio ja sitä vastaava data esille. Piirrä kuva ensin ilman mitään täsmennyksiä. Kokeile sitten miten kuva muuttuu vaiheittain, kun kirjoitat kenttään seuraavia täsmennyksiä ja aktivoit GPLOT-komennon (käytä apuna kyselysysteemiä):

HEADER	Kuvan otsikko
XLABEL	X-akselin nimi
YLABEL	Y-akselin nimi
SCALE	Numeerisen muuttujan asteikko (X tai Y)
XDIV	Kuvan mittasuhteet vaakasuunnassa (oletus 3,10,2)
YDIV	Kuvan mittasuhteet pystysuunnassa (oletus 3,10,2)
SHADING	Pylväiden värit
LEGEND	Pylväiden värien selitys
GAP	Pylväiden välit
VALUES	Havaintoarvot pylväisiin
NAMES	Pylväiden nimet

Muista että täsmennysten pitää olla yhtenäisiä merkkijonoja. Käytä välilyönnin sijasta alaviivaa sanojen välissä (esim. Suomi_kolmena_alueena).

Tehtävä 18: Kuvan tulostaminen paperille

Kuvia tehdään Survossa kahdella eri operaatiolla. Näistä toinen on GPLOT, joka siis piirtää Windows-ikkunoihin. Ikkunoita saa venytellä mielin määrin. Koot on mahdollista määrätä tarkasti etukäteenkin sopivilla täsmennyksillä (ks. GPLOT? - kohta 3). Kuvat voidaan tallettaa metatiedoistoina (*Enhanced Meta File*), jolloin ne ovat helposti siirrettävissä muihin Windows-ohjelmiin.

Toinen operaatio on nimeltään PLOT. Sillä piirretyt kuvat tulostetaan suoraan *PostScript*-kirjoittimella tai ne voidaan tallettaa PostScript-tiedostoina, joista niitä voidaan katsella kuvaruudulla tai liittää tulostettavaan dokumenttiin.

Kuvien tulostukseen palataan vielä myöhemmin, mutta harjoitellaan sitä ennen kuvan yksityiskohtien korostamista eri kirjasinlajeilla eli **fonteilla**. Kun piirretään kuvia GPLOT-komennolla, voidaan valita mitä tahansa Windowsissa käytössä olevia fontteja (ks. FONT? - kohta 2). Yleensä kannattaa pitäytyä tavallisimmissa:

Times-Roman	[Times(x)]	(x = fontin koko)
Helvetica	[Swiss(x)]	(oletus PostScript-grafiikassa)
Courier	[Courier(x)]	(oletus kuvaruutugrafiikassa)

Fonttimäärityksiä voi liittää yksittäisiin täsmennyksiin, Tällöin on yleensä myös paikallaan määritellä kuvan fonteille oletusarvot täsmennyksillä PEN ja LINETYPE. Ilman näitä täsmennyksiä saattaa tulla yllätyksiä, koska asetukset periytyvät kuvan eri kohtien välillä.

Kokeile fonttimäärittelyitä täsmennyksiin liitettyinä (ks. oheinen esimerkki). Huomaa että täsmennykset saa sijoitella täysin vapaasti, kunhan muistaa niihin liittyvät säännöt (ks. tehtävä 8). Ei ole mikään pakko pyrkiä oheisenlaiseen tiiviiseen esitykseen. Voi vaikka kirjoittaa joka täsmennyksen omalle rivilleen ja perään vielä selityksen mitä se tarkoittaa.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003          R:\DA1\    1000  100 0
 16 *
 17 *GPLOT VEROÄYRI1  /   (data tehtävästä 16)   TYPE=%VBAR XDIV=1.5,13,0.5
 18 *PEN=[Times(16)]   LINETYPE=[Courier(13.5)]  YLABEL=%   GAP=0.2,0.1,0.1
 19 *HEADER=[SwissB(27.3)],Veroäyrin_jakautuminen_alueittain SCALE=0(10)100
 20 *LEGEND=Veroäyrin_luokitus: XLABEL=Suomen_kolmijako NAMES=[Swiss(15)],1
 21 *.....................................................................
```

Kun saat kuvan mielestäsi valmiiksi, tulosta se paperille seuraavasti: Lisää piirtokaavioosi täsmennys `DEVICE=PS` ja korvaa `GPLOT` pelkällä `PLOT`:illa (ts. poista `G`-kirjain). Aktivoi `PLOT` ja vertaa paperitulostetta graafiseen ikkunaan tekemääsi kuvaan.

Tehtävä 19: Muuttujien välinen riippuvuus

Tarkastellaan aiemmin käsiteltyä `PALKKA`-aineistoa. Selvitä, miten palkka riippuu palvelusvuosista. Piirrä palkan ja palvelusvuosien välinen hajontakuvio komennolla

```
GPLOT PALKKA,YRS,SALARY
```

Lisää täsmennys `POINT=RANK` . Kokeile myös muita `POINT`-täsmennyksen vaihtoehtoja (ks. `POINT`?), esimerkiksi `POINT=3,5,RANK,4` tai `POINT=3,5,RANK,0` .

Rajoita aineisto varsinaisiin professoreihin täsmennyksellä `IND=RANK,4` .

Lisää kuvaan aiheeseen sopivat otsikkotekstit, asteikkonimet ja asteikkorajat.

Tehtävä 20: Regressiosuora

Tutkitaan seuraavaksi maailman maita. Miten urbanisoituminen eli kaupungistuminen ja miesten odotettavissa oleva keskimääräinen elinikä riippuvat toisistaan?

Kopioi aluksi käyttöösi maailman maita koskeva Survon esimerkkiaineisto `WORLD99`:

```
>COPY <Survo>\U\D\WORLD99.SVO
```

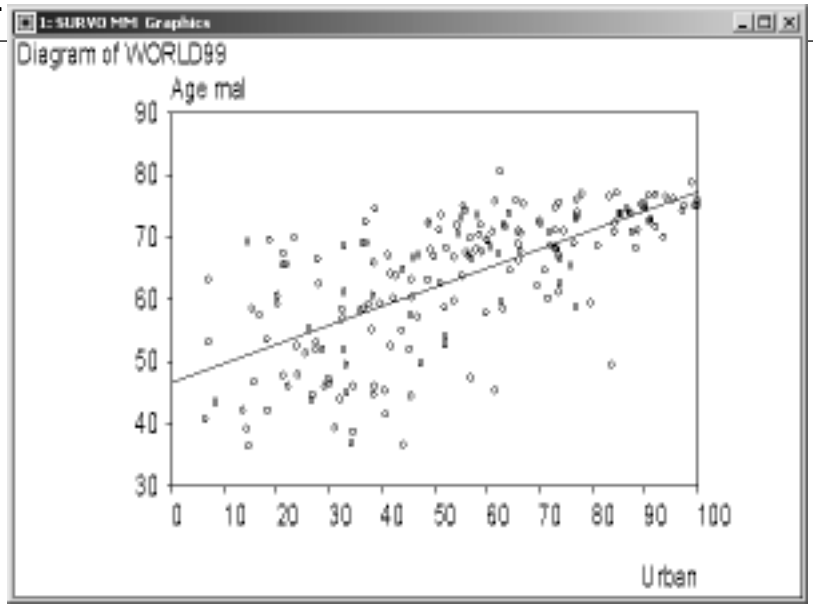
Tutustu tiedoston rakenteeseen (`FILE STATUS`) ja sisältöön (`FILE SHOW`).

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003          R:\DA1\    1000  100 0
  1 *SAVE DA1T20 / regressiosuora
  2 *
  3 *FILE SHOW WORLD99
  4 *
  5 *GPLOT WORLD99,Urban,Age_mal / TREND=0
  6 *.....................................................................
```

Piirrä tutkittavien muuttujien välinen hajontakuva eli korrelaatiodiagrammi.

Lisää siihen PNS-suora eli regressiosuora `TREND`-täsmennyksellä.

Katso oheista esimerkkiä.



Selvitä suoran yhtälö regressioanalyysilla. **Selitettävä muuttuja** sijoitetaan kuvassa **pystyeli Y-akselille** ja **selittäjä vaaka- eli X-akselille.** Vastaavasti Age_mal merkitään Y:llä ja Urban X:llä VARS-täsmennyksessä, jolla tässä ilmaistaan muuttujien roolit. Regressioanalyysi tehdään esimerkiksi LINREG-operaatiolla seuraavasti:

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003           R:\DA1\    1000  100 0
  7 *
  8 *LINREG WORLD99 END+2 / VARS=Age_mal(Y),Urban(X)
  9 *
 10 *Means, std.devs and correlations of WORLD99  N=190
 11 *Variable  Mean        Std.dev.
 12 *Urban     54.19474    23.97570
 13 *Age_mal   63.26105    10.92106
 14 *Correlations:
 15 *             Urban   Age_mal
 16 * Urban        1.0000  0.6718
 17 * Age_mal      0.6718  1.0000
 18 *
 19 *Linear regression analysis: Data WORLD99, Regressand Age_mal   N=190
 20 *Variable Regr.coeff.    Std.dev.    t     beta
 21 *Urban     0.306014       0.024608  12.44  0.672
 22 *constant  46.67673       1.457652  32.02
 23 *Variance of regressand Age_mal=119.2694787 df=189
 24 *Residual variance=65.78760843 df=188
 25 *R=0.6718 R^2=0.4513
 26 *
```

Regressiosuoran yhtälöhän on tässä yhden selittäjän tapauksessa muotoa Y=a0+a1*X. Tästä saat kysytyn suoran yhtälön selville sijoittamalla tilalle tehtävässä käytetyt muuttujat ja LINREG:in antamat numeeriset arvot. Tutustu tarvittaessa suomenkielisiin opetusohjelmiin (DEMO → OPETUS → A. Tilastolliset menetelmät → 9. Lineaarinen regressioanalyysi). Mitä sanoisit kuvan ja analyysin perusteella urbanisoitumisen ja miesten odotettavissa olevan eliniän riippuvuudesta?

Tehtävä 21: LOWESS-tasoitus

Tutkitaan vielä urbanisoitumisen ja miesten keskimääräisen eliniän riippuvuutta (ks. edellinen tehtävä). Tehtävänä on tasoittaa muuttujien välinen riippuvuus LOWESS-operaatiolla ja piirtää kuva tasoituksesta.

Tee ensin aineistoon uusi muuttuja tasoitettuja arvoja varten. Lisää se VARS-täsmennykseen S:llä merkittynä. Komento LOWESS ei tulosta mitään toimituskenttään vaan laskee tasoitetut arvot ja tallettaa ne S:llä merkittyyn muuttujaan.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003           R:\DA1\    1000  100 0
  1 *SAVE DA1T21 / LOWESS-tasoitus
  2 *
  3 *FILE SHOW WORLD99
  4 *
  5 *VAR TASOITUS=MISSING TO WORLD99        / uuden muuttujan perustaminen
  6 *
  7 *LOWESS WORLD99 / tasoitus      VARS=Age_mal(Y),Urban(X),TASOITUS(S)
  8 *
  9 *FILE SORT WORLD99 BY Urban TO SORTDATA / lajittelu X-muuttujan suhteen
 10 *
 11 *GPLOT SORTDATA,Urban,Age_mal,TASOITUS  / tasoituksen piirtäminen
 12 *
 13 *Täsmennyksellä TASOITUSLINE=1 yhdistetään tasoitetut pisteet viivalla.
 14 *
```

Mitä tulokset mielestäsi kertovat? Onko muuttujien välinen riippuvuus lineaarista? Vallitseeko muuttujien välillä syy-seuraus-yhteyttä?

Tehtävä 22: Riippuvuustarkastelu alueittain

Tarkastellaan edelleen urbanisoitumisen ja miesten keskimääräisen eliniän riippuvuutta (ks. edellinen tehtävä). Ilmiön luonne vaihtelee maasta toiseen, joten yksi ainoa regressiomalli ei välttämättä kuvaa tilannetta tyydyttävästi. Siirrytään siis tutkimaan riippuvuutta alueittain (tässä tapauksessa maanosittain).

Piirrä alueittaiset hajontakuvat kaikkien maanosien osalta oheisen mallin mukaan. Maanosien koodit näet tutustumalla tiedoston rakennekuvaukseen (`FILE STATUS`).

Jotta hajontakuvia voisi vertailla keskenään, niissä täytyy olla samat asteikot. Nämä kuten muutkin täsmennykset voidaan kirjoittaa jokaiseen kuvanpiirtokaavioon erikseen, mutta kätevämpää on kerätä yhteiset määrittelyt toimituskentän ensimmäiseen osakenttään ja varustaa se avainsanalla `*GLOBAL*` . Tällöin Survo etsii täsmennyksiä myös kentän alusta rajariveistä huolimatta (vrt. tehtävä 8).

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003           R:\DA1\    1000  100 0
  1 *SAVE DA1T22 / riippuvuustarkastelu alueittain
  2 *
  3 *    *GLOBAL*
  4 *                                          TREND=0 (PNS-suora)
  5 *YLABEL=Miesten_keskimääräinen_elinikä  XLABEL=Kaupungistumisprosentti
  6 *HEADER=Väestötietoja_maailman_maista   YSCALE=30(10)90 XSCALE=0(10)100
  7 *.....................................................................
  8 *GPLOT WORLD99,Urban,Age_mal / kaikki maat
  9 *.....................................................................
 10 *GPLOT WORLD99,Urban,Age_mal / IND=Contint,4  (ks. tiedoston kuvaus)
 11 *HEADER=Väestötietoja_Afrikan_maista
 12 *.....................................................................
 13 *GPLOT WORLD99,Urban,Age_mal / IND=Contint,1
 14 *HEADER=Väestötietoja_Euroopan_maista
 15 *.....................................................................
```

Onko riippuvuus samanlaista eri maanosissa? Mitä eroja havaitset?

Hyvin havainnollista on piirtää hajontakuviot samaan kuvaan. Se tapahtuu yksinkertaisesti kerrostamalla eli piirtämällä useampia kuvia päällekkäin. Tällä tekniikalla voidaan itse asiassa piirtää miten monimutkaisia kuvia tahansa yksinkertaisempien kuvien yhdistelmänä. Sama periaate koskee `PLOT`-komennolla tehtäviä PostScript-kuvia, vaikka niissä toimitaankin hieman eri tavalla. Siihen aiheeseen palataan myöhemmin.

Pienillä muutoksilla edelliseen kaavioon pääset kokeilemaan kuvien kerrostamista:

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003           R:\DA1\    1000  100 0
  6 *.....................................................................
  7 *GPLOT WORLD99,Urban,Age_mal / IND=Contint,4 POINT=[BLUE],5
  8 *OUTFILE=MAAILMA (kuva tallettuu Windowsin metatiedostoksi MAAILMA.EMF)
  9 *.....................................................................
 10 *GPLOT WORLD99,Urban,Age_mal / IND=Contint,1 POINT=[RED],5
 11 *OUTFILE=MAAILMA INFILE=MAAILMA (talletettu kuva haetaan pohjaksi)
 12 *.....................................................................
```

Huomaa määrittelyjen periytyminen täsmennysten välillä: regressiosuorat piirretään automaattisesti `POINT`-täsmennysten väreillä. Jos näin ei haluttaisi käyvän, voitaisiin `TREND`-täsmennys muuttaa esimerkiksi muotoon `TREND=[BLACK],0` . Kokeile!

Tehtävä 23: Aikasarjojen piirtäminen

Siirrytään seuraavaksi ajassa taaksepäin tutkimaan teollisuustyöntekijöiden palkkakehitystä vuosina 1920-1938.

Tehtävässä tarvittava aineisto on talletettu tavallisena tekstitiedostona tilastotieteen laitoksen atk-luokan hakemistoon `C:\DA1` . Hae se toimituskenttään komennolla

```
LOADP C:\DA1\ANSIOT.DAT
```

ja tee siitä `/DATACOPY`-sukrolla työhakemistoosi Survon havaintotiedosto `ANSIOT` . Ota huomioon rahan arvossa tapahtunut muutos eli tee aineistoon uusi muuttuja `RP` seuraavan kaavan perusteella:

$$\text{Reaalipalkka} = 100 \times \frac{\text{Nimellispalkka}}{\text{Elinkustannusindeksi}}$$

Piirrä samaan kuvaan reaalipalkan ja nimellispalkan kehitystä kuvaavat aikasarjat:

```
GPLOT ANSIOT TIME(V),RP,NP  /  LINE=1
```

Käytä eri aikasarjoille erilaista viivatyyppiä (ks. `LINETYPE`?) ja nimeä sarjat:

```
RPLINE=[line_type(0)],1,1,Reaalipalkka
NPLINE=[line_type(4)],1,1,Nimellispalkka
```

Laita tarvittaessa kuvioon paremmat asteikot (`XSCALE` ja `YSCALE`) ja muuta kuvan mittasuhteet (`XDIV`) sellaisiksi, että aikasarjojen nimet pysyvät kehikon sisällä.

Tehtävä 24: Koepisteiden histogrammi

Tarkastellaan aikaisemmissa harjoituksissa käsiteltyä `PISTEET`-aineistoa. Selvitä ensin `Scores`-muuttujan vaihteluväli luokitusta varten. Piirrä sen jälkeen muuttujan jakaumasta histogrammi komennolla

```
GHISTO PISTEET Scores
```

johon tarvitset myös `Scores`-muuttujan luokituksen muodossa `Scores=x(y)z` , jossa `x` = ensimmäisen luokan alaraja, `y` = luokkavälin pituus ja `z` = viimeisen luokan yläraja.

Kokeile erilaisia (tasavälisiä) luokituksia. Mikä antaa parhaan kuvan aineistosta? Lisää komennon perään tulostusrivi, niin saat tietoja histogrammista myös toimituskenttään.

Tehtävä 25: Logit-muunnos ja sen histogrammi

Tee `PISTEET`-aineiston muuttujalle `Scores` logit-muunnos kaavan

$$\texttt{LOGITS} = \log\left(\frac{\text{Scores}}{100 - \text{Scores}}\right)$$

perusteella ja kuvaa uuden muuttujan jakauma histogrammilla.

Tehtävä 26: Tilastolliset funktiot

Tilastollisten jakaumien hallinnassa turvauduttiin aiemmin taulukoihin, joita edelleenkin näkee monien oppikirjojen liitteinä. Taulukoiden käyttöön ei yleensä ole enää mitään syytä, sillä tarvittavat asiat voidaan laskea suoraan tilastollisten funktioiden avulla. Survossa on käytettävissä seuraavat jatkuviin jakaumiin liittyvät tilastolliset funktiot (lisätietoja löytyy kyselyllä FUNCSTAT?):

Jakauma	**Tiheysfunktio** (tf)	**Kertymäfunktio** (kf)	**Kertymäfunktion käänteisfunktio**
Normaalijakauma	`N.f(m,s²,x)`	`N.F(m,s²,x)`	`N.G(m,s²,p)`
t-jakauma	`t.f(n,x)`	`t.F(n,x)`	`t.G(n,p)`
χ^2-jakauma	`Chi2.f(n,x)`	`Chi2.F(n,x)`	`Chi2.G(n,p)`
F-jakauma	`F.f(n1,n2,x)`	`F.F(n1,n2,x)`	`F.G(n1,n2,p)`

Taulukon symbolien selitykset:

`m` = keskiarvo
s^2 = varianssi
`x` = muuttujan arvo
`p` = kertymäfunktion arvo
χ^2 = khi-toiseen (khi-neliö)
`n`, n_1 ja n_2 = vapausasteiden (*degrees of freedom*) lukumäärä (df)

Määrää em. funktioiden avulla

a) P(x < 1.96) kun x ~ N(0,1)
b) P(x > 17.5) kun x ~ N(10,5^2)
c) y siten, että P(x < y)=0.995 kun x ~ N(15,3^2)
d) y siten, että P(x < y)=0.95 kun x ~ t (df=20)
e) P(|x| > 2.03) kun x ~ t (df=7)
f) y siten, että P(x < y) =0.9 kun x ~ χ^2(df=4)
g) P(x > 5.6) kun x ~ χ^2(df=14)
h) P(x < 1.66) kun x ~ F(df=14 ja 20)
i) y siten, että P(x > y)=0.8 kun x ~ F(df=4 ja 20)

Tulostesi pitäisi löytyä alla olevasta lukujoukosta:

7.7794403397349	0.08191277644461	2.1842415712582	0.97500210485178
0.97558938047566	0.11122304887982	0.06680720126886	1.7247182429208
0.85366601309805	0.40900301928175	22.727487910647	

Mukana on kaksi ylimääräistä lukua. Mitkä ne ovat?

Tehtävä 27: Histogrammi ja normaalijakauman sovitus

Tarkastellaan jälleen jo tutuksi tullutta PISTEET-aineistoa.

Esitä Scores ja LOGITS histogrammin avulla. Sovita molempiin normaalijakauma. Laita GHISTO-operaatioon tulostusrivi, jotta saat kenttään yhteensopivuustestin tulokset.

Esimerkki Scores-muuttujan osalta:

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003          R:\DA1\    1000  100 0
  1 *SAVE DA1T27 / histogrammi ja normaalijakauman sovitus
  2 *
  3 *GHISTO PISTEET,Scores,END+2 / FIT=NORMAL
  4 *Scores=0(10)120 (riittävästi ylimääräisiä luokkia molempiin reunoihin)
  5 *
```

Mitkä ovat yhteensopivuustestin hypoteesit? Miten tulkitset lopputuloksen?

Tehtävä 28: Histogrammi ja beta-jakauman sovitus

Jatketaan edelleen PISTEET-aineiston parissa. Esitä muuttuja Scores histogrammin avulla. Sovita aineistoon beta-jakauma. Oheisen esimerkin DENSITY-kaaviossa jakauman tiheysfunktion ydin on määritelty skaalattuna välille (0,100).

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003            R:\DA1\    1000  100 0
  1 *SAVE DA1T28 / histogrammi ja beta-jakauman sovitus
  2 *
  3 *GHISTO PISTEET,Scores,END+2 / Scores=0(5)100 FIT=Beta
  4 *
  5 *DENSITY Beta(a,b)
  6 *  Y(X)=if(X<=0)then(0)else(Y2)
  7 *    Y2=if(X<100)then(tf)else(0)
  8 *    tf=X^(a-1)*(100-X)^(b-1)
  9 *
```

Miten tulkitset yhteensopivuustestin tuloksen?

Tehtävä 29: Normaalisuuden testaus COMPARE-operaatiolla

Testaa PISTEET-aineiston muuttujien Scores ja LOGITS normaalisuutta COMPARE-operaatiolla, esimerkiksi

```
COMPARE PISTEET(LOGITS),#NORMAL,END+2
```

Tulkitse tulokset. Vastaavatko ne tehtävän 27 tuloksia?

Tehtävä 30: Normaalisuuden testaus todennäköisyyspaperilla

Testaa PISTEET-aineiston muuttujien Scores ja LOGITS normaalisuutta piirtämällä muuttujien kertymäfunktiot ns. todennäköisyyspaperille (ks. PROBIT?).

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003            R:\DA1\    1000  100 0
  1 *SAVE DA1T30 / normaalisuuden testaus todennäköisyyspaperilla
  2 *
  3 *Aineisto on lajiteltava suuruusjärjestykseen:
  4 *
  5 *FILE SORT PISTEET BY Scores TO SPISTEET
  6 *
  7 *Kuva piirretään lajitellusta aineistosta:
  8 *
  9 *GPLOT SPISTEET,Scores,PROBIT
 10 *.....................................................................
```

Tehtävä 31: Normaalisuuden testaus NSCORES-operaatiolla

Testaa PISTEET-aineiston muuttujien Scores ja LOGITS normaalisuutta piirtämällä muuttujien kertymäfunktiot ja niille simuloimalla estimoidut luottamusvälit todennäköisyyspaperille. Laskelmat tehdään Juha Purasen laatiman lisämodulin NSCORES avulla. Aktivoi NSCORES ilman parametreja, niin saat käyttöohjeen. Oheisessa esimerkissä olevan mallin saat kenttään painamalla plus-nappia.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003            R:\DA1\    1000  100 0
  1 *SAVE DA1T31 / normaalisuuden testaus NSCORES-operaatiolla
  2 *
  3 *NSCORES / katsotaan miten NSCORES toimii
  4 *    Käyttöesimerkki (luottamusväli empiiriselle kertymäfunktiolle):
  5 *.....................................................................
  6 *   NSCORES <data>,SCORES.M     / vaihda tähän aineiston nimi
  7 *   VARS=<Muuttuja>(X)          / ja tähän tutkittavan muuttujan nimi
  8 *.....................................................................
  9 *   GPLOT SCORES.M,S,L,Z,U      / piirto: aktivoi tämä
 10 *   LLINE=[line_width(3)],1,1   ZLINE=[line_width(1)],1,1
 11 *   ULINE=[line_width(3)],1,1   SCALE=-4(1)4
 12 *.....................................................................
```

Tehtävä 32: Normaalisuuden testaus Lillieforsin testillä

Testaa PISTEET-aineiston muuttujien normaalisuutta Lillieforsin testillä. Se on vähemmän konservatiivinen versio Kolmogorovin ja Smirnovin yhteensopivuustestistä.

Testiä varten tarvitaan muuttujan empiirinen ja teoreettinen kertymäfunktio. Lillieforsin testisuure on näiden kertymäfunktioiden arvojen suurin erotus. Kertymäfunktio lasketaan järjestetystä aineistosta. Laske empiirinen ja teoreettinen kertymäfunktio sekä niiden erotuksen itseisarvo. Teoreettista kertymäfunktiota varten tarvitset muuttujan keskiarvon ja varianssin. Katso mallia oheisesta esimerkistä.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003            R:\DA1\    1000  100 0
  1 *SAVE DA1T32 / normaalisuuden testaus Lillieforsin testillä
  2 *
  3 *FILE SORT PISTEET BY Scores TO SPISTEET    /  aineiston järjestäminen
  4 *CORR SPISTEET CUR+1 / VARS=Scores
  5 *SPISTEET  N=111
  6 *Variable  Mean        Std.dev.
  7 *Scores    75.00000    12.36417
  8 *
  9 *VAR EKF=(ORDER-0.5)/N TO SPISTEET    /      empiirinen kertymäfunktio
 10 *VAR TKF=N.F(75,12.36417^2,Scores)   /     teoreettinen kertymäfunktio
 11 *VAR Z=ABS(TKF-EKF)                 / funktioiden erotuksen itseisarvo
 12 *.....................................................................
 13 *STAT SPISTEET CUR+1 / VARS=Z         RESULTS=0 (vain tunnusluvut)
 14 *Basic statistics: SPISTEET N=111
 15 *Variable: Z        ~ABS(TKF-EKF)
 16 *min=0.00078  in obs.#93
 17 *max=0.084324 in obs.#36
 18 *mean=0.032655 stddev=0.021768 skewness=0.629958 kurtosis=-0.528479
 19 *autocorrelation=0.8587
 20 *lower_Q=0.016417 median=0.028889 upper_Q=0.047292
 21 *
 22 *Testin kriittinen arvo 5 %:n tasolla on 0.886/sqrt(N)=
 23 *ja edellä havaittu itseisarvoltaan suurin erotus  max=
 24 *joten...
 25 *
```

Lillieforsin testin kriittinen arvo 5 %:n merkitsevyystasolla on $\frac{0.886}{\sqrt{n}}$.

Käytä hyväksesi esimerkin rivejä 22-23. Aktivoi lausekkeet aivan yhtäsuuruusmerkin oikealta puolelta. Tulkitse tulos (eli täydennä rivin 24 teksti).

Tehtävä 33: t-testi vaiheittain

Pikaruokaravintola **R** väitti, että heidän ranskanperuna-annoksessaan on enemmän ranskanperunoita kuin kilpailevan ravintolan **M** annoksessa. Asian tutkimiseksi molemmista paikoista käytiin hakemassa joukko annoksia ja laskettiin niiden perunalukumäärät.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003            R:\DA1\    1000  100 0
  1 *SAVE DA1T33 / t-testi vaiheittain
  2 *
  3 *Kun merkitään otoskokoja n1 ja n2, keskiarvoja m1 ja m2 sekä
  4 *keskihajontoja s1 ja s2, niin t-testisuure on t=(m1-m2)/sd
  5 *jossa sd=s*sqrt(1/n1+1/n2) ja yhteisen tuntemattoman hajonnan
  6 *estimaatti on s=sqrt(((n1-1)*s1^2+(n2-1)*s2^2)/(n1+n2-2))
  7 *
  8 *Tällöin t=
  9 *
 10 *Testin havaittu merkitsevyystaso (p-arvo) on p=1-t.F(n1+n2-2,t)
 11 *eli p=
 12 *
 13 *DATA R: 41, 48, 43, 57, 41, 43, 60, 51, 49, 55, 43, 48, 40, 48 END
 14 *DATA M: 52, 39, 43, 35, 53, 55 END
 15 *
 16 *CORR R END+2 / lasketaan ensin tarvittavat otossuureet ja
 17 *CORR M END+2 / merkitään ne kaavassa tarvittavilla symboleilla
 18 *
 19 *R  n1=14
 20 *Variable  Mean        Std.dev.
 21 *R         m1=47.64286 s1=6.319775
 22 *
 23 *M  n2=6
 24 *Variable  Mean        Std.dev.
 25 *M         m2=46.16667 s2=8.304617
```

Kirjoita aineistot toimituskenttään (ks. oheinen esimerkki). Muotoile testattavat hypoteesit sille, onko pikaruokaravintolassa **R** suuremmat annokset kuin pikaruokaravintolassa **M**. Suorita testaus t-testillä toimituskentässä editoriaalisen laskennan avulla.

Numeeriset tulokset saat aktivoimalla laskukaavat aivan rivien 8 ja 11 yhtäsuuruusmerkkien oikealta puolelta. Laske kuitenkin ensin tarvittavat otossuureet ja merkitse ne kaavassa tarvittavilla symboleilla kuten esimerkissä on tehty.

Miten tulkitset lopputuloksen? Onko pikaruokaravintolan **R** väite oikeutettu?

Tehtävä 34: t-testi COMPARE-operaatiolla

Tutustu kyselysysteemin avulla Survon tilastollisten testien valikoimaan (ks. TEST?). Testaa edellisen tehtävän ranskanperunahypoteesi COMPARE-operaatiolla. Vertaile tuloksia.

Tehtävä 35: t-testi Cooperin testin aineistolle

Cooperin testi lienee tuttu useimmille. Tutkitaan 17- ja 18-vuotiaiden poikien Cooperin testin tuloksia. Onko eri ikäisten välillä eroja? Miten muotoilet hypoteesit?

Kopioi valmis aineisto COOPER käyttöösi:

```
>COPY C:\DA1\COOPER.SVO
```

Siirrä testausta varten ikäryhmien tulokset erillisiin havaintotiedostoihin. Testaa tilanne ensin suoraan COMPARE-komennolla. Käytä sen jälkeen /VERTAA-sukroa.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003            R:\DA1\    1000  100 0
  1 *SAVE DA1T35 / t-testi Cooperin testin aineistolle
  2 *
  3 *>COPY C:\DA1\COOPER.SVO
  4 *
  5 */DATACOPY COOPER TO C17 / IND=Ikä,17
  6 */DATACOPY COOPER TO C18 / IND=Ikä,18
  7 *
  8 *COMPARE C17(Cooper),C18(Cooper),END+2
  9 *
 10 */VERTAA C17(Cooper),C18(Cooper) / sukro keskustelee suomeksi
 11 *
```

Tulkitse tulokset. Mitä mieltä olet niiden yleistettävyydestä?

Tehtävä 36: Havaintoaineiston aggrekointi

Usein on syytä siirtyä alkuperäisten havaintojen tasolta jollekin yhdistetylle tasolle, esimerkiksi kuntatasolta läänitasolle. Tällöin uusi yhdistetty aineisto muodostetaan alkuperäisen (painotetuista) summista tai keskiarvoista. Menettelyä kutsutaan aggrekoinniksi.

Tutkitaan tehtävässä 14 käsiteltyä KUNNAT-aineistoa eräiden väestömuuttujien osalta laskemalla kuntakohtaisten tietojen keskiarvot lääneittäin /AGGRE-sukrolla:

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003            R:\DA1\    1000  100 0
  1 *SAVE DA1T36 / havaintoaineiston aggrekointi
  2 *
  3 *Valitaan vain seuraavat muuttujat:   VARS=Lääni,Väestö,Synt.,SYNT
  4 *
  5 *FILE STATUS KUNNAT
  6 *  Suomen kunnat aakkosjärjestyksessä
  7 *  Tiedot ovat pääosin vuosilta 1978-80.      5.2.84/SM
  8 *            COND:KUNNAT       SORT:Kunta
  9 *FIELDS: (active)
 10 *   2 SA-   3 Lääni    UUS,TUR,AHV,HÄM,KYM,MIK,KAR,KUO,KES,VAA,OUL,LAP
 11 *   3 NA-   4 Väestö   Arvioitu maassa asuva väestö 1.1.1980 (######)
 12 *   4 NA-   4 Synt.    Elävänä syntyneet v.1978 (####)
 13 *  12 NA-   4 SYNT     1000*Synt./Väestö (##.###)
 14 *END
 15 *Survo data file KUNNAT: record=128 bytes, M1=30 L=64  M=14 N=464
 16 *
 17 */AGGRE KUNNAT BY Lääni TO KKARVOT
```

Tutustu näin saatuun läänitason aineistoon (FILE SHOW). Laske siihen muuttujaksi SYNTKA **läänikohtaiset syntyvyysluvut** ts. läänissä syntyneiden lasten lukumäärä 1000 asukasta kohden. Vertaa SYNT-muuttujaan. Ovatko muuttujien arvot samoja? Miksi?

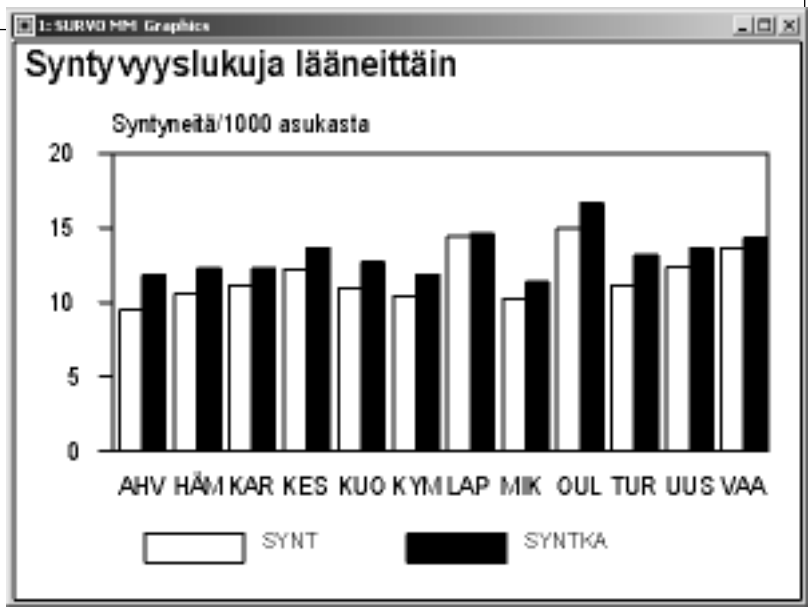
Tehtävä 37: Kuvan talletus tiedostoon

Palautetaan mieleen pylväskuvien piirtäminen, jota harjoiteltiin tehtävässä 18. Nyt tavoitteena on tallettaa kuva PostScript-muodossa tiedostoon ja liittää se myöhemmin osaksi tulostettavaa pienoisraporttia. Tällöin kannattaa GPLOT-komennon yhteydessä käyttää täsmennystä MODE=PS, jolla kuvan metriikka, oletusfontit yms. tulevat automaattisesti vastaamaan PostScript-puolen (PLOT-komennon) asetuksia.

Survon PostScript-kuvien oletuskoko on 1500 × 1500 (yksikkönä on millimetrin kymmenesosa). Koon voi valita vapaasti SIZE-täsmennyksellä. Kuvat voi suunnitella hyvin lopulliseen muotoon GPLOT:in avulla ennen kuin siirtyy PLOT:in puolelle. Tietenkin voi myös tallettaa kuvan **metatiedostona** (ks. tehtävä 22), jolloin GPLOT riittää mainiosti. Tuttu millimetripohjainen mittakaava auttaa joka tapauksessa kuvan hienosäädössä.

Piirrä edellisen tehtävän molemmat läänikohtaisia syntyvyyslukuja kuvaavat muuttujat monipylväskuviona (TYPE=MVBAR). Lisää kuvaan tarvittavat täsmennykset (ainakin otsikko ja asteikkonimet). Oheisessa esimerkissä on samalla näytetty miten pylväiden selitystekstit voidaan asetella oletuksesta poikkeavasti. Se ei ole mitenkään välttämätöntä (voit hyvin jättää rivin 7 pois kaaviostasi) mutta jos kiinnostaa, katso parametrien selitykset kyselysysteemistä (LEGEND?).

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003          R:\DA1\    1000  100 0
  1 *SAVE DA1T37 / kuvan talletus tiedostoon
  2 *
  3 *HEADER=[SwissB(15)],Syntyvyyslukuja_lääneittäin
  4 *GPLOT KKARVOT / TYPE=MVBAR  VARS=Lääni,SYNT,SYNTKA
  5 *YLABEL=Syntyneitä/1000_asukasta YSCALE=0(5)20
  6 *MODE=PS  SIZE=1200,750  XDIV=150,1000,50  YDIV=200,400,150
  7 *LEGEND=[Swiss(9)],200,50,2 LEGEND_BOX=400,0,150,40 LEGEND_TEXT=180,10
  8 *.....................................................................
```



Kun saat kuvan valmiiksi, talleta se **PostScript-muodossa**: lisää piirroskaavioon täsmennys DEVICE=PS,KKARVOT.PS ja aktivoi sitten GPLOT:in sijasta PLOT (ota G pois).

Työhakemistossasi pitäisi nyt olla tiedosto KKARVOT.PS. Tarkista DD:llä että näin on.

Tulosta kuva paperille Survon PRINT-operaatiolla oheisen esimerkin mukaisesti. Huomaa että rivin 11 edessä, ns. **kontrollisarakkeessa** on nyt miinus-merkki. Kontrollisarake on hyvin tähdellinen alue, jota käytetään mm. toimituskentän rivien nimeämiseen ja tulostuksen ohjaukseen. Pääset sinne siirtymällä rivin alkuun ja painamalla vasemman nuolinäppäimen hetkeksi pohjaan.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003          R:\DA1\    1000  100 0
  9 *
 10 *PRINT CUR+1,CUR+1 TO K.PS    / tulostetaan vain seuraava rivi
 11 - picture KKARVOT.PS,400,600 / kuvan vasemman alakulman sijainti A4:llä
 12 *
 13 */GV-SHOW K.PS / esikatselu ja lopullinen tulostus
 14 *
```

Ajan ja kustannusten säästämiseksi kannattaa PostScript-tulostukset aina ensin esikatsella kuvaruudulla esimerkiksi *GSview*-nimisellä ohjelmalla, jota varten on käytettävissä sukro /GV-SHOW. Ohjelma on asennettu valmiiksi tilastotieteen laitoksen atk-luokkaan, ja sen voi hakea myös itselleen verkosta, osoitteesta **www.cs.wisc.edu/~ghost/gsview** .

Paperitulosteen saa kätevästi esimerkiksi GSview:n File-valikon Print-valinnalla. Näin voi Survosta tulostaa millä tahansa Windows-kirjoittimella.

GSview:lla voi myös muuntaa PostScript-tiedostot PDF-muotoon esim. verkkojulkaisemista varten. Vastaava GSview:ta monipuolisempi kaupallinen ohjelma on *Distiller*, joka on osa Adoben *Acrobat*-pakettia. Malliksi hieman laajemmasta Survolla laaditusta painotuotteesta käy eräs vuonna 2000 julkaistu tilastotieteen väitöskirja, johon voi tutustua verkon kautta osoitteessa **ethesis.helsinki.fi/julkaisut/val/tilas/vk/vehkalahti** .

Tehtävä 38: Raportin tulostus

Survon toiminta-ajatukseen kuuluu, että sillä voidaan suorittaa tutkimusprojektin vaiheet alusta loppuun, suunnitelmista ja kustannuslaskelmista lähtien aineiston tallentamiseen, tilastolliseen käsittelyyn ja kuvantamiseen sekä tutkimusraportin saattamiseen kirjapainovalmiiseen tai verkossa julkaistavaan muotoon. Tällaisen työskentelytavan etuna on, ettei tiedon sirpaleita tarvitse siirrellä edestakaisin useamman ohjelman välillä. Myös työn dokumentointi (ks. tehtävä 3) pysyy paremmin yhtenäisenä.

Survon oma tulostuskieli tarjoaa hyvin joustavat keinot julkaisujen laadintaan. Sillä voi PostScript-muodon lisäksi tehdä myös HTML-tiedostoja (ks. `HTML?`). Mm. Survon omat verkkosivut (**www.survo.fi**) laaditaan ja ylläpidetään Survolla.

Tämän kurssin tehtävät koostuvat pääasiassa tilastollisesta käsittelystä ja kuvantamisesta. Jotta saataisiin tuntumaa myös tulosten julkaisemisesta, tehdään yksinkertainen pienoisraportti, joka sisältää tyypilliset tutkimusraportin ainekset: tekstiä, taulukon ja kuvan selityksineen. Halutessaan sitä voi käyttää mallina esimerkiksi tämän kurssin harjoitustyölle.

Tehtävänäsi on laatia yhden sivun pituinen raportti tehtävän 36 tuloksista. Raportin ytimen muodostaa läänikohtaisten syntyvyyslukujen taulukko, jota säestää tehtävässä 37 piirretty pylväskuvio. Perusta aluksi uusi toimituskenttä ja poimi taulukko siihen komennolla

```
FILE LOAD -KKARVOT CUR+2  /  VARS=Lääni,FREQ,SYNT,SYNTKA
```

Kopioi seuraavan sivun esimerkki muilta osin itsellesi. Voit mielellään otsikoida omalla tavallasi ja kirjoittaa lennokkaampaa tekstiä mutta jos et jaksa niin tyydy näihin. Pääasia on että saat raportin rakenteen haltuusi. Eräät yksityiskohdista on selitetty kommentteina. Loput selviävät varmasti kokeilemalla tai tutustumalla kyselysysteemiin esimerkiksi avainsanojen `PRINT?` ja `PSPICT?` kautta.

Tasaa lopuksi teksti valittuun palstanleveyteen. Tämä on niitä asioita, joissa saatat tarvita *"eiku"*-tekniikkaa (ks. tehtävä 3), joten **talleta toimituskenttäsi** ennen kuin jatkat eteenpäin.

Tehtävässä 3 kokeiltiin jo alustavasti `TRIM`-komentoa sellaisenaan. Käytetään nyt valmista sukrokomentoa, jolla saadaan valittua leveyttä ja fonttikokoa vastaava suhteutetuille fonteille sopiva `TRIM`-komento. Aktivoi siis tasattavien tekstikappaleiden yläpuolella (oheisessa esimerkissä riveillä 18 ja 41) komento

```
/TRIMP 65,Times(12)
```

jolloin teksti tasataan siten että tulostettaessa sen oikea reuna on suora. Tavutus tapahtuu suomen kielen sääntöjen mukaisesti. Joskus pieniä yksityiskohtia joutuu säätämään käsin, koska jotkut yhdyssanat ovat mahdottomia hallita ohjelmallisesti (mm. *kaivosaukko*).

Lienee makuasia, tarvitseeko oikeaa reunaa lainkaan tasata. Jotkut ovat sitä mieltä, että "liehureuna" on silmille mukavampi. Ainakin sellainen väkisin tasattu teksti jossa sanojen väliin jää isoja aukkoja, on todella rasittava. Vältä sellaisia ja pidä vain huoli että *vasen* reuna on suorassa. Siistin liehureunan oikealle saat yleensä käyttämällä yksinkertaista `TRIM`-komentoa, jonka voi vielä lyhentää pelkäksi `T`:ksi.

Kun olet tyytyväinen raporttisi ulkoasuun, tulosta se paperille ja ihaile.

Tehtävään 38 liittyvä esimerkkikenttä:

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003            R:\DA1\    1000  100 0
  1 *SAVE DA1T38 / raportin tulostus
  2 *
  3 *
  4 *PRINT CUR+2,E TO TEKSTI.PS                    / tulostus tiedostoon
  5 */GV-SHOW TEKSTI.PS                            / esikatselu
  6 - define [TEKSTI] [Times(12)][trim(65)]        / leipätekstin tyyli
  7 - define [OTSIKKO] [Times(18)][trim(65)]       / otsikkotyyli
  8 - define [TAULUKKO] [Courier(11)][trim(0)] / taulukon tyyli
  9 - [margin(400)][line_spacing(13)]              / vasen marginaali, riviväli
 10 - [TEKSTI]                                     / tyylin valinta
 11 R                          Mika Lahdenkari 19.7.1999
 12 R                        (Kimmo Vehkalahti 19.7.2001)
 13 *
 14 *
 15 - [OTSIKKO]
 16 C              Syntyvyyslukuja eri lääneissä
 17 - [TEKSTI]
 18 *
 19 *Oheisessa taulukossa on tietoja syntyvyydestä lääneittäin (syntyneitä
 20 *1000 asukasta kohden). Muuttuja SYNT on kuntakohtainen keskiarvo
 21 *lääneittäin ja SYNTKA läänikohtainen syntyvyysluku. Lisäksi taulukosta
 22 *ilmenee kunkin läänin kuntien lukumäärä. Tiedot ovat vuosilta 1978-80.
 23 *
 24 - [TAULUKKO]  /  hae nämä tiedot kenttään aineistostasi ja otsikoi!
 25 *Lääni  Kuntia     SYNT       SYNTKA
 26 * AHV      16       9.513      11.861
 27 * HÄM      49      10.597      12.251
 28 * KAR      19      11.186      12.271
 29 * TUR      96      11.198      13.168
 30 * UUS      40      12.420      13.606
 31 * KES      32      12.248      13.697
 32 * KUO      24      10.976      12.714
 33 * KYM      28      10.437      11.898
 34 * MIK      29      10.233      11.398
 35 * VAA      57      13.661      14.367
 36 * LAP      22      14.417      14.614
 37 * OUL      52      15.045      16.654
 38 *
 39 - [TEKSTI]
 40 *Taulukko 1: Syntyvyyslukuja lääneittäin
 41 *
 42 *Allaoleva kuva kertoo taulukon 1 sisältämän informaation pylväskuvion
 43 *muodossa. Vaaleat pylväät kuvaavat kuntakohtaisten syntyvyyslukujen
 44 *keskiarvoa (muuttuja SYNT) ja mustat pylväät läänikohtaisia
 45 *syntyvyyslukuja (muuttuja SYNTKA).
 46 *
 47 % 750 (75 mm tilaa kuvalle)
 48 - picture KKARVOT.PS,*+100,* / kuvan sijoitus ja siirto 10 mm oikealle
 49 *
 50 C Kuva 1: Syntyvyyslukuja lääneittäin
 51 *
 52 *Lisää tekstiä, esim. yhteenveto. Keksi itse.
 53 E
 54 *.................................................................
```

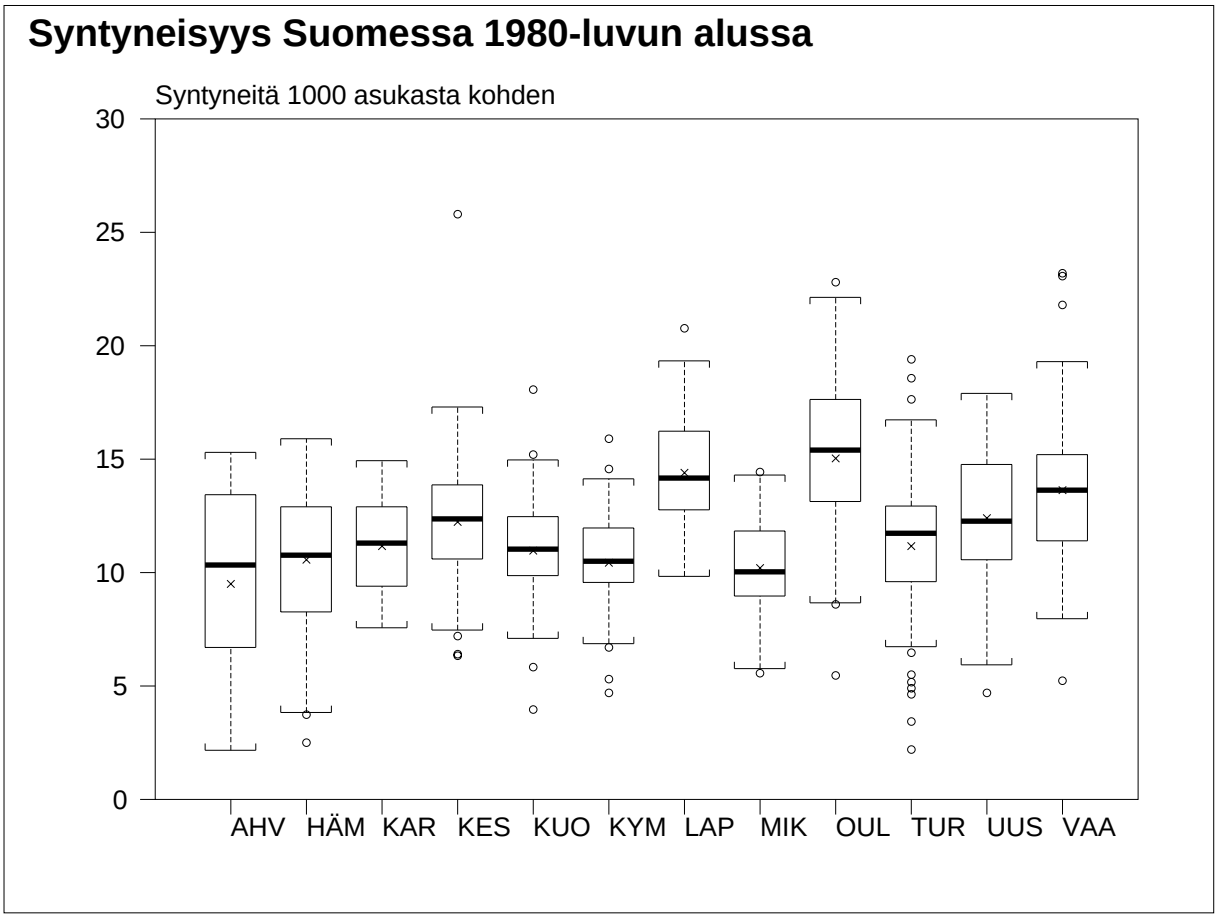
Survon tulostuskieli sisältää edellä käytettyjen lisäksi muitakin käteviä keinoja, joilla yksinkertaistetaan tyylien hallintaa sekä paperi- että verkkojulkaisuja rakennettaessa. Ovelimpia näistä ovat ns. **varjomerkit** (ks. SHADOWS?), jotka näkyvät toimituskentässä erilaisina väreinä. Suomenkielinen opetussarja kertoo aiheesta lisää (DEMO → OPETUS → 4. Tekstinkäsittelyn alkeita). Myös Survon englanninkieliseen käyttöoppaaseen sisältyy toistakymmentä esimerkkiä tekstin, taulukoiden ja kuvien tulostamisesta (START → 6. Survo-kirjan esimerkit).

Tehtävä 39: Box plot -kuvio

Box plot on tilastollinen kuvatyyppi, jossa esitetään tiiviisti jakaumaa koskevat tunnusluvut ryhmiteltyinä luokittelevan muuttujan arvojen mukaan.

Tehtävänä on esittää aiemmin käsitellyt Suomen kuntien läänikohtaiset syntyvyysluvut box plot -kuvana. Survon valikoimaan ei suoraan kuulu box plot -kuvatyyppiä vaan kuva tehdään sukrolla /BOXPLOT. Se on tyypillinen esimerkki yhdistelmäkuvia piirtävistä sukroista, joita Survossa on muitakin (ks. PLOTSUC?).

Tavoitteena on siis saada aikaan suurin piirtein seuraavanlainen kuva:



Lähdetään liikkeelle kuntatason aineistosta KUNNAT (ks. tehtävä 14). Sen aggrekoinnista läänitasolle (vrt. tehtävä 36) huolehtii nyt /BOXPLOT-sukro. Aktivoi siis

```
/BOXPLOT KUNNAT,Lääni,SYNT
```

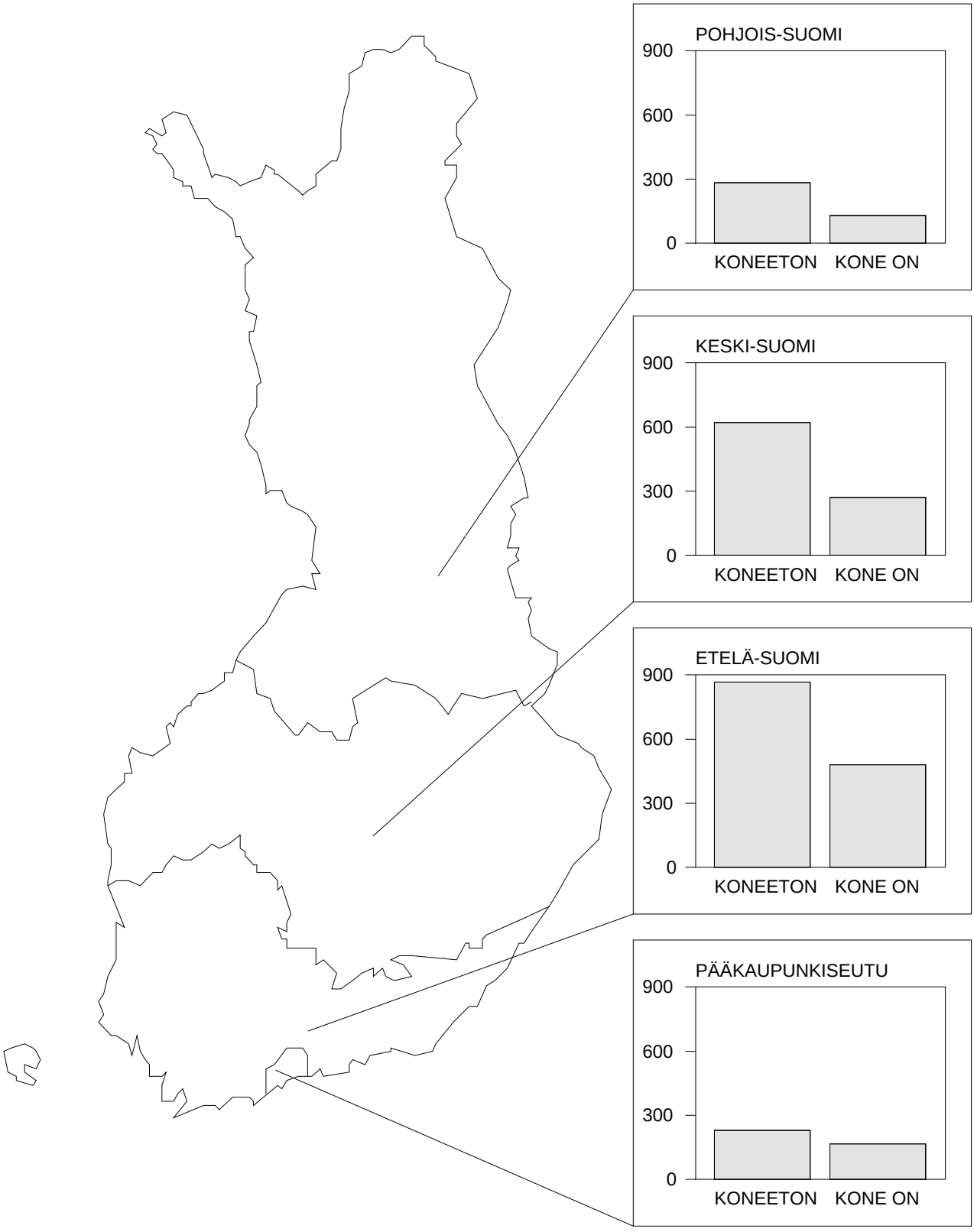
jolloin /BOXPLOT-sukro perustaa uuden työkentän kuvanpiirtoa varten. Kuva ilmestyy lopuksi usean kuvan yhdistelmänä omaan kuvaikkunaansa. Työkentässä (lähinnä ensimmäisessä osakentässä) olevia täsmennyksiä voi muutella vapaasti. Kentän alussa olevilla riveillä on valmiita sukrokomentoja, joilla kuvan voi piirtää uudelleen joko kuvaikkunaan tai PostScript-tiedostoon. Ensimmäisen rivin komennolla työkenttä unohdetaan ja palataan alkuperäiseen toimituskenttään. Työkentän lopussa on komennot PostScript-muotoisen kuvan esikatselua ja työhakemistoon kopiointia varten.

Tehtävä 40: Kartta

Viimeisen tehtävän aiheena on piirtää Suomen kartta. Sitä tarvitaan harjoitustyön alueellisissa tarkasteluissa. Survoon ei kuulu mitään erityistä kartanpiirto-operaatiota vaan kartta syntyy tavallisena "hajontakuvana" kun koordinaattipisteet yhdistetään viivoilla toisiinsa.

Koordinaatit on valmiiksi talletettu havaintotiedostoon SKARTTA. Kopioi se itsellesi:

```
>COPY C:\DA1\SKARTTA.SVO
```



Oheisessa esimerkissä on kuvanpiirtokaavio, jolla kartta piirretään kuvaikkunaan. Samalla se tallettuu metatiedostoon SKARTTA.EMF. Kerrostamistekniikalla (ks. tehtävä 22) voit hyödyntää sitä pohjana alueellisia tarkasteluja tehdessäsi. Samaan kaavioon upotetulla PLOT-komennolla kuva tallettuu PostScript-tiedostoksi SKARTTA.PS .

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003           R:\DA1\    1000  100 0
  1 *SAVE DA1T40 / kartta
  2 *
  3 *>COPY C:\DA1\SKARTTA.SVO
  4 *.....................................................................
  5 *GPLOT SKARTTA,X,Y /   OUTFILE=SKARTTA    MODE=PS XDIV=0,1,0 XLABEL=
  5 *       IND=SA2      HOME=200,100          SCALE=0,30 FRAME=0 HEADER=
  6 * PLOT SKARTTA,X,Y / DEVICE=PS,SKARTTA.PS LINE=1  YDIV=0,1,0 YLABEL=
  7 *.....................................................................
```

Harjoitustyöhön tehtävään karttaan pitää lisäksi liittää alueita kuvaavat pylväs- tai piirakkadiagrammit. Katsotaan malliksi miten edellisen sivun kuva on laadittu. Se koostuu kuudesta osakuvasta, jotka on kerrostettu päällekkäin. PostScript on täysin "läpinäkyvää", joten kuvia voi rakentaa mukavasti pala kerrallaan.

Aluksi on piirretty alueittaiset pylväsdiagrammit ja talletettu ne omiin kuvatiedostoihinsa. Tässä ei ole sinänsä mitään uutta; näitä asioita on harjoiteltu jo tehtävissä 14, 16–18 ja 37. Muuttujat on oheisessa kaaviossa aktivoitu MASK-täsmennyksellä, koska se on tässä yhteydessä kätevämpi kuin VARS.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003           R:\DA1\    1000  100 0
  8 *
  9 *Piirretään malliksi kuvat koko aineistosta tehdystä ristiintaulukosta:
 10 *
 11 *TAB KDATA,CUR+4 / VARIABLES=SUUR,VA07U (mikrotietokone)
 12 *SUUR=1,1(PKS),2(ETELÄ),3(KESKI),4(POHJOIS)
 13 *VA07U=0,0(KONEETON),1(_KONE_ON)
 14 *
 15 *TABLE KDATA1 A,B,F   N=3052
 16 A         SUUR     PKS   ETELÄ   KESKI POHJOIS
 17 *VA07U    ****
 18 *KONEETON          229     867     622     283
 19 B_KONE_ON          167     481     272     131
 20 *Chi_square=19.22 df=3 P=0.0002
 21 *
 22 *DATA KONEET,A+2,B,A
 23 *PLOT KONEET / MASK=AA--- DEVICE=PS,KONEET1.PS YLABEL=PÄÄKAUPUNKISEUTU
 24 *PLOT KONEET / MASK=A-A-- DEVICE=PS,KONEET2.PS YLABEL=ETELÄ-SUOMI
 25 *PLOT KONEET / MASK=A--A- DEVICE=PS,KONEET3.PS YLABEL=KESKI-SUOMI
 26 *PLOT KONEET / MASK=A---A DEVICE=PS,KONEET4.PS YLABEL=POHJOIS-SUOMI
 27 *
 28 *TYPE=VBAR SIZE=650,550 XDIV=120,480,50 YDIV=90,370,90 PEN=[Swiss(10)]
 29 *GAP=0.2,0.2,0.2 HEADER= LEGEND=- YSCALE=0(300)900
 30 *.....................................................................
```

Lisäksi on piirretty viivat pylväsdiagrammeista kartan vastaaville alueille selventämään kokonaiskuvaa. Tätä varten on määritelty pieni kahden muuttujan data, johon on arvioimalla ja kokeilemalla haettu sopivat koordinaatit. Pisteet yhdistävä viiva saadaan katkeamaan sopivasti laittamalla dataan puuttuvia tietoja (-).

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003           R:\DA1\    1000  100 0
 31 *
 32 *DATA VIIVAT
 33 *X    Y
 34 *70   50    POHJOIS-SUOMI
 35 *100  72
 36 *-    -
 37 *60   30    KESKI-SUOMI
 38 *100  48
 39 *-    -
 40 *50   15    ETELÄ-SUOMI
 41 *100  24
 42 *-    -
 43 *45   12    PÄÄKAUPUNKISEUTU
 44 *100  0
 45 *
 46 *PLOT VIIVAT,X,Y / LINE=1 SCALE=0,100 FRAME=0  HEADER= XLABEL= YLABEL=
 47 *SIZE=1250,2500 HOME=200,100 DEVICE=PS,VIIVAT.PS XDIV=0,1,0 YDIV=0,1,0
 48 *.....................................................................
```

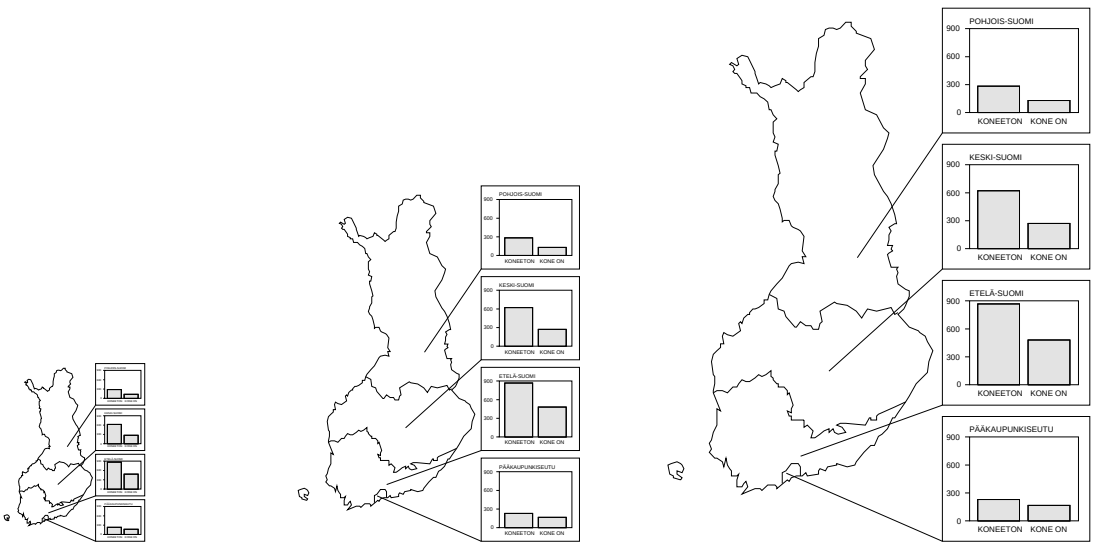
Edellä piirretyt kuvat on lopuksi yhdistetty yhdeksi kuvatiedostoksi KONEET.PS, joka voidaan liittää raporttiin vastaavalla tavalla kuin tehtävässä 38.

```
  1  1 SURVO MM  Wed Sep 17 16:15:14 2003          R:\DA1\    1000  100 0
 49 *
 50 *Lopuksi yhdistetään koko komeus yhdeksi PostScript-tiedostoksi:
 51 *
 52 *EPS JOIN KONEET.PS,SUOMI,K1,K2,K3,K4,MITKÄ / tekee tiedoston KONEET.PS
 53 *                              joka koostuu täsmennyksinä annetuista osista:
 54 *SUOMI=SKARTTA.PS,-150,400,1.6,1.6  (suurennetaan tässä karttaa hieman)
 55 *   K1=KONEET1.PS,1100,200
 56 *   K2=KONEET2.PS,1100,800
 57 *   K3=KONEET3.PS,1100,1400
 58 *   K4=KONEET4.PS,1100,2000
 59 *MITKÄ=VIIVAT.PS,-150,200
 60 *
 61 *Kuva on valmis ja se voidaan liittää raporttiin picture-ohjauksella
 62 *(sitä voidaan edelleen pienentää tai suurentaa).
 63 *
 64 */GV-SHOW KONEET.PS / ihaillaan lopputulosta kuvaruudulla
 65 *
```

Siinä olivat kaikki Survo-kurssin tehtävät! Jos jotain jäi epäselväksi tai kaipaat lisätietoja, käänny ensisijaisesti kurssin pitäjien puoleen.

Tämä kurssimateriaali kattaa Survon toimintojen valikoimasta vain pienen osan. Survon koko laajuudesta antaa kuvan mm. sen START-kentässä esiintyvä sanasto, joka löytyy myös verkosta osoitteesta **www.survo.fi/sanasto** .

PS. Harjoitustyön tekeminen kannattaa aloittaa samantien. Ohjeet löytyvät seuraavilta sivuilta.



Data-analyysi I -kurssin harjoitustyö

Seuraavassa on lyhyet ohjeet kurssin harjoitustyön laatimiseksi. Tarkemmat ohjeet löytyvät kurssin kotisivuilta osoitteesta **noppa5.pc.helsinki.fi/uudet/da1htm/hohje2000.html** .

Kulutustutkimus

Harjoitustyössä tutkitaan kotitalouksien kulutuskäyttäytymistä. Työn tekemiseen tarvitset oman 300 havainnon tutkimusaineiston (ks. tehtävä 11). Tarkoituksena on harjoittaa itsenäistä tilastollista työskentelyä, ja siksi annetut ohjeet ovat varsin ylimalkaisia.

Aineiston käyttö on rajoitettu tälle kurssille. Aineiston voi hakea harjoitustyön tekoa varten kurssin harjoitusten yhteydessä tai sopimuksen mukaan kurssin pitäjältä.

Harjoitustyö koostuu viidestä osasta:

1. Taulukointi ja graafiset esitykset
2. Jakaumatarkastelut
3. Alueelliset tarkastelut
4. Riippuvuustarkastelut
5. Testaus

1. Taulukointi ja graafiset esitykset

Taulukoi aineisto suuraluemuuttujan SUUR ja **kahden** alla mainitun muuttujan mukaan (siis yksi taulukko, kolme muuttujaa):

VA01U	Väritelevisio
VA03U	Videonauhuri
VA07U	Mikrotietokone
PSU	Pesukone
VA17U	Astianpesukone
VA18U	Mikroaaltouuni
VA22U	Lankapuhelin
VA23U	Matkapuhelin
VA28U	Henkilö- tai pakettiauto, lkm
AJOKITU	Oma- tai työsuhdeauto
VANETKU	Internetin käyttömahdollisuus

Tulkitse taulukon sisältö sanallisesti ja piirrä kuva, joka havainnollistaa asiaa mahdollisimman selkeästi. Esitä työssäsi sekä taulukko että kuva.

2. Jakaumatarkastelut

Valitse tai johda jakaumatarkasteluja varten kaksi jatkuvaa muuttujaa. Sopivan tyyppisiä ovat esimerkiksi

- tiettyyn ryhmään kuuluvat kulutusmenot
- tiettyyn ryhmään kuuluvat kulutusmenot henkilöä kohden
- tiettyyn ryhmään kuuluvien kulutusmenojen osuus tuloista
- ikä
- asunnon pinta-ala
- velat
- tulot

Kuvaa muuttujien jakaumaa histogrammin avulla. Yritä sovittaa aineistoon sopiva teoreettinen jakauma, esim. normaali- tai lognormaalijakauma, gammajakauma tms.

Eräissä tapauksissa aineisto kannattaa jakaa osiin (esimerkiksi suuraluemuuttujan SUUR tai asuinkunnan taajama-asteen TAAJAMA perusteella) ennen jakauman sovitusta. Joissakin tapauksissa kannattaa suorittaa sopiva muuttujamuunnos (esim. logaritmointi tai logit-muunnos).

Kulutusmenoissa on myös runsaasti sellaisia ryhmiä, joissa esiintyy paljon nollia, ts. tutkittavat eivät ole tarkastelujakson (2 viikkoa) aikana ostaneet näitä tuotteita. Nollahavainnot on syytä jättää jakaumatarkastelujen ulkopuolelle – muuten aineistoon on vaikea sovittaa mitään teoreettista jakaumaa.

3. Alueelliset tarkastelut

Alueellisissa tarkasteluissa tutkitaan karttapohjalla millaisia alueellisia eroja kulutuksessa esiintyy. Jokaisesta havainnosta on tieto siitä mihin suurpiiriin se kuuluu. Jaa aineisto muuttujan SUUR perusteella osa-aineistoihin. Muodosta joko histogrammi sopivasta muuttujasta tai pylväs- tai piirakkadiagrammi kahden luokittelumuuttujan taulukosta.

Piirrä kuvat karttapohjalle. Yhden esimerkin näet tehtävästä 40. Tulkitse sanallisesti millaisia eroja suuralueiden välillä esiintyy. Tarkemmassa harjoitustyöohjeessa kurssin kotisivuilla on lisää erilaisia malleja kartan ja kuvien yhdistelyyn. Valitse niistä sopiva.

4. Riippuvuustarkastelut

Riippuvuustarkasteluja tehdään sekä havaintotasolla että aluetasolla.

Valitse aineistosta kaksi muuttujaa. Sopivia ovat jälleen

- tiettyyn ryhmään kuuluvat kulutusmenot
- tiettyyn ryhmään kuuluvat kulutusmenot henkilöä kohden
- tiettyyn ryhmään kuuluvien kulutusmenojen osuus tuloista
- ikä
- asunnon pinta-ala
- velat
- tulot

Suorita tarkastelut seuraavasti:

1. Tasoita riippuvuutta sopivalla tasoitusmenetelmällä (esimerkiksi LOWESS-tasoitus tai luokittaiset keskiarvot)

2. Arvioi silmämääräisesti, voisiko riippuvuus olla lineaarista. Onko regressiosuora siis likimain sama kuin tasoitus? Mikäli on, sovita aineistoon PNS-suora (regressiosuora). Mikä on kyseisen regressiosuoran yhtälö?

3. Tutki onko aineistossa havaintoja, jotka vaikuttavat poikkeuksellisilta. Etsi kyseiset havainnot aineistosta ja yritä löytää syy poikkeavuudelle.

Riippuvuustarkasteluissa on erityisesti mietittävä, kumpi on riippuva (selitettävä) ja kumpi riippumaton (selittävä) muuttuja. Valittu tarkastelutapa on jollain tavalla perusteltava.

Peruspiiritasolla suoritettavia tarkasteluja varten yhdistetään samaan NUTS3-maakuntaan liittyvät havainnot laskemalla niistä painotettu muuttujasumma (FILE AGGRE). Painotuksella otetaan huomioon se, että havaintoja valittaessa eri osaryhmistä on poimittu aineistoon (otokseen) talouksia eri suhteessa. Painomuuttujana käytetään muuttujaa K0R8U.

Näin saadaan 20 summamuuttujaa, joiden välisiä riippuvuuksia voidaan tutkia kuten havaintotasolla. Pari mielekästä riippuvuustarkastelua kummankin osalta riittää.

5. Testaukset

Testauksessa tutkitaan jotakin ryhmään *Elintarvikkeet ja alkoholittomat juomat* sisältyvää osaryhmää kahdessa eri NUTS3-maakunnassa.

Etsi järkevä tutkimusongelma, muotoile siihen liittyvät hypoteesit ja suorita testaus. Esimerkiksi voisit tutkia kulutetaanko suklaaseen yhtä paljon rahaa Etelä- ja Pohjois-Suomessa. Tutkimusta varten valitsisit kaksi NUTS3-maakuntaa, esimerkiksi 4 (Satakunta) ja 18 (Kainuu), ja esittäisit hypoteesin *"Suklaaseen käytetään yhtä paljon rahaa Satakunnassa ja Kainuussa"* (sekä siihen liittyvän vastahypoteesin).

Tässä harjoitustyössä testaus suoritetaan seuraavalla (ehkä hieman virheellisellä) tavalla: käytetään sopivaa kahden riippumattoman otoksen keskiarvotestiä (havainto = talous). Jätä testauksessa pois kaikki ne taloudet, joissa tarkasteltavassa tuoteryhmässä ei ole ollut menoja (todellisuudessa ihan näin ei saisi menetellä). Tutki ensin jakauman normaalisuus ja valitse sen perusteella keskiarvotestissä käytettävä testi, joko t-testi tai Mann-Whitney-testi.