SUOMEN PANKIN

TILASTO-OSASTON TYÖPAPEREITA

29.12.1999 **5/99**

Heikki Hella ja Maria Sorsa

STATISTICA-ohjelman käyttö

SISÄLLYSLUETTELO

Statistica-ohjelman käyttö

1. Johdanto

Tämän työpaperin tavoitteena on johdattaa lukijaa tilastollisen selvitystyön ja analyysin käytännön tehtäviin, toimia "niksivihkona" käytettäessä tunnettua Statistica-ohjelmapakettia ja tukea tilastollisten aineistojen laadunvalvontatyötä. Lisäksi monisteen loppuun on valikoitu kirjallisuutta tilastomenetelmien soveltamisen tarpeisiin.

Tilastollinen ohjelmapaketti Statistica on vuosia sitten USA:ssa kehitetty ja laajasti yrityksissä ja yliopistoissa käytetty tilastollisten menetelmien monipuolisen arsenaalin PC-paketti. Ohjelman "vetovoima" on sen integroitu modulijärjestelmä, helppokäyttöinen tiedostorakenne, monipuolinen grafiikka ja Excel-yhteensopivuus. Ohjelmapaketin internet-sivulla:

http://www.statsoft.com/

on mm. elektroninen tilastollisten menetelmien oppikirja esimerkkeineen ja ohjeineen. Nettisivulla on myös paikka (FAQ´s) käyttäjien kysymyksille, ohjelman ylläpitäjien vastauksille (ongelmaratkaisuja), Statistica-uutisia jne.[1]

2. Soveltuvuus osaston erilaisiin työtehtäviin

2.1 Havaintoaineiston luominen

Havaintoaineisto voidaan joko syöttää suoraan Statisticaan muuttujittain[2] tai ladata valmis havaintomatriisi Excel-muodosta. Pankissa on otettu käyttöön syksylla 1999 Statistica -versio 5.5, joka pystyy lukemaan päinvastoin kuin useimmat muut tilastolliset ohjelmat myös Excel-97 -formaattia.

2.2 Editoriaaliset linkit

2.2.1 Havaintomatriisin haku Excel-muodosta

Haetaan havaintomatriisi komennolla *File, Import,Data,Quick.* Kun tiedosto on valittu, ohjelma kertoo Excelistä lukemiensa muuttujien lukumäärän ja tiedostossa olevien tapausten lukumäärän. Samassa yhteydessä pyydetään ruksaamaan kohta *Get variable names from 1st row of specified range* , jolloin ohjelma lukee Excel-tiedostossa ensimmäisellä rivillä olevien muuttujien nimistä kahdeksan merkkiä suoraan Statisticaan muuttujien nimiksi. Statistica hyväksyy vain kirjaimella alkavan muuttujanimen. Mikäli Excel-tiedoston aikamuuttujan havainnot tulevat Statisticaan viisinumeroisena lukuarvona, voidaan ne Statisticassa palauttaa selväkieliseksi aikamuuttujaksi kaksoisnapauttamalla kyseistä muuttujaa ja valitsemalla saadusta valikosta *Date*-määritys.

---

[1] Ohjelmaan on saatavissa erityinen "Help Menu – Animated Overviews", jonka TK asentanee rauhoitusajan jälkeen maaliskuussa 2000.

[2] Statcon Oy: Statisticaharjoitukset, 1998

2.2.2 Linkitys Excelistä Statisticaan

Excelin alaisuudessa olevan tiedoston havainnot (ei muuttujien nimiä) kopioidaan leikepöydälle. Siirrytään vastaavaan Statisticatiedostoon *Data Management* -moduliin. Klikataan kerran ensimmäisen muuttujan ensimmäistä havaintoa ja valitaan valikkoriviltä *Edit, Paste link.* Molempien linkattujen tiedostojen ollessa auki yhtä aikaa Excelissä tehdyt muutokset näkyvät välittömästi myös vastaavassa Statistican tiedostossa. Jos Statistican havaintomatriisi avataan vasta tehtyjen muutosten jälkeen, ohjelma kysyy käyttäjältä, haluaako hän pitää linkkaukset voimassa. Vastaamalla myöntävästi Statistica-tiedosto päivittyy samassa yhteydessä viimeisimmillä havainnoilla. Linkitys voidaan tehdä myös osalinkityksenä, jos halutaan päivitykset vain tiettyyn/tiettyihin muuttujaan/muuttujiin. Tällöin linkitys on syytä tehdä muuttujittain.

Linkitys voidaan korvata *Copy/Paste* -toiminnolla leikepöydän avulla. Tämä on perusteltua erityisesti silloin, kun tiedostojen säännöllinen päivittäminen ei ole tarpeen. On huomattava, että Excel-tiedostosta kopioidaan ainoastaan dataa ei muuttujan nimeä[3]. Statistican puolella varmistutaan siitä, että kopioitu aineisto liitetään alkavaksi täsmälleen samasta kohdasta kuin se on vastaavassa Excel-taulukossa.

2.2.3 Havaintoaineiston tallentaminen muihin ohjelmiin

Valitaan perusmoduli *Data Management.* Statisticasta voidaan tallentaa mm. seuraaviin tiedostomuotoihin: *Excel, Lotus, Teksti ja HTML.* Valitaan ylävalikosta *File, Export data, Quick.* Oletusarvona on tallennus Excel-muotoon. Ennen OK-painallusta napauta *Options*-painiketta ja ruksaa vaihtoehto *Put variable names in the first row*, jolloin Statisticassa määritellyt muuttujien nimet tallentuvat Exceliin ensimmäiselle riville.

2.2.4 Kuvion vieminen Wordiin

Piirretään kuvio Statisticassa. Sen muotoilu ei tarvitse olla lopullinen, vaan kuviota voidaan editoida myöhemmin aktivoimalla Wordissa oleva Statistica-kuvio. Jotta kuvio saataisiin mahdollisimman saman näköisenä siirretyksi tekstinkäsittelyohjelmaan, avataan Statisticassa kuvion piirtämisen jälkeen *File, Page/Output setup* -valinnalla ikkuna ja varmistetaan, että *Screen resolution* on valittuna. Napautetaan vielä advanced -painiketta ja tarkastetaan, että myös täällä *Screen resolution* on voimassa.

Valitaan valikkoriviltä *Edit, Copy* ja siirretään kuva leikepöydälle. Siirrytään Wordiin ja asetetaan kursori oikeaan paikkaan ja valitaan *Edit, Paste special.* Seuraavaksi Word ehdottaa kuvion liitettäväksi tiedostoon Statistica-kuviona, mikä hyväksytään. Kaksoisnapauttamalla kuviota Wordissa saadaan se avautumaan Statisticaan, jolloin voidaan tehdä tarvittavia muutoksia. Wordiin palataan komennolla *File, Exit & Return to Microsoft Word.* Ohjelma pyytää vielä vastaamaan kysymykseen *Update Microsoft Word before proceeding.* Vastaamalla *Yes* muutokset päivittyvät Word-asiakirjassa olevaan Statistica-kuvaan.

2.3 Puuttuvien havaintojen "paikkaus"

Käytännössä kaikki aikasarja-analyysit edellyttävät, että sarjaan ei jää puuttuvia havaintoja. Tutkija valitsee analysoitavan aineiston kannalta aukkojen paikkaamiseksi sopivimmat keinot. Statistican *Time series & Forecasting* -modulissa ovat seuraavat vaihtoehdot:

---

[3] Vrt. 2.2.1 Havaintomatriisin haku Excel-muodosta

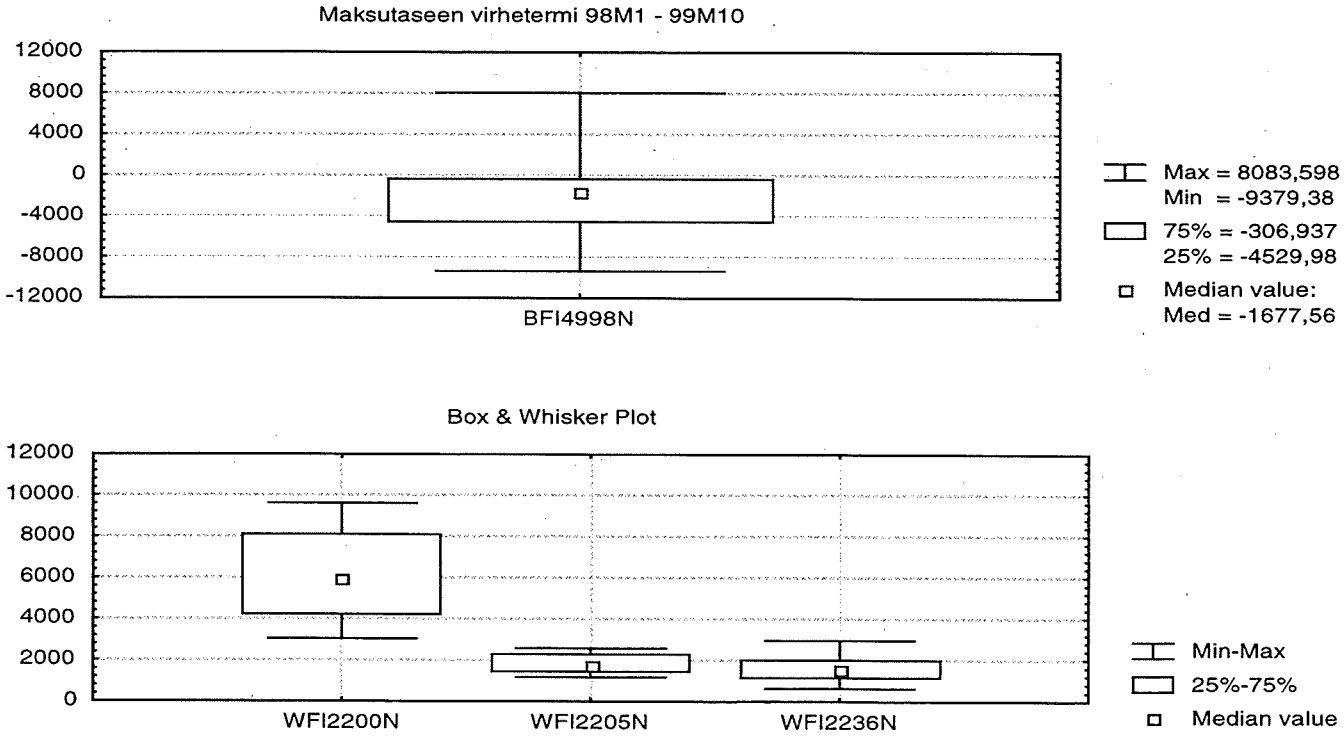
a) lasketaan aikasarjasta keskiarvo
b) interpoloidaan vierekkäisistä havainnoista
c) määritetään N kpl vierekkäisiä havaintoja ja lasketaan niiden keskiarvo tai mediaani
d) estimoidaan lineaarisella mallilla

2.4 Sovellutusesimerkkejä

2.4.1 Viiksikuviot

Ns. viiksikuvioista[4] voidaan selkeästi hahmottaa tietyn sarjajoukon keskinäiset tasoerot. Piirrettäessä vain yhtä sarjaa kuvion oikeaan reunaan tulostuu joitakin tarkkoja tunnuslukuja: maksimi, minimi, ylä- ja alakvartiili sekä mediaani (KUVA 1, ylempi kuvio). Tulostettaessa useita sarjoja samaan kuvaan (alempi kuvio) edellä mainitut tunnusluvut nähdään vain graafisesti. Vastaavat numeroarvot saadaan *More statistics* -vaihtoehdon avulla.

KUVA 1



[4] Ks. Peruskäytön "niksejä": 4.2 Viiksikuvio

2.4.2 Kyselyaineistot

Tässä käytetään esimerkkinä UMAn[5] poikkileikkausaineistoa 'palvelutulot yrityksittäin v. 1997'. Aineistoa voidaan graafisesti tarkastella monella eri tavalla. Histogrammi on ehkä yleisimpiä esitystapoja poikkileikkausaineiston jakauman havainnollistamiseksi[6]. *Data Management* -modulista valitaan vaihtoehto *Graphs/2D Histograms.* Avautuneesta ikkunasta voidaan määrittää sopivaksi katsottava luokkien lukumäärä. Liitteestä 1 (alakuva) nähdään aineiston vinous; vaikka frekvenssihistogrammi -kuvaan on määritelty 20 luokkaa, silti lähes kaikki havainnot asettuvat pienimpään luokkaan.

Jatkoanalyysien vuoksi vinoa/huipukasta jakaumaa usein muokataan sopivalla tavalla lähemmäksi normaalijakaumaa. Transformoiminen tehdään hyvin usein logaritmoinnin avulla[7]. Liitteen 1 ylempi kuvio on piirretty UMAn poikkileikkaushavaintojen logaritmoiduista havainnoista ja alempi kuva saman aineiston perushavainnoista.

2.4.3 Aikasarja-analyysi

Osastolla on parisen vuotta sovellettu aikasarja-analyysin menetelmiä maksutaseen palveluerien (pääerät) tilastoinnin apuna. Aikasarja-analyysin monista menetelmistä sovelletaan myös eksponentiaalisen tasoituksen aikasarjamallia, jossa aikasarjan eri komponentteja, kuten trendiä ja kausivaihtelua tasoitetaan yhtaikaisesti siten, että tuoreimmalla havainnolla on suurin painokerroin pienentyen geometrisen sarjan mukaisesti aikasarjassa taaksepäin (ks. A. Nyströmin artikkeli kirjallisuusvalikoiman julkaisussa Blåfield et al. (1977)). Menetelmällä voidaan laskea haluttu määrä ennusteita, jotka perustuvat aikasarjan havainnoista estimoituun malliin. Liitteessä 2 on esitetty kuukausittain tehtävä rutiiniajo maksutaseen palveluviennistä (WFI2200) ja -tuonnista (WFI3200).

Liitteen 2 esimerkkiajojen tulostuksesta nähdään, miten ohjelma listaa alkuperäisen aikasarjan ohella tasoitetun sarjan ja niiden erotuksen= residuaalin sekä tallentaa nämä uudet sarjat automaattisesti työskentelyn ajaksi (pysyvä tallentaminen mahdollista). Tuloksista voidaan lisäksi helposti saada graafiset esitykset. Lisäksi jäännössarjaa voidaan analysoida kätevästi laskemalla sille jakaumatunnuslukuja, autokorrelaatiokertoimia jne. Jäännössarjalle voidaan myös piirtää ns. normaalisuuskuvio, josta helposti nähdään onko aihetta epäillä alkuperäisessä sarjassa olevan poikkeavia havaintoja, outliereita (ks. Liite 4: Peruskäytön niksit, kohta 4.3).

Statisticalla voidaan laskea aikasarjalle myös robusteja muunnoksia, tasoituksia, joiksi ohjelmapakettiin on valittu tavallinen liukuvan mediaanin tasoitus (termien määrä valittavissa) ja filtteri "4253H", mikä on tiettyjen liukuvien mediaanitasoitusten yhdistelmä.[8]

Aikasarjan puuttuvien havaintojen paikkausta varten Statisticassa on valittavissa eri vaihtoehtoja, kuten vierekkäisten havaintojen keskiarvo ja lukusarjan mediaani. Aikasarjojen mahdollista kausivaihtelua voidaan monipuolisesti tutkia X11 –kausivaihteluproseduurilla.[9]

---

[5] Ulkomaisten maksujen aineisto (laadinta lopetettiin osastolla 31.12.1998)
[6] Liite 1
[7] Ks. Liite 4: Peruskäytön Niksit, kohta 2.2
[8] Sovellettaessa 4253H:ta saadaan tasoitettu arvo myös aikasarjan loppuun tuoreimmille havainnoille; liukuvan keskiarvon tai mediaanin tapauksessa tällaista tasoitusta ei saada.
[9] Edellä mainittu eksponentiaalisen tasoituksen malliversio (ns. Holt-Winters –malli) tulostaa automaattisesti tasoitusprosessin tuloksena syntyneiden kausikomponenttien arvot kullekin perusajanjaksolle esim. kullekin vuosineljänneksel-

3. Statistica ja maksutaseen laadunvalvonta

Maksutaseen (BOP) laadunvalvontatyössä Statisticalla on yksin ja Excel-taulukkolaskentaohjelman kanssa monia potentiaalisia soveltamiskohteita, mitkä voidaan karkeasti jakaa kolmeen ryhmään:

- kuvaileva analyysi: jakaumatunnusluvut, luokittelut jne
- graafiset tarkastelut (frekvenssihistogrammi, viiksikuviot, sirontakuviot jne.)
- aikasarja-analyysin estimoinnit, tasoitukset, muunnokset yms.

Kun kysymyksessä on kyselyaineistojen ja aikasarjojen tilastollinen laadunvalvonta, tilastotieteen ns. robusteilla menetelmillä on aivan erityinen soveltamisala poikkeavien ja puuttuvien havaintojen etsimisessä ja käsittelyssä (esim. editing ja imputation, ks. myös kohta 2.3 edellä) Statistica tarjoaa tähän monia tekniikoita ja proseduureja.

4. Kehittämisnäkymiä

Joustavien Excel-yhteyksien vuoksi Statistican integroitua käyttöä Patun ja pankin muiden tietokantojen välillä on mahdollista kehittää. Ohjelman uusi versio on 32-bittinen, mikä on tärkeä teknisen kytkennän kannalta. Kuten edellä todettiin maksutaseen rahoituskyselyaineistojen laadunvalvonnassa Statisticaa on mahdollista hyödyntää etenkin yhdistettynä Excelin toimintoihin. Näin on tarkoitus tehdä BOP:n rahoituskyselyjen suunnitelluissa rutiiniraporteissa.

Tilasto-osastolla onkin ohjelmoitu joukko Excel-pohjaisia robusteja tilastollisia tunnuslukuja lähinnä BOPn laadunvalvonnan tarpeisiin (ks. liite 3).

Tarvittaessa Statisticalla voidaan toki tehdä vaativampia laadunvalvontaa ja tilastointia palvelevia case-tutkimuksia ja –selvityksiä graafisine yms. osineen.

---

le. Maksutaseen palvelupääerien (WFI2200, WFI3200) vuoden sisäinen vaihtelu nähdään liitteen 2 kausikomponenteista.

Kirjallisuutta tilastotieteen perusteista ja tilastollisen selvitys- ja tutkimustyön käytännöstä

1. Statistica-ohjelmapaketin esittely- ja käyttömoniste. Statcon Oy, Salo 1998.

2. Blåfield, E., Leskinen, E. ja Teräsvirta, T. (toim.): Aikasarja-analyysin menetelmiä. Suomen Tilastoseuran julkaisuja 4. Helsinki 1977.

3. T. Heikkilä: Tilastollinen tutkimus. Edita 1998.

4. H. Helenius: Tilastollisten menetelmien perustiedot. Statcon Oy, Salo 1989.

5. M. Holopainen - P. Pulkkinen: Tilastolliset menetelmät. Perusteet. Weilin+Göös, Porvoo 1995.

6. G. K. Kanji: 100 Statistical Tests. SAGE Publications, London 1995.

7. L. Karjalainen - A. Ruuskanen: Tilastomatematiikka. Pii-Kirjat, Jyväskylä 1994.

8. H. Karttunen: Datan käsittely, CSC-Tieteellinen laskenta Oy, Otaniemi 1994.

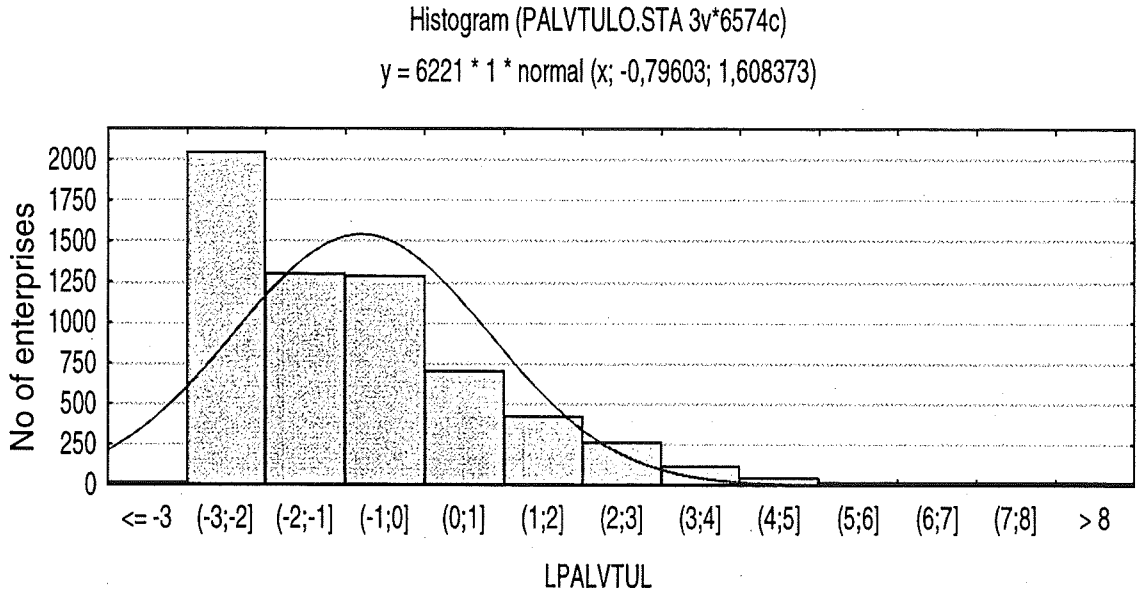
9. S. Mattila: Tilastotiede I – II, Kauppakorkeakoulu, Helsinki 1969

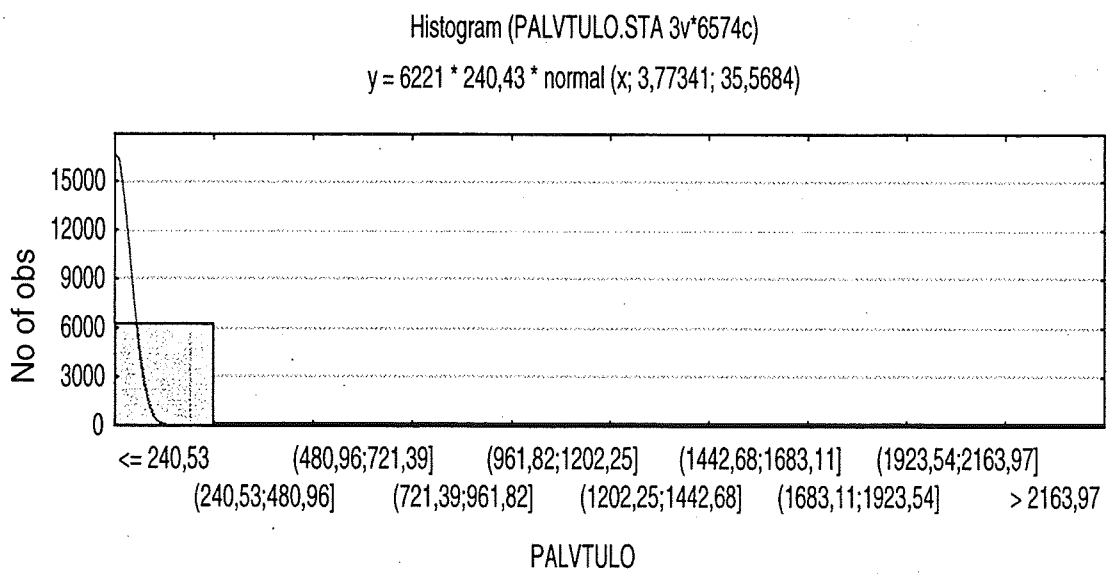
10. H. Niemi - K. Tourunen (toim.): Tilastoista tiedoiksi. Tilastokeskus, Helsinki 1996.

11. L. Törnqvist: Aikasarjojen analyysi ja ennustaminen, toimittanut P. Tavaila. Gaudeamus, Helsinki 1974.

12. Y. Vartia: Tilastotieteen perusteet. Gaudeamus, Helsinki 1989.

LIITE 1


Histogram (PALVTULO.STA 3v*6574c)
y = 6221 * 1 * normal (x; -0,79603; 1,608373)
No of enterprises
2000
1750
1500
1250
1000
750
500
250
0
<= -3
(-3;-2]
(-2;-1]
(-1;0]
(0;1]
(1;2]
(2;3]
(3;4]
(4;5]
(5;6]
(6;7]
(7;8]
> 8
LPALVTUL

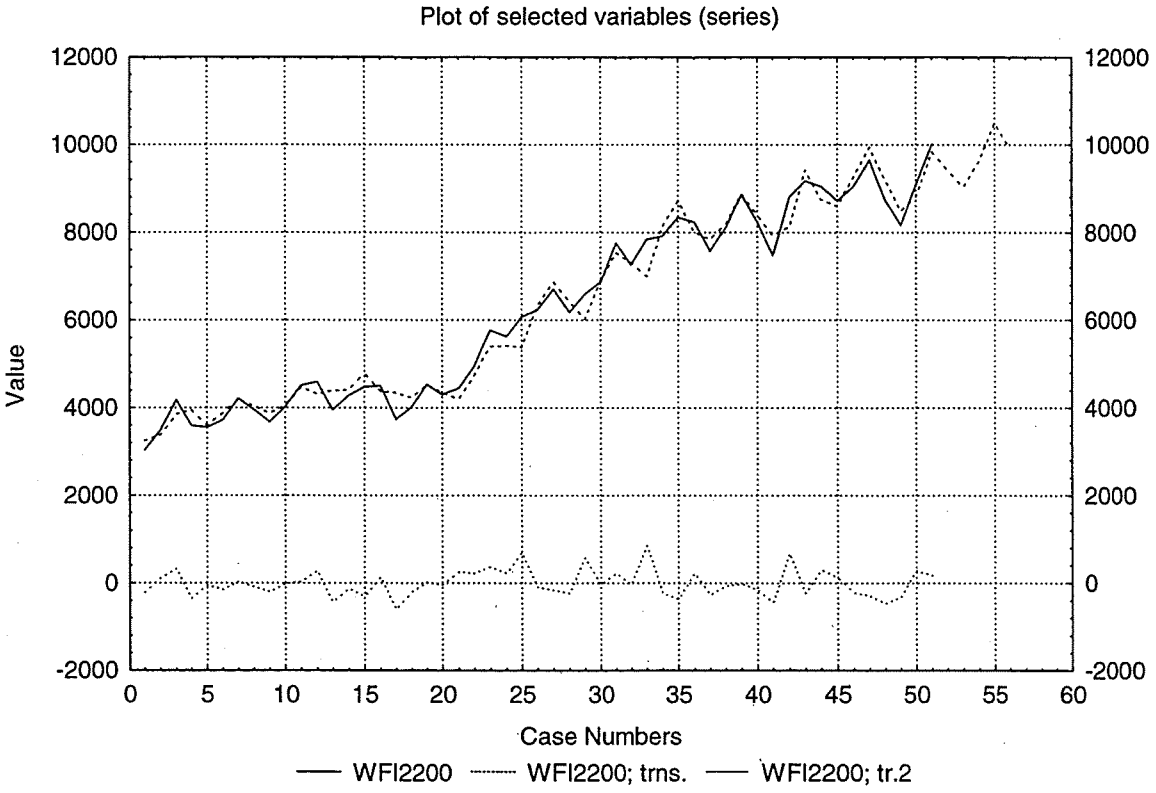

Histogram (PALVTULO.STA 3v*6574c)
y = 6221 * 240,43 * normal (x; 3,77341; 35,5684)
No of obs
15000
12000
9000
6000
3000
0
<= 240,53
(240,53;480,96]
(480,96;721,39]
(721,39;961,82]
(961,82;1202,25]
(1202,25;1442,68]
(1442,68;1683,11]
(1683,11;1923,54]
(1923,54;2163,97]
> 2163,97
PALVTULO

Sarja WFI2200, milj. mk

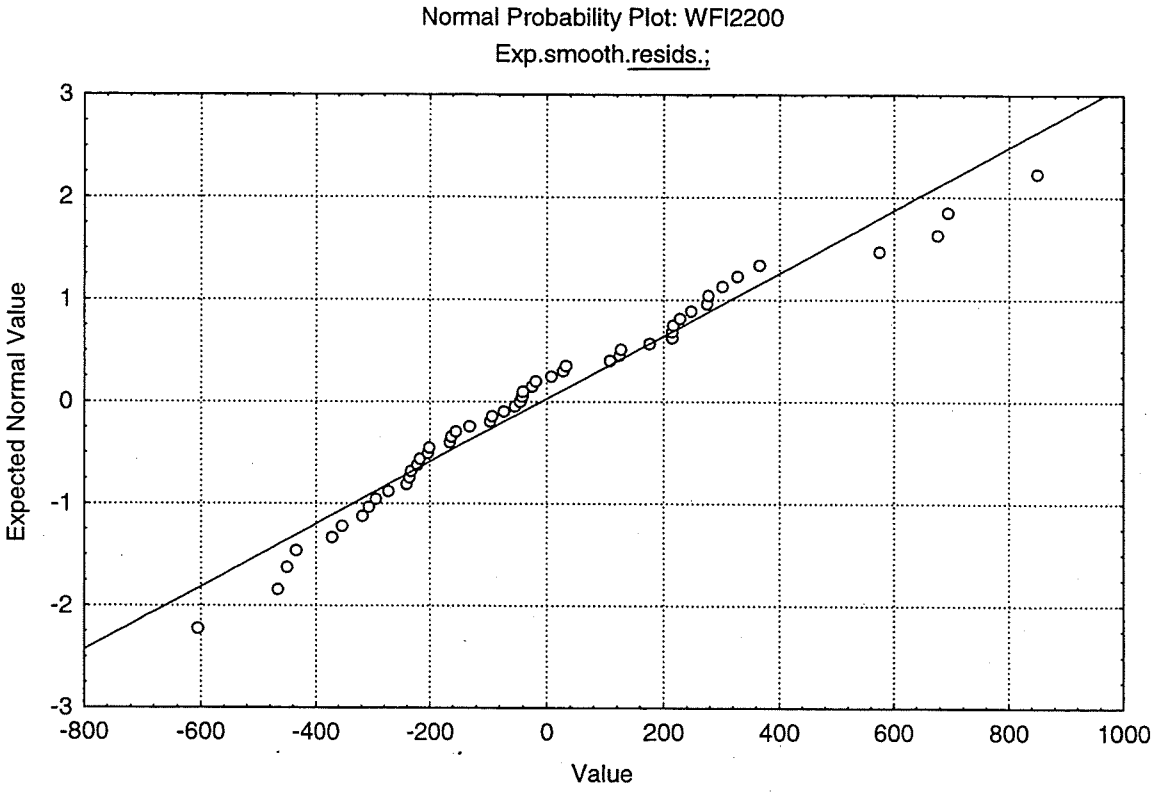
LIITE 2

STAT. TIME SERIES	Exp. smoothing: Multipl. season (4) S0=3328, T0=124,1 Lin.trend, mult.season; Alpha=,735 Delta=0,00 Gamma=0,00 WFI2200			
Case	WFI2200	Smoothed Series	Resids	Seasonal Factors
1	3035,26	3253,05	-217,787	94,2342
2 1987	3489,47	3380,83	108,635	99,2517
3	4185,26	3857,17	328,087	106,8211
4	3594,77	3948,56	-353,789	99,6930
5	3550,15	3603,50	-53,354	
6 1988	3722,97	3877,24	-154,278	
7	4211,88	4183,46	28,418	
8	3954,64	4047,52	-92,879	
9	3674,74	3878,31	-203,567	
10 1989	4025,96	4050,40	-24,435	
11	4505,17	4472,53	32,634	
12	4597,68	4320,19	277,482	
13	3959,23	4393,36	-434,133	
14 1990	4283,42	4414,39	-130,970	
15	4473,07	4780,01	-306,939	
16	4501,35	4374,22	127,134	
17	3735,13	4339,97	-604,846	
18 1991	4024,55	4226,00	-201,455	
19	4529,65	4521,50	8,156	
20	4304,69	4349,10	-44,412	
21	4444,67	4197,05	247,626	
22 1992	4949,91	4735,39	214,518	
23	5764,62	5398,79	365,822	
24	5627,82	5413,20	214,628	
25	6076,16	5382,85	693,314	
26 1993	6233,45	6329,36	-95,902	
27	6706,29	6868,76	-162,469	
28	6181,50	6422,69	-241,191	
29	6595,31	6020,38	574,926	
30 1994	6868,03	6909,19	-41,158	
31	7752,57	7536,12	216,457	
32	7265,16	7305,44	-40,278	
33	7843,94	6994,38	849,554	
34 1995	7913,91	8147,65	-233,738	
35	8345,79	8716,69	-370,893	
36	8232,09	8004,34	227,757	
37	7568,48	7841,23	-272,750	
38 1996	8098,29	8170,77	-72,484	
39	8850,86	8869,13	-18,275	
40	8223,15	8388,49	-165,342	
41	7481,39	7931,24	-449,849	
42 1997	8803,78	8128,48	675,309	
43	9178,62	9415,16	-236,541	
44	9050,24	8748,36	301,878	
45	8721,36	8596,01	125,349	
46 1998	9051,43	9273,92	-222,485	
47	9643,34	9937,75	-294,408	
48	8730,58	9196,39	-465,812	
49	8168,28	8486,15	-317,867	
50 1999	9090,40	8815,10	275,300	
51	10013,70	9837,72	175,983	
52		9425,69 E		
53		9026,52 E		
54		9630,32 E		
55 2000		10497,33 E		
56		9920,58 E		

LIITE 2 (jatkoa)



LIITE 2 (jatkoa)

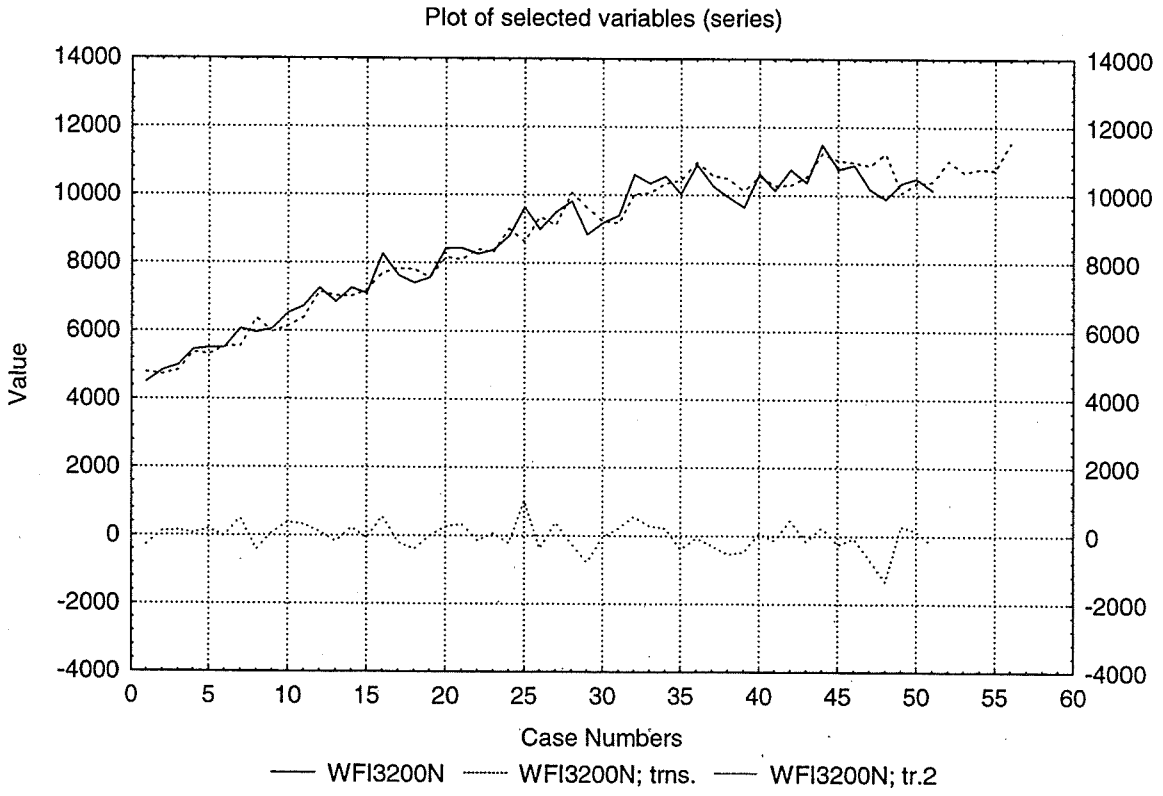


Sarja WFI3200, milj. mk

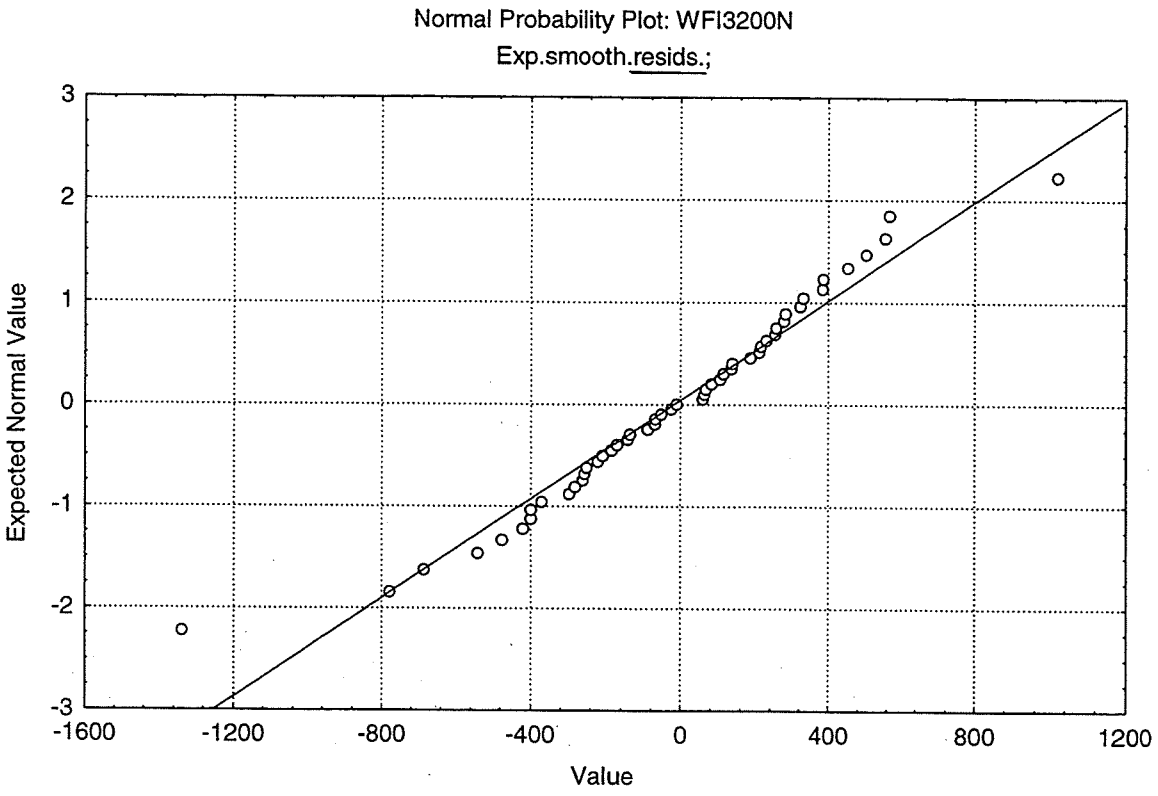
LIITE 2 (jatkoa)

STAT. TIME SERIES	Exp. smoothing: Multipl. season (4) S0=4689, T0=124,7 Lin.trend, mult.season; Alpha=,626 Delta=0,00 Gamma=0,00 WFI3200N			
Case	WFI3200N	Smoothed Series	Resids	Seasonal Factors
1	4503,20	4784,47	-281,27	99,3842
2	4836,17	4720,86	115,31	99,1428
3 /1987	4980,85	4842,31	138,54	97,6427
4	5435,17	5370,90	64,27	103,8304
5	5490,48	5303,38	187,09	
6	5506,41	5531,00	-24,59	
7 /1988	6058,34	5553,94	504,40	
8	5951,25	6371,17	-419,92	
9	6053,62	5970,69	82,93	
10	6516,58	6131,63	384,95	
11 /1989	6720,22	6397,98	322,24	
12	7253,98	7147,44	106,54	
13	6844,96	7029,17	-184,21	
14	7250,45	7020,72	229,73	
15 /1990	7089,21	7177,92	-88,71	
16	8259,35	7703,24	556,11	
17	7623,54	7830,55	-207,01	
18	7407,44	7805,92	-398,48	
19 /1991	7554,33	7563,92	-9,60	
20	8422,97	8166,37	256,60	
21	8424,28	8094,39	329,89	
22	8262,16	8404,40	-142,24	
23 /1992	8371,10	8311,33	59,77	
24	8787,59	9007,32	-219,73	
25	9633,99	8613,91	1020,07	
26	8983,07	9353,67	-370,60	
27 /1993	9491,75	9105,44	386,31	
28	9812,92	10069,13	-256,21	
29	8830,99	9608,39	-777,40	
30	9171,53	9223,24	-51,72	
31 /1994	9389,23	9173,59	215,64	
32	10595,23	10027,98	567,25	
33	10340,64	10062,42	278,22	
34	10546,31	10335,39	210,92	
35 /1995	10032,01	10430,83	-398,83	
36	10887,37	10955,86	-68,49	
37	10272,62	10569,63	-297,02	
38	9941,81	10482,14	-540,33	
39 /1996	9637,31	10112,20	-474,88	
40	10634,41	10566,41	68,01	
41	10142,66	10278,65	-135,99	
42	10745,80	10292,42	453,38	
43 /1997	10367,16	10538,00	-170,84	
44	11475,65	11221,59	254,06	
45	10756,87	11017,25	-260,38	
46	10881,88	10951,56	-69,68	
47 /1998	10179,67	10864,68	-685,01	
48	9889,40	11226,70	-1337,31	
49	10350,96	10068,61	282,35	
50	10479,54	10344,14	135,40	
51 /1999	10142,59	10392,89 E	-250,30	
52		11014,39 E		
53		10666,69 E		
54		10764,45 E		
55 2000		10723,36 E		
56		11532,41 E		

LIITE 2 (jatkoa)



LIITE 2 (jatkoa)



LIITE 3

Classic and robust statistics to support editing processes

There exists a numerous amount of different kind of statistics of which can be calculated from empirical business survey data. They include as well as robust statistics and non-robust statistics. Many of those statistics are relatively easy to calculate by a computer. In practice, it is very important to know something about the underlying probability distribution of the data. On the other hand, the probability distribution of the data will tell something about possible outliers and with the help of these statistics it can be concluded, whether the data contains exceptional observations. Below are listed some common non-robust and robust statistics which can be applied in statistical quality control:

- Arithmetic Mean
- Standard Deviation
- Pseudo Deviation (*Robust*)
- Lower Quartile
- Quartile Range (*Robust*)
- Coefficient of Variation
- Excess Kurtosis
- Min
- Geometric Mean
- Max – Min
- Trimmed Mean (*Robust*)
- Trimmed Standard Deviation (*Robust*)
- Median (*Robust*)
- Upper Quartile
- MAD (Minimum Absolute Deviation) (*Robust*)
- Skewness
- JB-Test Statistic and its P-value
- Max
- Winsorized Mean (*Robust*)
- Harmonic Mean

Lähde:
Kohta 4.2 julkaisusta: Heikki Hella and Hannu Viertola
"On quality control methods in business surveys
The pilot project of BOP surveys in the Bank of Finland
Report No 1"

Statistics Department
Working Papers 3/99

LIITE 4

PERUSKÄYTÖN NIKSIT

Peruskäytön "niksit"

1. Ohjelmassa liikkuminen

1.1 Liikkuminen modulien välillä

Statistica-sovellus sisältää useita eri tilastollisia ohjelmia eli moduleita, joista jokaisesta on lyhyt kuvaus *Module Switcher*-ikkunan oikeassa reunassa. Yläpalkissa olevasta *Module switcher*-kuvakkeesta klikataan ikkuna auki, ja valitaan haluttu moduli *Switch To* -painikkeen avulla. Useita moduleja voidaan avata käyttöön samanaikaisesti. Kannattaa ottaa tavaksi mennä aina aluksi *Data Management*-moduliin. Sieltä voidaan siirtyä muihin moduleihin valitsemalla *Analysis/Other Statistics*, jolloin *Module Schwitcher* avautuu. Avoinna olevat modulit nähdään näytön alareunassa olevasta ohjelmapalkista, josta haluttu moduli saadaan aktiiviseksi klikkaamalla kyseistä kohtaa. Modulit suljetaan yksitellen *File/Exit* -komennolla.

1.2 Liikkuminen yksittäisen modulin sisällä

Moduliin ensimmäisen kerran tultaessa avautuu automaattisesti kyseiseen moduliin liittyvä alkuvalikko. Muulloin alkuvalikon saamiseksi näytölle klikkaa yläpalkista *Analysis/Resume analysis.* Modulin sisällä liikkuminen on eräänlaista "surfailua". Valitsemalla ensimmäisestä ikkunasta haluttu vaihtoehto usein seuraava(t) ikkuna(t) avautuu automaattisesti jatkomäärityksiä varten ja samalla edelliset ikkunat tuhoutuvat. *Exit-* ja *Cancel*-vaihtoehdoilla voidaan peruuttaa ikkuna ikkunalta lähtöasetelmaan saakka.

2. Tiedoston käsittely

2.1 Muuttujamääritykset

Muuttujien määrityksiä(nimi, desimaalit, kaava, sarakkeen leveys...) voi muuttaa ikkunasta, joka avautuu kaksoisklikkaamalla kyseistä muuttujaa. Lukumuuttujien oletusmääritys on *Number.* Havainnot saadaan nopeasti prosenteiksi klikkaamalla vaihtoehtoa *Percentage.* Excelistä tuoduissa tiedostoissa päiväykset muuttuvat Statisticassa oletusarvon mukaisesti numeroiksi. Ne palautetaan päivämääriksi valitsemalla vaihtoehto *Date.*

2.2 Uuden muuttujan lisääminen

Klikkaa yläpalkista *Vars*-painiketta. Valitse *Add,* jolloin ohjelma ehdottaa muuttujan lisättäväksi sen muuttujan perään, minkä kohdalla kursori sijaitsee. Hyväksymisen jälkeen saadaan uusi muuttuja, jolle annetaan nimi. Havainnot voidaan tuoda kopioinnin avulla, yksitellen syöttämällä tai laskemalla transformaatio tiedostossa jo olevista muuttujista. Kaava määritellään seuraavasti:

- kaksoisklikkaa uutta muuttujaa
- lisää avautuneen ikkunan alareunaan = -merkki.
- klikkaa *Functions*-painiketta, jolloin saadaan esille operaattorilista
- valitse listasta operaattori (esim. *log(x)*)ja paina *Insert*-näppäintä. Operaattori siirtyy = -merkin perään.
- korvaa *x* sen muuttujan numerolla, josta transformaatio tehdään (esim.*log(v2)*) [1]. *OK*-hyväksymisen jälkeen ohjelma varmistaa vielä, että laskulauseke on oikein ja laskee sen jälkeen uudelle muuttujalle arvot.

---

[1] Statistica antaa muuttujille automaattisesti järjestysnumeron *1, 2, 3,.....n,* joihin voidaan viitata muuttujatransformaatioita tehtäessä. Järjestysnumeron eteen lisätään aina kirjain *v,* joka tulee sanasta *Variable.*

2.3 Uusien tapausten lisääminen/poistaminen

Klikkaa yläpalkista *Cases* -painiketta. Valitse *Add,* jolloin ohjelma pyytää määrittämään lisättävien tapausten lukumäärän sekä sen, mihin kohtaan tapaukset halutaan lisätä. Tapausten poistaminen tehdään avaamalla *Delete*-ikkuna.

2.4 Aineiston lajittelu

Aineisto voidaan lajitella tietyn muuttujan mukaan suuruus/aakkosjärjestykseen *Data Management* -modulissa. Klikkaa *Analysis/Sort* ja määritä avautuvaan ikkunaan haluamasi ehdot.

3. Kuvan editointi

Kaksoisklikkauksella (tai hiiren oikealla näppäimellä) saadaan avatuksi ikkunoita aina sen mukaan, mitä kursori osoittaa. Otsikkoa klikkaamalla aukeaa ikkuna, jossa voidaan kuvan otsikko muuttaa, lisätä alaotsikoita ja muuttaa myös Y-akselin otsikko.

Y-akselin skaalaa voidaan muuttaa asettamalla y-akselin kohdasta avatusta ikkunasta manuaalimääritys päälle ja määrittämällä minimi ja maksimi sekä asteikkoväli.

Tiedoston tapaukset (cases) tulostuvat automaattisesti X-akselille. Jos x-akselille halutaan päivämääreet, ne täytyy sinne erikseen määritellä. Eri tilanteissa päiväykset määritellään hieman eri tavalla.

3.1 Aikajänne

3.1.1 Perustilanne kaikissa moduleissa

- Piirrä kuva *Graph*-valikosta
- Valitse muuttuja
- Klikkaa samassa ikkunassa olevaa *Options*-painiketta
- Ruksaa *Case labels* kohdasta *VAR* ja kirjoita sen muuttujan nimi, jonka haluat x-akselille
- *OK*

3.1.2 Time Series/Forecasting -moduli

Modulin keskeinen anti on mahdollistaa erilaisten transformaatioiden tekeminen aikasarjoille, joista tulostuu kuvia. Jos kuviin halutaan paivämääritykset x-akselille, se vaatii tiettyjä toimenpiteitä. Statistica-tiedostossa olevassa aikamuuttujassa on oltava *Number* -määrityksen sijasta ns. *Date*-määritys. Päiväykset saadaan tulostumaan kuvaan vasta *Options*-painikkeen takaa valitsemalla *Date* -kohtaan haluttu aikamuuttuja. Jos muuttuja ei tällöin muutu Statistican tuntemaksi päivämääreeksi, x-akselille tulevat päiväykset eivät ole oikeita. Yleensä Excelistä tuoduissa tiedostoissa ei päivämääreen kanssa ole ongelmaa, mikäli päiväykset ovat ohjelman mukaan oikein luotuja.

3.1.3. Aikajänteen tallennus

Statisticassa voidaan aikamääritykset tallentaa myöhempää käyttöä varten. Tämä voi joissain tapauksissa olla perusteltua, kun halutaan tulostaa useiden erilaisten analyysien tuloksia, joissa on oleellista korostaa tiettyjä tarkkoja aikamäärityksiä. Hiiren oikealla näppäimellä avataan *Scaling style* -ikkuna. Täällä voidaan vielä aikamääreitä editoida manuaalisesti. Kun lopputulos miellyttää, nimetään ja talletetaan se ikkunassa oikealla olevan *Save as*-painikkeen kautta. Päivämääreet tallentuvat *axd*-loppuisiksi tiedostoiksi. Päivämääreet haetaan klikkaamalla samassa ikkunassa olevaa *Open*-painiketta, ja samalla ohjelma tuo kuvaan kaikki kyseiseen tiedostoon tallennetut ominaisuudet.

3.1.4 Aikavälin muuttaminen

Scaling style -ikkunassa voidaan määrittää, mikä on ensimmäinen ja viimeinen kuvioon tulostettava havainto. Aikamääritykset voidaan muuttaa tulostuvaksi kuvaan sopivin välimatkoin (esim. joka toinen kuukausi, puolivuosittain jne...)

3.2 Yhdistelmäkuvan tekeminen

- Klikkaa ylävalikosta *Graph*
- Valitse *Multible Graph layouts*
- *Wizard*
- Avautuu ikkuna, josta nähdään *Add graphs* -vaihtoehdot
- Hae aiemmin talletetut kuvat yksitellen tai yhdessä
- *OK*
- Ohjelmawizard ehdottaa niin monen kuvan yhdistelmämalleja, kuin olet edellä hakenut kuvia
- Valitse mieleisesi
- *OK*

3.3 Usean aikasarjan piirtäminen samaan kuvaan

- Klikkaa ylävalikosta *Graph*
- Valitse esim. *Line plots (variables)*
- Määrityksistä vaihtoehto *Multible*
- Merkitse hiirellä *ctrl*-näppäin alas painettuna halutut aikasarjat *Variables* -painikkeen takaa
- *OK*

4. Havaintojen jakauma

Havaintomatriisin normaalijakautuneisuutta voidaan yksinkertaisesti tutkia *Normal Probability Plots*-kuvion avulla. Jos havainnot noudattavat jokseenkin normaalijakaumaviivaa, aineisto on kunnossa jatkoanalyyseja varten.

4.1 Normaalijakaumakuvio (Normal Probability Plots)

- Kuva piirretään *Basic Statistics* -modulissa
- *Descriptive Statistics*
- Valitse piirrettävä muuttuja *variable*-painikkeen alta
- *Normal Probability Plots*

4.2 Viiksikuvio (Box & Whisker Plot)

- Viisikuvio voidaan piirtää esim. *Basic Statistics* -modulissa.
- Analysis
- Desriptive statistics
- Box & Whisker Plot

Viiksikuviossa näkyvät vastaavat tarkat arvot ja lukuisa joukko muitakin tunnuslukuja voidaan tulostaa seuraavasti:

- *Basic Statistics* -moduli
- Analysis
- Desriptive statistics
- *More statistics*

4.3 Poikkeavien havaintojen merkitseminen normaalisuuskuviosta

- Klikkaa yläpalkista *Brushing Tool* -kuvaketta, jolloin kuvaruudulle ilmaantuu *Brushing*-ikkuna.
- Valitse *Lasso* ja hiiren vasenta näppäintä painamalla ympyröi yhtenäisellä viivalla halutut havaintopisteet esim. normaalijakaumakuvasta
- Valitse *Layouts*-valikosta vaihtoehto *Edit data.*
- Ruudulle tulostuu kuvan taustalla oleva data, jossa "lassolla" merkityt havainnot näkyvät punaisina.
- Mikäli toteat, että punaisella merkitty havainto on virheellinen, se on mahdollista tässä yhteydessä poistaa. Piirrä *Redraw*-painikkeella uusi kuva muuttuneesta tilanteesta. Havainnon/havaintojen poistaminen ei vaikuta perusaineistoon, mutta mahdollista myöhempää käyttöä varten voidaan tallentaa editoitu havaintoaineisto tietokantaan erillisenä tiedostona.

4.4 Poikkeavien havaintojen tutkiminen residuaalien avulla

Regressioanalyysin yhteydessä voidaan selittäjien joukossa mahdollisesti olevia poikkeavia havaintoja tutkia residuaalien kautta.

- Valitse *Multible Regression* -moduli
- Valitse *Selitettävä* ja *Selittäjä*(t)
- Hyväksymisen jälkeen avautuu ikkuna, josta nähdään regression tulokset
- Klikkaa *Residual analysis*
- *Plots of residuals*
- Valitse esim. *Mahalanobis distances,* jonka avulla nähdään selittäjien joukossa olevat poikkeavat havainnot