

Mat-2.2104 Tilastollisen analyysin perusteet

Tentti 15.5.2007/Virtanen

Kirjoita selvästi jokaiseen koepaperiin alla mainitussa järjestyksessä:

- Mat-2.2104 Tap 15.5.2007
- opiskelijanumero + kirjain
- TEKSTATEN sukunimi ja kaikki etunimet
- koulutusohjelma ja vuosikurssi
- mahdolliset entiset nimet ja koulutusohjelmat
- nimikirjoitus

OHJEITA

- (i) **Tehtäviä on 5 kpl.**
- (ii) **Yhden tehtävistä saa korvata tämän kevään harjoitustyöllä.
Korvattava tehtävä on ilmaistava vastauspaperissa selvästi kokonaislukuna.**
- (iii) **Vastaa lyhyesti ja ytimekkäästi, mutta esitä niin paljon perusteluita, että vastauksestasi saa selville mitä ja miksi olet tehnyt.**
- (iv) **Tentissä saa käyttää laskinta ja Lainisen tai Mellinin kaava- ja taulukko-kokoelmaa.**

1. Komeltasadalta mieheltä ja kolmeltasadalta naiselta kysyttiin mielipidettä siitä, että voittaako Suomen edustaja Hönnä Mäkäräinen ensi kesän Tallinnan laulujuhlakilpailun. Miehistä 169 ja naisista 125 uskoi Mäkäräisen voittoon (Kyllä), miehistä 102 ja naisista 144 ei uskonut Mäkäräisen voittoon (Ei) ja miehistä 29 ja naisista 31 ei ottanut kantaa Mäkäräisen menestykseen (Eios). Kyselyn tekijä halusi tutkia tilastollisesti sitä, että onko miesten ja naisten mielipidejakaumissa eroa. Alla on annettu tähän ongelmaan liittyen Statistix-tulostus.

Statistix 8.1
11:14:01 AM

5/9/2007,

Chi-Square Test

| Case | | Variable | | | |
|--------------------|-------------|---------------|--------|--------|-----|
| | | Ei | Eios | Kyllä | |
| 1 | Observed | 31 | 144 | 125 | 300 |
| | Expected | 30.00 | ?????? | 147.00 | |
| | Cell Chi-Sq | 0.03 | 3.59 | 3.29 | |
| 2 | Observed | 29 | 102 | 169 | 300 |
| | Expected | 30.00 | 123.00 | 147.00 | |
| | Cell Chi-Sq | ???? | 3.59 | 3.29 | |
| | | 60 | 246 | 294 | |
| Overall Chi-Square | | ????? | | | |
| P-Value | | 0.0010 | | | |
| Degrees of Freedom | | ??? | | | |
| Cases Included | 6 | Missing Cases | 0 | | |

Huomautus:

Paholainen halusi estää vastaamisesi ja korvasi osan tulostuksen luvuista kysymysmerkeillä. Paholainen ei kuitenkaan tiennyt, että puuttuvat luvut voidaan laskea jäljelle jääneistä luvuista.

Puuttuvat luvut ovat *havaintojen kokonaislukumäärä*, solun (Naiset, Eios) *odotettu frekvenssi*, solun (Miehet, Ei) χ^2 -arvo, koko frekvenssitaulukon vastaava χ^2 -testisuureen arvo ja vapausasteiden lukumäärä.

Tehtävät:

- (a) Mitä testiä sovellettiin? Kuvaa testiä ja sen käyttöä lyhyesti.
- (b) Laske puuttuvat luvut.
- (c) Tee johtopäätökset tilastollisen analyysin tuloksista. Erosiko naisten ja miesten mielipiteet toisistaan tilastollisesti merkittävästi?
- (d) Eräs toinen testi tehdään teknisesti samaan tapaan kuin tehtävässä sovellettu testi. Mikä tämän toisen testin nimi on ja mitä tässä toisessa testissä testataan?

2. Tallinnan laulujuhlien Suomen delegaation jäsenten keskuudessa syntyi erimielisyyttä siitä, että onko alkoholilla vaikutusta autonkuljettajan reaktioaikaan hätäjarrutustilanteessa. Delegaatio päätti selvittää asian siten, että Mäkäraisen henkilökohtainen assari Jollen Purokangas lähetettiin ostamaan iso pullo alkoholia, delegaatiosta poimittiin satunnaisesti 21 henkilöä ja paikalliselta taksisuharilta vuokrattiin auto. Tämän jälkeen kunkin koehenkilön reaktioaika mitattiin hotelli Virun edessä suoritettussa ajokokeessa sekä ennen alkoholin nauttimista (tulostuksissa ”Ennen”) että painoon verrannollisen alkoholimäärän nauttimisen jälkeen (tulostuksissa ”Jälkeen”).

Delegaation jäsenillä ei ollut yllättäen mitään käsitystä tilastomenetelmistä, joten he kääntyivät asiassa sinun puoleen. Ongelmanasi on testata 5 %:n merkitsevyystasoa käyttäen nollahypoteesia H_0 , jonka mukaan alkoholilla ei ole vaikutusta reaktioaikaan, kun vaihtoehtoisena hypoteesina on, että alkoholilla on vaikutusta reaktioaikaan.

Alla on annettu yllä esitettyyn ongelmaan liittyen neljä Statistix-ohjelman tulostusta.

Tulostus 2.1:

Statistix 8.1
11:56:30 AM

Jarrutusaika, 5/9/2007,

Two-Sample T Tests for Ennen vs Jalkeen

| Variable | Mean | N | SD | SE |
|------------|---------|----|--------|----------|
| Ennen | 0.6967 | 21 | 0.0450 | 9.82E-03 |
| Jalkeen | 0.7433 | 21 | 0.0734 | 0.0160 |
| Difference | -0.0467 | | | |

Null Hypothesis: difference = 0.05
Alternative Hyp: difference <> 0.05

| Assumption | T | DF | P | 95% CI for Difference | |
|-------------------|-------|------|--------|-----------------------|-----------|
| | | | | Lower | Upper |
| Equal Variances | -5.15 | 40 | 0.0000 | -0.0846 | -8.71E-03 |
| Unequal Variances | -5.15 | 33.2 | 0.0000 | -0.0849 | -8.47E-03 |

| Test for Equality of Variances | F | DF | P |
|--------------------------------|------|-------|--------|
| | 2.66 | 20,20 | 0.0170 |

Cases Included 42 Missing Cases 2

Tulostus 2.2:

Statistix 8.1
11:54:33 AM

Jarrutusaika, 5/9/2007,

Paired T Test for Ennen - Jalkeen

Null Hypothesis: difference = 0.05
Alternative Hyp: difference <> 0.05

| | |
|--------------|---------|
| Mean | -0.0467 |
| Std Error | 0.0132 |
| Mean - H0 | -0.0967 |
| Lower 95% CI | -0.1242 |
| Upper 95% CI | -0.0692 |
| T | -7.33 |
| DF | 20 |
| P | 0.0000 |

Cases Included 21 Missing Cases 1

Tulostus 2.3:

Statistix 8.1
12:04:31 PM

Jarrutusaika, 5/9/2007,

Wilcoxon Rank Sum Test for Ennen VS Jalkeen

| Variable | Rank Sum | N | U Stat | Mean Rank |
|----------|----------|----|--------|-----------|
| Ennen | 368.00 | 21 | 137.00 | 17.5 |
| Jalkeen | 535.00 | 21 | 304.00 | 25.5 |
| Total | 903.00 | 42 | | |

Normal Approximation with Corrections for Continuity and Ties 2.094
Two-tailed P-value for Normal Approximation 0.0363

Total number of values that were tied 28
Maximum difference allowed between ties 0.00001

Cases Included 42 Missing Cases 2

Tulostus 2.4:

Statistix 8.1
12:06:53 PM

Jarrutusaika, 5/9/2007,

Wilcoxon Signed Rank Test for Ennen - Jalkeen

Sum of Negative Ranks -181.00
Sum of Positive Ranks 29.000

Exact probability of a result as or more extreme
than the observed ranks (one-tailed p-value) 0.0016

Normal Approximation with Continuity Correction 2.819
Two-tailed P-value for Normal Approximation 0.0048

Total number of values that were tied 15
Number of zero differences dropped 1
Max. diff. allowed between ties 0.00001

Cases Included 20 Missing Cases 2

Tehtävät:

- (a) Tulostuksessa 2.1 on sovellettu t -testiä (josta on kaksi versiota) ja F -testiä. Esittele testit: Kerro mitä on testattu ja mitkä olivat testien tulokset.
- (b) Tulostuksessa 2.2 on sovellettu t -testiä. Esittele testi: Kerro mitä on testattu ja mikä oli testi tulos.
- (c) Vain toinen tulostuksissa 2.1 ja 2.2 sovelletuista t -testeistä sopii tehtävän tilanteeseen. Kumpi? Perustele valintasi.
- (d) Tulostuksessa 2.3 on sovellettu Wilcoxonin rankisummatestiä (Mannin ja Whitneyyn testi). Esittele testi: Kerro mitä on testattu ja mikä oli testin tulos.
- (e) Tulostuksessa 2.4 on sovellettu Wilcoxonin rankitestiä. Esittele testi: Kerro mitä on testattu ja mikä oli testi tulos.
- (f) Vain toinen tulostuksissa 2.3 ja 2.4 sovelletuista testeistä sopii tehtävän tilanteeseen. Kumpi? Perustele valintasi.

3. Tallinnan laulujuhlien Suomen delegaatio halusi selvittää laululavalla käytettävissä olevien PA-laitteistovaihtoehtojen vaikutuksen Hönnä Mäkäräisen äänen möreyteen. Selvitystä varten delegaatio hiipi hämärän turvin laululavalle ja valitsi testattavaksi kaksi mikseriä (Miksi ja Möksi) ja kaksi mikrofonia (Mikki ja Mökki). Mäkäräinen lauloi jokaisella mikseri-mikrofoni-kombinaatiolla saman biisin kolme kertaa, joten jokaisesta kombinaatiosta saatiin kolme Mäkäräisen äänen möreyshavaintoa.

Tulokset kokeesta (Mäkäräisen äänen möreys; GHz) on annettu alla olevassa taulukossa.

| Mäkäräisen äänen möreys (GHz) | | Mikrofoni | |
|-------------------------------|-------|-----------|-------|
| | | Mikki | Mökki |
| Mikseri | Miksi | 30 | 16 |
| | | 26 | 9 |
| | | 16 | 11 |
| | Möksi | 22 | 6 |
| | | 12 | 10 |
| | | 14 | 8 |

120

60

46

35

27

28

22

22

10³8

72

180

Koetulosten perusteella haluttiin siis selvittää millaisia vaikutuksia mikrofonilla ja mikserillä on Mäkräisen äänen möreyteen.

Statistix-tulostus tehdystä tilastollisesta analyysistä on annettu alla.

Huomautus:

Painovirhepaholainen halusi estää vastaamisesi ja korvasi osan tulostuksen luvuista kysymysmerkeillä. Paholainen ei kuitenkaan tiennyt, että osat kyllä määrätä puuttuvat luvut.

Puuttuvat luvut ovat *jäännösneliösumma, kaikkien neliösummien vapausasteet, keskineliövirheet (MS) sekä F-testisuureiden arvot.*

Tulostus 3.1:

| ANALYSIS OF VARIANCE TABLE FOR MAKARAINEN | | | | | |
|---|----|---------|---------|-------|--------|
| SOURCE | DF | SS | MS | F | P |
| MIKSERI (A) | ?? | 108.000 | ??????? | ????? | 0.0678 |
| MIKROFONI (B) | ?? | 300.000 | ??????? | ????? | 0.0079 |
| A*B | ?? | 12.000 | ??????? | ????? | 0.5017 |
| RESIDUAL | ?? | ??????? | ??????? | | |
| TOTAL | ?? | 614.000 | | | |

Tehtävät:

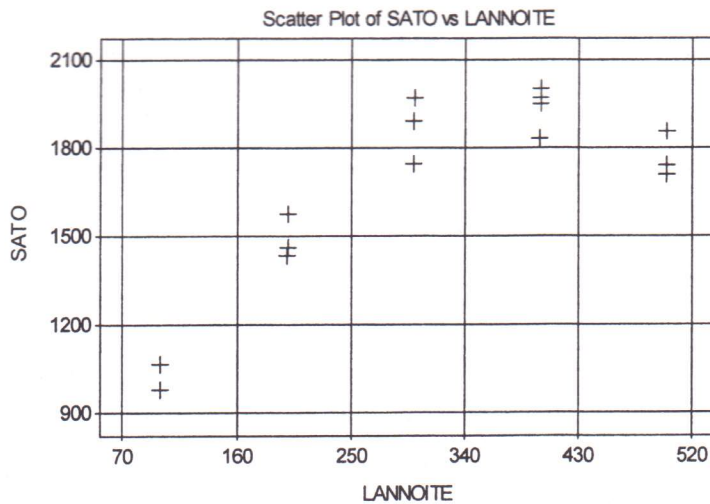
- (a) Mitä tilastollista menetelmää on käytetty?
Kuvaa käytetyn menetelmän tavoitetta lyhyesti.
- (b) Miksi käytetyn menetelmän nimi on harhaanjohtava? Mistä menetelmän nimi johtuu?
- (b) Mitkä ovat menetelmällä testatut nollahypoteesit?
- (c) Laske tulostuksen 3.1 puuttuvat luvut.
- (e) Tee johtopäätökset tulostuksesta 3.1.
- (f) Tehtävässä tarkasteltu koeastelma voitaisiin analysoida myös lineaarisen regressiomallin avulla. Jos näin tehtäisiin, mikä tekijä olisi regressiomallin selitettävä muuttuja ja mitkä tekijät selittäviä muuttujia? Mitä erikoista tällaisen regressiomallin selittäviin muuttujiin liittyy?

4. PA-kaluston testausyön jälkeisenä päivänä Suomen delegaation piti valita Mäkäräisen esiintymisasu. Delegaation jäsenet eivät päässet yksimielisyyteen sopivasta mekosta ja tämän takia Mäkäräisen osallistuminen laulujuhla kilpailuun peruttiin. Tästä suivaantuneena Mäkäräinen päätti lopettaa muusikon uran ja alkaa maanviljeliäksi. Ensi töikseen Mäkäräinen tutki eräässä vehnänviljelykokeessa käytetyn lannoiteaineen määrän (LANNOITE; kg/ha) vaikutusta vehnän satoon (SATO; kg/ha). Kokeessa oli mukana 15 samanlaista peltolohkoa, joille käytetyt lannoiteaineen määrät arvottiin. Lohkoihin kohdistettiin kaikissa muissa suhteissa samanlaiset käsittelyt.

Tutkimustulokset ja sadon riippuvuutta lannoiteaineen määrästä kuvaava pistediagrammi on annettu alla.

| CASE | LANNOITE | SATO |
|------|----------|-----------|
| 1 | 100 | 980.64580 |
| 2 | 100 | 1061.6804 |
| 3 | 200 | 1573.1315 |
| 4 | 200 | 1462.7448 |
| 5 | 200 | 1435.8953 |
| 6 | 300 | 1740.4902 |
| 7 | 300 | 1969.8052 |
| 8 | 300 | 1885.9458 |
| 9 | 400 | 1947.5419 |
| 10 | 400 | 1829.8867 |
| 11 | 400 | 1971.5341 |

| | | |
|----|-----|-----------|
| 12 | 400 | 2002.9903 |
| 13 | 500 | 1851.4186 |
| 14 | 500 | 1740.7020 |
| 15 | 500 | 1706.3952 |



Sadon riippuvuutta käytetyn lannoiteaineen määrästä tutkittiin lineaarisella regressioanalyysillä, jossa muuttujan SATO selittäjinä käytettiin lannoiteaineen määrää (LANNOITE) ja lannoiteaineen määrän neliötä (LANNOITE²) ja vakiota. Estimointitulokset on annettu seuraavalla sivulla.

Huomautus:

Painovirhepaholainen, joka halusi estää vastaamisen, korvasi osan tulostuksen luvuista kysymysmerkeillä.

Onneksi paholainen ei osannut tilastotiedettä ja ei siksi tiennyt, että puuttuvat luvut voidaan helposti laskea jäljelle jääneistä luvuista.

Puuttuvat luvut ovat estimoidun mallin muuttujaa LANNOITE vastaava *t-testisuureen arvo*, *selitysaste*, *mallineliösumma* (regressioneliösumma) ja sitä vastaava *keskineliövirhe* sekä *F-testisuureen arvo*.

Unweighted Least Squares Linear Regression of SATO

| Predictor Variables | Coefficient | Std Error | T | P | VIF |
|---------------------|-------------|-----------|-------|--------|------|
| Constant | 199.687 | 130.603 | 1.53 | 0.1522 | |
| LANNOITE | 9.34623 | 0.95789 | 9.76 | 0.0000 | 30.9 |
| LANNOITE2 | -0.01244 | 0.00153 | -8.12 | 0.0000 | 30.9 |

| | | | |
|--------------------|--------|--------------------------|---------|
| R-Squared | ?????? | Resid. Mean Square (MSE) | 8224.18 |
| Adjusted R-Squared | 0.9199 | Standard Deviation | 90.6873 |

| Source | DF | SS | MS | F | P |
|------------|----|---------|--------|-------|--------|
| Regression | 2 | ?????? | ?????? | ????? | 0.0000 |
| Residual | 12 | 98690 | 8224 | | |
| Total | 14 | 1436706 | | | |

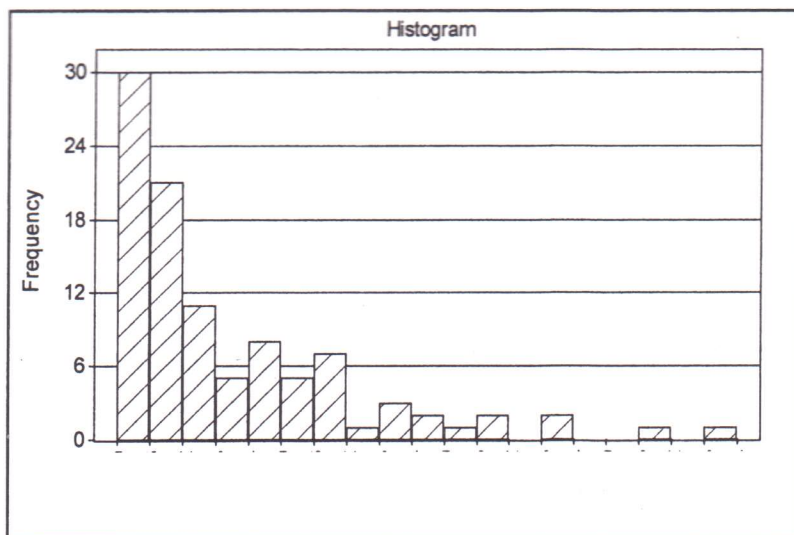
Cases Included 15 Missing Cases 0

Tehtävät:

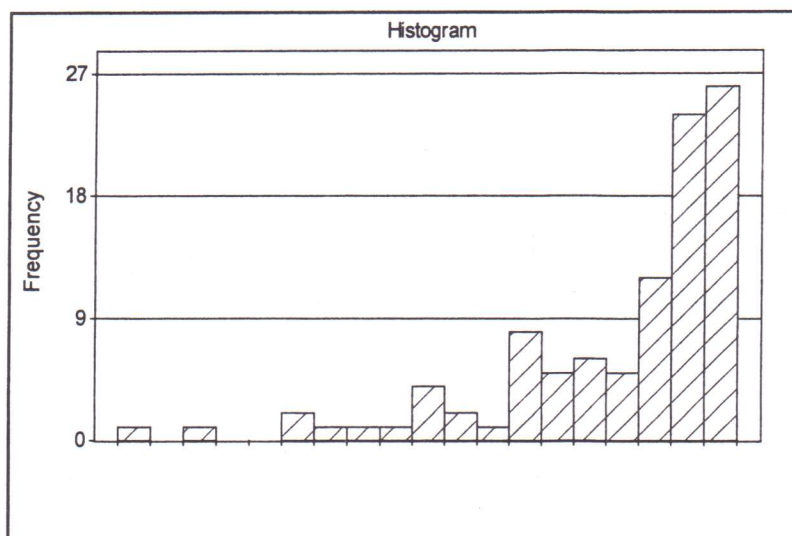
- Mistä Mäkäräinen keksi ottaa mallin selittäjäksi lannoitteen määrän lisäksi myös lainnoitteen määrän neliön?
- Laske tulostuksesta puuttuvat luvut.
- Missä tulostuksessa on esitetty estimoidun mallin varianssianalyysi-hajotelma? Esitä myös hajotelman tulkinta.
- Mitä tarkoittavat tulostuksessa esiintyvät VIF-luvut?
- Ovatko kaikki mallin regressiokertoimet merkitseviä 1 %:n merkitsevyystasolla?
- Mikä on suureiden R-SQUARED ja ADJUSTED R-SQUARED ero? Mitä johtopäätöksiä voit tehdä tulostuksen F-testistä?

5.1. Alla on esitetty histogrammit kahdesta aineistosta:

Kuvio 1



Kuvio 2



Histogrammi 1 on vino *oikealle*, Histogrammi 2 on vino *vasemmalle*.

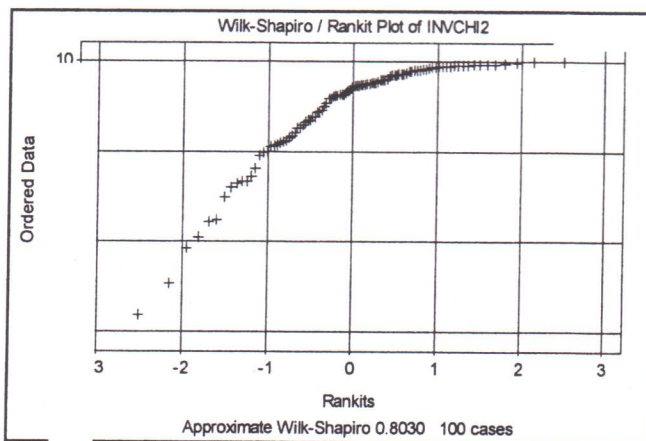
- (a) Aineistoista laskettiin myös niiden aritmeettiset keskiarvot, mediaanit ja vinoudet. Tulokset on annettu alla olevassa taulukossa.

| STATISTIX FOR WINDOWS | | | |
|------------------------|--------|--------|---------|
| DESCRIPTIVE STATISTICS | | | |
| VARIABLE | MEAN | MEDIAN | SKEW |
| A | 2.0878 | 1.1429 | 1.7359 |
| B | 7.9122 | 8.8571 | -1.7359 |

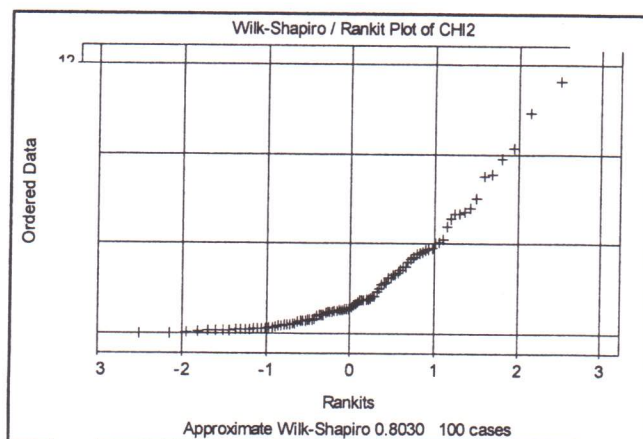
Tehtävä: Päätele tulostuksista kumpi muuttujista A ja B liittyy Kuvion 1 histogrammiin ja kumpi Kuvion 2 histogrammiin.

- (b) Aineistoista muodostettiin myös ns. Rankit Plot –kuviot:

Kuvio 3



Kuvio 4



Tehtävä: Kumpi kuvioista liittyy Kuvion 1 histogrammiin ja kumpi Kuvion 2 histogrammiin?

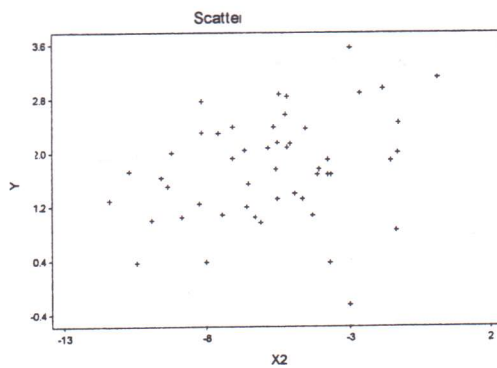
5.2. Alla on esitetty neljä kahden muuttujan pistediagrammia, joihin liittyvät otoskorrelaatiokertoimet ovat *umpimähkäisessä järjestyksessä*

-0.79, -0.003, 0.95, 0.29

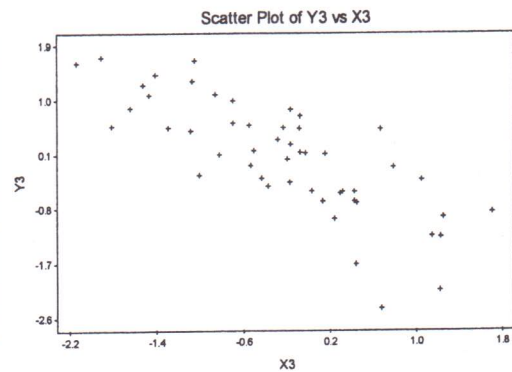
Tehtävä:

Liitä yo. korrelaatiot kuvioiden 5 – 8 aineistoihin.

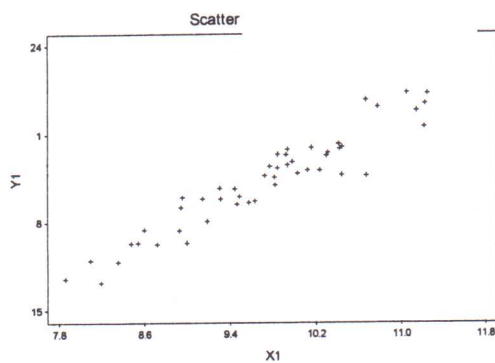
Kuvio 5



Kuvio 6



Kuvio 7



Kuvio 8

