MS-C2104 Tilastollisen analyysin perusteet

Tentti 3.9.2014/Virtanen

Kirjoita selvästi jokaiseen koepaperiin alla mainitussa järjestyksessä:

- MS-C2104 Tap 3.9.2014
- opiskelijanumero + kirjain
- TEKSTATEN sukunimi ja kaikki etunimet
- koulutusohjelma/tutkinto-ohjelma/kandidaattiohjelma ja vuosikurssi
- mahdolliset entiset nimet ja koulutusohjelmat
- nimikirjoitus

OHJEITA

(i) **Tehtäviä on 5 kpl.**

(ii) **Yhden tehtävistä saa korvata kevään 2014 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 taulukkokokoelmaa.**

1. Kokeessa verrattiin kahden tulostimen, A ja B, tulostusnopeuksia tulostamalla molemmalla samat 14 tehtävää. Tulokset kokeesta (kunkin tehtävän tulostusaika tunteina) on annettu alla.

	Tulostin	
Tehtävä	A	B
1	73	68
2	56	73
3	95	89
4	64	73
5	68	66
6	94	87
7	55	75

	Tulostin	
Tehtävä	A	B
8	84	88
9	73	85
10	92	96
11	99	91
12	68	86
13	44	59
14	53	67

Ongelmanasi on testata 5 %:n merkitsevyystasoa käyttäen nollahypoteesia H_0, jonka mukaan tulostimien A ja B tulostusnopeudet ovat yhtä suuria, kun vaihtoehtoisena hypoteesina on, että tulostusnopeudet eivät ole yhtä suuria.

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

Tulostus 1.1:

```
TWO-SAMPLE T TESTS FOR A VS B

                          SAMPLE
VARIABLE        MEAN      SIZE        S.D.         S.E.
---------   ---------   ------   ---------   ---------
A             72.714       14       17.687      4.7271
B             78.786       14       11.383      3.0422
DIFFERENCE   -6.0714

NULL HYPOTHESIS: DIFFERENCE =  0
ALTERNATIVE HYP: DIFFERENCE <> 0

ASSUMPTION          T     DF    P     95% CI FOR DIFFERENCE
----------------- -----  ---- -----  --------------------
EQUAL VARIANCES   -1.08 26   0.290    (-17.626, 5.4835)
UNEQUAL VARIANCES -1.08 22.2 0.292    (-17.724, 5.5808)

                     F    NUM DF DEN DF  P
TESTS FOR EQUALITY -----  ------ ------ ------
      OF VARIANCES  2.41      13     13 0.0624

CASES INCLUDED 28    MISSING CASES 0
```

Tulostus 1.2:

```
PAIRED T TEST FOR A - B

NULL HYPOTHESIS: DIFFERENCE =  0
ALTERNATIVE HYP: DIFFERENCE <> 0

MEAN         -6.0714
STD ERROR     2.7265
LO 95% CI    -11.962
UP 95% CI    -0.1812
T              -2.23
DF                13
P             0.0543

CASES INCLUDED 14     MISSING CASES 0
```

Tehtävät:

(a) Tulostuksessa 1.1 on sovellettu *t*-testiä (josta on kaksi versiota) ja *F*-testiä. Esittele testit: Kerro mitä on testattu ja mitkä olivat testien tulokset.

(b) Tulostuksessa 1.2 on sovellettu *t*-testiä.
Esittele testi: Kerro mitä on testattu ja mikä oli testi tulos.

(c) Vain toinen tulostuksissa 1 ja 2 sovelletuista *t*-testeistä sopii tehtävän tilanteeseen. Kumpi? Perustele valintasi.

(d) Tarkastellaan tulostuksessa 1.2 sovellettua *t*-testiä ja oletaan, että vaihtoehtoinen hypoteesi on "tulostimen A tulostusnopeus on suurempi kuin tulostimen B tulostusnopeus". Mikä on testin tulos nyt? Käytettävä riskitaso on edelleen 5%.

(e) Tarkastellaan tulostuksessa 1.2 sovellettua *t*-testiä ja oletaan, että vaihtoehtoinen hypoteesi on "tulostimen B tulostusnopeus on suurempi kuin tulostimen A tulostusnopeus". Mikä on testin tulos nyt? Käytettävä riskitaso on edelleen 5%.

(f) Jos tehtävänä olisi ollut tulostusnopeuksien mediaanien vertaaminen, niin mitä testiä olisit käyttänyt?

2. Helsingin kaupungin puhtaanapitolaitoksen puhdistaja halusi poistaa lokit kauppatorilta. Puhdistajalla oli käytössään neljää erilaista myrkkyä lokkien likvidointiin. Myrkkyjen toimivuuden testaamiseksi puhdistaja nappasi torilta kiinni 20 lokkia. Tämän jälkeen puhdistaja jakoi lokit viiden hengen ryhmiin ja juotti kullekin ryhmälle yhtä myrkkylaatua. Yhteenveto koetuloksista (lokin elinikä millisekunneisssa myrkyn nauttimisen jälkeen) on annettu alla olevassa taulukossa.

MYR1	MYR2	MYR3	MYR4
70.6	70.3	67.7	62.4
68.4	67.6	68.9	63
71.8	68.4	63.8	64.3
71.4	69.4	64.9	65.1
67	70.0	66.3	65.0

Koetulosten perusteella haluttiin siis selvittää onko myrkkylaadulla vaikutusta lokkien elinikään.

Statistix-tulostukset tehdystä tilastollisesta analyysista on annettu seuraavalla sivulla.

Huomautus:

Eräs viisaampi lokki halusi estää vastaamisesi ja korvasi osan tulostuksen 2.1 luvuista kysymysmerkeillä.

Lokki ei kuitenkaan tiennyt, että osaat kyllä määrätä puuttuvat luvut.

Puuttuvat luvut ovat *ryhmien sisäistä vaihtelua kuvaava neliösumma*, *kaikkien neliösummien vapausasteet*, *keskineliövirheet* (MS) sekä *F-testisuureen arvo*.

Tulostus 2.1:

```
ONE-WAY AOV FOR: MYR1 MYR2 MYR3 MYR4

SOURCE     DF     SS          MS        F       P
-------   ---  ---------   -------   -----   ------
BETWEEN    ??    94.7299   ???????   ?????   0.0017
WITHIN     ??    ???????   ???????
TOTAL      ??    134.589

                      CHI-SQ     DF       P
BARTLETT'S TEST OF    ------   ------   ------
   EQUAL VARIANCES      2.18        3    0.5349

COCHRAN'S Q                        0.4993
LARGEST VAR / SMALLEST VAR         5.1429

COMPONENT OF VARIANCE FOR BETWEEN GROUPS     7.13811
EFFECTIVE CELL SIZE                             4.0

                          SAMPLE      GROUP
VARIABLE        MEAN       SIZE     STD DEV
---------   ----------   ------   ----------
MYR1          69.840         5       2.0611
MYR2          68.925         5       1.1758
MYR3          66.800         5       2.6665
MYR4          63.700         5       1.2247
TOTAL          67.506        20        1.8225

CASES INCLUDED 20    MISSING CASES 0
```

Tulostus 2.2:

```
BONFERRONI COMPARISON OF MEANS

                          HOMOGENEOUS
VARIABLE        MEAN      GROUPS
---------   ----------   ----------
MYR1          69.840     I
MYR2          68.925     I
MYR3          66.800     I I
MYR4          63.700    .. I

THERE ARE 2 GROUPS IN WHICH THE MEANS ARE
NOT SIGNIFICANTLY DIFFERENT FROM ONE ANOTHER.

CRITICAL T VALUE 3.153   REJECTION LEVEL 0.050
STANDARD ERRORS AND CRITICAL VALUES OF DIFFERENCES
VARY BETWEEN COMPARISONS BECAUSE OF UNEQUAL SAMPLE
SIZES.
```

Tehtävät:

(a) Mitä tilastollista menetelmää on käytetty?
Mistä menetelmän nimi johtuu ja miksi nimi on hassu?

(b) Mikä on menetelmällä testattu nollahypoteesi?
Mikä on vaihtoehtoinen hypoteesi?

(c) Mikä on tulostuksessa 2.1 mainitun Bartlettin testin rooli menetelmän soveltamisessa.

(d) Laske tulostuksen 2.1 puuttuvat luvut.

(e) Tee johtopäätökset tulostuksesta 2.1.

(f) Tee johtopäätökset tulostuksesta 2.2.

3. Erästä tappavaa tautia vastaan on kehitetty rokote. Rokotuksen tehon selvittämiseksi järjestettiin seuraava rokotuskoe. Kokeen kohteiksi valitut henkilöt jaettiin satunnaisesti kahteen ryhmään:

Ryhmä 1 (CASE = 1): Rokotetut

Ryhmä 2 (CASE = 2): Ei-rokotetut

Kokeessa rekisteröitiin rokotusta seuranneen vuoden aikana sairastuneiden ja ei-sairastuneiden lukumäärät.

Kokeen tulokset on annettu alla olevassa 2×2-frekvenssitaulukossa.

	VARIABLE	
CASE	SAIRASTUI	TERVE
1	8	42
2	20	30

Kokeen tekijät halusivat tutkia tilastollisesti ovatko rokotus ja sairastuminen riippumattomia tekijöitä. Tulokset tehdystä tilastollisesta analyysista on annettu tehtävän alla.

Huomautus:

Painovirhepaholainen 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 (CASE = 1, TERVE) *odotettu frekvenssi*, *solun* (CASE = 2, TERVE) χ^2*-arvo*, koko frekvenssitaulua 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.
Olisitko halukas suosittelisitko rokotusta analyysituloksen perusteella?
Pohdi asiaa siinä valossa, että ko. tauti on vakava.

(d) Eräs toinen testi tehdään teknisesti samaan tapaan kuin tehtävässä sovellettu testi. Mikä tämän toisen testin nimi on ja mitä tässä toisessa testissä testataan?

```
Statistix 8.1                                                    5/4/2013,
12:40:54 PM

Chi-Square Test for Heterogeneity or Independence

                          Variable
Case                   sairastui     terve
                  +-----------+-----------+
1       Observed  |      8    |     42    |      50
        Expected  |    14.00  |    ?????  |
    Cell Chi-Sq   |     2.57  |     1.00  |
                  +-----------+-----------+
2       Observed  |     20    |     30    |      50
        Expected  |    14.00  |    36.00  |
    Cell Chi-Sq   |     2.57  |     ????  |
                  +-----------+-----------+
                       28          72          ???

Overall Chi-Square      ????
P-Value                0.0075
Degrees of Freedom          ?

Cases Included 4    Missing Cases 0
```

4. STATISTIX-tiedostossa CITYDAT on seuraavat muuttujat:

HSEVAL = Omakotitalojen hintojen keskiarvo

SIZEHSE = Talojen mediaanikoko

TAXRATE = Kiinteistöverosuhde

TOTEXP = Kunnallispalveluihin käytetty rahamäärä

COMPER = Vuokratalojen osuus

Aineisto koostuu 90 USA:n kuntaa koskevista tiedoista.

Havainnoista on estimoitu lineaarinen regressiomalli

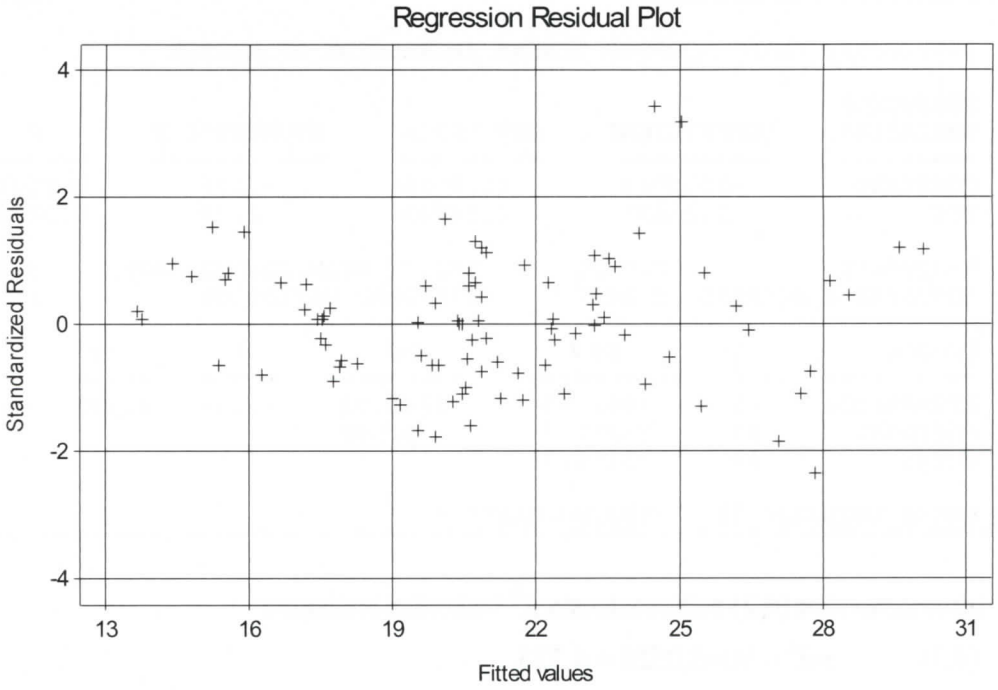
(4.1) $$HSEVAL = \beta_0 + \beta_1 SIZEHSE + \beta_2 TAXRATE + \beta_3 TOTEXP + \beta_4 COMPER + \varepsilon$$

Mallin tavoitteena on selvittää erilaisten taustatekijöiden vaikutus omakotitalojen keskimääräiseen hintaan.

Estimointitulokset mallista (4.1) on annettu alla:

```
STATISTIX FOR WINDOWS                                          CITYDAT

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF HSEVAL

PREDICTOR
VARIABLES     COEFFICIENT     STD ERROR      STUDENT'S T       P        VIF
---------     -----------     ---------      -----------     ------     -----
CONSTANT        -23.4328        8.98561         -2.61        0.0108
SIZEHSE          9.21014        1.56419          5.89        0.0000       1.1
TAXRATE         -177.534        39.8668         -4.45        0.0000       1.0
TOTEXP        1.423E-06      2.963E-07          4.80        0.0000       1.1
COMPER          -20.3704        6.19937         -3.29        0.0015       1.2

R-SQUARED           0.5505      RESID. MEAN SQUARE (MSE)     11.5623
ADJUSTED R-SQUARED  0.5294      STANDARD DEVIATION           3.40033

SOURCE          DF        SS           MS          F         P
----------     ---    ----------   ----------    -----    ------
REGRESSION      4      1203.84       300.960     26.03    0.0000
RESIDUAL       85      982.792       11.5623
TOTAL          89      2186.63

CASES INCLUDED 90   MISSING CASES 0
```

Kuva alla esittää estimoidun mallin (4.1) standardoituja residuaaleja:



Residuaaleihin on sovitettu *apuregressio*

(4.2) $$e_j^2 = \alpha_0 + \alpha_1 \hat{y}_j + \delta_j$$

jossa

e_j = estimoidun mallin residuaali

$\hat{y}_j$ = estimoidun mallin sovite

Estimointitulokset *apuregressiosta* (4.2) on annettu seuraavalla sivulla.

```
STATISTIX FOR WINDOWS                                          CITYDAT

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF RESSQR

PREDICTOR
VARIABLES     COEFFICIENT     STD ERROR      STUDENT'S T        P
---------     -----------     ---------      -----------     ------
CONSTANT        -15.0943        11.9630          -1.26       0.2104
FIT              1.23697        0.56043           2.21       0.0299

R-SQUARED           0.0525     RESID. MEAN SQUARE (MSE)     378.099
ADJUSTED R-SQUARED  0.0417     STANDARD DEVIATION           19.4448

SOURCE          DF       SS           MS          F         P
----------     ---   ----------   ----------   -----   ------
REGRESSION      1      1841.99      1841.99     4.87    0.0299
RESIDUAL       88      33272.7      378.099
TOTAL          89      35114.7

CASES INCLUDED 90   MISSING CASES 0
```

Apuregression (4.2) selitysasteesta R^2 laskettiin testisuure

(4.3) $nR^2 = 90 \times 0.0525 = 4.725$

jossa on n on havaintojen lukumäärä.

Erään nollahypoteesin pätiessä

$$nR^2 \sim \chi^2(1).$$

Testisuureen (4.3) arvoa 4.725 vastaava p-arvo on 0.02973.

Tehtävät:

(a) Ovatko kaikki mallin (4.1) regressiokertoimet merkitseviä 1 %:n merkitsevyystasolla?

(b) Mikä on estimoidun mallin (4.1) selitysaste? Mitä johtopäätöksiä voit tehdä tulostuksen F-testistä?.

(c) Mikä on suureiden `R-SQUARED` ja `ADJUSTED R-SQUARED` ero?

(d) Onko multikollineaarisuus ollut estimoinnissa ongelma?

(e) Miksi alkuperäisen regressiomallin (4.1) residuaaleihin on sovitettu apuregressio (4.2)?

(f) Mitä nollahypoteesia testisuureella (4.3) on testattu? Mikä on testin tulos?

5. Tutkimuksessa haluttiin selvittää tietokoneen prosessorin nopeuden ja RAM-muistin koon vaikutus laskenta-aikaan monimutkaisissa matemaattisissa laskutoimituksissa.

Kokeeseen valittiin kaksi prosessoria (144 MHz ja 400 MHz) ja kaksi muistikokoa (128 MB ja 256 MB). Sama matemaattinen ohjelma ajettiin jokaisella nopeus-muistikoko-kombinaatiolla kolme kertaa niin, että jokaisesta kombinaatiosta saatiin 3 havaintoa.

Tulokset kokeesta (suoritusajat; 1/1000 s) on annettu alla olevassa tulostuksessa.

Suoritusaika (1/1000 s)		Prosessorin nopeus	
		144 MHz	400 MHz
RAM	128 MB	30	16
		26	9
		16	11
	256 MB	22	6
		12	10
		14	8

Koetulosten perusteella haluttiin selvittää millaisia vaikutuksia prosessorin nopeudella ja RAM-muistin koolla on ko. tehtävän suoritusaikaan.

Statistix-tulostus tehdystä tilastollisesta analyysista on annettu alla.

Huomautus:

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

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

Tulostus 5.1:

```
ANALYSIS OF VARIANCE TABLE FOR AIKA

SOURCE          DF     SS        MS         F        P
------------- ---   -------   --------   ------   ------
RAM (A)         ??   108.000    ???????    ?????   0.0678
PROSNOP (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) Mitkä ovat menetelmällä testatut nollahypoteesit?

(c) Laske tulostuksen 5.1 puuttuvat luvut.

(d) Tee johtopäätökset tulostuksesta 5.1.