MS-C2104 Tilastollisen analyysin perusteet

Tentti 7.4.2014/Virtanen

Kirjoita selvästi jokaiseen koepaperiin alla mainitussa järjestyksessä:

- MS-C2104 TAP 7.4.2014
- 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 2013 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. 20 ranskan opettajaa osallistui neljän viikon kurssille, jonka tavoitteena oli puhetaidon parantaminen. Opettajien puhetaito mitattiin ennen kurssia ja kurssin jälkeen kokeella, jossa maksimipistemääränä oli 36 pistettä.

   Tulokset (koepisteet) kokeista ennen ja jälkeen kurssin on annettu alla (korkeampi pistemäärä osoittaa parempaa puhetaitoa).

CASE	ENNEN	JALKEEN
1	32	34
2	31	31
3	29	32
4	10	15
5	30	33
6	33	35
7	22	24
8	25	28
9	32	29
10	20	24
11	30	34
12	20	24
13	24	27
14	24	25
15	31	30
16	30	33
17	15	19
18	32	34
19	23	26
20	23	25

Haasteenasi on testata 1 %:n merkitsevyystasoa käyttäen nollahypoteesia H_0, jonka mukaan opettajien puhetaito ei ole (keskimäärin) parantunut kurssin aikana, kun vaihtoehtoisena hypoteesina on, että puhetaito on parantunut.

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

Tulostus 1.1:

```
TWO-SAMPLE T TESTS FOR JALKEEN VS ENNEN

                          SAMPLE
VARIABLE        MEAN      SIZE        S.D.        S.E.
---------   ---------   ------   ---------   ---------
JALKEEN       28.100       20      5.4183      1.2116
ENNEN         25.800       20      6.3046      1.4097
DIFFERENCE     2.3000

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

ASSUMPTION             T      DF       P     95% CI FOR DIFFERENCE
-----------------   -----   -----   ------   ---------------------
EQUAL VARIANCES      1.24      38   0.2236     (-1.4630, 6.0630)
UNEQUAL VARIANCES    1.24    37.2   0.2237     (-1.4658, 6.0658)

                          F      NUM DF   DEN DF      P
TESTS FOR EQUALITY   -------   ------   ------   ------
      OF VARIANCES      1.35       19       19   0.2577

CASES INCLUDED 40     MISSING CASES 0
```

Tulostus 1.2:

```
PAIRED T TEST FOR JALKEEN - ENNEN

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

MEAN          2.3000
STD ERROR     0.4236
LO 95% CI     1.4133
UP 95% CI     3.1867
T               5.43
DF                19
P             0.0000

CASES INCLUDED 20     MISSING CASES 0
```

Tehtävät:

(a) Tulostuksessa 1.1 on sovellettu *t*-testiä (josta on esitetty 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 testin tulos.

(c) Vain toinen tulostuksissa 1.1 ja 1.2 sovelletuista *t*-testeistä sopii tehtävän tilanteeseen. Kumpi? Perustele valintasi.

2. Kokeessa verrattiin kolmean automerkin A, B ja C bensiinin kulutusta.

Koejärjestely oli seuraava: 20 koeajajaa jaettiin satunnaisesti kolmeen ryhmään siten, että ajajista 7 sai ajettavakseen merkin A auton, ajajista 7 sai ajettavakseen merkin B auton ja ajajista 6 sai ajettavakseen merkin C auton. Kaikilla autoilla ajettiin sama matka pyrkien käyttämään samaa nopeutta ja autojen bensiinin kulutukset (maili/gallona) rekisteröitiin.

Tulokset kokeesta on annettu alla.

A	B	C
22.2	24.6	22.7
19.9	23.1	21.9
20.3	22	23.2
21.4	23.5	24.1
21.2	23.6	22.1
21	22.1	23.4
20.3	23.5	M

Koetulosten perusteella haluttiin selvittää onko automerkillä vaikutusta bensiinin kulutukseen.

Statistix-tulostukset tehdyistä tilastollisista analyyseista on annettu alla.

Huomautus:

Painovirhepaholainen halusi estää vastaamisen ja korvasi osan tulostuksen 2.1 luvuista kysymysmerkeillä.

Paholainen ei kuitenkaan tiennyt, että osaat määrätä puuttuvat luvut.

Puuttuvat luvut ovat *ryhmien välistä vaihtelua kuvaava neliösumma, osa vapausasteista, keskineliöt* (MS) sekä *F-testisuureen arvo*.

Tulostus 2.1:

ONE-WAY AOV FOR: A B C

SOURCE	DF	SS	MS	F	P
BETWEEN	??	???????	???????	?????	0.0002
WITHIN	??	12.1800	???????		
TOTAL	19	33.7295			

	CHI-SQ	DF	P
BARTLETT'S TEST OF EQUAL VARIANCES	0.11	2	0.9464

COCHRAN'S Q 0.3853
LARGEST VAR / SMALLEST VAR 1.3191

COMPONENT OF VARIANCE FOR BETWEEN GROUPS 1.51252
EFFECTIVE CELL SIZE 6.7

VARIABLE	MEAN	SAMPLE SIZE	GROUP STD DEV
A	20.900	7	0.7916
B	23.200	7	0.9092
C	22.900	6	0.8319
TOTAL	22.305	20	0.8464

CASES INCLUDED 20 MISSING CASES 1

Tulostus 2.2:

BONFERRONI COMPARISON OF MEANS

VARIABLE	MEAN	HOMOGENEOUS GROUPS
B	23.200	I
C	22.900	I
A	20.900	.. I

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

CRITICAL T VALUE 2.655 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?
Kuvaa käytetyn menetelmän tavoitteita lyhyesti.

(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. Tutkimuksessa haluttiin verrata viiden erilaisen lannoiteaineseoksen vaikutus viiden maissilajin satoon. Kokeessa jokaista lannoiteaine-maissilajike yhdistelmää (25 kpl) kokeiltiin 6:lla peltoalalla.

Koetulosten perusteella haluttiin selvittää millaisia vaikutuksia lannoiteaineseoksella ja lajikkeella on maissin satoon.

Statistix-tulostus tehdystä tilastollisesta analyysista on annettu alla.

Huomautus:

Painovirhepaholainen halusi estää vastaamisen ja korvasi osan tulostuksen 3.1 luvuista kysymysmerkeillä.

Paholainen ei kuitenkaan tiennyt, että osaat määrätä puuttuvat luvut.

Puuttuvat luvut ovat *jäännösneliösumma, osa vapausasteista, keskineliöistä* (MS) ja *F-testisuureiden arvoista.*

Tulostus 3.1:

ANALYSIS OF VARIANCE TABLE FOR SATO

SOURCE	DF	SS	MS	F	P
LANNOITE (A)	??	807.667	???????	?????	0.0000
LAJIKE (B)	??	3003.47	???????	?????	0.0000
A*B	??	221.200	13.8250	0.84	0.6374
RESIDUAL	125	???????	???????		
TOTAL	149	6087.33			

Tehtävät:

(a) Mitä tilastollista menetelmää on käytetty?
Kuvaa käytetyn menetelmän tavoitteita lyhyesti.

(b) Mitkä ovat menetelmällä testatut nollahypoteesit?

(c) Laske tulostuksen 3.1 puuttuvat luvut.

(e) Tee johtopäätökset tulostuksesta 3.1.

4. Kulutusmenojen tutkimuksessa yksityiset kulutusmenot jaetaan useaan eri osaan, joista yksi on kulutusmenot alkoholiin. Talousteorian mukaan alkoholin kulutus riippuu alkoholin hinnasta ja kokonaiskulutusmenoista.

Alla on estimointitulokset regressiomallista

$$LQ1C_t = \beta_0 + \beta_1 LR1C_t + \beta_1 LQTOTAL_t + \varepsilon_t$$

jossa

LQ1C = Alkoholin kokonaiskulutusmenot (kiinteisiin hintoihin)

LR1C = Alkoholin reaalihintaindeksi

LQTOTAL = Kokonaiskulutusmenot (kiinteisiin hintoihin)

Havaintoina oli Suomea koskevat tiedot vuosilta 1950-1981 (32 vuotta).

Huomautus:

Painovirhepaholainen halusi estää vastaamisen ja korvasi osan tulostuksen 4.1 luvuista kysymysmerkeillä.

Paholainen ei kuitenkaan tiennyt, että osaat määrätä puuttuvat luvut.

Puuttuvat luvut ovat *jäännösneliösumma*, *kaikkien neliösummien vapausasteet* ja *keskineliöt* (MS), *selitysaste* sekä *F-testisuureen arvo*.

Tulostus 4.1:

```
UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF LQ1C

PREDICTOR
VARIABLES   COEFFICIENT   STD ERROR   STUDENT'S T   P        VIF
---------   -----------   ---------   -----------   ------   ----
CONSTANT      -2.47490      2.08702       -1.19     0.2453
LR1C          -1.07549      0.39199       -2.74     0.0103    1.1
LQTOTAL        1.39033      0.05395       25.77     0.0000    1.1

R-SQUARED                   ??????
RESID. MEAN SQUARE (MSE)    0.01116
STANDARD DEVIATION          0.10563

SOURCE       DF        SS         MS         F        P
----------   ---   -------   --------   ------   ------
REGRESSION   ??    9.25113   ???????    ??????   0.0000
RESIDUAL     ??    ???????   ???????
TOTAL        ??    9.57471

CASES INCLUDED 32   MISSING CASES 0
```

Tehtävät:

(a) Mitä tilastollista menetelmää on käytetty?
Kuvaa käytetyn menetelmän tavoitetta lyhyesti.

(b) Laske tulostuksen 4.1 puuttuvat luvut.

(c) Mitä johtopäätöksiä voit tehdä tulostuksen *F*-testistä?.

(d) Mitä johtopäätöksiä voit tehdä tulostuksen *t*-testeistä?

(e) Tulkitse hintamuuttujan LR1C ja LQTOTAL regressiokertoimet.

(f) Onko multikollineaarisuus ollut estimoinnissa ongelma?

5. Tehtävässä 5 tutkitaan tehtävässä 4 estimoidun mallin residuaaleja. Huomaa, että tehtävä 5 voidaan ratkaista, vaikka ei olisi ratkaissut tehtävää 4.

Tehtävät:

(a) Tulostus 5.1 esittää tehtävässä 4 estimoidun regressiomallin residuaalien rankit plot -kuviota. Siihen liittyvä Wilkin ja Shapiron testisuureen arvoa vastaava *p*-arvo on 0.043.

Kerro mitä on testattu ja mitä johtopäätöksiä testin tuloksesta voi tehdä.

(b) Tulostus 5.2 esittää tehtävässä 4 estimoidun regressiomallin residuaaleista määrättyä Durbinin ja Watsonin testisuureen arvoa.

Kerro mitä on testattu ja mitä johtopäätöksiä testin tuloksesta voi tehdä.

(c) Tulostus 5.3 esittää tehtävässä 4 estimoidun regressiomallin residuaaleille estimoitua apuregressiota

$$e_t^2 = \alpha_0 + \alpha_1 \hat{y}_t + \delta_t$$

jossa

e_t = tehtävän 4 estimoidun regressiomallin residuaali

$\hat{y}_t$ = tehtävän 4 estimoidun regressiomallin sovite

Apuregression selitysasteesta R^2 voidaan laskea χ^2-testisuure

$$\chi^2 = nR^2$$

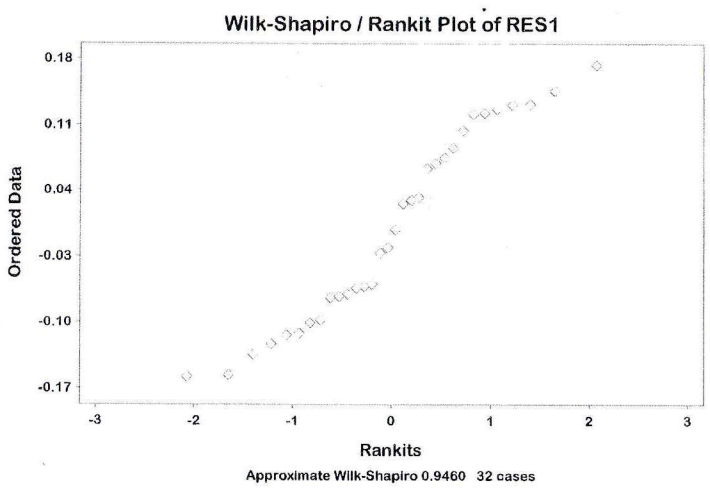
jossa n on apuregression havaintojen lukumäärä. Nollahypoteesin (kerro mikä on nollahypoteesina) pätiessä

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

Tee testi ja kerro mitä tällöin testataan ja mitä johtopäätöksiä testin tuloksesta voidaan tehdä. Vinkki: $P(\chi^2(1) > 32*0.2064)=0.01$.

(d) Mitä sanoisit tehtävän 4 regressiomallin hyvyydestä tämän tehtävän kohdissa (a), (b) ja (c) saatujen tulosten perusteella?

Tulostus 5.1:



Tulostus 5.2:

```
DURBIN-WATSON TEST FOR AUTOCORRELATION

DURBIN-WATSON STATISTIC  0.2367

P-VALUES, USING DURBIN-WATSON'S BETA APPROXIMATION:
P (POSITIVE CORR) = 0.0000,   P (NEGATIVE CORR) = 1.0000

EXPECTED VALUE OF DURBIN-WATSON STATISTIC      2.1043
EXACT VARIANCE OF DURBIN-WATSON STATISTIC      0.11539

CASES INCLUDED 32      MISSING CASES 0
```

Tulostus 5.3:

```
UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF RES1SQRD

PREDICTOR
VARIABLES     COEFF   STD ERROR   STUDENT'S T   P
---------   -------   ---------   -----------   ------
CONSTANT    0.05842     0.01734          3.37   0.0021
FIT1       -0.00679     0.00243         -2.79   0.0090

R-SQUARED    0.2064   RESID. MEAN SQUARE (MSE)   5.460E-05
ADJUSTED R-SQUARED    0.1799
STANDARD DEVIATION    0.00739

SOURCE          DF      SS           MS           F       P
----------      ---   ---------   ---------   -----   ------
REGRESSION       1   4.260E-04   4.260E-04    7.80   0.0090
RESIDUAL        30     0.00164   5.460E-05
TOTAL           31     0.00206

CASES INCLUDED 32   MISSING CASES 0
```