Mat-2.2104 Tilastollisen analyysin perusteet

Virtanen/Tentti 31.8.2006

Kirjoita *selvästi jokaiseen koepaperiin* alla mainitussa järjestyksessä: - Mat-2.2104 Tap/Tentti 31.8.2006 - opiskelijanumero + kirjain, TEKSTATEN sukunimi, kaikki etunimet - koulutusohjelma, vuosikurssi - mahdolliset entiset nimet ja koulutusohjelmat - nimikirjoitus

OHJEITA

(i) **Tehtäviä on 5 kpl.**

(ii) **Osa tehtävien tulostuksista on tuotettu STATISTIX-ohjelmalla.**

(iii) **Vastaa lyhyesti ja ytimekkäästi, mutta esitä perustelut.**

(iv) **Yhden tehtävistä saa korvata tämän vuoden kevään harjoitustyöllä. Korvattava tehtävä on ilmaistava vastauspaperissa selvästi kokonaislukuna.**

(v) **Tentissä saa käyttää laskinta ja Lainisen tai Mellinin kaava- ja taulukkokokoelmaa.**

1. Määrittele lyhyesti seuraavat käsitteet (oikeasta vastauksesta 1p., täysin järjettömästä vastauksesta -1p.):

   a) Luottamusväli.

   b) Vinous ja huipukkuus.

   c) Hylkäysvirhe.

   d) Tilastollisten hypoteesien testauksessa käytettävä **testisuure**. Kysytään siis em. testisuureen määritelmää.

   e) Homoskedastisuus.

   f) Yksisuuntaisen varianssianalyysin nollahypoteesi.

2. Ihmiskehon rasvaprosentin määräämiseen käytetään yleisesti kahta menetelmää, A ja B. Menetelmä A perustuu kehon ominaispainon mittaamiseen ja on tarkka, mutta sitä on työlästä soveltaa. Menetelmä B perustuu kehon sähköisen ominaisvastuksen mittaamiseen ja se ei ole yhtä tarkka kuin menetelmä A, mutta sen soveltaminen on helppoa.

Menetelmien vertaamiseksi rasvaprosentit mitattiin molemmilla menetelmillä 10 koehenkilöltä. Tavoitteena oli selvittää antavatko menetelmät keskimäärin samat tulokset.

Tutkijat X ja Y sovelsivat kerättyyn aineistoon erilaista testausmenetelmää. Tulokset tutkijoiden X ja Y tekemistä testeistä on annettu seuraavalla sivulla.

Tehtävät:

(a) Sekä tutkija X että Y sovelsivat *t*-testiä, mutta testit olivat erilaisia. Kuvaa testejä ja niiden käyttöä lyhyesti.

(b) Toinen tutkijoista X ja Y sovelsi testiä, jota *ei saa* soveltaa tehtävän testausasetelmassa. Kumpi? Perustele valintasi.

(c) Tee johtopäätökset siitä testistä, jonka valitsit (b)-kohdassa.

Tutkija X:

```
STATISTIX FOR WINDOWS                                    RASVAPROS

PAIRED T TEST FOR A - B

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

MEAN          -1.5000
STD ERROR      0.5217
LO 95% CI     -2.6803
UP 95% CI     -0.3197
T               -2.87
DF                  9
P              0.0183

CASES INCLUDED 10     MISSING CASES 0
```

Tutkija Y:

```
STATISTIX FOR WINDOWS                                    RASVAPROS

TWO-SAMPLE T TESTS FOR A VS B

                          SAMPLE
VARIABLE        MEAN      SIZE        S.D.        S.E.
---------    ---------    ------    ---------   ---------
A              17.400      10        6.9634      2.2020
B              18.900      10        6.2619      1.9802
DIFFERENCE    -1.5000

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

ASSUMPTION              T        DF        P      95% CI FOR
DIFFERENCE
-----------------    ------   ------   ------   --------------------
EQUAL VARIANCES      -0.51     18      0.6186     (-7.7217, 4.7217)
UNEQUAL VARIANCES    -0.51     17.8    0.6187     (-7.7267, 4.7267)

                        F      NUM DF   DEN DF      P
TESTS FOR EQUALITY   -------   ------   ------   ------
      OF VARIANCES     1.24      9        9      0.3784

CASES INCLUDED 20     MISSING CASES 0
```

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	16	34

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

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, SAIRASTUI) *odotettu frekvenssi*, *solun* (CASE = 2, SAIRASTUI) 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.

```
STATISTIX FOR WINDOWS

CHI-SQUARE TEST FOR HETEROGENEITY OR INDEPENDENCE

                              VARIABLE
CASE                  SAIRASTUI      TERVE
                     +-----------+-----------+
1        OBSERVED    |     8     |    42     |      50
         EXPECTED    |   ?????   |   38.00   |
     CELL CHI-SQ     |    1.33   |    0.42   |
                     +-----------+-----------+
2        OBSERVED    |    16     |    34     |      50
         EXPECTED    |   12.00   |   38.00   |
     CELL CHI-SQ     |    ????   |    0.42   |
                     +-----------+-----------+
                          24          76          ???

OVERALL CHI-SQUARE      ????
P-VALUE               0.0610
DEGREES OF FREEDOM         ?

CASES INCLUDED 4     MISSING CASES 0
```

4. Tehtaalla on kolme sähkölamppuja valmistavaa konetta, kone A, kone B ja kone C. Koneiden valmistamien lamppujen paloajat vaihtelevat satunnaisesti jonkin verran noudattaen normaalijakaumaa.

Tehtaan laadunvalvontaosasto halusi tutkia koneiden toimintaa ja poimi koneen A, koneen B ja koneen C valmistamien lamppujen joukosta toisistaan riippumattomat yksinkertaiset satunnaisotokset, joiden koko oli 15. Jokaisesta lampusta mitattiin sen paloaika (yksikkö = 1 h).

Otosten perusteella haluttiin selvittää palavatko eri koneiden tekemät lamput keskimäärin yhtä kauan.

Tulostukset tilastollisesta analyyseista on annettu seuraavalla sivulla.

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 *ryhmien välistä vaihtelua kuvaava neliösumma* ja sitä *vastaava keskineliövirhe*, *ryhmien sisäiseen vaihteluun liittyvä vapausasteiden lukumäärä* sekä menetelmässä käytetyn *F-testisuureen arvo*.

Tehtävät:

(a) Mitä tilastollista menetelmää on käytetty?
Kuvaa käytettyä menetelmää lyhyesti.

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

(c) Mikä on tulostuksessa 1 mainitun Bartlettin testin rooli menetelmän soveltamisessa.

(d) Laske tulostuksen 1 puuttuvat luvut.

(e) Tee johtopäätökset tulostuksesta 1.

(f) Tee johtopäätökset tulostuksesta 2.

Tulostus 1

```
STATISTIX FOR WINDOWS

ONE-WAY AOV FOR: KONEA KONEB KONEC

SOURCE     DF        SS          MS         F        P
-------   ----   ---------   ---------   ------   ------
BETWEEN     2     ???????     ??????     ????   0.0136
WITHIN     42     7077204     168505
TOTAL      44     8685820

                      CHI-SQ     DF        P
BARTLETT'S TEST OF   ------   ------   ------
   EQUAL VARIANCES     0.71      2     0.7007

COCHRAN'S Q                     0.4146
LARGEST VAR / SMALLEST VAR      1.5785

COMPONENT OF VARIANCE FOR BETWEEN GROUPS     42386.9
EFFECTIVE CELL SIZE                             15.0

                          SAMPLE      GROUP
VARIABLE        MEAN       SIZE     STD DEV
---------   ----------   ------   ----------
KONEA          10103       15       457.81
KONEB          10009       15       403.91
KONEC          9663.5      15       364.39
TOTAL          9925.3      45       410.49

CASES INCLUDED 45     MISSING CASES 0
```

Tulostus 2

```
BONFERRONI COMPARISON OF MEANS

                          HOMOGENEOUS
VARIABLE        MEAN      GROUPS
---------   ----------   -----------
KONEA          10103     I
KONEB          10009     I I
KONEC         9663.5    .. I

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

CRITICAL T VALUE                        2.494   REJECTION LEVEL   0.050
CRITICAL VALUE FOR COMPARISON          373.78
STANDARD ERROR FOR COMPARISON          149.89
```

5. Kulutusmenojen tutkimuksessa yksityiset kulutusmenot jaetaan useaan eri osaan, joista yksi on kulutusmenot alkoholiin. Talousteorian mukaan kulutus riippuu 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 vastaamisesi ja korvasi osan tulostuksen 5 luvuista kysymysmerkeillä.

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

Puuttuvat luvut ovat *mallineliösumma*, *kaikkien neliösummien vapausasteet*, *keskineliövirheet* (MS), *selitysaste* sekä *F-testisuureen arvo*.

Tulostus 5:

```
UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF LQ1C

PREDICTOR
VARIABLES    COEFF   STD ERROR   STUDENT'S T   P        VIF
---------   ------   ---------   -----------   ------   ---
CONSTANT    -2.475      2.087         -1.19    0.2453
LR1C        -1.075      0.392         -2.74    0.0103   1.1
LQTOTAL      1.390      0.054         25.77    0.0000   1.1

R-SQUARED    ??????    RESID. MEAN SQUARE (MSE)   0.01116
ADJUSTED R-SQUARED     0.9639
STANDARD DEVIATION     0.10563

SOURCE         DF      SS         MS         F        P
----------    ---    -------    -------    ------   ------
REGRESSION     ??    ???????    ???????    ??????   0.0000
RESIDUAL       ??    0.32358    ???????
TOTAL          ??    9.57471

CASES INCLUDED 32   MISSING CASES 0
```

Tehtävät:

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

(b) Laske tulostuksen 5 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?