Oulun yliopiston matemaattisten tieteiden laitos/tilastotiede

806112P DATA-ANALYYSIN PERUSMENETELMÄT (Jari Päkkilä)

Loppukoe, ma 27.9.2010 klo 14-18, L6

Mukana saa olla KAKSI molemmin puolin täytettyä A4-kokoista lunttilappua sekä tavanomainen funktiolaskin tai graafinen laskin. Tarvittavat taulukot ovat liitteinä.

Vastaa kaikkiin kysymyksiin.

1. Epäilet, että tilastotieteen (biostatistiikan) lukeminen on lääkäreille niin tuskaista, että heidän hikoilemansa nestemäärä on merkittävästi suurempi sitä lukiessa kuin lääketiedettä tai sanomalehtiä lukiessa. Teet kliinisen kokeen 15 vapaaehtoiselle lääkärille satunnaistamalla heidät kolmeen viiden lääkärin ryhmään. Ensimmäinen ryhmä lukee tunnin ajan Biostatistiikan taskutietoa, toinen ryhmä saman ajan lääketieteen alan tieteellistä julkaisua (British Medical Journalia) ja kolmas ryhmä Helsingin Sanomia. Koehenkilöt punnitaan ennen kokeen alkua tarkalla vaa'alla, ja punnitus toistetaan lukukokeen päätyttyä. Punnitustulosten erotus mittaa hikoilun määrää kokeen aikana. Kokeen päätyttyä saat tietoosi seuraavat tulokset (hikoilun määrä grammoina):

	hikoilun		
luettava lehti	havaitut arvot	keskiarvo	keskihajonta
Biostatistiikka	142, 118, 120, 131, 128	127.8	9.6
British Medical Journal	12, 21, 30, 26, 22	22.2	6.7
Helsingin Sanomat	12, 11, 18, 21, 16	15.6	4.2
Koko aineisto		55.2	53.6

Aineiston lähde: Uhari, M. ja Nieminen, P. Epidemiologia ja biostatistiikka, 2001.

Tutkimuskysymyksenä on se vaikuttaako lääkäreillä luettavan tekstin aihe hikoilun määrään.

(a) Kuvaile aineisto sopivalla graafisella esityksellä.

(b) Muotoile tilanteeseen sopiva tilastollinen malli ja luettele malliin liittyvät oletukset. Analysoi aineistoa asetelmaan sopivin menetelmin. Raportoi tuloksista mallin systemaattiseen osaan liittyvien, odotusarvojen erotusta kuvaavien, parametrien piste-estimaatit, 95 %:n luottamusvälit ja merkitsevyystestaukset.

(c) Muodosta mallin Anova-taulu ja tulkitse antamat tulokset lyhyesti.

(6 p)

2. Tutkitaan erään rauhoittavan lääkkeen vaikutusta muistiin. Joukolle vapaaehtoisia opiskelijoita ($n = 7$) näytettiin 8-numeroisia lukuja. Kokeessa opiskelija sai pisteitä sen mukaan, kuinka hyvin hän muisti näytetyt numerosarjat. Kun opiskelija oli nauttinut rauhoittavaa lääkettä (ja riittävä aika oli odotettu, jotta lääkkeen vaikutus alkaisi), koe tehtiin uudelleen. Saadut pistemäärät olivat:

	Opiskelija							
Lääke	1	2	3	4	5	6	7	keskiarvo
Kyllä	106	95	102	87	72	83	64	87.0
Ei	122	118	100	95	84	83	80	97.4

Aineiston lähde: Uhari, M. ja Nieminen, P. Epidemiologia ja biostatistiikka, 2001.

(a) Minkä niminen tai tyyppinen koeasetelma on kyseessä? Määrittele tutkimusongelman ("onko rauhoittavalla lääkkeellä vaikutusta muistiin?") kuvaamiseen soveltuva vertailuparametri ja sitä koskeva nollahypoteesi.

(b) Testaa kohdan (a) nollahypoteesia ja laske ao. parametrin 95 %:n luottamusväli. Mitä päätelmiä voit tehdä tulosten perusteella?

(6 p)

3. Kalabiologi Pekka Brofelt keräsi Längelmävedellä vuonna 1916 aineistoa tutkimuksiinsa eri kalojen ruumiinmitoista ja niiden yhteyksistä toisiinsa. Tässä tehtävässä tarkastelemme erityisesti sitä, miten lahnoilla (*Abramis brama*) ruumiin paino (dekagrammoina, deka=10) riippuu maksimipituudesta (muuttuja `pituus` senttimetreinä mitattuna suusta pyrstön päähän) ja korkeudesta (`korkeus`, cm kalan korkeimmalta kohdalta.)

Aineiston lähde: Brofelt, P. Bidrag till kännedom om fiskbeståendet in våra sjöar. Längelmävesi. Kirjassa Järvi, T.H. Finlands Fiskeriet Band 4, Meddelanden utgivna av fiskeriföreningen i Finland. Helsingfors 1917.

Liitteessä 1 annetaan havaintoaineisto sekä otteita R-ohjelman tulostuksesta (osa tuloksista on peitetty). Vastaa niiden pohjalta seuraaviin kysymyksiin.

(a) Muuttujaa korkeus koskevan regressiokertoimen estimaatti on erilainen mallissa 1 (`malli.1`), jossa se on ainoana selittäjänä, kuin mallissa 3 (`malli.3`), jossa selittäjänä on lisäksi pituus. Millainen tulkinta ao. kertoimella ja estimaatilla on mallissa 1 ja millainen se on mallissa 3?

(b) Laske mallissa 3 muuttujan `korkeus` regressiokertoimen 95 % luottamusväli sekä arvioi P-arvoa sitä nollahypoteesia vastaan, jonka mukaan ao. kerroin olisi todellisuudessa arvoltaan nolla.

(c) Kalastaja saa lahnan, jonka pituudeksi hän mittaa 40 cm. Laske Brofeltin mittauksia apuna käyttäen ennuste kalan painolle sekä ennusteen 95 % ennusteväli.

(6 p)

4. Tutkimuksessa haluttiin selvittää sitä eroavatko kahden perunalajikkeen (Fambo ja Van Gogh) tärkkelyspitoisuudet toisistaan. Tutkimusryhmän käytössä oli 16 koealaa, joista kahdeksalla satunnaisesti valitulla koealalla viljeltiin Fambo-lajiketta ja lopuilla koealoilla Van Gogh-lajiketta. Koealoittain saatiin tärkkelyspitoisuuksien osalta seuraavat tulokset (%):

Fambo: 16.1, 16.4, 16.7, 16.8, 17.2, 17.2, 17.4, 17.8
Van Gogh: 17.0, 17.0, 17.4, 17.5, 17.6, 17.8, 19.1, 19.2

Laske kiinnostavan parametrin Δ

(a) piste-estimaatti,

(b) luottamusväli likimääräisellä 95 % luottamustasolla,

(c) sekä p-arvo kaksisuuntaiseen merkitsevyystestiin, jossa testataan nollahypoteesia $H_0 : \Delta = 0$.

kun estimointikriteerinä on järjestysluvuin painotettu itseisarvopoikkeamien summa. Tulkitse tulokset lyhyesti. Laskelmissa saattaa auttaa liitteen 2 tiedot.

(6 p)

5. Eräässä tutkimuksessa seurattiin viittätoista tupakoitsijaa, jotka olivat ilmoittaneet haluavansa lopettaa tupakoinnin. Kunkin koehenkilön seuruu alkoi tupakoinnin lopettamispäivänä. Vastemuuttujana oli tupakoinnin uudelleen aloittamiseen kulunut aika (ts. "tupakkalakon pituus"). Osalla koehenkilöistä vaste jäi havaitsematta, ja heidän kohdallaan tiedetään vain, kuinka kauan he ainakin olivat tupakoimatta ennen kuin seuruu lopetetiin ja tulokset analysoitiin. Näiden henkilöiden seuruuaikoja merkitään seuraavassa $^+$:lla. Seuruuajat (viikkoina) olivat ikäryhmittäin seuraavat:

alle 40-vuotiaat: 2, 2^+, 3^+, 4, 7^+, 9. 10, 12^+
vähintään 40-vuotiaat: 1, 2^+, 3, 5, 5, 6, 8

(a) Tee tarvittavat laskelmat ja piirrä elinaikojen jakauman kuvaamisessa tavallisesti käytettävä, Kaplan–Meier -menetelmällä estimoitu, käyrä kummallekin ikäryhmälle erikseen samaan koordinaatistoon.

(b) Tarkastellaan vasteen jakauman alakvartiilia, mediaania ja yläkvartiilia. Mitkä ko. tunnusluvuista voidaan estimoida alle 40-vuotiaille ja mitkä vähintään 40-vuotiaille, kun aineiston analysoinnissa käytetään KM-menetelmää? Arvioi ko. parametrien arvot.

(c) Estimoi vasteen jakauman odotusarvo **vähintään 40-vuotiailla**, kun aineiston analysoinnissa, käytetään

(c1) Kaplan–Meier -menetelmää. (c2) eksponenttimallia.

(d) Mikä on eksponenttimallin antama estimaatti todennäköisyydelle, että vähintään 40-vuotiaan tupakkalakko kestää korkeintaan viisi viikkoa?

(6 p)

LIITE 1

```
1      24   30.0    11.5
2      29   31.2    12.5
3      34   31.1    12.4
4      36   33.5    12.7
5      43   34.0    12.4
6      45   34.7    13.6
7      50   34.5    14.2
8      39   35.0    12.7
9      45   35.1    14.0
10     50   36.2    14.2
11     48   36.2    14.3
12     50   36.2    14.4
13     50   36.4    13.8
14     60   37.2    15.0
15     60   37.2    15.4
16     70   38.3    14.9
17     70   38.5    14.9
18     61   38.6    15.6
19     65   38.7    14.5
20     58   39.5    15.1
21     68   39.2    16.0
22     62   39.7    15.5
23     68   40.6    15.5
24     70   40.5    16.2
25     72   40.9    16.4
26     72   40.6    16.4
27     71   41.5    16.5
28     85   41.6    16.9
29    100   42.6    19.0
30     92   44.1    18.0
31     96   44.0    18.1
32     92   45.3    18.8
33     98   45.9    18.6
34     95   46.5    17.6
>
> round(mean(lahna),2)  # keskiarvot
  paino  pituus korkeus
  62.59   38.39   15.22
> round(sd(lahna),2)    # keskihajonnat
  paino  pituus korkeus
  20.68    4.22    1.98
> round(cor(lahna),3)   # korrelaatiokertoimet
        paino pituus korkeus
paino   1.000  0.964   0.970
pituus  0.964  1.000   0.954
korkeus 0.970  0.954   1.000
>
>
>
```

```
> malli.1 <- lm(paino ~ korkeus) ; summary(malli.1)

Residuals:
   Min     1Q Median     3Q    Max
-6.789 -3.374 -1.100  1.921 10.687

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -91.5034             -13.38
korkeus      10.1219              22.72

Residual standard error: 5.074 on 32 degrees of freedom
Multiple R-squared: 0.9416,     Adjusted R-squared: 0.9398
F-statistic: 516.2                 p-value:

>
>
> malli.2 <- lm(paino ~ pituus) ; summary(malli.2)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8588 -3.2623 -0.6116  2.8599 17.4836

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -118.9071             -13.35
pituus         4.7283              20.50

Residual standard error: 5.587 on 32 degrees of freedom
Multiple R-squared: 0.9292,     Adjusted R-squared: 0.927
F-statistic: 420.1                 p-value:

>
>
> malli.3 <- lm(paino ~ pituus + korkeus) ; summary(malli.3)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1860 -2.7604 -0.8489  1.5306  9.4953

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -107.1522             -14.416
pituus         2.0860               3.461
korkeus        5.8901               4.596

Residual standard error: 4.378 on 31 degrees of freedom
Multiple R-squared: 0.9579,     Adjusted R-squared: 0.9552
F-statistic: 352.7                 p-value:
```

LIITE 2

```
> y_F
[1] 16.1 16.4 16.7 16.8 17.2 17.2 17.4 17.8
> y_V
[1] 17.0 17.0 17.4 17.5 17.6 17.8 19.1 19.2
> y <- c(y_F,y_V)
> n1 <- length(y_F); n2 <- length(y_V)
> d <- c(matrix(rep(y_V,n1),n2)-t(matrix(rep(y_F,n2),n1)))
> d <- sort(d) ; d
 [1] -0.8 -0.8 -0.4 -0.4 -0.4 -0.3 -0.2 -0.2 -0.2 -0.2 -0.2  0.0  0.0  0.1  0.2
[16]  0.2  0.2  0.2  0.2  0.3  0.3  0.3  0.3  0.4  0.4  0.4  0.6  0.6  0.6  0.6
[31]  0.6  0.7  0.7  0.8  0.8  0.9  0.9  0.9  1.0  1.0  1.1  1.1  1.2  1.3  1.3
[46]  1.4  1.4  1.4  1.5  1.7  1.7  1.8  1.9  1.9  2.0  2.0  2.3  2.4  2.4  2.5
[61]  2.7  2.8  3.0  3.1
> n <- length(y)
> y. <- sort(y) ; y.
 [1] 16.1 16.4 16.7 16.8 17.0 17.0 17.2 17.2 17.4 17.4 17.5 17.6 17.8 17.8 19.1
[16] 19.2
> apu <- matrix(rep(y.,n),n)
> VV <- (apu+t(apu))/2
> V <- c(VV[row(VV) <= col(VV)])
> V <- sort(V) ; V
  [1] 16.10 16.25 16.40 16.40 16.45 16.55 16.55 16.55 16.60 16.65 16.65 16.70
 [13] 16.70 16.70 16.75 16.75 16.75 16.80 16.80 16.80 16.80 16.85 16.85 16.85
 [25] 16.90 16.90 16.90 16.90 16.95 16.95 16.95 16.95 16.95 17.00 17.00 17.00
 [37] 17.00 17.00 17.00 17.05 17.05 17.10 17.10 17.10 17.10 17.10 17.10 17.10
 [49] 17.10 17.10 17.15 17.15 17.20 17.20 17.20 17.20 17.20 17.20 17.20 17.20
 [61] 17.25 17.25 17.25 17.25 17.30 17.30 17.30 17.30 17.30 17.30 17.30 17.30
 [73] 17.35 17.35 17.40 17.40 17.40 17.40 17.40 17.40 17.40 17.40 17.40 17.45
 [85] 17.45 17.50 17.50 17.50 17.50 17.50 17.50 17.50 17.55 17.60 17.60 17.60
 [97] 17.60 17.60 17.60 17.65 17.65 17.65 17.70 17.70 17.75 17.80 17.80 17.80
[109] 17.80 17.90 17.95 17.95 18.00 18.05 18.05 18.10 18.10 18.15 18.15 18.20
[121] 18.20 18.25 18.25 18.30 18.30 18.30 18.35 18.35 18.40 18.45 18.45 18.50
[133] 18.50 19.10 19.15 19.20
>
```