806109P TILASTOTIETEEN PERUSMENETELMÄT I
Loppukoe 14.3.2011 (Marjatta Mankinen ja Jari Päkkilä)

VALITSE KUUDESTA TEHTÄVÄSTÄ VIISI JA VASTAA VAIN NIIHIN!

1. Valitse kohdissa A-F oikea (vain yksi) vaihtoehto. Oikeasta vastauksesta saat +1 pistettä, väärästä et menetä pisteitä.

A) Perheen lasten lukumäärän ($= x$) jakauma eräässä populaatiossa on seuraava:

Lasten lkm perheessä	frekvenssi
0	3
1	4
2	6
3	2
4	1
Yhteensä	16

Lasten lukumäärän aritmeettinen keskiarvo on yhden desimaalin tarkkuudella

a1) 1.0 a2) 1.2 a3) 1.6 a4) 2.0 a5) 3.0 a6) 3.2

B) Ylipeitto otantatutkimuksessa tarkoittaa sitä, että

b1) käytettäessä yksinkertaista satunnaistotantaa palauttamatta otoskoko tulee liian suureksi,

b2) kehikkopopulaatio sisältää sellaisia havaintoyksiköitä, jotka eivät kuulu kohdepopulaatioon,

b3) kohdepopulaatio sisältää sellaisia havaintoyksiköitä, jotka eivät ole mukana kehikkopopulaatiossa,

b4) ositettua otantaa käytettäessä jokin ositteista tulee yliedustetuksi ositekohtaisten otoskokojen laskennassa tapahtuvien pyöristysvirheiden takia,

b5) haastattelututkimukseen osallistuva henkilö valitsee tarjolla olevista vastausvaihtoehdoista useamman kuin yhden,

b6) samaan populaatioon kohdistuu samanaikaisesti useampia tutkimuksia.

C) Luokitteluasteikollisen muuttujan x mahdolliset arvot ovat A, B, C ja D. Sata havaintoa sisältävässä havaintoaineistossa arvo A esiintyy 25 kertaa, arvo B 35 kertaa, arvo C 20 kertaa ja arvo D 20 kertaa. Muuttujan x

c1) jakauma voidaan esittää pistekuviona,

c2) jakauman moodi on 35,

c3) hajontaa voidaan kuvailla vaihteluvälin avulla,

c4) summajakauma on mielekäs muodostaa,

c5) havaintoarvot voidaan standardoida.

c6) Mikään edellä esitetyistä kohdista c1)–c5) ei pidä paikkaansa.

D) Kun välimatka-asteikollisen muuttujan x jakauma on selvästi vino vasemmalle, pätee, että

d1) $\bar{x} < Md$, d2) $\bar{x} = Md$, d3) $\bar{x} > Md$,
d4) $\bar{x} < s$, d5) $\bar{x} = s$, d6) $\bar{x} > s$.

Väitteet E–F liittyvät alla esitettyyn R-ohjelman tulostukseen. Analysoitavassa aineistossa on 300 havaintoa ja kaksi muuttujaa: lapsen syntymäpaino (SYNTPAIN, grammoina) ja syntymäpituus (SYNTPIT, senttimetreinä). Syntymäpainon keskihajonta on 524.9 ja syntymäpituuden keskihajonta 2.5.

```
Call:
lm(formula = SYNTPAIN ~ SYNTPIT, data = kohortti)

Residuals:
   Min     1Q Median     3Q    Max
-998.2 -220.0  -14.1  184.1 3432.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4190.98     429.44   -9.76   <2e-16 ***
SYNTPIT       152.96       8.54   17.90   <2e-16 ***

Residual standard error: 363 on 295 degrees of freedom
Multiple R-squared: x.xxx,Adjusted R-squared: x.xxx
```

E) Tulostuksen regressioyhtälön determinaatiokerroin eli selitysaste on likimain

e1) 0.005 e2) 0.73 e3) 0.53 e4) <2e-16 e5) 0.81 e6) 8.5.

F) Yhden aineistoon kuuluvan lapsen syntymäpaino on 3300 grammaa ja pituus 49 cm. Kyseisen havaintoyksikön kohdalla regressioyhtälöön liittyvä residuaali on likimain

f1) 3304 f2) -4190.98 f3) 3432.2 f4) -4 f5) 363 f6) 4.

2. Yrityksen työntekijöiden iän (vuosina) jakauma on seuraava:

Ikä	frekvenssi
20–29	5
30–39	12
40–49	30
50–59	18
60–69	2
Yhteensä	67

a) Esitä iän jakauma graafisesti. *(1.5 p)*

b) Laske iän aritmeettinen keskiarvo ja keskihajonta. *(2 p)*

c) Yrityksen osastolla A työskentelee 25 työntekijää, joiden kuukausipalkan (euroina) havaituista arvoista on laskettu R-ohjelmalla seuraavat tunnuslukujen arvot:

```
> numSummary(palkka, statistics=c("mean", "sd", "quantiles"),
  quantiles=c( 0,.25,.5,.75,1 ))
  mean  sd   0%  25%  50%  75% 100%  n
  2920 516 1550 2890 3150 3250 3300 25
```

Osastolla B työskentelee 9 työntekijää, joiden kuukausipalkat ovat:

2450 2250 2600 2850 2650 2950 2900 2700 2500

Vertaile osastojen A ja B kuukausipalkkoja laatikko-jana -kuvion avulla. Kommentoi saamaasi tulosta. *(2.5 p)*

3. a) Kokkikilpailuun osallistuu kuusi kilpailijaa: A, B, C, D, E ja F. Tuomari 1 asetti kilpailijat paremmuusjärjestykseen C, F, E, A, D, B ja tuomarin 2 paremmuusjärjestys oli C, D, F, A, E, B. Tutki tuomareiden arvostelun yhdenmukaisuutta tilanteeseen sopivan riippuvuustunnusluvun avulla. *(2 p)*

b) Kopioi alla oleva ristiintaulukko vastauspaperiisi ja täydennä se solufrekvensseillä siten, että muuttujat x ja y ovat täysin riippumattomia toisistaan. *(2 p)*

		x			
		A	B	C	Yhteensä
y	0				20
	1				30
Yhteensä		25	10	15	50

c) Havaintoaineistossa on kolme muuttujaa: x, y ja z. R-ohjelmalla laskettu ko. havaintoaineistoon liittyvä kovarianssimatriisi on

```
>cov(aineisto) #kovarianssimatriisi

          x      y      z
x      81.9   57.2  -13.4
y      57.2  159.4   40.7
z     -13.4   17.5   10.7
```

Laske x:n ja z:n välisen korrelaatiokertoimen r_{xz} arvo ja tulkitse tulos lyhyesti. *(2 p)*

4. Erään suuren varusmiesryhmän Cooperin testin tulokset noudattivat likimain normaalijakaumaa odotusarvolla (keskiarvolla) 2494 metriä ja keskihajonnalla 264 metriä.

Cooperin testi on 12 minuutin aikana tasaisella alustalla, täydellä teholla juostu matka.

a) Monellako prosentilla ko. ryhmästä testin tulos oli korkeintaan 2000 metriä? *(2 p)*

b) Jos valitaan satunnaisesti viisi varusmiestä ko. ryhmästä, mikä on todennäköisyys, että ainakin kahdella tulos oli vähintään 2700 metriä? *(2 p)*

c) Määrää se Cooperin testin tulos, jota huonomman arvon sai viidennes ko. ryhmän varusmiehistä? *(2 p)*

5. a) Eräässä USA:ssa tehdyssä tutkimuksessa 678:sta satunnaisesti valitusta 20-34 -vuotiaasta miehestä 58:lla todettiin korkea (=suositukset ylittävä) verenpaine.

Laske ja tulkitse

a1) piste-estimaatti, a2) 95%:n luottamusväli korkeasta verenpaineesta kärsivien suhteelliselle osuudelle ko. otosta vastaavassa populaatiossa. *(3 p)*

b) Tietynmerkkisistä muropakkauksista on valittu satunnaisesti sata ja punnittu pakkaukset (paino grammoina). Aineistosta on sitten R:llä saatu seuraava tulostus:

```
One Sample t-test

data:  murot$paino
t = 1.003, df = 99, p-value = 0.3183
alternative hypothesis: true mean is not equal to 520
95 percent confidence interval:
 519.0726 522.8234
sample estimates:
mean of x
  520.948
```

Tulkitse sekä testiä että luottamusväliä koskevat tulokset. *(3 p)*

6. 506:lle satunnaisesti valitulle naiselle ja 399:lle satunnaisesti valitulle miehelle USA:ssa esitettiin seuraava kysymys: "Kuinka monta tuntia keskimäärin katsot TV:tä päivässä?". Vastauksista saatiin seuravat tulokset:

	Naiset	Miehet
Katseluaikojen keskiarvo	3.06	2.89
Katseluaikojen keskihajonta	2.12	2.63

Tutki sopivalla merkitsevyystestillä, katsovatko naiset ja miehet keskimäärin yhtä paljon TV:tä USA:ssa. Kirjoita kaikki testauksen vaiheet näkyviin. Katseluaikojen varianssien oletetaan olevan yhtä suuret naisilla ja miehillä. *(6 p)*