806109P TILASTOTIETEEN PERUSMENETELMÄT I
Loppukoe 9.1.2012 (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. Perusteluita ei vaadita.

A) Erään ryhmän jäsenet punnitaan jokaisen kuukauden alussa. Muuttuja, joka kertoo henkilön painonmuutoksen joulukuun ja tammikuun punnitusten välillä, on

a1) suhdeasteikkoa ja diskreetti,

a2) suhdeasteikkoa ja jatkuva,

a3) järjestysasteikkoa ja diskreetti,

a4) järjestysasteikkoa ja jatkuva,

a5) välimatka-asteikkoa ja diskreetti,

a6) välimatka-asteikkoa ja jatkuva.

B) Kausaali- eli syy-seuraussuhteita tutkittaessa

b1) havaintoaineisto tulisi aina hankkia yksinkertaisella satunnaisotannalla,

b2) ankaran kausaalisuuden periaatteen mukaan syy-seuraussuhteista voidaan tehdä päätelmiä vain kokeita tekemällä,

b3) vastemuuttujan tulisi olla vähintään välimatka-asteikkoa,

b4) kehikkopopulaatiossa ei saa esiintyä alipeittoa,

b5) selittävien muuttujien tulee olla jatkuvia muuttujia,

b6) analyysissä käytettävät muuttujat täytyy standardoida.

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, mutta sitä ei voi esittää graafisesti,

c5) havaintoarvot voidaan standardoida,

c6) jakaumalle ei ole mielekästä määrätä mediaania.

D) Olli osallistui soveltuvuuskokeeseen, joka koostui kahdesta osiosta (I ja II). Soveltuvuuskokeeseen osallistuneiden pistemäärien keskiarvot ja keskihajonnat ovat seuraavat:

	keskiarvo	keskihajonta
Osio I	100	10
Osio II	120	15

Olli sai osiosta I 120 pistettä ja hän menestyi osiossa II suhteellisesti yhtä hyvin kuin osiossa I. Ollin saama pistemäärä osiosta II on

d1) 150, d2) 140, d3) 130, d4) 120, d5) 110, d6) 100.

E) Diskreetin satunnaismuuttujan X kertymäfunktio $F(x)$ on

$$F(x) = \begin{cases} 0, & \text{kun } x < -3 \\ 0.2, & \text{kun } -3 \leq x < -1 \\ 0.5, & \text{kun } -1 \leq x < 0 \\ 0.6, & \text{kun } 0 \leq x < 2 \\ 0.7, & \text{kun } 2 \leq x < 3 \\ 1, & \text{kun } x \geq 3. \end{cases}$$

Todennäköisyys $P(-1 < X \leq 2)$ on

e1) 0.4, e2) 0.7, e3) 0.5, e4) 0.2, e5) 0.6, e6) 0.1.

F) Satunnaismuuttuja $X \sim Exp(0.25)$. Tällöin satunnaismuuttujan $Y = 2X + 5$ varianssi on

f1) 13, f2) 16, f3) 32, f4) 37, f5) 64, f6) 89.

2. Tarkastellaan havaintoaineistoa, joka sisältää tietoja Suomen kunnista vuosilta 2009 ja 2010. Aineistosta ($n = 326$) on jätetty pois Ahvenanmaan kunnat ja se on kerätty vuoden 2010 kuntajaon mukaisesti. Tässä tehtävässä tarkastellaan kuntien vuonna 2010 keräämiä kunnallisveroja. (`Aineiston lähde: http://www.kunnat.net`)

a) Alla olevassa taulukossa on esitetty kerätyn kunnallisveron (euroa/asukas) frekvenssijakauma.

Kerätty kunnallisvero (euroa/asukas)	Frekvenssi
1710 – 2200	105
2210 – 2700	129
2710 – 3200	74
3210 – 5200	18
Yhteensä	326

a1) Esitä kerätyn kunnallisveron prosenttinen summajakauma graafisesti.

a2) Arvioi piirtämäsi kuvion perusteella tarkasteltavan jakauman 90% fraktiili.

(2 p)

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

c) Suomen kunnat voidaan jakaa taajamaväestön osuuden ja suurimman taajaman väkiluvun perusteella kolmeen luokkaan: kaupunkimaisiin, taajaan asuttuihin ja maaseutumaisiin kuntiin. Alla esitetyssä R-ohjelman tulostuksessa on esitetty kyseisen kuntajaon mukaisille kuntaryhmille laskettuja tunnuslukuja kerätyistä kunnallisveroista.

```
> numSummary(kunnat[,"kvero2010"], groups=kunnat$kuntaryhmitys2,
+  statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
                mean    sd   0%  25%  50%  75% 100% data:n
Kaupunkimaiset  3043 414.2 2525 2837 2924 3174 5203     61
Taajaan asutut  2600 318.6 2008 2369 2519 2851 3722     66
Maaseutumaiset  2237 330.7 1710 2024 2206 2355 4587    199
```

Piirrä (samaan kuvioon!) eri kuntaryhmien keräämien kunnallisverojen laatikko-jana -kuviot. Onko eri kuntaryhmien keräämien kunnallisverojen jakaumien sijainnissa eroa? Kommentoi lyhyesti. *(2 p)*

3. Tämän tehtävän aineisto on esitelty tehtävässä 2.

a) Aineiston perusteella muodostettu kuntaluokan (kaupunkimainen/taajaan asuttu/maaseutumainen) ja vuoden 2010 kunnallisveroprosentin välinen ristiintaulukko on seuraava:

	Kunnallisveroprosentti			
Kuntaluokka	16.50 – 19.25	19.50 – 20.00	20.25 – 21.00	Yhteensä
Kaupunkimainen	27	28	6	61
Taajaan asuttu	38	107	54	199
Maaseutumainen	13	36	17	66
Yhteensä	78	171	77	326

a1) Tutki kuntaluokan ja kunnallisveroprosentin välistä riippuvuutta ehdollisten prosenttijakaumien avulla. *(1.5 p)*

a2) Kuinka monella aineiston taajaan asutulla kunnalla voidaan odottaa kunnallisveroprosentin olevan välillä 20.25 – 21.00, jos ko. muuttujien oletetaan olevan täysin toisistaan riippumattomia? *(0.5 p)*

b) Liitteen 1 R-ohjelman tulostuksessa on analysoitu kunta-aineiston muuttujien kvero2009 ja kvero2010 välistä riippuvuutta. Muuttuja kvero2009 kertoo kunnan keräämän kunnallisveron määrän vuonna 2009 (euroa/asukas) ja muuttuja kvero2010 vastaavasti kunnan keräämän kunnallisveron määrän vuonna 2010 (euroa/asukas). Käytä hyväksesi liitteen 1 tulostusta vastatessasi seuraaviin kysymyksiin.

b1) Määrää tulostuksen hajontakuvioon liittyvän Pearsonin tulomomenttikorrelaatiokertoimen arvo ja tulkitse se. *(1.5 p)*

b2) Tulkitse tulostuksen regressioanalyysin tulokset (regressioyhtälö ja sen kertoimien selväkielinen tulkinta). *(1.5 p)*

b3) Vuonna 2009 kuntien keräämien kunnallisverojen aritmeettinen keskiarvo oli 2454.17 euroa/asukas. Määrää kuntien vuonna 2010 keräämien kunnallisverojen aritmeettinen keskiarvo. *(1 p)*

4. Perunoita pussittava kone on säädetty toimimaan siten, että perunapussin paino (perunat + pussi) noudattaa likimain normaalijakaumaa odotusarvolla 2.030 kg ja keskihajonnalla 80 g.

a) Mikä on todennäköisyys, että satunnaisesti valitun perunapussin paino on alle 1.850 kg? *(2 p)*

b) Mikä on todennäköisyys, että kymmenen satunnaisesti valitun perunapussin yhteispaino on välillä 20.000 kg - 20.500 kg? *(2 p)*

c) Määrää a siten, että kymmenen satunnaisesti valitun perunapussin painojen keskiarvo on korkeintaan a todennäköisyydellä 0.05. *(2 p)*

5. TNS Gallup teki joulukuussa 2011 Helsingin Sanomien toimeksiannosta tutkimuksen suomalaisten suhtautumisesta sotilasliitto Naton jäsenyyteen. Tutkimuksessa puhelinhaastateltiin 1002 suomalaista, jotka edustavat Suomen 15 vuotta täyttänyttä väestöä Ahvenanmaata lukuun ottamatta. Vastaajista 160 kannatti Suomen Nato-jäsenyyttä.

a) Tee tarvittavat oletukset ja määrää Nato-jäsenyyttä joulukuussa 2011 kannattavien suhteelliselle osuudelle Suomessa

   a1) piste-estimaatti,

   a2) 99%:n luottamusväli.

   a3) Tulkitse a1) ja a2) -kohdissa saadut tulokset. *(3 p)*

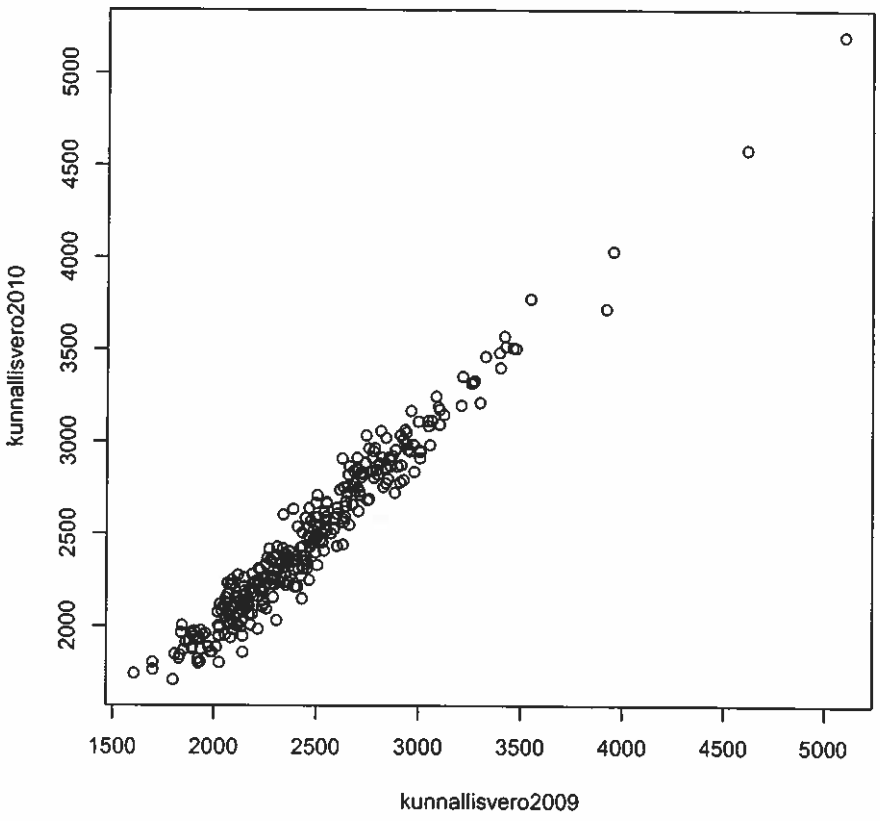
b) Alla on edellä esitetystä TNS Gallupin aineistosta R-ohjelmalla saatu tulostus:

```
> prop.test (160,1002,p=0.20,alternative="less", correct=FALSE)

1-sample proportions test without continuity correction

data:  160 out of 1002, null probability 0.2
X-squared = xxxxxx, df = 1, p-value = 0.0007096
```

Aseta tulostuksen merkitsevyystestin hypoteesit, laske peitetty testisuureen arvo ja esitä lyhyesti perustellen, millaiseen johtopäätökseen testissä päädytään. *(3 p)*

6. Erään pikkukaupan juustovalikoima käsittää vain viisi erilaista juustoa (A, B, C, D ja E). Kaupan omistaja tietää aiempien aikojen myynnin perusteella, että asiakkaista 20% on suosinut A:ta, 35% B:tä, 18% C:tä, 12% D:tä ja loput E:tä. Omistaja on kuitenkin viime aikoina pannut merkille, että juustojen suosiossa on lisääntyneen terveysvalistuksen myötä ehkä tapahtunut muutos. Tehdyssä tutkimuksessa 80 satunnaisesti valitusta asiakkaasta 12 valitsi nyt suosikikseen A:n, 26 B:n, 16 C:n, 18 D:n ja loput E:n. Tutki sopivalla merkitsevyystestillä, näyttääkö saadun aineiston perusteella omistajan epäily aiheelliselta. Kirjoita kaikki testauksen vaiheet näkyviin. *(6 p)*

Liite 1:




```
Call:
lm(formula = kvero2010 ~ kvero2009, data = kunnat)

Residuals:
     Min       1Q   Median       3Q      Max
-285.444  -62.534    2.023   63.921  273.311

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -91.67006   31.22481  -2.936  0.00356
kvero2009         1.04004    0.01252  83.044  < 2e-16

Residual standard error: 98.78 on 324 degrees of freedom
Multiple R-squared: 0.9551,Adjusted R-squared: 0.955
F-statistic:  6896 on 1 and 324 DF,  p-value: < 2.2e-16
```