806109 TILASTOTIETEEN PERUSMENETELMÄT I
Harjoitus 7, kevät 2010

1. Pienen ala-asteen kakkosluokan liikuntaryhmän pojat (n=11) ottelivat neliottelun, jossa lajeina olivat 60 metrin juoksu, 100 metrin juoksu, pallonheitto ja pituushyppy.

   a) Pallonheitossa lopputulokset (metreinä) olivat
   18.7 23.4 15.9 19.6 21.0 19.7 18.5 19.7 21.8 18.4 20.1
   Laske pallonheiton lopputulosten aritmeettinen keskiarvo ja keskihajonta.

   b) Liitteessä 1 on esitetty tarkasteltavasta aineistosta saatu R:n tulostus. Kommentoi 60 metrin juoksun ja 100 metrin juoksun välistä riippuvuutta korrelaatiokertoimen r perusteella. Missä korrelaatiodiagrammin kuvassa/kuvissa (1-12) kuvataan 60 metrin juoksun ja 100 metrin juoksun välistä riippuvuutta?

   c) Määrää regressioyhtälö $y = a + bx$ ja tulkitse kertoimet, kun vastemuuttujana on 100 metrin juoksu ja selittävänä muuttujana 60 metrin juoksu. Määrää lisäksi regressioyhtälön determinaatiokerroin (eli selitysaste) ja tulkitse se.

   Matin tulokset neliottelussa olivat: 60 metrin juoksu = 12.5 sekuntia, 100 metrin juoksu = 23.2 sekuntia, pallonheitto = 19.6 metriä ja pituushyppy 2.20 metriä. Ennusta Matin vastemuuttujan arvo regressioyhtälön avulla. Laske ennustetun arvon ja todellisen arvon erotus eli residuaali.

2. Liitteessä 2 on R:llä saatuja tuloksia aineistosta, joka sisältää seuraavat tiedot 50:stä Oulussa keväällä 2002 myytävänä olleesta rivitaloasunnosta: hintapyyntö (1000 euroina), neliömäärä, ikä (vuosia, väh. yksi) ja etäisyys keskustasta (km).

   a) Mikä mukana olevista muuttujista on paras selittävä muuttuja hintapyynnölle? Perustele vastauksesi. Määrää kyseinen regressioyhtälö ja tulkitse yhtälön kertoimet. Määrää myös regressioyhtälön determinaatiokerroin.

   b) Tulkitse regressioanalyysin tulokset liitteen kohdasta b) (regressioyhtälö, kertoimien tulkinta, determinaatiokerroin).

   c) Laske sekä a)- että b)-kohdan yhtälöä käyttäen ennustearvo sellaisen rivitaloasunnon hinnaksi v. 2002, joka on ollut silloin 10 vuotta vanha, kooltaan 100 neliötä ja sijainnut 5 km:n päässä keskustasta.

3. Paperisilppurin jäljiltä paperiarkki on silputtu 264 palaan, joista jokaisessa on korkeintaan yksi kirjain. Kirjaimia on seuraavasti:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	R	S	T	U	V	Ä	Ö
12	1	0	3	8	0	0	5	15	4	7	2	6	12	8	6	6	4	6	12	2	3	1

   Valitset satunnaisesti yhden lapun. Mikä on todennäköisyys, että

   a) se on tyhjä, b) siinä on vokaali, c) siinä on konsonantti,

   d) siinä on kirjain, jota ei ole missään muussa lapussa,

   e) siinä on F,

   f) siinä on joku sanassa "tilastotiede" oleva kirjain?

4. a) Mikä on todennäköisyys, että satunnaisesti valittu viikonpäivä on
   a1) sunnuntai a2) lauantai tai sunnuntai?

   b) Herätyskello pysähtyy pariston loputtua. Millä todennäköisyydellä kello pysähtyy
   b1) klo 01.00 ja 06.00 välisenä aikana,
   b2) lauantaina tai sunnuntaina klo 01.00 ja 06.00 välisenä aikana?

5. Eräänä päivänä lääkärin vastaanotolle tuli 18 potilasta, joista 6 sairasti influenssaa. Millä todennäköisyydellä kahdesta satunnaisesti valitusta potilaasta

   a) kumpikaan valituista ei sairastanut influenssaa?

   b) ainakin toinen sairasti influenssaa?

6. Seuraavassa taulukossa on vuonna 2007 valittujen kansanedustajien lukumäärät sukupuolen ja syntymävuoden mukaan:

	1930-39	1940-49	1950-59	1960-69	1970-79	1980-89	Yhteensä
Miehet	4	39	36	26	11	1	117
Naiset	1	7	16	38	21	0	83
Yhteensä	5	46	52	64	32	1	200

Valitaan satunnaisesti (umpimähkään) yksi kansanedustaja. Mikä on todennäköisyys, että valittu kansanedustaja on

a) mies,

b) syntynyt 1960-luvulla,

c) mies ja syntynyt 1960-luvulla,

d) mies tai syntynyt 1960-luvulla,

e) nainen, kun tiedetään, että hän on syntynyt 1940-luvulla,

f) syntynyt 1940-luvulla, kun tiedetään, että hän on nainen,

g) syntynyt 1930- tai 1980-luvulla,

h) syntynyt 1930- tai 1980-luvulla, kun tiedetään, että hän on mies?

Huom. Muista mikroluokkaharjoitukset viikolla 9:

Harjoitusryhmät:

MA KLO 12.15 - 13.45 (M304)
MA KLO 14.15 - 15.45 (M304)
MA KLO 16.00 – 17.30 (M302) *(ryhmä suunnattu biologeille)*

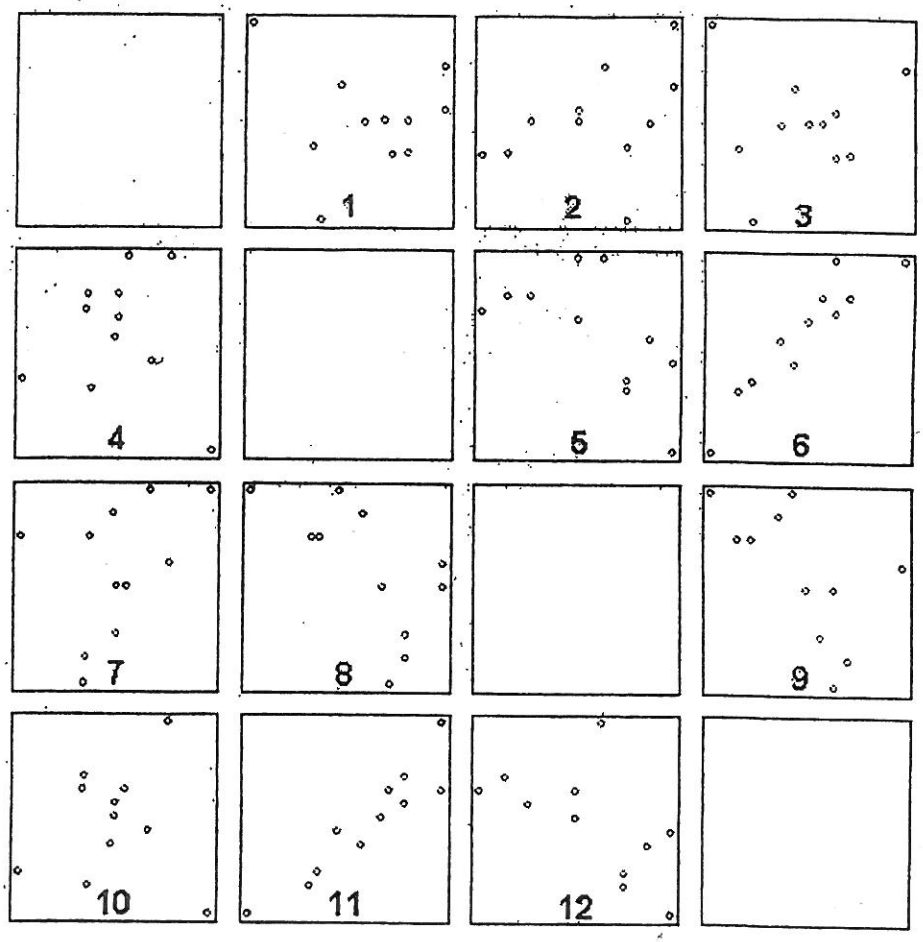
TI KLO 14.15 - 15.45 (M304)

KE KLO 12.15 - 13.45 (M302)

TO KLO 10.15 - 11.45 (M304)
TO KLO 14.15 - 15.45 (M302) *(ryhmä suunnattu biologeille)*
TO KLO 14.15 - 15.45 (M304) *(muut kuin biologit)*

PE KLO 08.15 - 9.45 (M304) *(ryhmä suunnattu biologeille)*
PE KLO 10.00 - 11.30 (M302) Huomaa aloitusaika!

LIITE 1



#keskiarvot:

```
> mean(juoksu.100m)
[1] 23.3182

> mean(juoksu.60m)
[1] 12.6636

> mean(pituushyppy)
[1] 2.1318
```

varianssit:

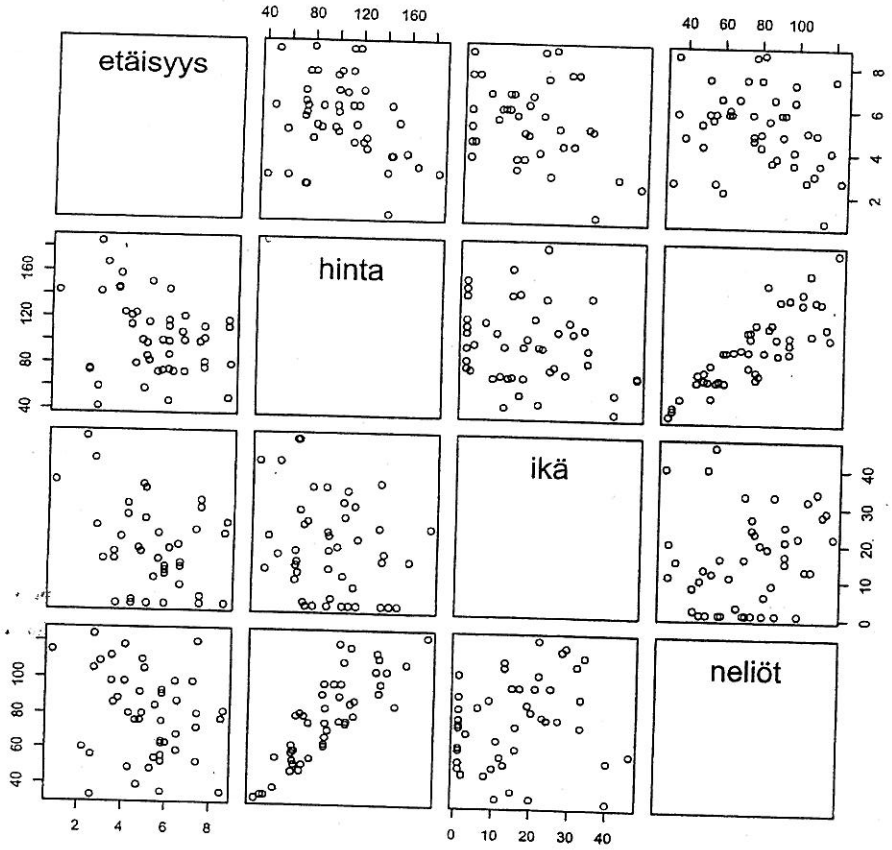
```
> var(juoksu.100m)
[1] 1.5796

> var(juoksu.60m)
[1] 0.1605

> var(pituushyppy)
[1] 0.0186
```

#korrelaatiomatriisi

```
> cor(neliottelu, use="complete.obs")
```

	juoksu.100m	juoksu.60m	pallonheitto	pituushyppy
juoksu.100m	1.0000	0.9328	-0.1067	-0.6302
juoksu.60m	0.9328	1.0000	0.0322	-0.5983
pallonheitto	-0.1067	0.0322	1.0000	0.3907
pituushyppy	-0.6302	-0.5983	0.3907	1.0000

LIITE 2



a)

```
> round(cor(rivitalot[,c("etäisyys","hinta","ikä","neliöt")],
use="complete.obs"),3)

         etäisyys  hinta    ikä neliöt
etäisyys    1.000 -0.285 -0.473 -0.218
hinta      -0.285  1.000 -0.218  0.850
ikä        -0.473 -0.218  1.000  0.139
neliöt     -0.218  0.850  0.139  1.000
```

```
> numSummary(rivitalot[,c("etäisyys", "hinta", "ikä", "neliöt")],
statistics=c("mean", "sd", "quantiles"), quantiles=c( 0,.25,.5,.75,1 ))

            mean        sd  0%    25%    50%     75%  100%  n
etäisyys   5.216  1.819549 0.8  4.100  5.50   6.300   8.6 50
hinta    100.606 33.225116 42.0 74.125 99.55 119.925 185.0 50
ikä       16.120 12.874021  1.0  3.750 14.00  23.750  46.0 50
neliöt    76.756 24.822004 33.0 56.475 77.00  96.750 124.0 50
```

b)

```
> malli <- lm(hinta~etäisyys+ikä+neliöt, data=rivitalot)

> summary(malli)

Call:
lm(formula = hinta ~ etäisyys + ikä + neliöt, data = rivitalot)

Residuals:
    Min      1Q  Median      3Q     Max
-17.886  -5.770  -2.727   5.180  23.068

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 65.94189    7.73437   8.526 5.02e-11
etäisyys    -6.08546    0.87227  -6.977 9.88e-09
ikä         -1.27212    0.12151 -10.469 9.28e-14
neliöt       1.13232    0.05691  19.898  < 2e-16
---

Residual standard error: 9.642 on 46 degrees of freedom
Multiple R-squared: 0.9209,     Adjusted R-squared: 0.9158
F-statistic: 178.6 on 3 and 46 DF,  p-value: < 2.2e-16
```