806109 TILASTOTIETEEN PERUSMENETELMÄT I
Harjoitus 7, viikko 9, kevät 2011
(Muut kuin taloustieteiden tiedekunnan opiskelijat)

MUISTA MIKROLUOKKAHARJOITUKSET VIIKOILLA 8 JA 9!

1. Eräässä suuressa yrityksessä haluttiin selvittää, onko työntekijän sukupuolella yhteyttä siihen, kannattaako työntekijä yrityksen johdon ehdottamaa uudistusta vai ei. Yrityksen koko henkilökunnan osalta tulokset olivat seuraavat:

Suhtautuminen	Sukupuoli		
uudistukseen	Mies	Nainen	Yhteensä
Kannattaa	60	65	125
Ei kannata	170	150	320
Yhteensä	230	215	445

Koska yrityksen henkilökunta oli jaettavissa kahteen eri organisaatioon tuloksia tarkasteltiin myös erikseen näillä organisaatiotasoilla (A ja B) ja saatiin seuraavat taulukot:

Organisaatio A:

Suhtautuminen	Sukupuoli		
uudistukseen	Mies	Nainen	Yhteensä
Kannattaa	10	45	55
Ei kannata	20	90	110
Yhteensä	30	135	165

Organisaatio B:

Suhtautuminen	Sukupuoli		
uudistukseen	Mies	Nainen	Yhteensä
Kannattaa	50	20	70
Ei kannata	150	60	210
Yhteensä	200	80	280

Tutki ristitulosuhteen ja ehdollisten prosenttijakaumien avulla sukupuolen ja uudistukseen suhtautumisen välistä riippuvuutta

a) koko aineistossa,

b) erikseen eri organisaatiotasoilla.

Mitä päätelmiä voit näistä tuloksista tehdä?

2. Jatkoa harjoituksen 6 tehtävään 4: Sähkölämmitteisen loma-asunnon sähkön kulutusta ja ulkoilman lämpötilaa seurattiin seitsemän vuorokautta. Tällöin saatiin seuraavat havainnot:

vuorokausi:	1	2	3	4	5	6	7
Sähkön kulutus (kWh):	32	28	23	21	30	28	22
Ulkoilman lämpötila (^{o}C):	5	8	12	10	-1	3	7

a) Sovita aineistoon regressiosuora $y = a + bx$, missä y = sähkön kulutus ja x = ulkoilman lämpötila. Tulkitse regressiokertoimet a ja b selväkielisesti. Määrää myös regressioyhtälön determinaatiokerroin eli selitysaste ja tulkitse se.

b) Paljonko regressioyhtälö ennustaa loma-asunnon lämmönkulutuksen olevan (=ennustearvo), jos ulkoilman lämpötila on 9 astetta?

3. Kevään 2010 välikoetehtävä: Erään pohjoisen luonnonkasvin ekofysiologiaa selvittävän kasvatuskokeen tulosten analysoinnissa saatiin R-ohjelmalla liitteessä 1 esitetty tulostus (osa tuloksista peitetty merkinnällä xx.xxx). Kokeessa mitatut muuttujat olivat kasvin sisältämä typpi (muuttuja N, milligrammoina) ja kasvin biomassa (muuttuja biomass, milligrammoina). Aineistosta laskettu Pearsonin tulomomenttikorrelaatiokerroin oli 0.85. Muuttujan N varianssi oli 0.0840 ja muuttujan biomass varianssi 40.4469.

a) Onko Pearsonin tulomomenttikorrelaatiokertoimen käyttäminen mielekästä/luvallista tälle aineistolle? Jos on, niin mistä syystä? Kumpi aineiston muuttujista on valittu regressioyhtälössä $y = a + bx$ vasteeksi ja kumpi selittäväksi muuttujaksi?

b) Määrää regressioyhtälön kertoimet a ja b ja tulkitse ne lyhyesti.

c) Määrää regressioyhtälön determinaatiokertoimen arvo (selitysaste) ja tulkitse se lyhyesti.

d) Laske mallin mukainen ennustearvo, kun kasvin biomassa on 40 milligrammaa ja typen määrä 1.8 milligrammaa.

4. 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 (grammoina) riippuu maksimipituudesta (muuttuja `pituus` senttimetreinä mitattuna suusta pyrstön päähän) ja korkeudesta (`korkeus`, senttimetreinä 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 2 on esitetty otteita R-ohjelman tulostuksesta. Käytä tulostuksen tietoja apunasi vastatessasi seuraaviin kysymyksiin.

a) Mallissa 1 vastemuuttujana on lahnan paino ja selittäjänä lahnan korkeus. Miksi lahnan korkeus näyttäisi olevan pituutta parempi selittäjä lahnan painolle?

b) Tulkitse regressioanalyysin tulokset, jotka liittyvät malliin 2 (määrää regressioyhtälö, kertoimien selväkielinen tulkinta, determinaatiokertoimen arvo ja sen tulkinta).

c) Erään lahnan paino on 500 grammaa, pituus 36.4 cm ja korkeus 13.8 cm. Ennusta regressioyhtälön 2 (malli2) avulla vastemuuttujan arvo ko. tilastoyksikölle. Laske myös vasteen todellisen arvon ja ennustetun arvon erotus eli residuaali.

5. Ville on päättänyt uusia talonsa ulkomaalaukset. Hän valitsee arpomalla sekä seinän värin että ikkunanpuitteiden värin. Seinän väriksi on tarjolla kolme vaihtoehtoa: rosa, sininen ja lila. Ikkunanpuitteiden väriksi tarjolla on joko valkoinen tai keltainen väri. Valittava ikkunanpuitteen väri ei riipu millään tavalla seinän väristä. Millä todennäköisyydellä talon

a) seinäväriksi tulee sininen?

b) seinäväriksi tulee rosa ja ikkunanpuitteiden väriksi valkoinen?

c) seinäväriksi tulee sininen tai ikkunanpuitteiden väriksi keltainen?

6. Ihmiset kuuluvat johonkin neljästä pääveriryhmästä A, B, AB ja O. Suomalaisten veriryhmäjakauma on seuraava: A 44%, B 17%, AB 8% ja O 31 %. Millä todennäköisyydellä suomalaisen avioparin aviopuolisot kuuluvat

a) A-veriryhmään,

b) samaan veriryhmään,

c) eri veriryhmiin?

7. 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.: Mikroluokkaharjoitusten harjoitusryhmät viikolle 9 ovat:

MA KLO 10.00–11.30 (M302)
MA KLO 13.15–14.45 (M302) (ryhmä suunnattu biologeille)
MA KLO 14.15–15.45 (M304)
TI KLO 8.30–10.00 (M304)
TI KLO 14.15–15.45 (M304) (ryhmä suunnattu biologeille)
KE KLO 10.15–11.45 (M304)
KE KLO 14.15–15.45 (M304)
TO KLO 8.15–9.45 (M302) (ryhmä suunnattu biologeille)
TO KLO 14.15–15.45 (M302)
PE KLO 12.15–13.45 (M304)

Vastauksia tehtäviin:

2. a) 30.6 -0.68 0.49 b) 24.5

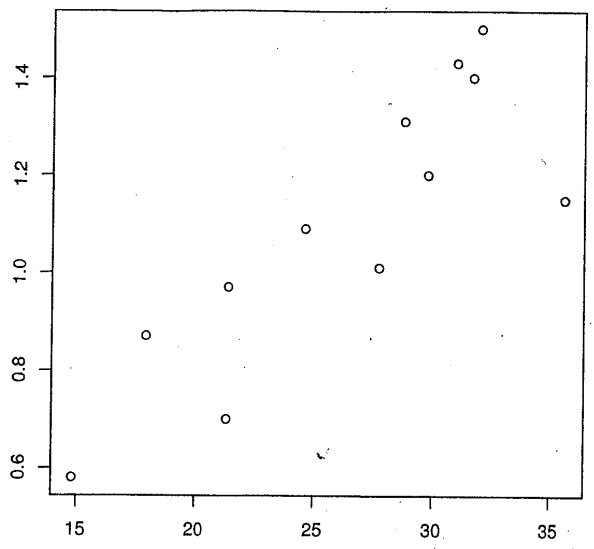
3. b) 0.078982 0.0387 c) 0.723 d) 1.6

4. c) 480

5. a) 0.3333 b) 0.1667 c) 0.6667

6. a) 0.1936 b) 0.3250 c) 0.6750

LIITE 1:




```
> summary(model)

Call:
lm(formula = N ~ biomass)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30731 -0.09022  0.06011  0.10151  0.18157

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.078982   0.206457   0.383 0.710052
biomass       xx.xxx   0.007616   5.079 0.000478 ***

Residual standard error: 0.1606 on 10 degrees of freedom
Multiple R-squared:  xx.xxx     , Adjusted R-squared: xx.xxx
F-statistic:  25.8 on 1 and 10 DF,  p-value: 0.0004783
```

LIITE 2:

```
>cor(lahna)  #korrelaatomatriisi
         paino  pituus korkeus
paino   1.0000  0.9639  0.9707
pituus  0.9639  1.0000  0.9540
korkeus 0.9707  0.9540  1.0000

> malli1 <- lm(paino~korkeus)
> summary(malli1)

Call:
lm(formula = paino ~ korkeus)

Residuals:
   Min     1Q Median     3Q    Max
 -62.8  -37.4   -9.9   19.0  106.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -913.94      67.91   -13.5    1e-14 ***
korkeus       101.16       4.42    22.9   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 50.4 on 32 degrees of freedom
Multiple R-squared: 0.942,      Adjusted R-squared: 0.941
F-statistic:  523 on 1 and 32 DF,  p-value: <2e-16

>
> malli2 <- lm(paino~korkeus+pituus)
> summary(malli2)

Call:
lm(formula = paino ~ korkeus + pituus)

Residuals:
   Min     1Q Median     3Q    Max
 -66.7  -27.2   -7.2   15.4   94.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1068.84      73.91  -14.46  2.5e-15 ***
korkeus        59.27      12.74    4.65  5.8e-05 ***
pituus         20.65       5.99    3.45   0.0017 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 43.5 on 31 degrees of freedom
Multiple R-squared: 0.958,      Adjusted R-squared: 0.956
F-statistic:  356 on 2 and 31 DF,  p-value: <2e-16
```