806109P TILASTOTIETEEN PERUSMENETELMÄT I
1. välikoe 4.3.2013 (Jari Päkkilä)

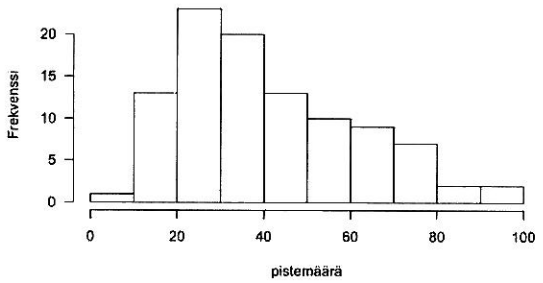
VALITSE VIIDESTÄ TEHTÄVÄSTÄ NELJÄ JA VASTAA VAIN NIIHIN!

1. Valitse kohdissa A–F oikea (vain yksi) vaihtoehto. Oikeasta vastauksesta saat 1 pistettä, väärästä et menetä pisteitä. Perusteluja ei vaadita.

A) Suuren lentoyhtiön johtoryhmä haluaa tutkia asiakkaittensa mielipidettä yhtiön palvelutasosta. Johtoryhmä valitsee yhtiön kaikkien lentojen joukosta satunnaisesti 30 lentoa, joiden matkustajat haastatellaan tutkimusta varten. Kyseessä on tällöin

a1) yksinkertainen satunnaisotanta, a2) ositettu otanta,
(a3) ryväsotanta. a4) systemaattinen otanta,
a5) kiintiöpoiminta. a6) harkintaotanta.

B) Alla olevassa histogrammissa on esitetty erään soveltuvuuskokeen pistemäärän jakauma.



Pistemäärän aritmeettiselle keskiarvolle $\bar{x}$, mediaanille Md ja vinoustunnusluvulle g_1 pätee, että

b1) $\bar{x} = 36.0$, $Md = 40.8$, $g_1 = -0.63$. b2) $\bar{x} = 36.0$, $Md = 40.8$, $g_1 = 0.63$.
b3) $\bar{x} = 40.8$, $Md = 36.0$, $g_1 = -0.63$, (b4) $\bar{x} = 40.8$, $Md = 36.0$, $g_1 = 0.63$.
b5) $\bar{x} = 36.0$, $Md = 40.8$, $g_1 = 0.00$, b6) $\bar{x} = 40.8$, $Md = 36.0$, $g_1 = 0.00$.

C) Laihdutusvalmisteen tehoa selvittävässä kokeellisessa tutkimuksessa koehenkilöt punnittiin (kg) sekä kokeen alussa että laihdutusvalmisteen kuukauden mittaisen käytön jälkeen. Havaittua painon muutosta kuvaava muuttuja on mitta-asteikoltaan

c1) luokitteluasteikkoa ja epäjatkuva, c2) järjestysasteikkoa ja epäjatkuva,
c3) välimatka-asteikkoa ja epäjatkuva, (c4) välimatka-asteikkoa ja jatkuva,
c5) suhdeasteikkoa ja epäjatkuva, c6) suhdeasteikkoa ja jatkuva.

D) Muuttujan x havaintoarvoista on laskettu $\bar{x} = 3$ ja $s_x = 3$. Havaintoaineistoon muodostetaan uusi muuttuja y siten, että $y = x + 1$. Tällöin muuttujalle y pätee, että

d1) $\bar{y} = 3$ ja $s_y^2 = 4$, d2) $\bar{y} = 4$ ja $s_y^2 = 3$, d3) $\bar{y} = 4$ ja $s_y^2 = 4$,
(d4) $\bar{y} = 4$ ja $s_y^2 = 9$, d5) $\bar{y} = 4$ ja $s_y^2 = 10$, d6) $\bar{y} = 4$ ja $s_y^2 = 16$.

E) 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

e1) jakauma voidaan esittää pistekuviona,
e2) jakauman moodi on 35,
e3) hajontaa voidaan kuvailla vaihteluvälin avulla,
e4) summajakauma on mielekäs muodostaa,
e5) havaintoarvot voidaan standardoida.
(e6) Mikään edellä esitetyistä kohdista e1)–e5) ei pidä paikkaansa.

F) Arvottujen lohkojen koejärjestelyssä

f1) lohkot jaetaan satunnaisesti eri käsittelyille,
f2) koeyksiköt jaetaan satunnaisesti eri lohkoihin,
f3) koeyksiköt valitaan satunnaisotannalla tarjolla olevasta perusjoukosta,
f4) lohkoja on aina yhtä monta kuin käsittelyjä,
(f5) jokaisessa lohkossa koeyksiköt jaetaan satunnaisesti eri käsittelyille,
f6) kunkin lohkon sisällä koeyksiköt ovat mahdollisimman heterogeenisia (erilaisia) sellaisten ominaisuuksien suhteen, joilla oletetaan olevan vaikutusta vastemuuttujaan.

2. Erään urheiluseuran nuorille järjestettiin pienimuotoiset urheilukilpailut, joiden yhtenä lajina oli vauhditon pituushyppy. Kisaan osallistuneiden poikien hyppyjen pituuksien (metreinä) frekvenssijakauma on seuraava:

Hyppytulos	Frekvenssi
2.00 - 2.09	5
2.10 - 2.19	7
2.20 - 2.29	12
2.30 - 2.39	10
2.40 - 2.49	3
Yhteensä	37

a) Laske poikien hyppytulosten aritmeettinen keskiarvo ja keskihajonta. *(2 p)*

b) Muodosta poikien hyppytulosten summajakauma ja esitä se graafisesti. Arvioi lisäksi, mikä on sellainen hypyn pituus, jonka ylitti 20 % pojista. *(2.5 p)*

(2)

b)

Hyppytulos	f	% f	F	% F	todelliset luokkarajat	luokka-keskus
2.00 – 2.09	5	14	5	14	$[1.995, 2.095[$	2.045
2.10 – 2.19	7	19	12	32 (tai 33)	$[2.095, 2.195[$	2.145
2.20 – 2.29	12	32	24	65	$[2.195, 2.295[$	2.245
2.30 – 2.39	10	27	34	92	$[2.295, 2.395[$	2.345
2.40 – 2.49	3	8	37	100	$[2.395, 2.495[$	2.445
Yhteensä	37	100				

a) Merk. X = (pojan) hyppytulos (metreinä)

Keskiarvo $\bar{X} = \frac{1}{n}\sum_{i=1}^{r} f_i X_i$, $n = 37$, $r = 5$, X_i = luokkakeskus (i:nnen luokan)

$\bar{X} = \frac{1}{37}\sum_{i=1}^{5} f_i X_i$ (Esim. $X_1 = \frac{1.995 + 2.095}{2} = 2.045$)

$= \frac{1}{37}(5 \cdot 2.045 + 7 \cdot 2.145 + 12 \cdot 2.245 + 10 \cdot 2.345 + 3 \cdot 2.445)$

$= \frac{82.965}{37} = 2.242297\ldots \approx 2.2423 \approx \underline{2.24\ \text{m}}$

Keskihajonta $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{r} f_i (X_i - \bar{x})^2} = \sqrt{\frac{1}{37-1}\sum_{i=1}^{5} f_i (X_i - \bar{x})^2}$

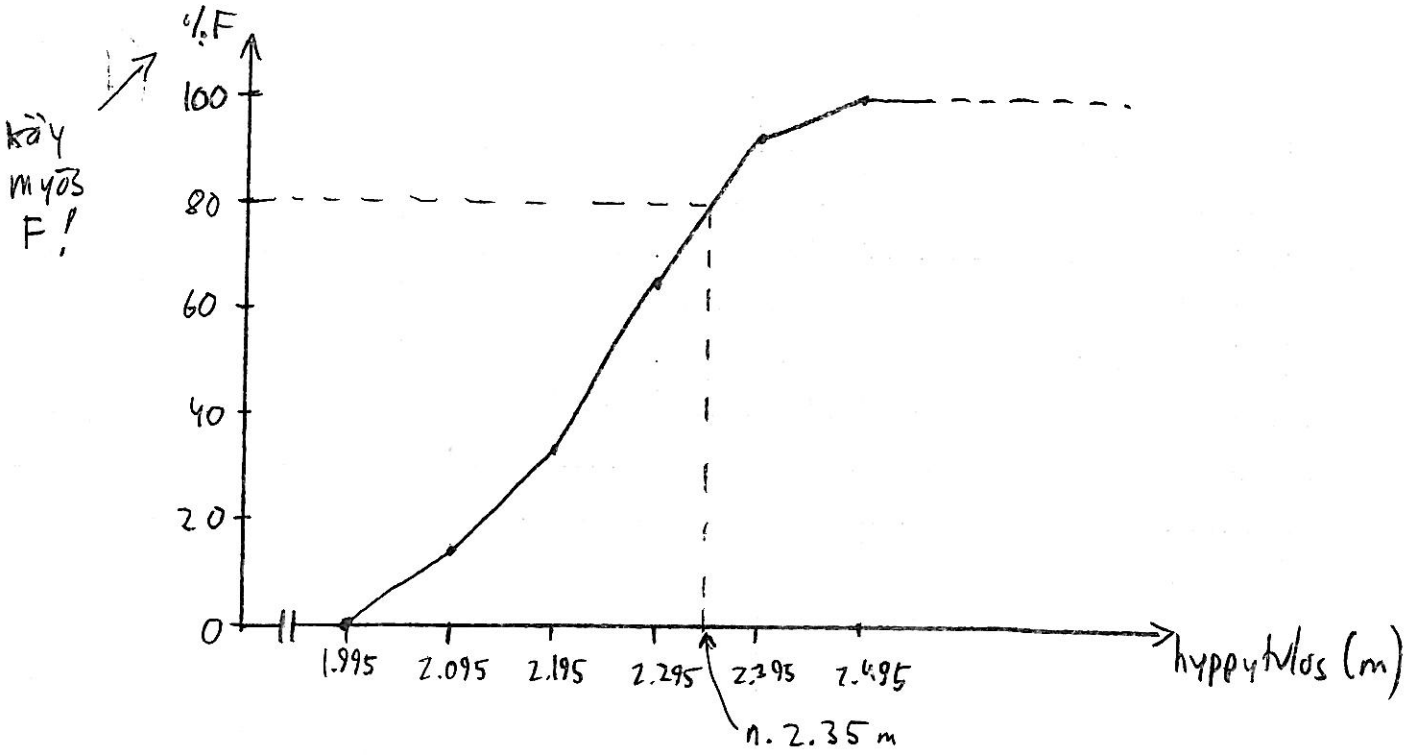
$= \sqrt{\frac{1}{37-1} \cdot [5 \cdot (2.045 - 2.2423)^2 + \cdots + 3 \cdot (2.445 - 2.2423)^2}$

$= \sqrt{\frac{1}{36} \cdot (0.19463645 + 0.06627103 + 0.00008748 + 0.1054729 + 0.12326187)}$

$= \sqrt{\frac{0.48972973}{36}} \approx \sqrt{0.0136036} = 0.11663\ldots \approx \underline{0.12\ \text{m}}$

b) Summajakauma on esitetty taulukossa tehtävän alussa!

Hyppytulos on jatkuva muuttuja, joten summajakauman graafinen esitys on summakäyrä.



20% pojista hyppäsi yli (noin) 2.35 metriä.

c) Merk. $\begin{cases} X = \text{koriheittojen lukumäärä (viiden heiton sarjassa)} \\ X_p = \text{--11--} \quad \text{--11--} \quad \text{pojilla} \\ X_t = \text{--11--} \quad \text{--11--} \quad \text{tytöillä} \end{cases}$

$n_p = 37$, $n_t = 19$, $n = 56$, $\bar{X}_t = 2.74$, $\bar{X}_p = ?$, $\bar{X} = ?$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{6} f_i X_i = \frac{1}{56}\sum_{i=1}^{6} f_i X_i$$

$$= \frac{1}{56}\cdot(7\cdot 0 + 9\cdot 1 + 18\cdot 2 + 12\cdot 3 + 5\cdot 4 + 5\cdot 5)$$

$$= \frac{126}{56} = 2.25$$

Palhotettu kaava: $\frac{n_p\cdot\bar{X}_p + n_t\cdot\bar{X}_t}{n} = \bar{X} \quad | \cdot n$

$$\Leftrightarrow n_p\cdot\bar{X}_p + n_t\cdot\bar{X}_t = n\cdot\bar{X} \quad | - n_t\cdot\bar{X}_t$$

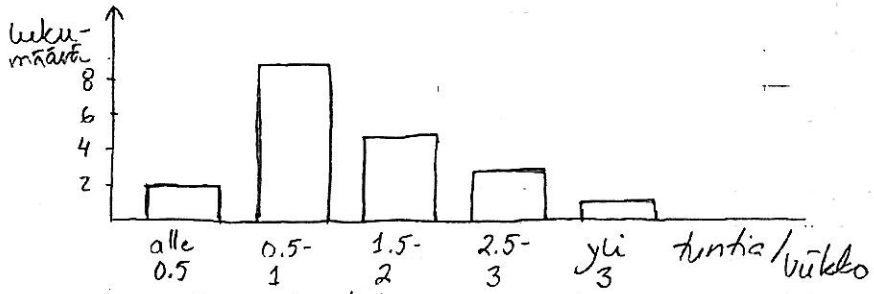
$$\Leftrightarrow n_p\cdot\bar{X}_p = n\bar{X} - n_t\bar{X}_t \quad | : n_p$$

$$\Leftrightarrow \bar{X}_p = \frac{n\bar{X} - n_t\cdot\bar{X}_t}{n_p} = \frac{56\cdot 2.25 - 19\cdot 2.74}{37} = 1.998378\ldots \approx \underline{2.00}$$

$\Rightarrow$ Poikien koriheittojen keskiarvo oli (noin) 2.00.

3. a1)

Tottelevaisuus-koulutukseen käytetty aika (h/vk)	luku-määrä	prosentti-osuus
alle 0.5	2	10
0.5 - 1	9	45
1.5 - 2	5	25
2.5 - 3	3	15
yli 3	1	5
Yhteensä	20	100



Erään koiraryhdistyksen jäsenten koiran tottelevaisuuskoulutukseen käyttämät viikkotunnit.

a2) tokoaika on järjestysasteikon muuttuja $\Rightarrow$ lasketaan moodi ja mediaani sekä vaihteluväli ja kvartiliväli

Moodi = 0.5 - 1 h/vk
Mediaani = 0.5 - 1 h/vk
vaihteluväli = [alle 0.5 , yli 3]
kvartiliväli
= [0.5 - 1 h/vk , 1.5 - 2 h/vk]

b1) Mg-pitoisuuden alakvartiili = 39.0 mg/l ja mediaani = 43.0 mg/l, joten 32:sta havainnosta tuolla välillä on

8.

b2) $x = Ca$, $y = Mg$

$r_{xy} = 0.48$ (R:in tulostuksesta)

b3) $x = Sato$

$$z = \frac{x_i - \bar{x}}{s_x} = \frac{27.0 - 28.46}{\sqrt{35.95}} \approx -0.2435$$

(tarkin -0.244233)

4. a)

Maakunta	Hukkakauraa ei	Hukkakauraa on	Yhteensä
Etelä.-Pohj.	87	13	100
Pohjanmaa	72	28	100
Pohjois-Pohj.	98	2	100

$100 \times 4143/5000 = 86.82$ $(659/5800) \cdot 100 = 13.18$
$100 \times 3581/4970 = 72.05$ $(1389/4970) \cdot 100 = 27.95$
$100 \times 5297/5400 = 98.09$ $(103/5400) \cdot 100 = 1.91$

Tulkinta: Hukkakauran esiintyvyys vaihtelee eri maakunnissa. (Muuttujien välillä on riippuvuutta.) Pohjanmaalla hukkakauraa on 28 prosentille peltohehtaareista, kun taas Pohjois-Pohjanmaalla vain 2 prosentilla.

b) odotusarvot: $e_{ij} = \frac{f_{i\cdot} f_{\cdot j}}{n}$

4300	699
4274	695
4644	755

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{r} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

$$= \frac{(4341 - 4300)^2}{4300} + \ldots + \frac{(103 - 755)^2}{755}$$

$$= 0.386 + 112.502 + 91.735 + 2.372 + 691.38 + 563.757$$

$$\approx 1462$$

$$C/C_{max} = \sqrt{\frac{\chi^2}{n+\chi^2}} \Big/ \sqrt{\frac{q-1}{q}} = \sqrt{\frac{1462}{15370+1462}} \Big/ \sqrt{\frac{1}{2}} = 0.4168$$

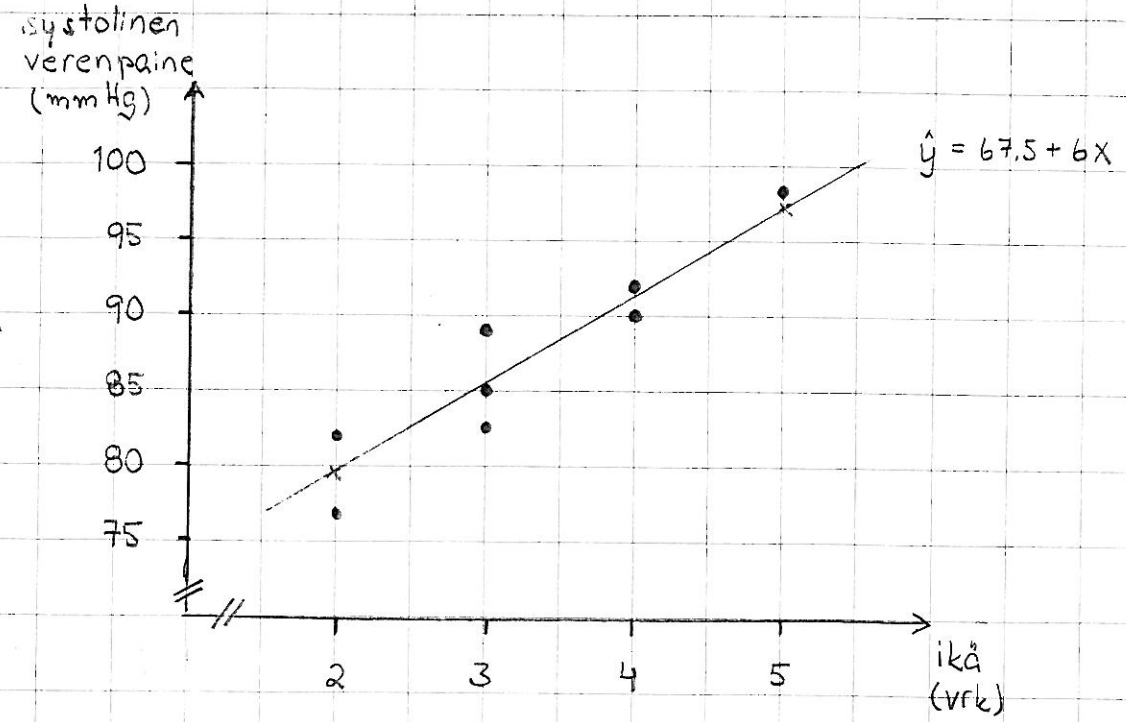
$$\text{Cramerin } V = \sqrt{\frac{\chi^2}{nq}} = \sqrt{\frac{1462}{15370 \cdot (2-1)}} = 0.3084$$

Tulkinta: Hukkakauraesiintymän ja maakunnan välillä vallitsee kohtalainen riippuvuus.

(myös melko heikko hyväksyttä ja jonkin verran riippuvuutta)

5.

a)



Vastasyntyneen lapsen iän ja systolisen verenpaineen välillä vallitsee vahva positiivinen lineaarinen riippuvuus

b) Korrelaatiokerroin r mittaa lineaarista riippuvuutta (todellu yliä). Lisäksi rin käytiö edellyttää, että tarkasteltavat muuttujat ovat vähintään välimatka-asteikkoja. Nyt molemmat muuttujat ovat suhdeasteikon muuttujia.

c) $y = a + bx$, missä y = systolinen verenpaine ja
x = lapsen ikä

Rin tulostuksesta: $b = \underline{6}$

$a = \bar{y} - b\bar{x}$ $\quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 3{,}25$ ja $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 87$

$= 87 - 6 \cdot 3{,}25 = \underline{67{,}5}$

$\Rightarrow \underline{\hat{y} = 67{,}5 + 6x}$

Kertoimien tulkinnat:

b: Kun vastasyntyneen ikä kasvaa yhdellä vuorokaudella, nousee systolinen verenpaine keskimäärin 6 mmHg.

a: Kun vastasyntyneen ikä on nolla vuorokautta, on systolinen verenpaine keskimäärin 67.5 mmHg.

Determinaatiokerroin $\hat{R}^2 = \underline{\underline{0.8882}}$ ← R-tulostus

$\hat{R}^2$:n tulkinta: regressiomallilla (eli vastasyntyneen iällä) voidaan selittää noin 88.8 % systolisen verenpaineen kokonaisvaihtelusta.

d) Lasketaan regressioyhtälön $\hat{y} = 67.5 + 6x$ avulla kaksi sovitearvoa:

$x = 2$: $\hat{y} = 67.5 + 6 \cdot 2 = 79.5$

$x = 5$: $\hat{y} = 67.5 + 6 \cdot 5 = 97.5$

Merkitään edellä saadut koordinaattipisteet (2, 79.5) ja (5, 97.5) a)-kohdan hajontakuvioon ja yhdistetään pisteet suoralla.