Luentoharjoitus: kontingenssikertoimen laskemisesta:

Onko lempi juomalla ja kotikunnalla yhteyttä?

Tehtiin otos ja tulokseksi saatiin seuraava aineisto:

	Vesi	Cola	Maito
Turku	50	70	70
Pori	200	60	80
Helsinki	100	250	120

Lasketaan Odotetut frekvenssit:

ODOTETUT FREKVENSSIT (kaavat)

	Vesi	Cola	Maito	
Turku	190*350/1000	190*380/1000	190*270/1000	190
Pori	340*350/1000	340*380/1000	340*270/1000	340
Helsinki	470*350/1000	470*380/1000	470*270/1000	470
	350	380	270	1000

ODOTETUT FREKVENSSIT (Tulokset)

	Vesi	Cola	Maito	
Turku	66.5	72.2	51.3	190
Pori	119	129.2	91.8	340
Helsinki	164.5	178.6	126.9	470
	350	380	270	1000

Lasketaan jäännökset eli havaittujen ja odotettujen havaintojen erotus:

JÄÄNNÖKSET (Kaavat)

	Vesi	Cola	Maito
Turku	50-66.5	70-72.2	70-51.3
Pori	200-119	60-129.2	80-91.8
Helsinki	100-164.5	250-178.6	120-126.9

JÄÄNNÖKSET (Tulokset)

	Vesi	Cola	Maito
Turku	**-16.5**	**-2.2**	**18.7**
Pori	**81**	**-69.2**	**-11.8**
Helsinki	**-64.5**	**71.4**	**-6.9**

Lasketaan Khiin neliösuure:

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

$$\chi^2 = \frac{(-16.5)^2}{66.5} + \frac{(-2.2)^2}{72.2} + \ldots + \frac{(-6.9)^2}{126.9} = 158.90$$

Tai hyödynnetään taulukointia:

f_{ij}	e_{ij}	$(f_{ij} - e_{ij})^2$	$(f_{ij} - e_{ij})^2 / e_{ij}$
50	66.5	272.3	4.1
70	72.2	4.8	0.1
70	51.3	349.7	6.8
200	119	6561.0	55.1
60	129.2	4788.6	37.1
80	91.8	139.2	1.5
100	164.5	4160.3	25.3
250	178.6	5098.0	28.5
120	126.9	47.6	0.4
			158.9

Khiin neliösuure normeerataan ja saadaan C eli kontingenssikerroin:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{158.90}{158.90 + 1000}} = 0.370$$

Kontingenssikertoimen laskennallinen maksimiarvo:

$$C_{MAX} = \sqrt{\frac{q-1}{q}} = \sqrt{\frac{3-1}{3}} = 0.816$$

Lasketaan muihin verrannollinen Kontingenssikertoin:

$\frac{C}{C_{MAX}}$ **= 0.370/0.816 = 0.453**

Tulkinta: 0.453 > 0.3 eli riippuvuutta on.

Riippuvuutta löytyy, tilannetta kuvaa riviprosenttijakauma :

RIVIPROSENTIT

	Vesi	Cola	Maito	
Turku	26 %	37 %	37 %	100 %
Pori	59 %	18 %	24 %	100 %
Helsinki	21 %	53 %	26 %	100 %

Porissa suositaan vettä, Helsingissä Colaa, Turussa kolmen juoman suosio on melko tasainen, joskin cola ja maito ovat vettä suositumpia.

Pearson esimerkki:

Onko iän ja Tilin saldon välillä korrelaatiota?

n=7

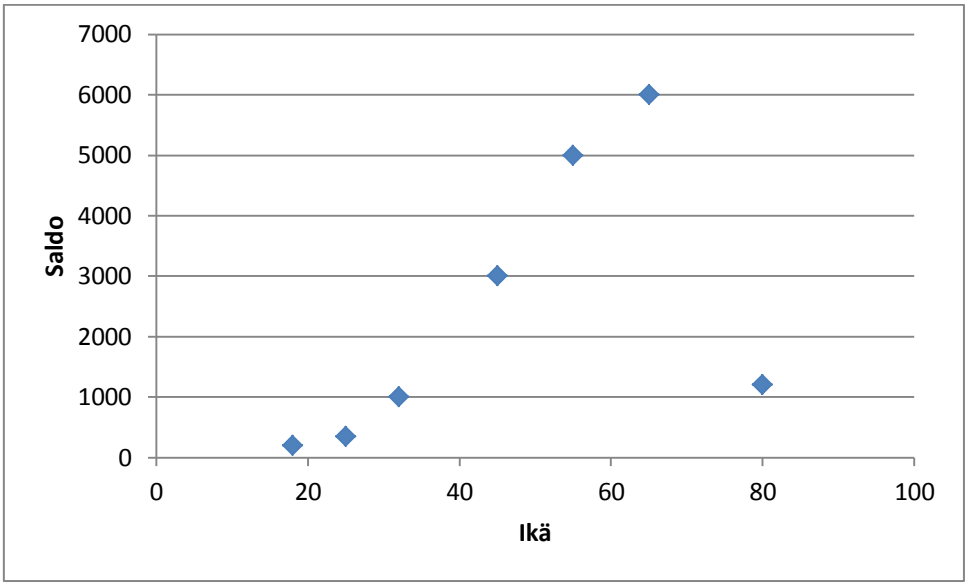
Ikä (X)	Tilin saldo (Y)	X^2	Y^2	XY
18	200	324	40000	3600
25	350	625	122500	8750
32	1000	1024	1000000	32000
45	3000	2025	9000000	135000
55	5000	3025	25000000	275000
65	6000	4225	36000000	390000
80	1200	6400	1440000	96000
320	**16750**	**17648**	**72602500**	**940350**

$$r_{XY} = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i) * (\sum_{i=1}^{n} y_i)}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2\right] * \left[n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2\right]}}$$

$$r_{XY} = \frac{\mathbf{7*940350-320*16750}}{\sqrt{(\mathbf{7*17648-320^2})*(\mathbf{7*72602500-16750^2})}} = 0{,}557$$

Tulkinta: Kohtalainen positiivinen lineaarinen riippuvuus iän ja tilin saldon välillä välillä.

Huom! Alussa voidaan myös kuvaajan avulla silmäillä onko yhteys lineaarista:



Näyttää siltä, että noin 20 -65v. yhteys on lineaarista, mutta 80-vuotiaan kohdalla on poikkeava havainto. Tämä voitaisiin poistaa aineistosta, koska ehkä eläkeläisillä ymmärrettävästi säästöt eivät nouse vaan ovat jo alkaneet laskea. Korrelaatiokerroin todennäköisesti kasvaisi jos eläkeläinen tiputettaisiin pois laskuista.

Luentotehtävä: laske muuttujien välinen Pearson korrelaatio:

Tutkitaan vaikuttaako tv-mainoksen pituus siihen kuinka pitkään keskimäärin mainos muistetaan.

X=Mainoksen pituus	Y= Muistitestin pistem.
20	10
24	8
28	10
32	11
36	14
40	16
44	12
48	13

Pearson-korrelaatio, esimerkki:

	X=Mainoksen pituus	Y= Muistitestin pistem.	x^2	y^2	x*y
	20	10	400	100	200
	24	8	576	64	192
	28	10	784	100	280
	32	11	1024	121	352
	36	14	1296	196	504
	40	16	1600	256	640
	44	12	1936	144	528
	48	13	2304	169	624
SUM	**272**	**94**	**9920**	**1150**	**3320**

n = 8

$$r_{XY} = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i) * (\sum_{i=1}^{n} y_i)}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2\right] * \left[n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2\right]}}$$

$$r_{XY} = \frac{8*3320-272*94}{\sqrt{(8*9920-272^2)*(8*1150-94^2)}} = 0.71$$

Tulkinta: Voimakas positiivinen lineaarinen riippuvuus mainoksen pituuden ja muistitestin pistemäärän välillä.

Tästäkin tehtävästä voitaisiin piirtää kuva ja varmistaa, että yhteys on lineaarista. Tässä yhteys näyttää melko lineaariselta, eikä poikkeavia havaintoja silmämääräisen tarkastelun avulla havaita.