[MTTTP1] TILASTOTIETEEN JOHDANTOKURSSI, kevät 2019
https://coursepages.uta.fi/mtttp1/kevat-2019/

HARJOITUS 3

Joitain ratkaisuja

1. $\bar{x} = (8+9+6+7+10)/5 = 8$, $s^2 = ((8-8)^2 + (9-8)^2 + (6-8)^2 + (7-8)^2) + (10-8)^2)/4 = 10/4$, $s \approx 1{,}58$.

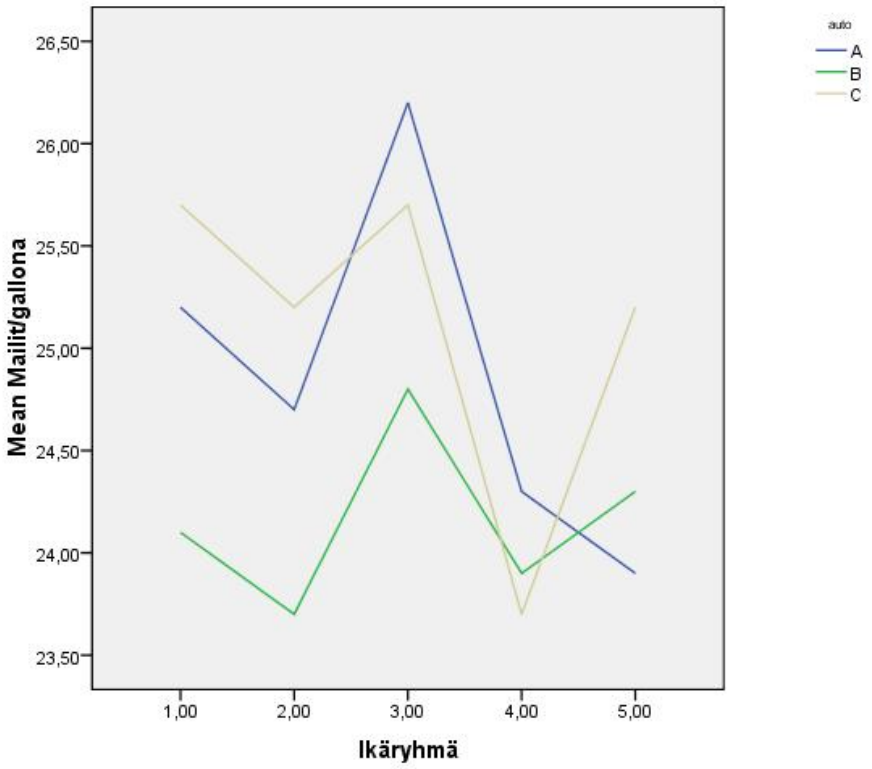
palamisaika	standardoitu arvo
8	$(8-8)/1{,}58 = 0$
9	$(9-8)/1{,}58 = 0{,}63$
6	$(6-8)/1{,}58 = -1{,}26$
7	$(7-8)/1{,}58 = -0{,}63$
10	$(10-8)/1{,}58 = 1{,}26$

Näistä standardoiduista arvoista laskettu keskiarvo on nolla ja varianssi 1, näin on standardoidulle muuttujalle aina.

2. Suoritetaan standardointi. Standardoidut arvot ovat
$z_1 = (-9{,}4 - 14{,}4583)/11{,}82505 \approx -2{,}02$,
$z_2 = (-7{,}8 - 17{,}1)/11{,}6666 \approx -2{,}14$,
joten -7,8 pistettä saanut on menestynyt huonommin ollen 2,14 hajonnan päässä ryhmänsä keskiarvon alapuolella. Tässä ratkaisussa ryhmittely on tehty samanaikaisesti sukupuolen ja opetustavan mukaan. Voi tietysti tarkastella vain opetustavan tai sukupuolen ryhmissä.

3. $\Sigma x_i = 261$ (ks. harj. 2 teht. 4) $\bar{x} \approx 3{,}7$, $s^2 \approx ((0-3{,}7)^2 + (1-3{,}7)^2 \cdot 6 + (2-3{,}7)^2 \cdot 10 + \ldots + (9-3.7)^2)/69 \approx 3{,}27$, $s \approx 1{,}8$, $\bar{x} + 1{,}8 = 5{,}5$; $\bar{x} - 1{,}8 = 1{,}9$.
Välille 1,9–5,5 jää 50 havaintoa (≈ 71 %). Vrt. normaalijakauma, jossa vastaava osuus 68 %.

4.



Keskiarvot muuttuvat jokin verran eri tavalla autotyypeittäin mentäessä ikäryhmästä toiseen, murtoviivat ”käyttäytyvät” eri tavalla. Keskimäärin vähiten kulutusta on autotyypillä A ikäryhmässä 3. Ikäryhmittäin on eroja keskimääräisissä kulutuksissa. Ikäryhmässä 3 kulutus keskimäärin pienintä, ikäryhmässä 4 suurinta. Autotyypeittäin on myös eroja keskimääräisissä kulutuksissa, autotyyppi B kuluttaa keskimäärin eniten.

Edellä tehtiin päätelmät vain kuvailevan analyysin keinoin, varsinainen tilastollinen päättely voitaisiin tehdä varianssianalyysin avulla (opintojakson MTTTA1 asiaa).

5.
a)
Ehdolliset

keskiarvot	5,11	9,72
mediaanit	5	9

(ks. SPSS-tulos alla)

b) Menetelmä näyttäisi paljastavan kopioijat, jotka tekevät keskimäärin enemmän virheitä. Voit verrata myös jakaumia. Jos frekvenssijakaumat haluttaisiin esittää graafisesti, niin esitykset olisivat histogrammeja, jotka piirrettäisiin samoilla luokituksilla ja prosentuaalisista frekvensseistä. Voi myös käyttää laatikko-jana-kuviota, ks. alla.

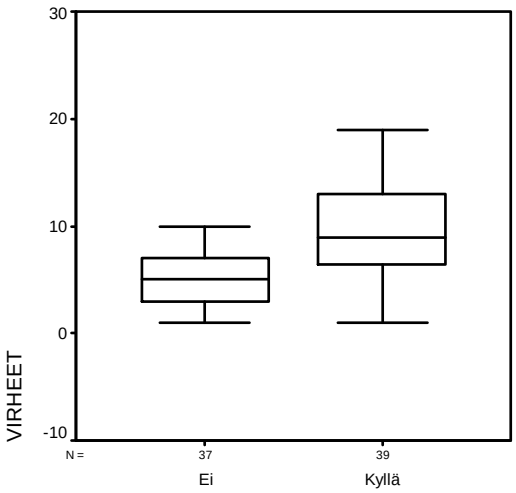
SPSS-tulostus ehdollisista tunnusluvuista

Report

VIRHEET

Kopiointi?	Mean	N	Std. Deviation	Median
Ei	5,11	37	2,777	5,00
Kyllä	9,72	39	4,696	9,00
Total	7,47	76	4,500	7,00

Laatikko-jana-kuvio



Edellä tehtiin päätelmät vain kuvailevan analyysin keinoin, varsinainen tilastollinen päättely voitaisiin tehdä luottamusvälin tai testin avulla (esillä myöhemmin opintojaksolla).

6. Selitettävä = tupakointi, selittäjä = sukupuoli

Koska ehdolliset prosenttijakaumat näyttävät poikkeavan toisistaan, niin tupakointikäyttäytyminen saattaisi olla erilaista miehillä ja naisilla (testattaessa p = 0,028).

		sukupuoli		
		mies	nainen	Total
tupakointi	ei koskaan polttanut	30,95	52,75	49,84
	entinen tupakoija	52,38	34,07	36,51
	nykyinen tupakoija	16,67	13,19	13,65
Total		100,00	100,00	100,00

Ei koskaan polttaneita on 0,4984x315 = 157, entisiä tupakoitsijoita on 0,3651x315 = 115 ja nykyisiä tupakoitsijoita 0,1365x315 = 43.

Olkoon x = miesten lukumäärä, tällöin naisten lukumäärä on 315-x. Ryhmästä ”ei koskaan polttanut” saadaan 0,3095x+0,5275(315-x)=157, josta x = 42.

Tämän jälkeen kaikki frekvenssit on laskettavissa. Saadaan

tupakointi * sukupuoli Crosstabulation

Count

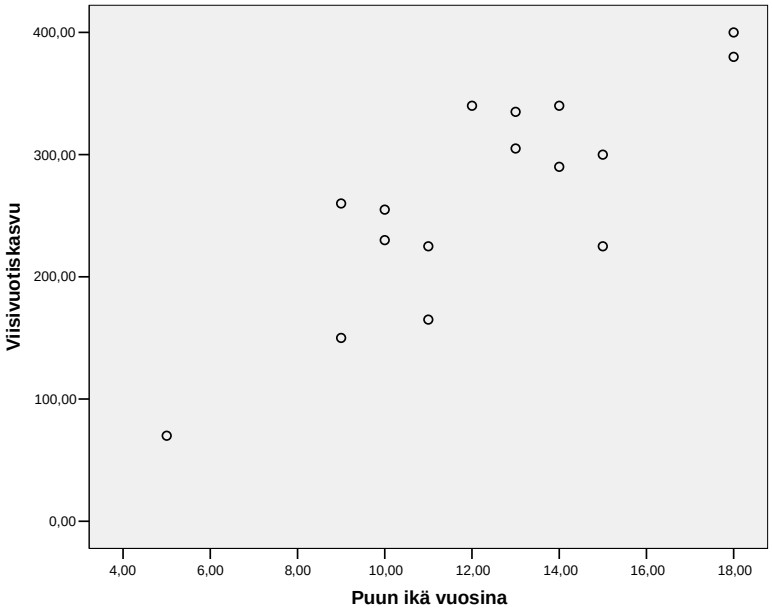
		sukupuoli		
		mies	nainen	Total
tupakointi	ei koskaan polttanut	13	144	157
	entinen tupakoija	22	93	115
	nykyinen tupakoija	7	36	43
Total		42	273	315

7. Selitettävä = Nuoren tupakointi, selittäjä = Vanhempien tupakointi

	Opiskelija polttaa	Opiskelija ei polta	yht.
Molemmat vanhemmat polttavat	100*400 / 1780 ≈22	100*1380 / 1780≈78	1780
Toinen vanhemmista polttaa	100*416 / 2239 ≈ 19	100*1823 / 2239≈81	2239
Vanhemmat eivät polta	100* 188 / 1356≈14	100*1168 / 1356≈86	1356
Yht.	1004	4371	5375

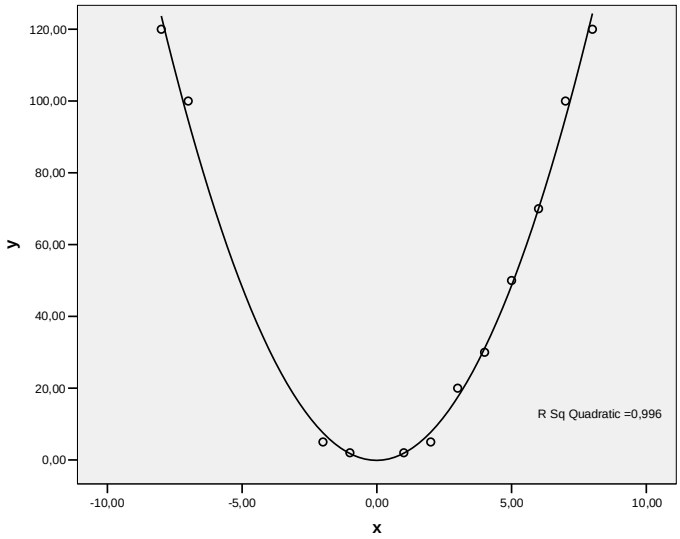
Koska ehdolliset prosenttijakaumat näyttävät poikkeavan toisistaan, niin riippuvuutta saattaisi olla (testattaessa p <0,0001). Näyttäisi siis siltä, että vanhempien tupakointitavat vaikuttavat lastensa tupakointiin. Esimerkiksi jos vanhemmat eivät polta, niin tarkasteltavassa aineistossa heidän lapsistaan poltti 14 %. Vastaava luku perheissä, jossa molemmat vanhemmat polttivat, oli 22 %. Toki koululaisten tupakointiin vaikuttaa monet muutkin tekijät!!

8. a)



Riippuvuus näyttää olevan lineaarista, koska pisteet ovat melko hyvin keskittyneet suoran ympärille. Koska pisteet ovat ryhmittyneet nousevan suoran ympärille, on kyse positiivisesta lineaarisesta riippuvuudesta. Korrelaatiokerroin 0,826. Jos ensimmäisen puun kohdalla viisivuotiskasvu olisi 700, niin korrelaatiokerroin olisi -0,133.

b)



Kyse hyvin voimakkaasta riippuvuudesta, joka ei kuitenkaan ole lineaarista. Pisteparveen voidaan sovittaa toisen asteen polynomi, joka kuvaa riippuvuutta.