MTTTA1 Tilastomenetelmien perusteet
Luento 17.1.2019

Kertausta ja täydennystä 1-VA

$H_0: \mu_1 = \mu_2 = \ldots = \mu_I$

H_1: kaikki μ:t eivät samoja

Oletetaan riippumattomat otokset:

$Y_{11}, Y_{12}, \ldots, Y_{1n1}$ satunnaisotos $N(\mu_1, \sigma_1^2)$:sta
$Y_{21}, Y_{22}, \ldots, Y_{2n2}$ satunnaisotos $N(\mu_2, \sigma_2^2)$:sta
.
.
.
$Y_{I1}, Y_{I2}, \ldots, Y_{InI}$ satunnaisotos $N(\mu_I, \sigma_I^2)$:sta

Oletetaan lisäksi, että $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2 = \sigma^2$.

Neliösummat:

$$\underbrace{\sum\sum(Y_{ij} - \bar{Y})^2}_{SST} = \underbrace{\sum n_i(\bar{Y}_i - \bar{Y})^2}_{SSB} + \underbrace{\sum\sum(Y_{ij} - \bar{Y}_i)^2}_{SSW}$$

vaihtelu	neliö-summat (SS)	vapaus-asteet (df)	keskineliö-summat (MS)	F-arvo	p-arvo
välinen	SSB	$I-1$	$MSB = \frac{SSB}{I-1}$	$F = \frac{MSB}{MSW}$	$P(F \geq F_{\text{hav.}})$
sisäinen (jäännös)	SSW	$n-I$	$MSW = \frac{SSW}{n-I}$	$\sim F(I-1, n-I)$ kun H_0 tosi	
kokonais	SST	$n-1$			

ks. kaavakokoelma
http://www.sis.uta.fi/tilasto/mttta1/kevat2019/kaavat.pdf

Esim. 2.1.4 Tutkitaan autotyyppien A, B, ja C kulutusta (miles per gallon), http://www.sis.uta.fi/tilasto/mttta1/kevat2019/luentorunko.pdf#page=15

A-autot	B-autot	C-autot
22.2	24.6	22.7
19.9	23.1	21.9
20.3	22.0	23.3
21.4	23.5	24.1
21.2	23.6	22.1
21.0	22.1	23.4
20.3	23.5	

	n	Mean	Std. Deviation
A	7	20.9000	.79162
B	7	23.2000	.90921
C	6	22.9167	.84004
Total	20	22.3100	1.33610

Test of Homogeneity of Variances

Kulutus (miles/gallon)

Levene Statistic	df1	df2	Sig.
,036	2	17	,965

$H_0: \sigma_A^2 = \sigma_B^2 = \sigma_C^2$

H_0 hyväksytään, koska p-arvo > 0,05

ANOVA

Kulutus (miles/gallon)

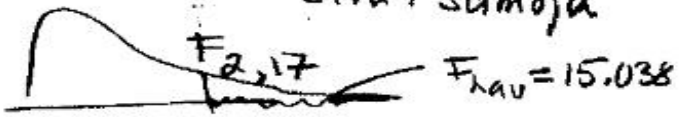
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	SSB 21,670	I-1 2	MSB 10,835	15,038	,000
Within Groups	SSW 12,248	n-I 17	MSW ,720	MSB/MSW	
Total	33,918	19			

$H_0: \mu_A = \mu_B = \mu_C$

H_1: kaikki odotusarvot eivät samoja

Post Hoc Tests

I = 3, n = 20



$6{,}11 = F_{0.01,2,17}$

H_0 hylätään

Multiple Comparisons

Kulutus (miles/gallon)
Bonferroni

(I) Autot yyppi	(J) Autot yyppi	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
A	B	-2,3000*	,4537	,000	-3,505	-1,095
	C	-2,0167*	,4722	,002	-3,270	-,763
B	A	2,3000*	,4537	,000	1,095	3,505
	C	,2833	,4722	1,000	-,970	1,537

*. The mean difference is significant at the 0.05 level.

B ja C eivät eroa toisistaan

Esim. 2.1.1 Tutkitaan golfpallojen keskimääräisiä lentomatkoja, saadaan tulokset:

Merkki	Keskiarvo	Keskihajonta	Lukumäärä
A	251,28	5,977	10
B	261,06	3,866	10
C	269,95	4,501	10

H_0: $\mu_A = \mu_B = \mu_C$

H_1: kaikki μ:t eivät samoja

$\bar{y}$= 260,76, n = 30, I = 3, $n_1 = n_2 = n_3 = 10$

$$SSW = \sum\sum (Y_{ij} - \bar{Y}_i)^2 = (n_1 - 1)s_1^2 + \cdots + (n_I - 1)s_I^2$$

$SSW = (10 - 1)5{,}977^2 + (10 - 1)3{,}866^2 + (10 - 1)4{,}501^2$
$= 638{,}36$

$$SSB = \sum n_i(\bar{Y}_i - \bar{Y})^2$$

$SSB = 10(251{,}28\text{-}260{,}76)^2 + 10(261{,}06\text{-}260{,}76)^2 +$
$10(269{,}95\text{-}260{,}76)^2 = 1744{,}17$

$MSB = SSB/(I-1)$

$MSB = 1744,17/2 = 872,08$

$MSW = SSW/(n-I)$

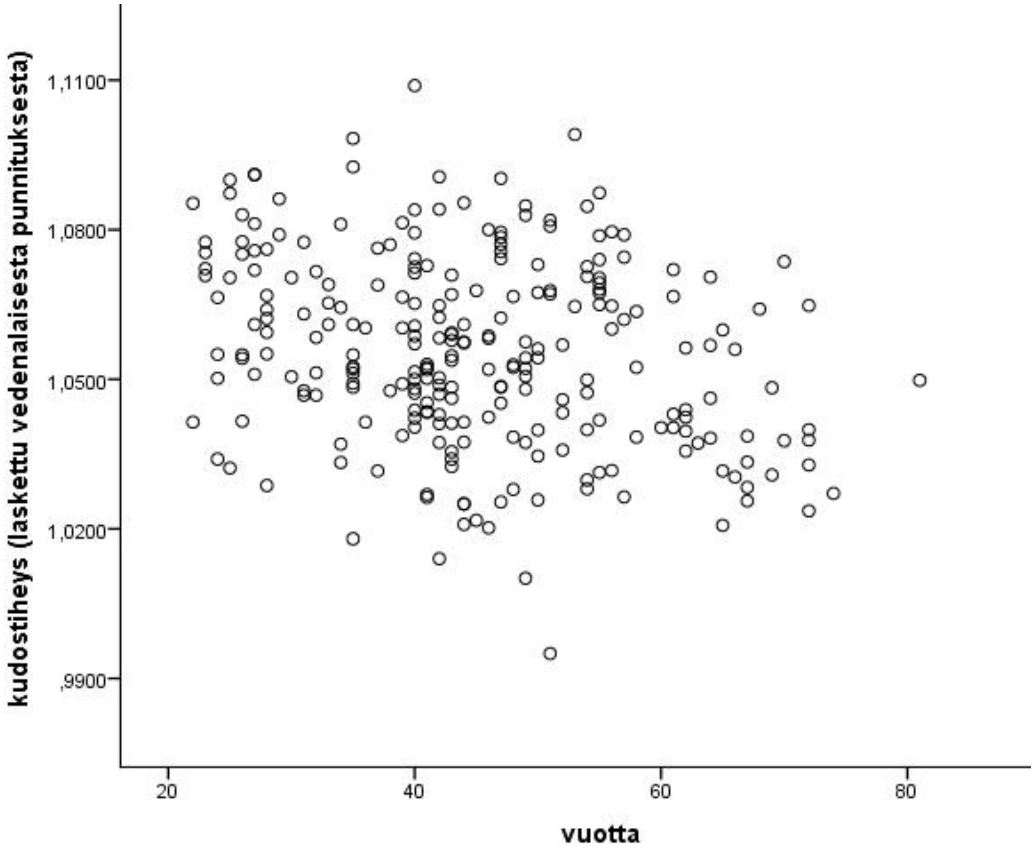
$MSW = 638,36/27 = 23,64$

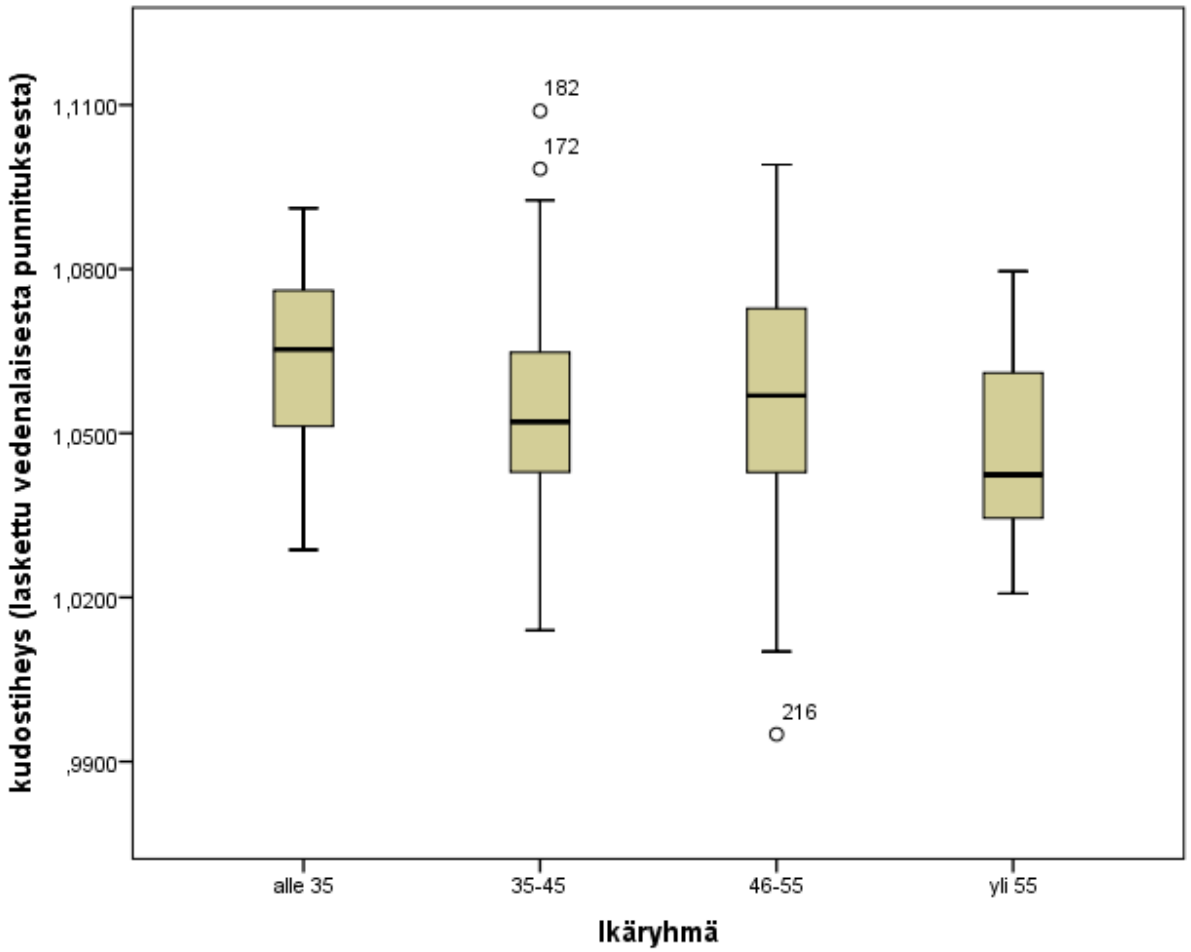
$F = MSB/MSW \sim F_{I-1,\ n-I}$, kun H_0 tosi

$F_{hav.} = 872,08/23,64 = 36,87 > F_{0,01;2,27} = 5,49$, joten H_0 hylätään 1 %:n riskitasolla. Päätellään odotusarvoissa olevan eroja. Voidaan sanoa, että p-arvo = $P(F_{2,27} > 36,87) < 0,01$.

Esim. Miehillä iän vaikutus kudostiheyteen

Aineisto rasvaprosentti.sav sivulta
https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/




kudostiheys (laskettu vedenalaisesta punnituksesta)
1,1100
1,0800
1,0500
1,0200
,9900
182
172
216
alle 35
35-45
46-55
yli 55
Ikäryhmä

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1: kaikki odotusarvot eivät samoja

ANOVA

kudostiheys (laskettu vedenalaisesta punnituksesta)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	,007	3	,002	7,247	,000
Within Groups	,084	248	,000		
Total	,091	251			

Koska p-arvo <0,001, niin H_0 hylätään, päätellään eroja olevan. Monivertailusta huomataan, että kaikkien ikäryhmien välillä ei kuitenkaan ole eroja.

Multiple Comparisons

Dependent Variable: kudostiheys (laskettu vedenalaisesta punnituksesta)
Bonferroni

(I) Ikäryhmä	(J) Ikäryhmä	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
alle 35	35-45	,0098402*	,0032131	,015	,001295	,018386
	46-55	,0072854	,0033748	,191	-,001690	,016261
	yli 55	,0168150*	,0036783	,000	,007032	,026598
35-45	alle 35	-,0098402*	,0032131	,015	-,018386	-,001295
	46-55	-,0025547	,0029992	1,000	-,010531	,005422
	yli 55	,0069748	,0033371	,226	-,001900	,015850
46-55	alle 35	-,0072854	,0033748	,191	-,016261	,001690
	35-45	,0025547	,0029992	1,000	-,005422	,010531
	yli 55	,0095295*	,0034930	,041	,000240	,018820
yli 55	alle 35	-,0168150*	,0036783	,000	-,026598	-,007032
	35-45	-,0069748	,0033371	,226	-,015850	,001900
	46-55	-,0095295*	,0034930	,041	-,018820	-,000240

*. The mean difference is significant at the 0.05 level.

Test of Homogeneity of Variances

kudostiheys (laskettu vedenalaisesta punnituksesta)

Levene Statistic	df1	df2	Sig.
1,254	3	248	,291

Populaatioiden varianssit voitiin olettaa samoiksi (H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ hyväksytään, koska p-arvo 0,291>0,05), joten varianssianalyysin käyttö sallittua.

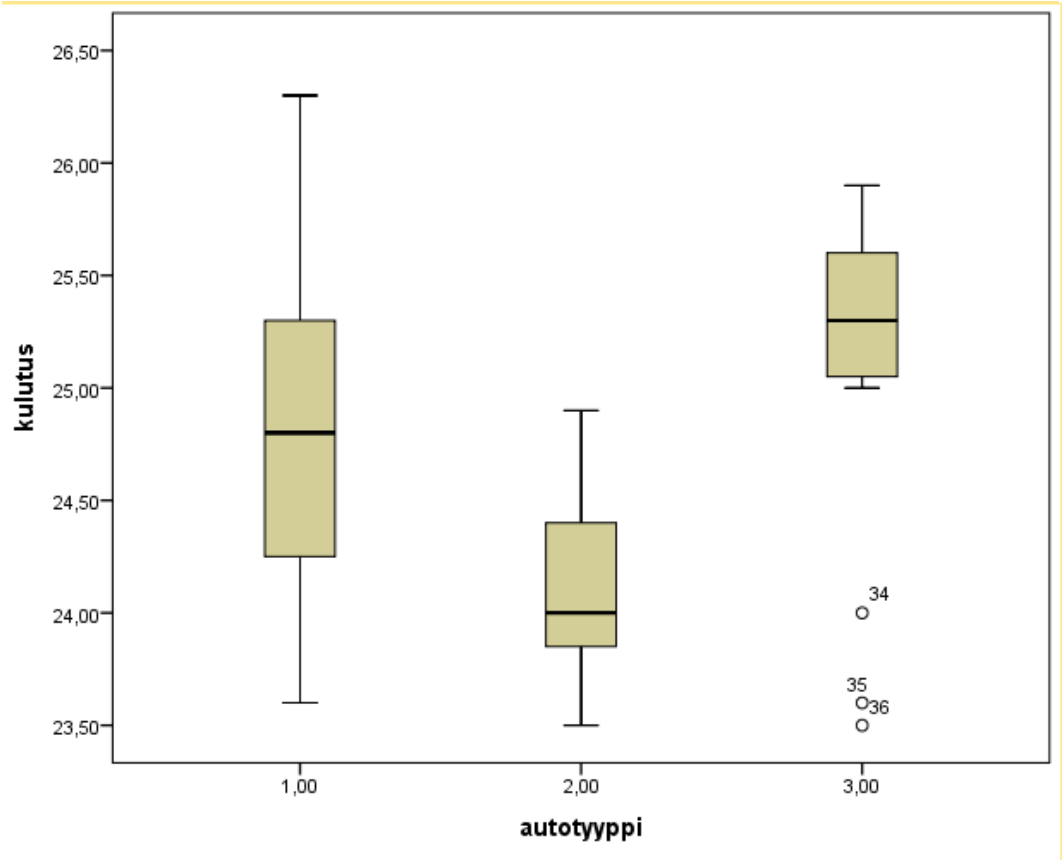
Varianssianalyysi nettilaskurilla:
http://vassarstats.net/ - > ANOVA ->
http://vassarstats.net/anova1u.html

2.2 Kaksisuuntainen varianssianalyysi

Esim. Tutkitaan kolmen autotyypin polttoaineen kulutusta (kulutus = mailit/gallona) huomioiden kuljettajan ikä (5 ikäryhmää), aineisto http://www.sis.uta.fi/tilasto/tiltp3/kevat2004/Aineistoja/autotNB2va.sav

Tehdään aluksi yksisuuntaiset varianssianalyysit.

y = kulutus
x = autotyyppi



Test of Homogeneity of Variances

KULUTUS

Levene Statistic	df1	df2	Sig.
2,302	2	42	,113

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2$

H_0 hyväksytään, koska $p = 0,113 > 0,05$

ANOVA

KULUTUS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	7,156	2	3,578	7,186	,002
Within Groups	20,912	42	,498		
Total	28,068	44			

$H_0: \mu_1 = \mu_2 = \mu_3$

H_0 hylätään, koska $p < 0,01$, kaikki odotusarvot eivät samoja

Post Hoc Tests

Multiple Comparisons

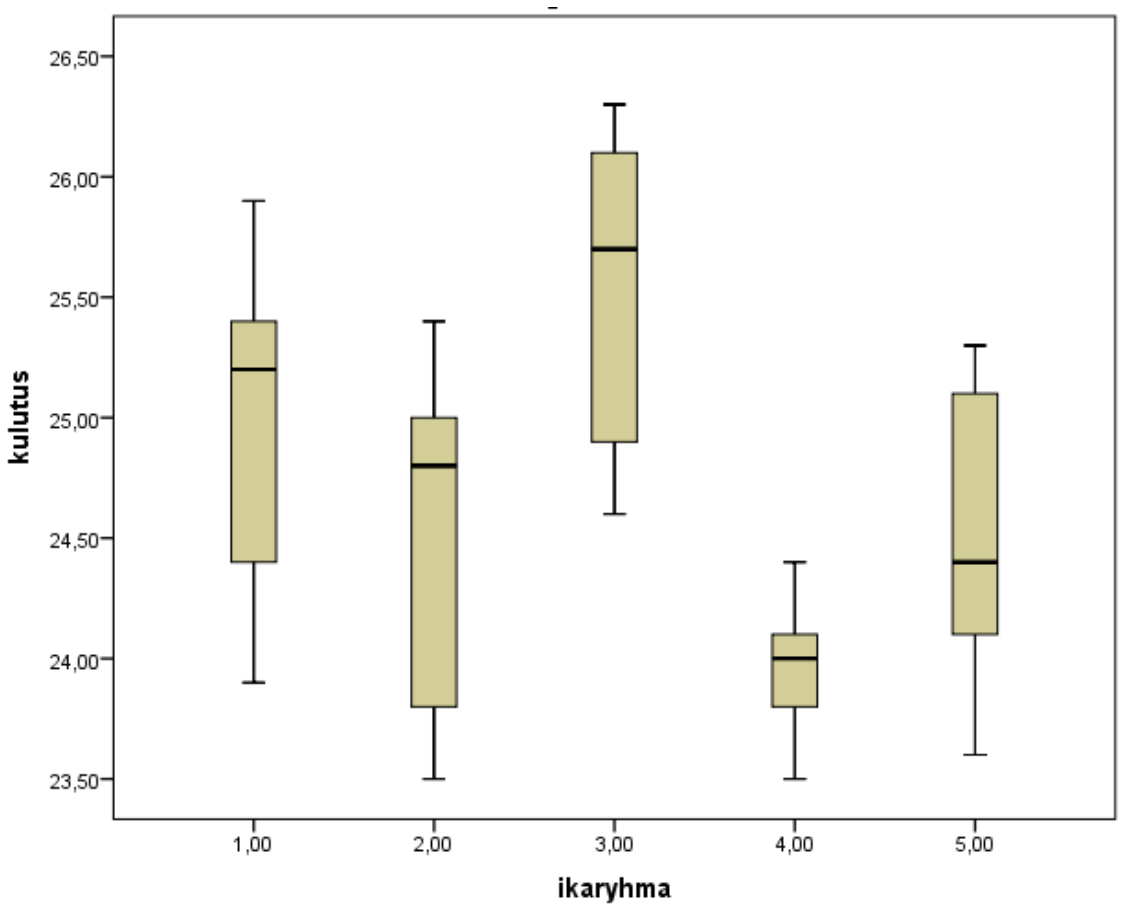
Dependent Variable: KULUTUS
Bonferroni

Autotyyppien 1 ja 2 sekä 2 ja 3 odotusarvot erisuuret

(I) AUTO	(J) AUTO	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1,00	2,00	,7000*	,25766	,029	,0575	1,3425
	3,00	-,2400	,25766	1,000	-,8825	,4025
2,00	1,00	-,7000*	,25766	,029	-1,3425	-,0575
	3,00	-,9400*	,25766	,002	-1,5825	-,2975
3,00	1,00	,2400	,25766	1,000	-,4025	,8825
	2,00	,9400*	,25766	,002	,2975	1,5825

*. The mean difference is significant at the .05 level.

y= kulutus
x = ikäryhmä



Oneway

Test of Homogeneity of Variances

KULUTUS

Levene Statistic	df1	df2	Sig.
2,016	4	40	,111

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$

H_0 hyväksytään, koska $p = 0,111 > 0,05$

ANOVA

KULUTUS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	13,148	4	3,287	8,812	,000
Within Groups	14,920	40	,373		
Total	28,068	44			

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_0 hylätään, päätellään, että kaikki odotusarvot ei samoja

Post Hoc Tests

Multiple Comparisons

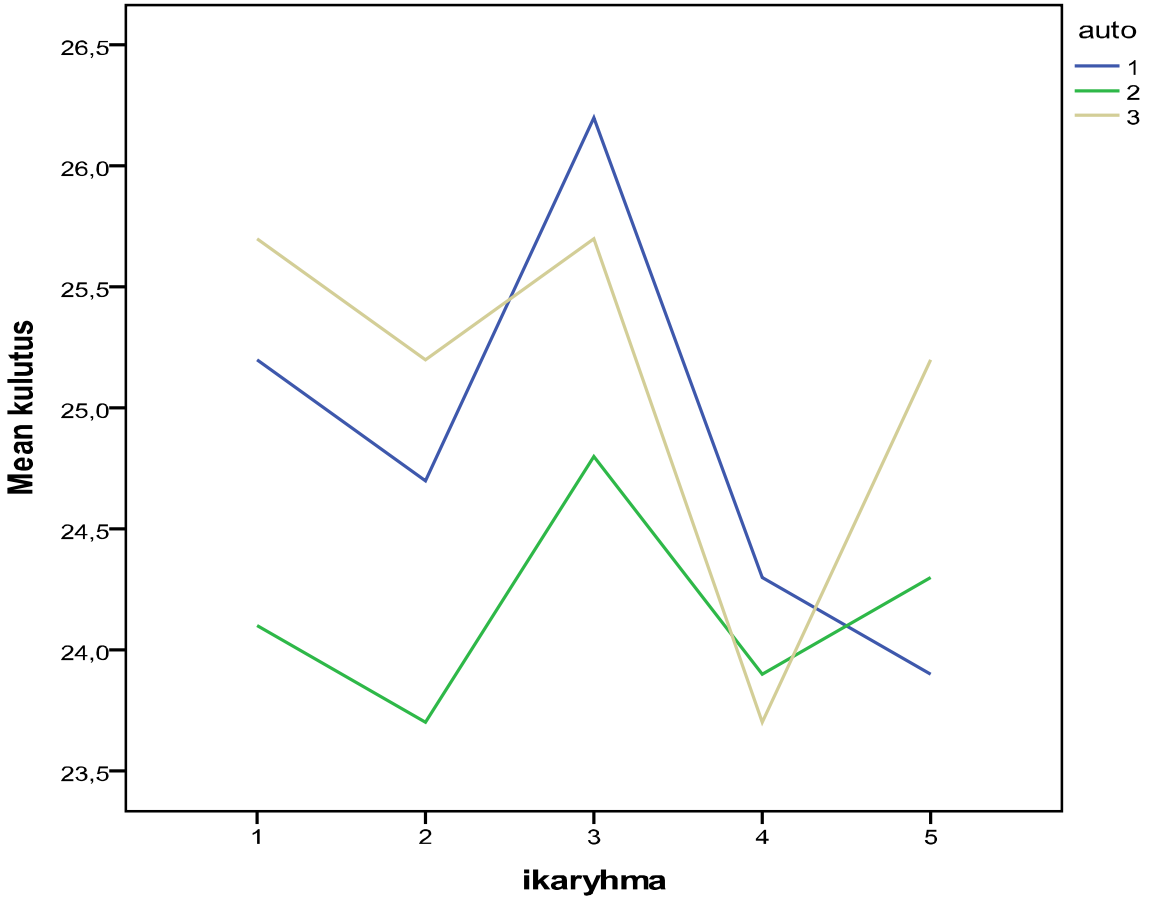
Dependent Variable: KULUTUS
Bonferroni

(I) IKARYHMA	(J) IKARYHMA	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1,00	2,00	,4667	,28790	1,000	-,3887	1,3221
	3,00	-,5667	,28790	,560	-1,4221	,2887
	4,00	1,0333*	,28790	,009	,1779	1,8887
	5,00	,5333	,28790	,713	-,3221	1,3887
2,00	1,00	-,4667	,28790	1,000	-1,3221	,3887
	3,00	-1,0333*	,28790	,009	-1,8887	-,1779
	4,00	,5667	,28790	,560	-,2887	1,4221
	5,00	,0667	,28790	1,000	-,7887	,9221
3,00	1,00	,5667	,28790	,560	-,2887	1,4221
	2,00	1,0333*	,28790	,009	,1779	1,8887
	4,00	1,6000*	,28790	,000	,7446	2,4554
	5,00	1,1000*	,28790	,005	,2446	1,9554
4,00	1,00	-1,0333*	,28790	,009	-1,8887	-,1779
	2,00	-,5667	,28790	,560	-1,4221	,2887
	3,00	-1,6000*	,28790	,000	-2,4554	-,7446
	5,00	-,5000	,28790	,901	-1,3554	,3554
5,00	1,00	-,5333	,28790	,713	-1,3887	,3221
	2,00	-,0667	,28790	1,000	-,9221	,7887
	3,00	-1,1000*	,28790	,005	-1,9554	-,2446
	4,00	,5000	,28790	,901	-,3554	1,3554

*. The mean difference is significant at the .05 level.

Eroja ikäryhmien 1 & 4, 2 & 3, 3 & 4, 3 & 5 välillä

Kulutuksen ehdolliset keskiarvot ryhmitellen sekä ikäryhmän että autotyypin mukaan



Nyt

y= kulutus

x_1 = autotyyppi

x_2 = ikäryhmä

Suoritetaan kaksisuuntainen varianssianalyysi. Halutaan selvittää miten autotyyppi ja ikäryhmä yhdessä vaikuttavat kulutukseen. Tutkitaan autotyypin ikäryhmästä riippumatonta vaikutusta (omavaikutusta), ikäryhmän autotyypistä riippumatonta vaikutusta (omavaikutusta) sekä autotyypin ja ikäryhmän yhdysvaikutusta.

Ks.
http://www.sis.uta.fi/tilasto/mttta1/kevat2019/luentorunko.pdf#page=21

Tests of Between-Subjects Effects

Dependent Variable: kulutus

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	26,908[a]	14	1,922	49,707	,000
Intercept	27468,872	1	27468,872	710401,862	,000
auto	7,156	2	3,578	92,534 F_A	,000
ikaryhma	13,148	4	3,287	85,009 F_B	,000
auto * ikaryhma	6,604	8	,826	21,349 F_{AB}	,000
Error	1,160	30	,039		
Total	27496,940	45			
Corrected Total	28,068	44			

a. R Squared = ,959 (Adjusted R Squared = ,939)

yhtysvaikutus

H_0: ei yhdysvaikutusta

H_1: on -"-

$F_{hav} = 21,349$, $p < 0,001$

H_0 hylätään

Päätellään: ikäryhmittäin kuljettajien väliset erot erilaiset eri autotyypeillä. Myös molemmilla selittäjillä on omavaikutusta (p-arvot <0,001).

SPSS-ohjeet

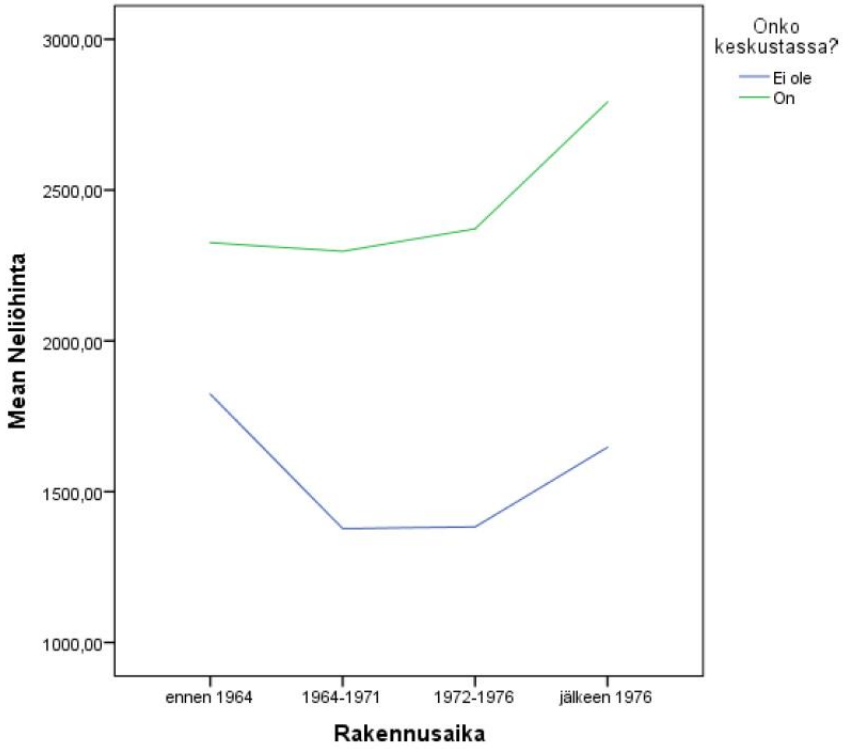
Ehdolliset keskiarvot graafisesti

Graphs-> Line-> Multiple-> Variable -> kulutus-> Category Axis-> ikaryhma -> Define line by->auto

2-VA

General Linear Model -> Univariate -> Dependent -> kulutus-> Fixed Factors ->auto, ikaryhma, Model -> auto, ikaryhma, interaction...

Esim. Rakennusajan ja sijainnin vaikutus keskineliöhintaan, SPSS-monisteen
http://www.uta.fi/sis/reports/index/R55_2017.pdf
esimerkki 19



Tests of Between-Subjects Effects

Dependent Variable: Neliöhinta

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	51473967,7[a]	7	7353423,957	67,519	,000
Intercept	699593758,9	1	699593758,9	6423,680	,000
Onko keskustassa?	34440038,88	1	34440038,88	316,229	,000
Rakennusaika	4400240,537	3	1466746,846	13,468	,000
Onko keskustassa * Rakennusaika	2298538,654	3	766179,551	7,035	,000
Error	24068790,94	221	108908,556		
Total	907041110,3	229			
Corrected Total	75542758,64	228			

a. R Squared = ,681 (Adjusted R Squared = ,671)

Aineisto
http://www.sis.uta.fi/tilasto/tiltp_aineistoja/Asunnot_2006.sav
sivulta https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/