MTTTA1 Tilastomenetelmien perusteet
Luento 15.1.2019

Luku 2

Varianssianalyysi

2.1 Yksisuuntainen varianssianalyysi

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

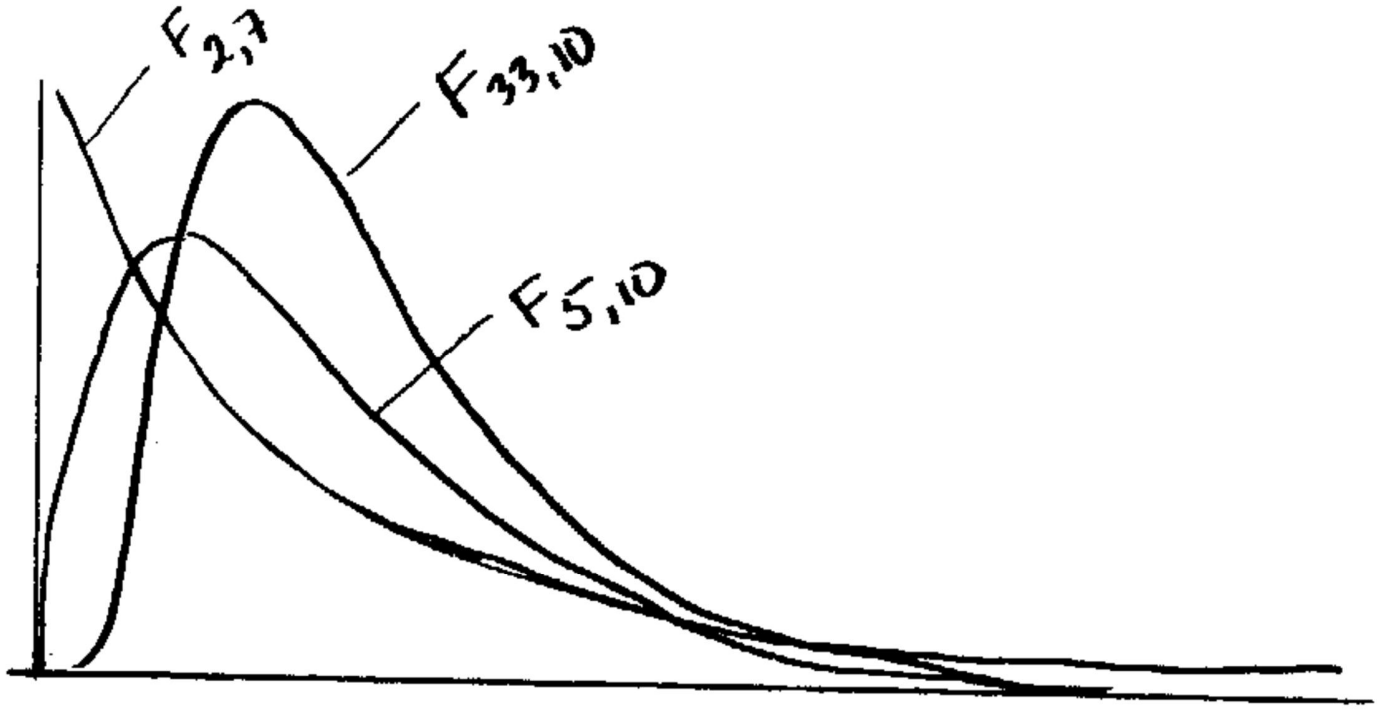
H_0: $\mu_A = \mu_B = \mu_C$
H_1: kaikki μ:t eivät samoja

F-testisuure H_0:n testaamiseksi

Annettujen lukujen perusteella voidaan laskea testisuureelle arvo, saadaan $F_{hav.} = 36{,}87$ ja p-arvo $< 0{,}0001$.

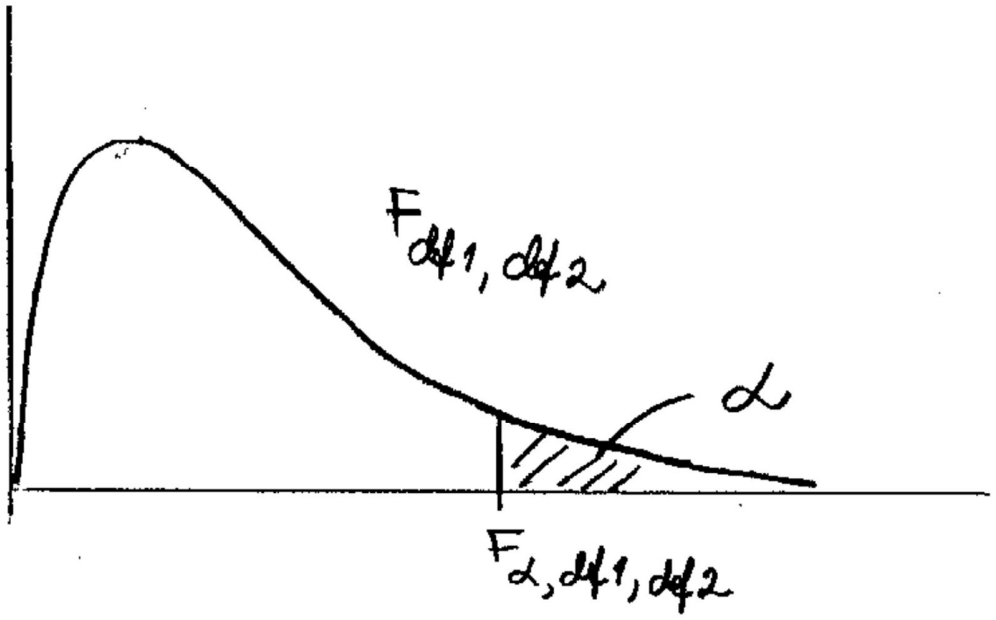
Hylätään H_0 ja päätellään odotusarvoissa olevan eroja.

Fisherin F-jakauman tiheysfunktion kuvaajia



F-jakauma määritellään kaksin vapausastein, $F_{df1,df2}$

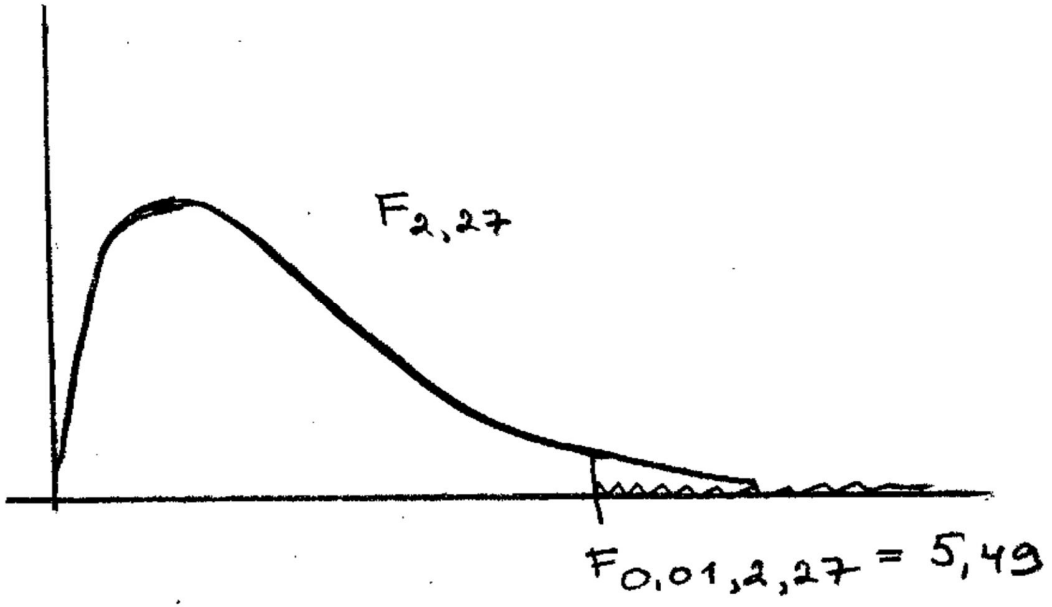
Määritellään $F_{\alpha;df1,\,df2}$ siten, että $P(F_{df1,df2} > F_{\alpha;df1,df2}) = \alpha$.



Näitä arvoja taulukosta
http://www.sis.uta.fi/tilasto/mttta1/kevat2019/F_jakauma.pdf,
kun α = 0,01 tai α = 0,05.

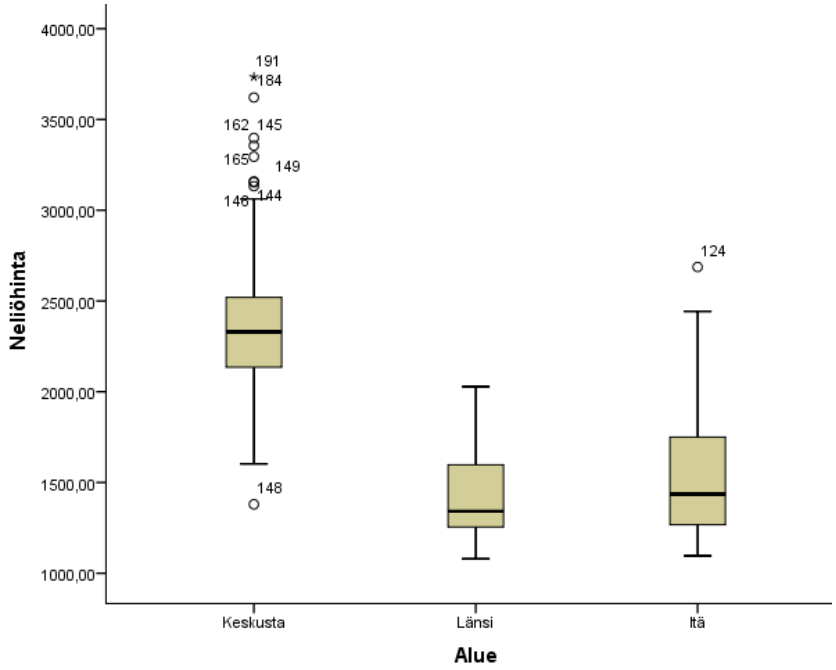
Esim. 2.1.1 Testisuure noudattaa H_0:n ollessa tosi F-jakaumaa vapausastein 2 ja 27.

$F_{0,01;2,27} = 5,49 < F_{hav.} = 36,87$, joten H_0 hylätään 1 %:n riskitasolla.



Esim. 2.1.6 Tutkitaan keskimääräisiä neliöhintoja Tampereen keskustassa, Länsi- ja Itä-Tampereella

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



H_0: $\mu_K = \mu_L = \mu_I$

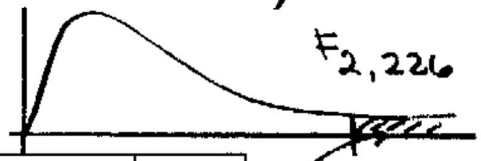
H_1: kaikki μ:t eivät samoja

Neliöhinta

Alue	Mean	N	Std. Deviation
Keskusta	2397,6072	103	408,02462
Länsi	1414,2870	34	260,39544
Itä	1536,1328	92	341,69439
Total	1905,5176	229	575,61088

$H_0: \mu_K = \mu_I = \mu_L$

H_1: kaikki μ:t eivät samoja



ANOVA

Neliöhinta

	Sum of Squares	*df*	Mean Square	*F*	Sig.
Between Groups	45699081	2	22849540.291	173.035	.000
Within Groups	29843678	226	132051.673		
Total	75542759	228			

Koska p-arvo < 0,001, H_0 hylätään ja päätellään eroja olevan. Päättely taulukkoarvon (http://www.sis.uta.fi/tilasto/mttta1/kevat2019/F_jakauma.pdf) perusteella: $F_{0,01;\ 2,\ 226} \approx 4,61 < F_{hav.} = 173,035$, joten H_0 hylätään 1 %:n riskitasolla.

Onko kaikkien alueiden välillä eroja?

Multiple Comparisons

Dependent Variable: Neliöhinta
Bonferroni

(I) Alue	(J) Alue	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Keskusta	Länsi	983,32022*	71,87439	,000	809,9641	1156,6764
	Itä	861,47438*	52,12868	,000	735,7435	987,2052
Länsi	Keskusta	-983,32022*	71,87439	,000	-1156,6764	-809,9641
	Itä	-121,84584	72,93296	,289	-297,7552	54,0635
Itä	Keskusta	-861,47438*	52,12868	,000	-987,2052	-735,7435
	Länsi	121,84584	72,93296	,289	-54,0635	297,7552

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

Länsi- ja Itä-Tampereen välillä ei eroja, muissa on. Tutkitaan odotusarvojen yhtäsuuruutta pareittain, päättely p-arvon tai luottamusvälin perusteella.

Varianssianalyysin liittyvät oletukset ja laskukaavat

$Y_{11}, Y_{12}, \ldots, Y_{1n1}$ 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, että $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2 = \sigma^2$ ja otokset riippumattomia.

H_0: $\mu_1 = \mu_2 = \ldots = \mu_I$
H_1: kaikki μ:t eivät samoja

$$SST = \sum\sum (Y_{ij} - \bar{Y})^2 \,,\ \bar{Y} = \frac{1}{n}\sum\sum Y_{ij} \,,\ n = n_1 + \cdots + n_I$$

$$\text{SSB} = \sum n_i (\bar{Y}_i - \bar{Y})^2 \quad ,\ \bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$

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

$SST = SSB + SSW$
$MSB = SSB/(I\text{-}1)$
$MSW = SSW/(n\text{-}I)$

E(MSW) = σ^2 aina
E(MSB) = σ^2, jos H_0 tosi

$F = MSB/MSW \sim F_{I\text{-}1,\ n\text{-}I}$, kun H_0 tosi
H_0 hylätään riskitasolla α, jos $F_{hav} > F_{\alpha;\ I\text{-}1,\ n\text{-}I}$.

Esim. 2.1.3 Valmennusmenetelmien vaikutus urheilusuoritukseen

H_0: $\mu_1 = \mu_2 = \mu_3$
H_1: kaikki odotusarvot eivät samoja

Urheilusuoritukset menetelmittäin

Menetelmä 1: 6, 4, 6, 4
Menetelmä 2: 14, 9, 10, 11
Menetelmä 3: 5, 11, 8, 8

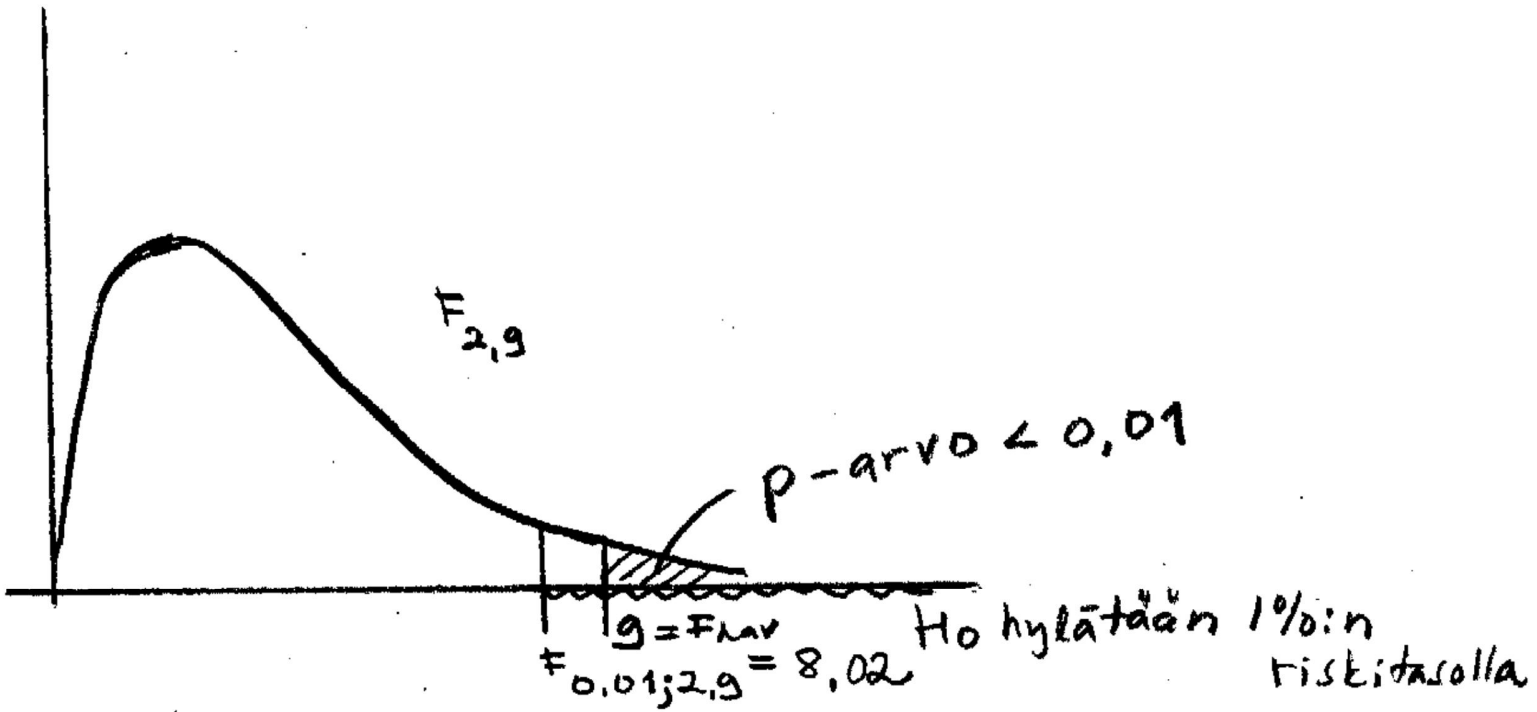
$$\overline{y}_1 = 5, \quad \overline{y}_2 = 11, \quad \overline{y}_3 = 8, \quad \overline{y} = 8,$$
$$n_1 = n_2 = n_3 = 4, \quad n = 12,$$

$$
\begin{aligned}
SST &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 \\
&= (6-8)^2+\cdots+(8-8)^2 = 108, \\
SSB &= \sum_{i=1}^{3} n_i(\overline{y}_i-\overline{y})^2 \\
&= 4(5-8)^2+4(11-8)^2+4(8-8)^2 = 72, \\
SSW &= \sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2 \\
&= (6-5)^2+(4-5)^2+(6-5)^2+(4-5)^2 \\
&\qquad +(14-11)^2+(9-11)^2+(10-11)^2+(11-11)^2 \\
&\qquad +(5-8)^2+(11-8)^2+(8-8)^2+(8-8)^2 = 36,
\end{aligned}
$$

$$MSB = SSB/(I-1) = 72/(3-1) = 36,$$
$$MSW = SSW/(n-I) = 36/(12-3) = 4,$$
$$F = MSB/MSW = 36/4 = 9,$$

$F_{0,01;\ 2,\ 9} = 8{,}02 < F_{hav.} = 9$, joten H_0 hylätään 1 %:n riskitasolla.

Voidaan sanoa, että p-arvo = $P(F_{2,9} > 9) < 0{,}01$.


$F_{2,9}$
p-arvo < 0,01
$g = F_{havainto}$
$F_{0,01;2,9} = 8,02$
Ho hylätään 1%:n riskitasolla

SPSS-tulos

ANOVA

Suoritus

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	72,000	2	36,000	9,000	,007
Within Groups	36,000	9	4,000		
Total	108,000	11			

Jos H_0: $\mu_1 = \mu_2 = \ldots = \mu_I$ hylätään, niin voidaan tutkia mitkä odotusarvot poikkeavat toisistaan. Tutkitaan odotusarvoja pareittain testin tai luottamusvälin avulla.

Esim. 2.1.3 Vain menetelmien 1 ja 2 välillä eroja.

Multiple Comparisons

Dependent Variable: Suoritus
Bonferroni

(I) Menetelmä	(J) Menetelmä	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-6,000*	1,414	,006	-10,15	-1,85
	3	-3,000	1,414	,189	-7,15	1,15
2	1	6,000*	1,414	,006	1,85	10,15
	3	3,000	1,414	,189	-1,15	7,15
3	1	3,000	1,414	,189	-1,15	7,15
	2	-3,000	1,414	,189	-7,15	1,15

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

Oletusta varianssien yhtäsuuruudesta voidaan myös testata (Levenen testi). Tällöin H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2$.

Jos variansseja ei voida olettaa samoiksi (Levenen testin p-arvo < 0,05), niin käytetään Welchin tai Brown-Forsythen testejä odotusarvojen yhtäsuuruuden testaamisessa.

Esim. 2.1.3 Varianssien yhtäsuuruuden testaaminen

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

Test of Homogeneity of Variances

Suoritus

Levene Statistic	df1	df2	Sig.
,214	2	9	,811

Hyväksytään H_0, koska p-arvo = 0,811 > 0,05.
Voidaan siis olettaa varianssit yhtä suuriksi.

Nimitys varianssianalyysi tulee siitä, että testisuure on kahden varianssiestimaattorin osamäärä.

Jos $I = 2$, niin $H_0: \mu_1 = \mu_2$. Tällöin $t^2 = F$.