Varianssianalyysi ja ei-parametriset menetelmät

Jyrki Möttönen

Sosiaalitieteiden laitos, Helsingin yliopisto

14.-15.2.2012

Johdanto

- Yksisuuntainen varianssianalyysi on kahden riippumattoman otoksen t-testin yleistys tilanteeseen, jossa perusjoukko on jaettu useampaan kuin kahteen ryhmään.
- Verrataan yhden ryhmittelymuuttujan vaikutusta jatkuvan muuttujan vaihteluun.
- Tutkitaan sekä havaintojen vaihtelua ryhmien sisällä että ryhmäkeskiarvojen vaihtelua koko populaatiossa.
- Ennen varianssianalyysin suorittamista on tutkittava varianssien yhtäsuuruus eri ryhmissä sekä normaalijakaumaoletuksen voimassaolo.

Esimerkki: Satoaineisto

Halutaan tutkia eri käsittelyiden vaikutusta sadon määrään. Kokeessa on mitattu sadon määrä kontrolliolosuhteissa (ei käsittelyä) ja kahden eri käsittelyn vallitessa. Olkoon μ_i sadon määrän odotusarvo ryhmässä i. Nollahypoteesina on nyt

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

Seuraavalla sivulla on havaintoaineisto.

Esimerkki: Satoaineisto

Kontrolli	Käsittely 1	Käsittely 2
4.17	4.81	6.31
5.58	4.17	5.12
5.18	4.41	5.54
6.11	3.59	5.50
4.50	5.87	5.37
4.61	3.83	5.29
5.17	6.03	4.92
4.53	4.89	6.15
5.33	4.32	5.80
5.14	4.69	5.26

Tilastollinen malli

- Oletetaan, että tutkimuksen kohteena oleva perusjoukko voidaan jakaa k ryhmään.
- Kustakin ryhmästä poimitaan toisistaan riippumattomat yksinkertaiset satunnaisotokset, joiden koot ovat $n_1, \ldots, n_k$. Aineiston koko on siis $n = n_1 + \ldots + n_k$.
- Olkoon

$$Y_{ij} = j.\ \text{havainto ryhmässä}\ i, \quad i = 1, \ldots, k,\ j = 1, \ldots, n_i$$

Oletetaan, että

$$\mathrm{E}(Y_{ij}) = \mu_i, \quad i = 1, \ldots, k,\ j = 1, \ldots, n_i$$

ja

$$\mathrm{Var}(Y_{ij}) = \sigma^2, \quad i = 1, \ldots, k,\ j = 1, \ldots, n_i$$

Tilastollinen malli

- Jakaumaoletus:

$$Y_{ij} \sim N(\mu_i, \sigma^2), \quad i = 1, \ldots, k, \; j = 1, \ldots, n_i.$$

- Huom! Samaan ryhmään i kuuluvilla oletetaan olevan sama odotusarvo μ_i. Ryhmien odotusarvot voivat erota toisistaan.
- Huom! Kaikilla havainnoilla sama varianssi σ^2.

Varianssianalyysin oletusten testaus

- Varianssien yhtäsuuruutta voidaan testata esim. Levenen testillä. Jos varianssit eivät ole yhtäsuuria eri ryhmissä, niin ryhmien odotusarvojen testaukseen voidaan käyttää esimerkiksi Welchin (1951) testiä.
- Normaalijakaumaoletusta voidaan tutkia erilaisten kuvioiden avulla (histogrammi, Q-Q plot eli kvantiilikuvio, ...) ja testien avulla (Kolmogorov-Smirnov, Shapiro-Wilk, ...).
- Varianssianalyysi on hyvin robusti normaalisuusoletuksen suhteen. Varianssianalyysi on käyttökelpoinen, vaikka muuttujan jakauma ei olisikaan aivan normaalinen.

Levenen testi

- Levenen testillä testataan hypoteesia

$$H_0^* : \ \sigma_1^2 = \ldots = \sigma_k^2$$

jossa $\sigma_i^2 = \mathrm{Var}(Y_{ij})$. Jos hypoteesi H_0^* on voimassa, niin kaikilla havainnoilla Y_{ij} on sama varianssi.

- Levenen testin testisuure W on likimain F-jakautunut nollahypoteesin vallitessa, ts.

$$W \sim F(k-1, n-k) \quad \text{likimain, kun } H_0^* \text{ tosi.}$$

Esimerkki: Satoaineisto

SPSS antaa seuraavat tulokset testattaessa varianssien yhtäsuuruutta:

Test of Homogeneity of Variances

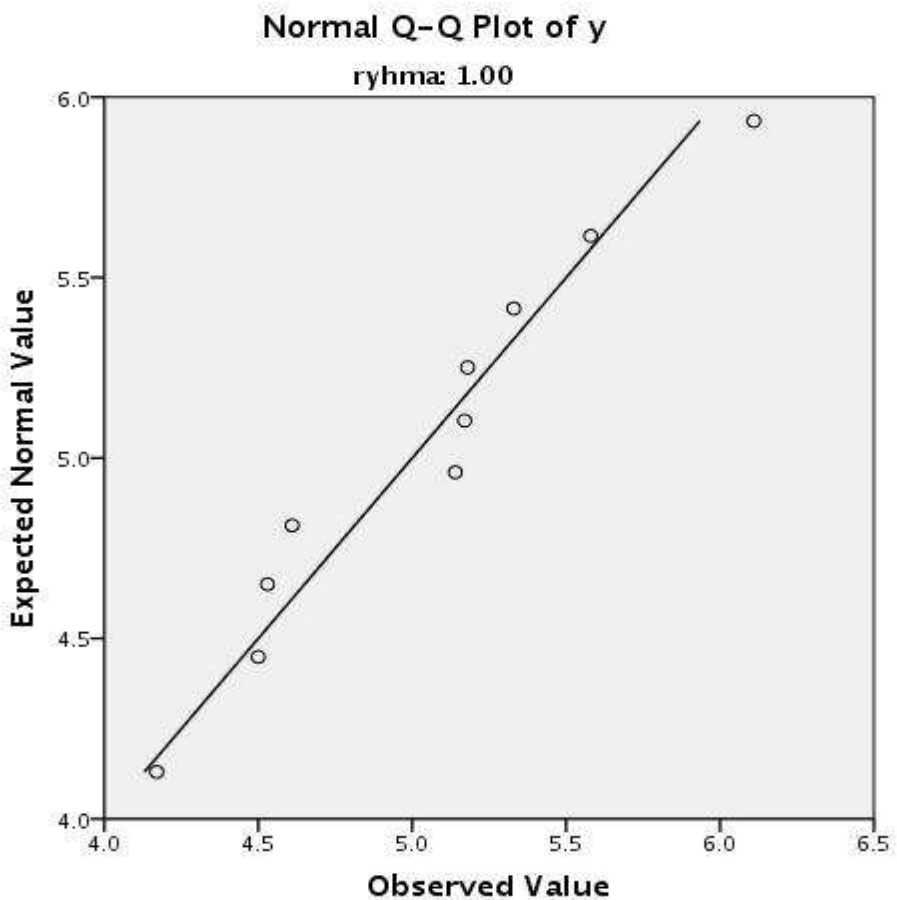
Levene Statistic	df1	df2	Sig.
1.237	2	27	.306

P-arvoksi saatiin $P\{F(2,27) \geq 1.237\} = 0.306$, joten nollahypoteesi voidaan jättää voimaan. Testin perusteella ryhmien variansseissa ei ole tilastollisesti merkitsevää eroa.

Esimerkki: Satoaineisto

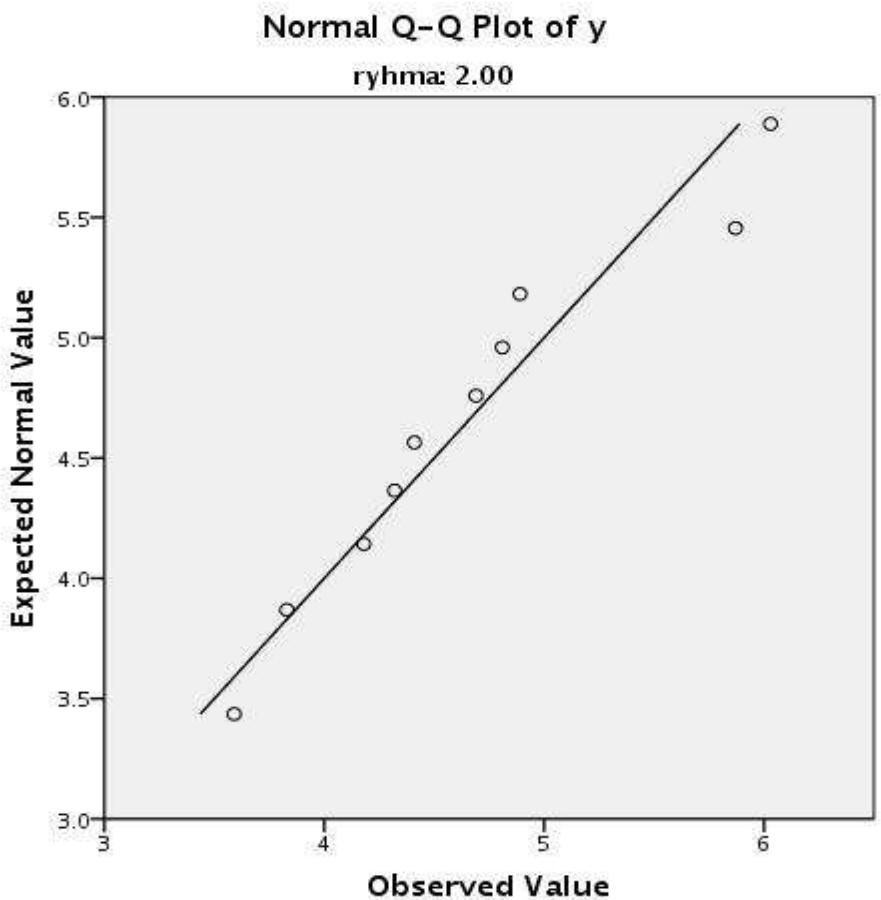
Kuvissa 1-3 on kvantiilikuviot kolmen eri ryhmän havainnoille.

Esimerkki: Satoaineisto



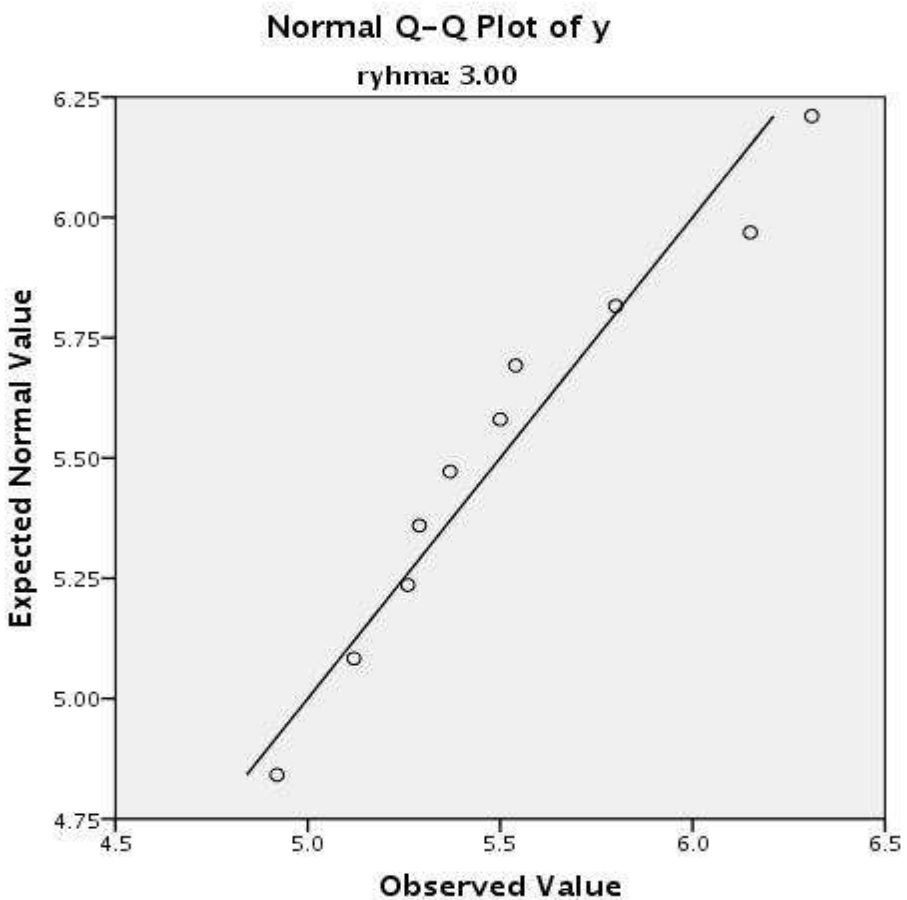
Kuva: Ryhmän 1 havaintojen kvantiilikuvio.

Esimerkki: Satoaineisto



Kuva: Ryhmän 2 havaintojen kvantiilikuvio.

Esimerkki: Satoaineisto



Kuva: Ryhmän 3 havaintojen kvantiilikuvio.

Yksisuuntaisen varianssianalyysin nollahypoteesi

- Nollahypoteesi on muotoa

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

- Jos nollahypoteesi H_0 pätee, niin havainnot noudattavat kaikissa ryhmissä samaa normaalijakaumaa ja ryhmät voidaan satunnaismuuttujaa Y koskevissa tarkasteluissa yhdistää yhdeksi ryhmäksi.

Neliösummat

- Määritellään ryhmäkohtaiset keskiarvot

$$\bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

ja kokonaiskeskiarvo

$$\bar{Y}_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \bar{Y}_{i\cdot},$$

missä $n = n_1 + \ldots + n_k$.

Neliösummat

Hypoteesin testauksessa käytetään seuraavia kolmea havaintoaineistosta laskettua neliösummaa:

$$\begin{aligned} SS_T &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2 \\ &= \text{Kokonaisneliösumma (T=Total).} \end{aligned}$$

Mittaa havaintoarvojen vaihtelua kokonaiskeskiarvon ympärillä

Neliösummat

$$\begin{aligned} SS_G &= \sum_{i=1}^{k} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \\ &= \text{Ryhmien välinen neliösumma (G=Group).} \end{aligned}$$

Mittaa ryhmäkeskiarvojen vaihtelua kokonaiskeskiarvojen suhteen ts. havaintoarvojen ryhmien välistä systemaattista vaihtelua.

Neliösummat

$$\begin{aligned} SS_E &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2 \\ &= \text{Jäännösneliösumma (E=Error).} \end{aligned}$$

Mittaa havaintoarvojen vaihtelua omien ryhmäkeskiarvojensa suhteen ts. havaintoarvojen vaihtelua ryhmien sisällä.

Neliösummat

Voidaan osoittaa, että edellä määritellyille neliösummille pätee seuraavanlainen varianssianalyysihajotelma:

$$SS_T = SS_G + SS_E.$$

Kokonaisvaihtelu voidaan siis jakaa ryhmien väliseen ja ryhmien sisäiseen vaihteluun.

Keskineliövirheet

Määritellään *keskineliövirheet*

$$MS_T = \frac{SS_G}{n-1},$$

$$MS_G = \frac{SS_G}{k-1}$$

ja

$$MS_E = \frac{SS_E}{n-k}.$$

Voidaan osoittaa, että $\mathrm{E}(MS_T) = \mathrm{E}(MS_E) = \sigma^2$ riippumatta siitä onko nollahypoteesi tosi vai epätosi. Keskineliövirheet MS_T ja MS_E estimoivat siis harhattomasti havaintoaineiston varianssia σ^2.

Keskineliövirheet

Kun H_0 on **tosi**, niin $\mathrm{E}(MS_G) = \sigma^2$. Tästä seuraa, että kun H_0 on tosi, niin on odotettavissa, että

$$MS_G \approx MS_E \quad \text{eli} \quad \frac{MS_G}{MS_E} \approx 1$$

Kun H_0 on **epätosi**, niin $E(MS_G) > \sigma^2$. Tästä seuraa, että kun H_0 on epätosi, niin on odotettavissa, että

$$MS_G > MS_E \quad \text{eli} \quad \frac{MS_G}{MS_E} > 1$$

F-testisuure

H_0:n testaamiseen voidaan nyt käyttää F-testisuuretta

$$F = \frac{MS_G}{MS_E} = \frac{SS_G/(k-1)}{SS_E/(n-k)},$$

joka noudattaa nollahypoteesin vallitessa F-jakaumaa vapausastein $k-1$ ja $n-k$.

Varianssianalyysitaulukko

Vaihtelun lähde	Neliö-summa	Vapaus-asteet	Varianssi-estimaatti	F-testisuure
Ryhmien välinen	SS_G	$k-1$	$MS_G = \frac{SS_G}{k-1}$	$F = \frac{MS_G}{MS_E}$
Ryhmien sisäinen	SS_E	$n-k$	$MS_E = \frac{SS_E}{n-k}$	
Kokonaisvaihtelu	SS_T	$n-1$	$MS_T = \frac{SS_T}{n-1}$	

- Huom! Keskineliövirheet MS_T ja MS_E ovat aina perusjoukon varianssin σ^2 harhattomia estimaattoreita mutta $\mathrm{E}(MS_G) = \sigma^2$ vain, jos H_0 pätee.

Esimerkki: Satoaineisto

Varianssianalyysitaulukko:

Vaihtelun lähde	Neliö-summa	Vapaus-asteet	Varianssi-estimaatti	F-testisuure
Ryhmien välinen	3.76634	2	1.88317	4.846088
Ryhmien sisäinen	10.49209	27	0.3885959	
Kokonaisvaihtelu	14.25843	29	0.49167	

p-arvo $= \mathrm{P}(F(2,27) \geq 4.846088) = 0.01590996 < 0.05$
Aineiston perusteella ainakin kahden ryhmän sadon määrät eroavat toisistaan (käytettäessä merkitsevyystasoa $\alpha = 0.05$.

Esimerkki: Satoaineisto

SPSS antaa varianssianalyysitaulun seuraavassa muodossa:

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3,766	2	1,883	4,846	,016
Within Groups	10,492	27	,389		
Total	14,258	29			

Esimerkki: Satoaineisto

Welchin testiä käytetään silloin, kun ryhmien varianssit eroavat toisistaan. Jos käytetään Welchin testiä satoaineiston tapaukseen, niin SPSS antaa seuraavanlaisen taulukon:

```
Robust Tests of Equality of Means
y
|-----|----------|---|------|----|
|     |Statistica|df1|df2   |Sig.|
|-----|----------|---|------|----|
|Welch|5.181     |2  |17.128|.017|
|-------------------------------|
 a Asymptotically F distributed.
```

Nähdään, että Welchin testi antaa lähes saman p-arvon kuin tavallinen yksisuuntainen varianssianalyysi. Näin piti käydäkin, koska satoaineistossa ryhmien varianssit ovat likimain yhtäsuuret.

Parittaiset vertailut

- Jos yksisuuntaisessa varianssianalyysissä H_0 hylätään, niin ainakin kahden ryhmän odotusarvot ovat erisuuria.
- Varianssianalyysi ei anna vastausta siihen mitkä odotusarvot eroavat toisistaan! Siihen kysymykseen voidaan etsiä vastausta parittaisten vertailujen testien avulla. Vertailuun voidaan käyttää esimerkiksi Bonferronin t-testiä tai Tukeyn HSD-testiä.

Esimerkki: Satoaineisto

SPSS antaa seuraavat tulokset Bonferronin testille:

Multiple Comparisons
Dependent Variable: y

	(I) ryhma	(J) ryhma	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Bonferroni	1,00	2,00	,37100	,27878	,583	-,3406	1,0826
		3,00	-,49400	,27878	,263	-1,2056	,2176
	2,00	1,00	-,37100	,27878	,583	-1,0826	,3406
		3,00	-,86500*	,27878	,013	-1,5766	-,1534
	3,00	1,00	,49400	,27878	,263	-,2176	1,2056
		2,00	,86500*	,27878	,013	,1534	1,5766

* The mean difference is significant at the 0.05 level.

Esimerkki: Satoaineisto

SPSS antaa seuraavat tulokset Tukeyn HSD-testille:

Multiple Comparisons
Dependent Variable: y

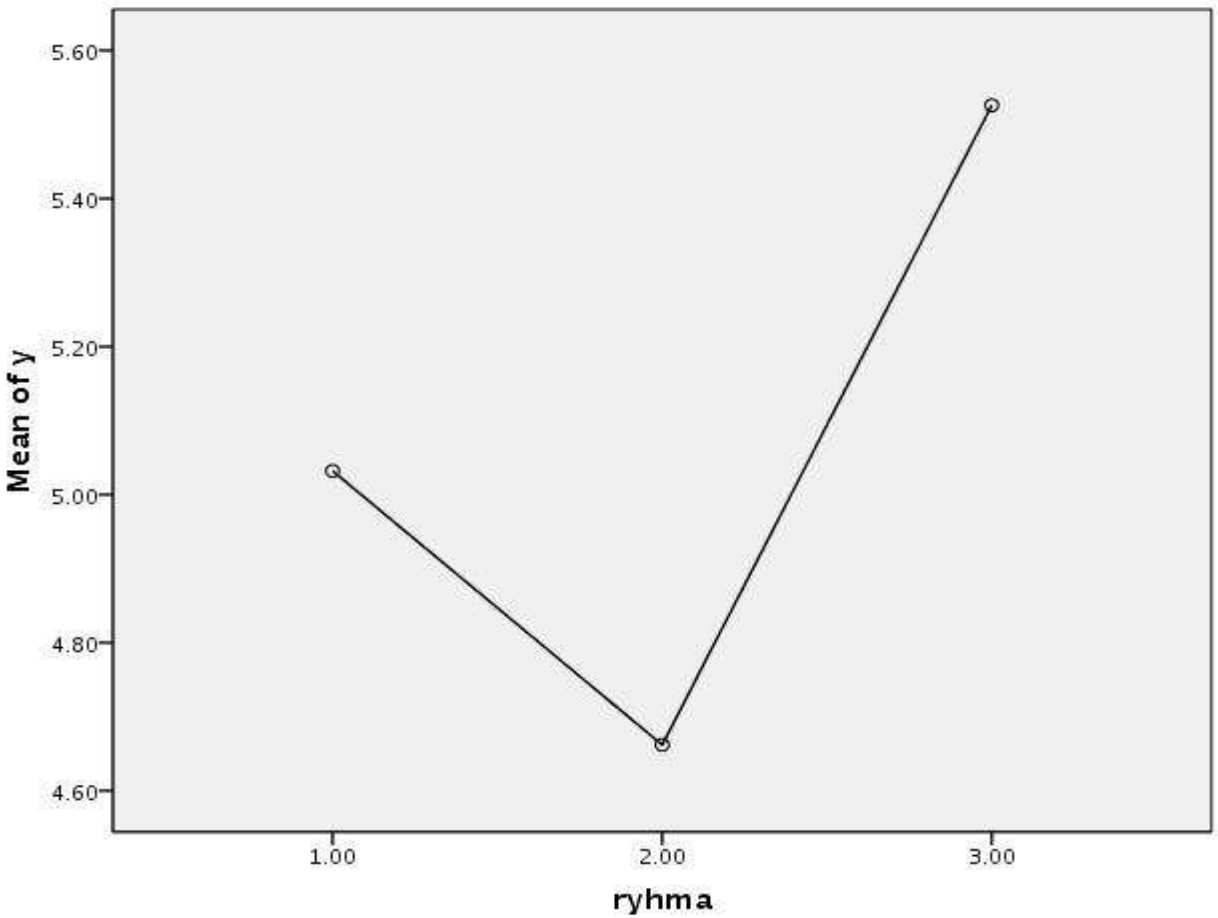
	(I) ryhma	(J) ryhma	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	1,00	2,00	,37100	,27878	,391	-,3202	1,0622
		3,00	-,49400	,27878	,198	-1,1852	,1972
	2,00	1,00	-,37100	,27878	,391	-1,0622	,3202
		3,00	-,86500*	,27878	,012	-1,5562	-,1738
	3,00	1,00	,49400	,27878	,198	-,1972	1,1852
		2,00	,86500*	,27878	,012	,1738	1,5562

* The mean difference is significant at the 0.05 level.

Esimerkki: Satoaineisto

Bonferronin testin ja Tukeyn HSD-testin perusteella ainoastaan ryhmien 2 ja 3 odotusarvot eroavat tilastollisesti merkitsevästi toisistaan. Kuvassa 4 on graafinen esitys ryhmien keskiarvoista.

Esimerkki: Satoaineisto



Kuva: Ryhmien 1-3 havaintojen keskiarvot.

Parametriton malli

- Jakaumaoletus:

$$Y_{ij} = \tau_i + e_{ij}, \quad i = 1, \ldots, k, \; j = 1, \ldots, n_i,$$

  jossa e_{ij} tulee jatkuvasta jakaumasta, jonka mediaani on 0.
- Huom! Jakaumaoletus tarkoittaa sitä, että ryhmän i havainnot Y_{ij}, $j = 1, \ldots, n_i$, tulevat jatkuvasta jakaumasta (voi olla myös vino jakauma), jonka mediaani on τ_i. Ryhmien mediaanit voivat erota toisistaan.
- Huom! Jakaumat oletaan samanmuotoisiksi - ne poikkeavat ainoastaan sijannin suhteen!!

Havainnoista järjestyslukuja

Kruskalin-Wallisin testisuureessa käytetään pelkästään havainnoista laskettuja järjestyslukuja. Yhdistetään kaikki ryhmät yhdeksi suureksi havaintoaineistoksi ja korvataan havainnot vastaavilla järjestysluvuillaan. Olkoon

$$R_{ij} = R(Y_{ij}) = \text{"Havainnon } Y_{ij} \text{ järjestysluku"}.$$

Esim. Meillä on havaintoaineisto $(1.3, 2.8, 4.7, 9.6)$, jossa havainnot on järjestetty pienimmästä suurimpaan. Havaintoja vastaavat järjestysluvut ovat silloin $(1, 2, 3, 4)$. Toisin sanottuna k:nneksi pienimmän havainnon järjestysluku on k.

Testisuure

Kruskalin-Wallisin testisuure nollahypoteesille

$$H_0 : \ \tau_1 = \ldots = \tau_k$$

konstruoidaan seuraavasti:

$$H = \left(\frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_{.j}^2}{n_j} \right) - 3(n+1),$$

jossa

$$R_{.j} = \sum_{i=1}^{n_j} R_{ij}$$

on j:nnettä ryhmää vastaavien havaintojen järjestyslukujen summa.

Testisuure

Kun nollahypoteesi on tosi, niin testisuure H on likimain χ^2-jakautunut vapausastein $k-1$, ts.

$$H \sim \chi^2_{k-1} \quad \text{likimain, kun } H_0 \text{ on tosi.}$$

Suuret testisuureen arvot ovat kriittisiä nollahypoteesin kannalta, joten P-arvoksi saadaan

$$p = P\{\chi^2_{k-1} \geq h\},$$

jossa h on havainnoista laskettu testisuureen arvo.

Testisuure

Esim. Jos $k = 6$, niin χ^2-jakauman taulukosta saadaan

$$P\{\chi^2_5 \geq 11.07\} \approx 0.05.$$

Tällöin testisuureen arvot, jotka ovat suurempia kuin 11.07 johtavat nollahypoteesin hylkäämiseen merkitsevyystasolla $\alpha = 0.05$.

Testisuureen H yhteys varianssianalyysin F-testisuureeseen

Huom! Jos käytetään tavallisen F-testisuureen laskemiseen alkuperäisten havaintojen sijasta niiden järjestyslukuja, niin huomataan, että saatu testisuure on yhtäpitävä testisuureen H kanssa. Laskettaessa (esim. SPSS:n avulla) järjestyslukuaineistolle yksisuuntaisen varianssianalyysin p-arvo, niin Kruskal-Wallisin testin pitäisi antaa suurinpiirtein sama p-arvo.

Esimerkki: Satoaineisto

```
Kruskal-Wallis Test
Ranks
|-|-----|--|---------|
| |ryhma|N |Mean Rank|
|-|-----|--|---------|
|y|1.00 |10|14.70    |
| |-----|--|---------|
| |2.00 |10|10.40    |
| |-----|--|---------|
| |3.00 |10|21.40    |
| |-----|--|---------|
| |Total|30|         |
|-|-----|--|---------|

Test Statistics a,b
|-----------|-----|
|           |y    |
|-----------|-----|
|Chi-Square |7.930|
|-----------|-----|
|df         |2    |
|-----------|-----|
|Asymp. Sig.|.019  |
|-----------|-----|
 a Kruskal Wallis Test
 b Grouping Variable: ryhma
```

Esimerkki: Satoaineisto

Kruskal-Wallisin testi antaa p-arvoksi 0.019, joten nollahypoteesi voidaan hylätä merkitsevyystasolla 0.05. Joidenkin ryhmien mediaanit näyttäisivät siis eroavan toisistaan.

Esimerkki: Satoaineisto

Pareittainen vertailu voidaan suorittaa (karkealla tavalla) seuraavasti. Testataan ryhmien eroja Mann-Whitney-Wilcoxonin testin (sama kuin Kruskal-Wallis kahdelle ryhmälle) avulla.

Ryhmä 1 vs 2: P-arvo ≈ 0.199.
Ryhmä 1 vs 3: P-arvo ≈ 0.059.
Ryhmä 2 vs 3: P-arvo ≈ 0.010.

Yksittäisten testien p-arvot pitää kertoa pareittaisten vertailujen lukumäärällä 3 (ns. Bonferronin korjaus). P-arvoiksi saadaan siis $3 \cdot 0.1999 = 0.60 > 0.05$, $3 \cdot 0.059 = 0.18 > 0.05$ ja $3 \cdot 0.010 = 0.03 < 0.05$. Siis ainoastaan ryhmien 2 ja 3 välillä on tilastollisesti merkitsevä ero.