Tilastolliset menetelmät: Varianssianalyysi

Sisällys

20. Yksisuuntainen varianssianalyysi

Yksisuuntaisessa varianssianalyysissa perusjoukko on jaettu ryhmiin *yhden ryhmittelevän* **tekijän** suhteen ja tavoitteena on testata hypoteesia, jonka mukaan **kiinnostuksen kohteena olevan muuttujan ryhmäkohtaiset odotusarvot ovat yhtä suuria**.

Kaksi- tai **useampisuuntaisessa varianssianalyysissa** perusjoukko on jaettu ryhmiin **kahden** tai **useamman ryhmittelevän tekijän** suhteen ja nytkin tavoitteena on testata hypoteesia, jonka mukaan **kiinnostuksen kohteena olevan muuttujan ryhmäkohtaiset odotusarvot ovat yhtä suuria**.

Tässä luvussa tarkastellaan **yksisuuntaista varianssianalyysia**. Tarkastelun kohteena ovat mm. **yksisuuntaisen varianssianalyysin malli** ja sen **parametrointi**, **parametrien estimointi**, **odotusarvojen yhtäsuuruuden testaaminen** ja **laskutoimitusten suorittaminen**.

Avainsanat:

Aritmeettinen keskiarvo, Bartlettin testi, Bonferronin epäyhtälö, Bonferronin menetelmä, Cochranin lause, Estimointi, *F*-testi, Faktori, Harha, Harhattomuus, Indikaattorimuuttuja, Jäännösvaihtelu, Jäännösvarianssi, χ^2-testi, Kokonaiskeskiarvo, Kokonaisvaihtelu, Kontrasti, Luottamuskerroin, Luottamustaso, Luottamusväli, Malli, Matriisi, Merkitsevyystaso, Muuttuja, Normaalijakauma, Neliösumma, Odotusarvo, Odotusarvojen parivertailu, Odotusarvojen simultaaninen vertailu, Otos, Otostunnusluku, Parametri, Parametrointi, Pienimmän neliösumman estimaattori, Pienimmän neliösumman menetelmä, Residuaali, Riippumattomuus, Ryhmien sisäinen vaihtelu, Ryhmien välinen vaihtelu, Ryhmittely, Ryhmä, Ryhmäkeskiarvo, Satunnaisuus, Side-ehto, Sovite, *t*-testi, Taso, Tekijä, Testi, Vapausaste, Varianssi, Varianssianalyysihajotelma, Varianssianalyysitaulukko, Yksisuuntainen varianssianalyysi, Yleinen lineaarinen malli, Yleiskeskiarvo

20.1. Varianssianalyysi: Johdanto

Kahden riippumattoman otoksen *t*-testi

Monisteen **Tilastotiede** luvun **Testejä suhdeasteikollisille muuttujille** tarkastellaan **kahden riippumattoman otoksen *t*-testiä.** Testin testausasetelma on seuraava:

(i) Perusjoukko koostuu *kahdesta* ryhmästä.

(ii) Havainnot noudattavat kummassakin ryhmässä *normaalijakaumaa*.

(iii) Kummastakin ryhmästä on poimittu *toisistaan riippumattomat satunnaisotokset*.

(iv) Tehtävänä on *testata ryhmäkohtaisten odotusarvojen yhtäsuuruutta*.

Varianssianalyysin perusasetelma

Varianssianalyysi on *kahden riippumattoman otoksen t-testin yleistys* tilanteisiin, jossa perusjoukko koostuu *kahdesta* tai *useammasta ryhmästä*:

(i) Perusjoukko koostuu *kahdesta* tai *useammasta* ryhmästä.

(ii) Havainnot noudattavat jokaisessa ryhmässä *normaalijakaumaa*.

(iii) Jokaisesta ryhmästä poimitaan *toisistaan riippumattomat satunnaisotokset*.

(iv) Tehtävänä on *testata ryhmäkohtaisten odotusarvojen samuutta*.

Perusjoukon *jako ryhmiin* voidaan tehdä *yhden* tai *useamman* **faktorin** eli **tekijän** (*muuttujan*) *arvojen perusteella*. Jos perusjoukon jako ryhmiin perustuu *yhteen tekijään*, puhutaan **yksisuuntaisesta varianssianalyysista**. Jos perusjoukon jako ryhmiin perustuu *m tekijään*, puhutaan ***m*-suuntaisesta varianssianalyysista**.

Huomautus:

- Tässä luvussa käsitellään **yksisuuntaista varianssianalyysia**; **kaksi-** ja **kolmisuuntaista varianssianalyysia** käsitellään seuraavissa luvuissa.

Varianssianalyysin nimi

On hyvä huomata heti näin alussa, että varianssianalyysin nimi saattaa *johtaa harhaan*. Varianssianalyysissa *testauksen kohteena ei ole* **varianssien** *yhtäsuuruus*, *vaan* **odotusarvojen** *yhtäsuuruus*. Varianssianalyysin nimi johtuu siitä, että odotusarvojen yhtäsuuruuden testaaminen perustuu eri periaatteilla määrättyjen *varianssien vertailuun*; ks. käsittelyä alla.

20.2. Yksisuuntaisen varianssianalyysin perusasetelma

Oletetaan, että tutkimuksen kohteena oleva perusjoukko voidaan *jakaa kahteen tai useampaan ryhmään jonkin* **faktorin** eli **tekijän** (*muuttujan*) *A arvojen suhteen* ja oletetaan, että tekijällä *A* on *k* **tasoa**, jolloin jaossa syntyy *k* **ryhmää**. Oletetaan edelleen, että ryhmistä on poimittu *toisistaan riippumattomat satunnaisotokset*, joiden koot ovat

$$n_1,\ n_2,\ \ldots,\ n_k$$

Olkoon

$$y_{ij} = i.\ \text{havainto ryhmässä}\ j\ ,\ i = 1, 2, \ldots, n_j\ ,\ j = 1, 2, \ldots, k$$

Käytetystä otantamenetelmästä seuraa, että havainnot y_{ij} voidaan olettaa *riippumattomiksi* (ja siten myös *korreloimattomiksi*) satunnaismuuttujiksi.

Oletetaan, että jokaisella havainnolla y_{ij} on ryhmässä *j sama odotusarvo*:

$$\mathrm{E}(y_{ij}) = \mu_j \,,\ i = 1, 2, \ldots, n_j \,,\ j = 1, 2, \ldots, k$$

ja kaikilla havainnoilla y_{ij} on (ryhmäjaosta riippumatta) *sama varianssi*:

$$\mathrm{Var}(y_{ij}) = \sigma^2 \,,\ i = 1, 2, \ldots, n_j \,,\ j = 1, 2, \ldots, k$$

Oletetaan lisäksi, että kaikki havainnot y_{ij} ovat *normaalijakautuneita*:

$$y_{ij} \sim \mathrm{N}(\mu_j \,,\ \sigma^2) \,,\ i = 1, 2, \ldots, n_j \,,\ j = 1, 2, \ldots, k$$

Haluamme testata **nollahypoteesia**, että *ryhmäkohtaiset odotusarvot* $\mathrm{E}(y_{ij}) = \mu_j$ *ovat yhtä suuria*:

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

Jos *nollahypoteesi* H_0 *ryhmäkohtaisten odotusarvojen yhtäsuuruudesta pätee, ryhmät voidaan yhdistää* kaikissa havaintojen keskimääräisiä arvoja koskevissa tarkasteluissa. Sen sijaan, *jos nollahypoteesi* H_0 *hylätään*, tiedetään, että *muuttujan y ryhmäkohtaiset odotusarvot eroavat toisistaan ainakin kahdessa ryhmässä*. *Jos nollahypoteesi* H_0 *on hylätty*, ryhmäkohtaisia *odotusarvoja voidaan verrata pareittain* tai *simultaanisesti toisiinsa*; ks. kappaletta **Odotusarvojen vertaaminen**.

Yksisuuntainen varianssianalyysi tarkoittaa siis nollahypoteesin

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

testaamista.

Yksisuuntainen varianssianalyysi ja koesuunnittelu

Yksisuuntaista varianssianalyysiä voidaan soveltaa koetulosten analyysiin seuraavassa **koeasetelmassa**:

(i) Oletetaan, että kokeen tavoitteena on verrata, miten **käsittelyt**

$$A_1, A_2 \,, \ldots, A_k$$

vaikuttavat kiinnostuksen kohteena olevan **vastemuuttujan** *y keskimääräisiin arvoihin*.

(ii) Valitaan käsittelyn A_j kohteeksi kaikkien kokeen kohteiksi valittujen yksilöiden joukosta *satunnaisesti* n_j yksilöä, $j = 1, 2, \ldots, k$ ja olkoon $n_1 + n_2 + \cdots + n_k = N$.

(iii) Mitataan **vasteet** eli kiinnostuksen kohteena olevan muuttujan y arvot y_{ij} , $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, k$.

Huomaa, että koeasetelma on **täydellisesti satunnaistettu**: *Sattuma määrää täydellisesti millaisen käsittelyn kohteeksi kokeen kohteiksi valitut yksilöt joutuvat*.

20.3. Yksisuuntaisen varianssianalyysin suorittaminen

Havainnot ja niiden keskiarvot

Havaintoarvot

$$y_{ij} \,,\ i = 1, 2, \ldots, n_j \,,\ j = 1, 2, \ldots, k$$

voidaan *ryhmitellä* seuraavalla tavalla:

Ryhmä 1: $y_{11}, y_{21}, \mathsf{K}, y_{n_1 1}$

Ryhmä 2: $y_{12}, y_{22}, \mathsf{K}, y_{n_2 2}$

...

Ryhmä *k*: $y_{1k}, y_{2k}, \mathsf{K}, y_{n_k k}$

Määritellään havaintoarvojen **ryhmäkeskiarvot** kaavoilla

Ryhmä 1: $\overline{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{i1}$

Ryhmä 2: $\overline{y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{i2}$

...

Ryhmä *k*: $\overline{y}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} y_{ik}$

On odotettavissa, että ryhmäkeskiarvot $\overline{y}_i$ *eivät poikkea paljon toisistaan*, jos nollahypoteesi

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

pätee.

Jos ryhmäkohtaiset otokset *yhdistetään* yhdeksi otokseksi, yhdistetyn otoksen havaintoarvojen **yleis-** eli **kokonaiskeskiarvo** saadaan kaavalla

$$\overline{y} = \frac{1}{N}\sum_{j=1}^{k}\sum_{j=1}^{n_j} y_{ij} = \frac{1}{N}\sum_{j=1}^{k} n_j \overline{y}_j$$

jossa

$$N = n_1 + n_2 + \mathsf{L} + n_k$$

on havaintojen kokonaislukumäärä yhdistetyssä otoksessa ja

$$\overline{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}\ ,\ j = 1, 2, \mathsf{K}, k$$

Varianssianalyysihajotelma

Kirjoitetaan identiteetti

$$y_{ij} - \overline{y} = (\overline{y}_j - \overline{y}) + (y_{ij} - \overline{y}_j)$$

jossa

$y_{ij} - \overline{y}$ = havaintoarvon y_{ij} poikkeama kokonaiskeskiarvosta $\overline{y}$

$\overline{y}_j - \overline{y}$ = ryhmäkeskiarvon $\overline{y}_j$ poikkeama kokonaiskeskiarvosta $\overline{y}$

$y_{ij} - \bar{y}_j$ = havaintoarvon y_{ij} poikkeama ryhmäkeskiarvosta $\bar{y}_j$

Yksisuuntaisen varianssianalyysin *testi nollahypoteesille*

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

perustuu *poikkeamien* $\bar{y}_j - \bar{y}$ ja $y_{ij} - \bar{y}_j$ *neliösummille*. Jos nollahypoteesi H_0 pätee, on odotettavissa, että ryhmäkeskiarvot $\bar{y}_j$ *eivät poikkea kovin paljon kokonaiskeskiarvosta* $\bar{y}$, jolloin poikkeamat $\bar{y}_j - \bar{y}$ eivät ole itseisarvoiltaan kovin suuria.

Määritellään **havaintoarvojen kokonaisvaihtelua kuvaava kokonaisneliösumma** *SST*:

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2$$

Jos ryhmäkohtaiset otokset yhdistetään *yhdeksi otokseksi*, saadun *yhdistetyn otoksen varianssi* on

$$s_y^2 = \frac{1}{N-1} SST = \frac{1}{N-1} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2$$

Määritellään **ryhmien välistä** (*systemaattista*) **vaihtelua kuvaava** (ryhmä-) **neliösumma** *SSG*:

$$SSG = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{y}_j - \bar{y})^2 = \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2$$

Määritellään **ryhmien sisäistä vaihtelua kuvaava** (jäännös-) **neliösumma** *SSE*:

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

Havaintoarvojen y_{ij} *ryhmävarianssit* eli *ryhmäkohtaiset varianssit* s_j^2 saadaan lausekkeista

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \ , \ j = 1, 2, \mathrm{K}, k$$

Siten ryhmien sisäistä vaihtelua kuvaavan neliösumman *SSE* lauseke voidaan esittää myös muodossa

$$SSE = \sum_{j=1}^{k} (n_j - 1) s_j^2$$

Korottamalla identiteetti

$$y_{ij} - \bar{y} = (\bar{y}_j - \bar{y}) + (y_{ij} - \bar{y}_j)$$

potenssiin kaksi ja laskemalla yhteen saadaan **varianssianalyysihajotelma**

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{y}_j - \bar{y})^2 + \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

joka voidaan edellä esitettyjen merkintöjen avulla kirjoittaa lyhyesti muotoon

$$SST = SSG + SSE$$

Perustelu:

Korottamalla identiteetti

$$y_{ij} - \overline{y} = (\overline{y}_j - \overline{y}) + (y_{ij} - \overline{y}_j)$$

potenssiin kaksi ja laskemalla yhteen saadaan ensin

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \overline{y})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{y}_j - \overline{y})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2 + 2\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{y}_j - \overline{y})(y_{ij} - \overline{y}_j)$$

Varianssianalyysihajotelma

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \overline{y})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{y}_j - \overline{y})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2$$

tulee siis todistetuksi, jos näytämme, että

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{y}_j - \overline{y})(y_{ij} - \overline{y}_j) = 0$$

Suoraan laskemalla saamme

$$\begin{aligned}
\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{y}_j - \overline{y})(y_{ij} - \overline{y}_j) &= \sum_{j=1}^{k}(\overline{y}_j - \overline{y})\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j) \\
&= \sum_{j=1}^{k}(\overline{y}_j - \overline{y})\left(\sum_{i=1}^{n_j} y_{ij} - n_j\overline{y}_j\right) \\
&= \sum_{j=1}^{k}(\overline{y}_j - \overline{y})\left(\sum_{i=1}^{n_j} y_{ij} - n_j \cdot \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}\right) \\
&= \sum_{j=1}^{k}(\overline{y}_j - \overline{y}) \cdot 0 = 0
\end{aligned}$$

kuten halusimme.

■

Varianssianalyysihajotelmassa

$$SST = SSG + SSE$$

havaintojen *kokonaisvaihtelua kuvaava neliösumma*

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \overline{y})^2$$

on hajotettu kahden osatekijän summaksi, jossa *neliösumma*

$$SSG = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{y}_j - \overline{y})^2 = \sum_{j=1}^{k} n_j(\overline{y}_j - \overline{y})^2$$

kuvaa ryhmien välistä (systemaattista) *vaihtelua* ja *neliösumma*

$$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2$$

kuvaa ryhmien sisäistä vaihtelua.

Testi odotusarvojen samuudelle

Jos *ryhmien välistä vaihtelua kuvaava neliösumma*

$$SSG = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{y}_j - \bar{y})^2 = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2$$

on suuri verrattuna ryhmien sisäistä vaihtelua kuvaavaan neliösummaan

$$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$$

nollahypoteesi

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

on syytä asettaa kyseenalaiseksi. Määritellään ***F*-testisuure**

$$F = \frac{N-k}{k-1}\cdot\frac{SSG}{SSE} = \frac{N-k}{k-1}\cdot\frac{\sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2}$$

Jos havainnot ovat *normaalijakautuneita* ja *nollahypoteesi*

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

pätee, testisuure F noudattaa *F-jakaumaa* vapausastein $(k-1)$ ja $(N-k)$:

$$F \quad F(k-1, N-k)$$

jossa

$$N = n_1 + n_2 + \mathrm{L} + n_k$$

Testisuureen F *normaaliarvo* on

$$\mathrm{E}(F) \underset{\mathrm{H}_0}{=} \frac{N-k}{N-k-2}$$

Huomautus:

- $\mathrm{E}(F) \approx 1$ suurille N.

Suuret testisuureen F arvot johtavat *nollahypoteesin* H_0 *hylkäämiseen*.

Perustelemme testisuureen *F* jakaumaa koskevan tuloksen alla useassa osassa tarkastelemalla erikseen *kokonaisneliösumman, ryhmäneliösumman ja jäännösneliösumman jakaumia* sekä soveltamalla osiin ns. *Cochranin lausetta.*

Kokonaisneliösumman *SST* jakauma ja vapausasteet

Oletetaan, että nollahypoteesi

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

pätee. Tällöin satunnaismuuttuja SST / σ^2 noudattaa χ^2*-jakaumaa* vapausastein $(N-1)$:

$$\frac{SST}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2 \quad \chi^2(N-1)$$

jossa

$$N = n_1 + n_2 + \mathsf{L} + n_k$$

Perustelu:

Oletetaan, että havainnot y_{ij} ovat *riippumattomia* ja

$$y_{ij} \sim \mathrm{N}(\mu\,,\,\sigma^2)\,,\; i = 1, 2, \ldots, n_j\,,\; j = 1, 2, \ldots, k$$

Tällöin satunnaismuuttujat

$$\frac{y_{ij} - \mu}{\sigma} \quad \mathrm{N}(0,1)\,,\; j = 1,2,\mathsf{K}\,,n_j\,,\; j = 1,2,\mathsf{K}\,,k$$

ovat *riippumattomia*, joten suoraan χ^2*-jakauman* määritelmän mukaan

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\frac{y_{ij} - \mu}{\sigma}\right)^2 \quad \chi^2(N)$$

jossa *vapausasteiden* lukumäärä

$$N = n_1 + n_2 + \mathsf{L} + n_k$$

Korvataan lausekkeessa

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\frac{y_{ij} - \mu}{\sigma}\right)^2 \quad \chi^2(N)$$

tuntematon parametri μ *estimaattorillaan*

$$\bar{y} = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij}$$

Voidaan osoittaa, että tällöin

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\frac{y_{ij} - \bar{y}}{\sigma}\right)^2 = \frac{SST}{\sigma^2} \quad \chi^2(N-1)$$

Menetämme siis *yhden vapausasteen* korvatessamme parametrin μ estimaattorillaan $\bar{y}$.

Tämä selittyy seuraavalla tavalla:

(i) Havaintoarvot y_{ij} (N kpl) *voivat varioida vapaasti*, joten niillä on N vapausastetta.

(ii) Erotukset $y_{ij} - \mu$ (N kpl) *voivat varioida vapaasti*, joten niillä on N vapausastetta.

(iii) Erotukset $y_{ij} - \bar{y}$ (N kpl) *eivät voi varioida vapaasti*, koska niitä sitoo *yksi lineaarinen side-ehto*:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}) = 0$$

(iv) Kohdan (iii) *yksi lineaarinen side-ehto* saa χ^2-jakauman *vapausasteiden lukumäärän vähenemään yhdellä*.

■

Ryhmien välisen vaihtelun neliösumman *SSG* jakauma ja vapausasteet

Oletetaan, että nollahypoteesi

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

pätee. Tällöin satunnaismuuttuja SSG / σ^2 noudattaa χ^2*-jakaumaa* vapausastein $(k - 1)$:

$$\frac{SSG}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2 \quad \chi^2(k-1)$$

Perustelu:

Oletetaan, että havainnot y_{ij} ovat *riippumattomia* ja

$$y_{ij} \sim N(\mu , \sigma^2) , \; i = 1, 2, \ldots , n_j , \; j = 1, 2, \ldots , k$$

Tällöin satunnaismuuttujat

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij} \quad N\left(\mu, \frac{\sigma^2}{n_j}\right), \; j = 1, 2, \mathsf{K}, k$$

ovat *riippumattomia*, jolloin myös satunnaismuuttujat

$$\frac{\bar{y}_j - \mu}{\sigma / \sqrt{n_j}} \quad N(0,1) , \; j = 1, 2, \mathsf{K}, k$$

ovat *riippumattomia* ja suoraan χ^2*-jakauman* määritelmän mukaan

$$\sum_{j=1}^{k}\left(\frac{\bar{y}_j - \mu}{\sigma / \sqrt{n_j}}\right)^2 \quad \chi^2(k)$$

Korvataan tässä lausekkeessa tuntematon parametri μ *estimaattorillaan*

$$\bar{y} = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij}$$

jossa siis

$$N = n_1 + n_2 + \mathsf{L} + n_k$$

Voidaan osoittaa, että tällöin

$$\sum_{j=1}^{k}\left(\frac{\bar{y}_j - \bar{y}}{\sigma / \sqrt{n_j}}\right)^2 = \frac{1}{\sigma^2}\sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2 = \frac{SSG}{\sigma^2} \quad \chi^2(k-1)$$

Menetämme siis *yhden vapausasteen* korvatessamme parametrin μ estimaattorillaan $\bar{y}$.

Tämä selittyy seuraavalla tavalla:

(i) Keskiarvot $\bar{y}_j$ (k kpl) *voivat varioida vapaasti*, joten niillä on k vapausastetta.

(ii) Erotukset $n_j(\bar{y}_j - \mu)$ (k kpl) *voivat varioida vapaasti*, joten niillä on k vapausastetta.

(iii) Erotukset $n_j(\bar{y}_j - \bar{y})$ (k kpl) *eivät voi varioida vapaasti*, koska niitä sitoo *yksi lineaarinen side-ehto*:

$$\sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y}) = 0$$

(iv) Kohdan (iii) *yksi lineaarinen side-ehto* saa χ^2-jakauman *vapausasteiden lukumäärän vähenemään yhdellä*.

Ryhmien sisäisen vaihtelun neliösumma *SSE* ja vapausasteet

Voidaan osoittaa, että *riippumatta siitä päteekö nollahypoteesi*

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

vai ei, niin satunnaismuuttuja SSE / σ^2 noudattaa χ^2*-jakaumaa* vapausastein ($N - k$):

$$\frac{SSE}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2 \quad \chi^2(N-k)$$

jossa

$$N = n_1 + n_2 + \mathrm{L} + n_k$$

Perustelu:

Oletetaan, että havainnot y_{ij} ovat *riippumattomia* ja

$$y_{ij} \sim \mathrm{N}(\mu\ ,\ \sigma^2)\ ,\ i = 1,\ 2,\ \ldots\ ,\ n_j\ ,\ j = 1,\ 2,\ \ldots\ ,\ k$$

Tällöin satunnaismuuttujat

$$\frac{y_{ij} - \mu_j}{\sigma} \quad \mathrm{N}(0,1)\ ,\ i = 1, 2, \mathrm{K}, n_j\ ,\ j = 1, 2, \mathrm{K}, k$$

ovat *riippumattomia* ja suoraan χ^2*-jakauman* määritelmän mukaan

$$\sum_{i=1}^{n_j}\left(\frac{y_{ij} - \mu_j}{\sigma}\right)^2 \quad \chi^2(n_j)\ ,\ j = 1, 2, \mathrm{K}, k$$

Korvataan tässä lausekkeessa tuntemattomat parametrit μ_i *estimaattoreillaan*

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}\ ,\ j = 1, 2, \mathrm{K}, k$$

Voidaan osoittaa, että tällöin

$$\sum_{i=1}^{n_j}\left(\frac{y_{ij} - \bar{y}_j}{\sigma}\right)^2 \quad \chi^2(n_j - 1)\ ,\ j = 1, 2, \mathrm{K}, k$$

Menetämme siis *yhden vapausasteen* korvatessamme parametrin μ_j estimaattorillaan $\bar{y}_j$.

Tämä selittyy seuraavalla tavalla:

(i) Havaintoarvot y_{ij} (n_j kpl) *voivat varioida vapaasti*, joten niillä on n_j vapausastetta.

(ii) Erotukset $y_{ij} - \mu_j$ (n_j kpl) *voivat varioida vapaasti*, joten niillä on n_j vapausastetta.

(iii) Erotukset $y_{ij} - \bar{y}_j$ (n_j kpl) *eivät voi varioida vapaasti*, koska niitä sitoo *yksi lineaarinen side-ehto*:

$$\sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j) = 0 \,,\; j = 1, 2, \mathsf{K}, k$$

(iv) Kohdan (iii) *yksi lineaarinen side-ehto* saa χ^2-jakauman *vapausasteiden lukumäärän vähenemään yhdellä*.

Koska satunnaismuuttujat

$$\sum_{i=1}^{n_j} \left(\frac{y_{ij} - \bar{y}_j}{\sigma} \right)^2 \quad \chi^2(n_j - 1) \,,\; j = 1, 2, \mathsf{K}, k$$

ovat *riippumattomia*, niin

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(\frac{y_{ij} - \bar{y}_j}{\sigma} \right)^2 = \frac{SSE}{\sigma^2} \quad \chi^2(N - k)$$

jossa

$$N - k = \sum_{j=1}^{k} (n_j - 1) = \sum_{j=1}^{k} n_j - k = N - k$$

■

Cochranin lause

Cochranin lause on tärkeä matemaattisen tilastotieteen teoreema, jonka avulla voidaan todistaa monet *yleisen lineaarisen mallin* ja *varianssianalyysin* testisuureita koskevat *jakaumatulokset*.

Oletetaan, että satunnaismuuttujat

$$z_i \,,\; i = 1, 2, \ldots, \nu$$

ovat *riippumattomia* ja noudattavat *standardoitua normaalijakaumaa* N(0,1):

$$z_1, z_2, \mathsf{K}, z_\nu \perp$$
$$z_i \quad \mathrm{N}(0,1) \,,\; i = 1, 2, \mathsf{K}, \nu$$

χ^2-jakauman määritelmän mukaan

$$Q = \sum_{i=1}^{\nu} z_i^2 \quad \chi^2(\nu)$$

Oletetaan, että

$$Q = \sum_{i=1}^{\nu} z_i^2 = Q_1 + Q_2 + \mathsf{L} + Q_s$$

jossa $s \le \nu$ ja Q_i on *neliömuoto*, jonka *aste* on

$$r(Q_i) = \nu_i \, , \; i = 1, 2, \ldots , s$$

Tällöin neliömuodot

$$Q_1, Q_2, \ldots , Q_s$$

ovat **Cochranin lauseen** mukaan *riippumattomia χ^2-jakautuneita satunnaismuuttujia*, joiden vapausteiden lukumäärä ovat $\nu_1, \nu_2, \ldots , \nu_s$, jos ja vain jos

$$\nu = \nu_1 + \nu_2 + \cdots + \nu_s$$

Cochranin lauseesta seuraa erityisesti: Oletetaan, että

$$Q = Q_1 + Q_2 + \mathrm{L} + Q_s \quad \chi^2(\nu)$$

jossa

$$Q_i \quad \chi^2(\nu_i) \, , i = 1, 2, \mathrm{K} \, , s$$

Tällöin

$$\nu = \nu_1 + \nu_2 + \cdots + \nu_s$$

on *välttämätön ja riittävä ehto* sille, että satunnaismuuttujat $Q_1, Q_2, \ldots , Q_s$ ovat *riippumattomia*.

Yksisuuntaisen varianssianalyysin *F*-testisuureen jakauma

Yksisuuntaisen varianssianalyysin testi nollahypoteesille

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

on muotoa

$$F = \frac{N-k}{k-1} \cdot \frac{SSG}{SSE}$$

jossa

SSG = ryhmien *välistä vaihtelua* kuvaava neliösumma

SSE = ryhmien *sisäistä vaihtelua* kuvaava neliösumma

ja

$$N = n_1 + n_2 + \mathrm{L} + n_k$$

Nollahypoteesin H_0 pätiessä

$$F \quad F(k-1, N-k)$$

Perustelu:

Tarkastellaan yksisuuntaisen varianssianalyysin *F-testisuuretta*

$$F = \frac{N-k}{k-1} \cdot \frac{SSG}{SSE}$$

jossa

SSG = ryhmien *välistä vaihtelua* kuvaava neliösumma

SSE = ryhmien *sisäistä vaihtelua* kuvaava neliösumma

ja

$$N = n_1 + n_2 + \mathrm{L} + n_k$$

Varianssianalyysihajotelman mukaan

$$SSG + SSE = SST$$

jossa

SST = ryhmien *kokonaisvaihtelua* kuvaava neliösumma

Edellä on todettu, että nollahypoteesin H_0 pätiessä

$$\frac{SST}{\sigma^2} \quad \chi^2(N-1)$$
$$\frac{SSG}{\sigma^2} \quad \chi^2(k-1)$$
$$\frac{SSE}{\sigma^2} \quad \chi^2(N-k)$$

Cochranin lauseen mukaan neliösummat *SSG* ja *SSE* ovat *riippumattomia*, koska varianssianalyysihajotelman neliösummia vastaavat vapausasteet toteuttavat yhtälön

$$N - 1 = (k - 1) + (N - k)$$

Siten suoraan *F-jakauman* määritelmän mukaan

$$F = \frac{N-k}{k-1} \cdot \frac{SSG}{SSE} = \frac{\dfrac{SSG}{(k-1)\sigma^2}}{\dfrac{SSE}{(N-k)\sigma^2}} \quad F(k-1, N-k)$$

jos nollahypoteesi

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$

pätee.

■

Yksisuuntaisen varianssianalyysin *F*-testisuureen tulkinta

Testisuure

$$F = \frac{N-k}{k-1} \cdot \frac{SSG}{SSE}$$

voidaan tulkita *varianssien vertailutestisuureeksi*, jossa havaintojen y_{ij} varianssin σ^2 estimaattoria

$$MSG = \frac{1}{k-1} SSG = \frac{1}{k-1} \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2$$

verrataan estimaattoriin

$$MSE = \frac{1}{N-k} SSE = \frac{1}{N-k} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

Estimaattori

$$MSE = \frac{1}{N-k} SSE$$

on *aina harhaton* havaintojen y_{ij} varianssille σ^2, mutta *estimaattori*

$$MSG = \frac{1}{k-1} SSG$$

on *harhaton* havaintojen y_{ij} varianssille σ^2 vain, *jos nollahypoteesi*

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

pätee.

Varianssiestimaattorin *MSE* harhattomuus

Estimaattori

$$MSE = \frac{1}{N-k} SSE$$

on *harhaton* havaintojen y_{ij} varianssille σ^2 :

$$E(MSE) = \sigma^2$$

Perustelu:

Oletetaan perustelun yksinkertaistamiseksi, että

$$n_1 = n_2 = \cdots = n_k = n$$

jolloin

$$N = kn$$

Estimaattori

$$MSE = \frac{1}{N-k} SSE = \frac{1}{N-k} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

on *harhaton* varianssille σ^2, jos

$$E(MSE) = \frac{1}{N-k} E(SSE) = \sigma^2$$

Todistamme siksi, että

$$E(SSE) = (N-k)\sigma^2$$

Huomautus:

- Todistuksessa ei oleteta, että nollahypoteesi $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$ pätee.

Yksisuuntaisen varianssianalyysin tilastollinen malli voidaan esittää seuraavassa muodossa (ks. tarkemmin seuraavaa kappaletta):

$$y_{ij} = \mu_j + \varepsilon_{ij}\ ,\ i = 1, 2, \mathrm{K}, n\ ,\ j = 1, 2, \mathrm{K}, k$$

jossa satunnaismuuttujat ε_{ij} ovat *riippumattomia* ja *normaalijakautuneita*:

$$\varepsilon_{ij} \quad \mathrm{N}(0, \sigma^2)\ ,\ i = 1, 2, \mathrm{K}, n\ ,\ j = 1, 2, \mathrm{K}, k$$

Todetaan ensin, että

$$\overline{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij} = \mu_j + \overline{\varepsilon}_j\ ,\ j = 1, 2, \mathrm{K}, k$$

jossa

$$\overline{\varepsilon}_j = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{ij}\ ,\ j = 1, 2, \mathrm{K}, k$$

Satunnaismuuttujat $\overline{\varepsilon}_j$ ovat *riippumattomia* ja *normaalijakautuneita*:

$$\overline{\varepsilon}_j \quad \mathrm{N}\left(0, \frac{\sigma^2}{n}\right),\ j = 1, 2, \mathrm{K}, k$$

Lisäksi

$$y_{ij} - \overline{y}_j = \varepsilon_{ij} - \overline{\varepsilon}_j\ ,\ i = 1, 2, \mathrm{K}, n\ ,\ j = 1, 2, \mathrm{K}, k$$

Edellä esitetyn mukaan

$$\begin{aligned}
SSE &= \sum_{j=1}^{k}\sum_{i=1}^{n}(y_{ij} - \overline{y}_j)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n}(\varepsilon_{ij} - \overline{\varepsilon}_j)^2 \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n}(\varepsilon_{ij}^2 - 2\varepsilon_{ij}\overline{\varepsilon}_j + \overline{\varepsilon}_j^2) \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n}\varepsilon_{ij}^2 - 2n\sum_{j=1}^{k}\overline{\varepsilon}_j^2 + n\sum_{j=1}^{k}\overline{\varepsilon}_j^2 \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n}\varepsilon_{ij}^2 - n\sum_{j=1}^{k}\overline{\varepsilon}_j^2 \\
&= \sum_{j=1}^{k}\left(\sum_{i=1}^{n}\varepsilon_{ij}^2 - n\overline{\varepsilon}_j^2\right)
\end{aligned}$$

Siten

$$\begin{aligned}
\mathrm{E}(SSE) &= \mathrm{E}\left[\sum_{j=1}^{k}\sum_{i=1}^{n}(y_{ij} - \overline{y}_j)^2\right] \\
&= \mathrm{E}\left[\sum_{j=1}^{k}\left(\sum_{i=1}^{n}\varepsilon_{ij}^2 - n\overline{\varepsilon}_j^2\right)\right] \\
&= \sum_{j=1}^{k}\left(\sum_{i=1}^{n}\mathrm{E}(\varepsilon_{ij}^2) - n\,\mathrm{E}(\overline{\varepsilon}_j^2)\right)
\end{aligned}$$

$$= \sum_{j=1}^{k} \left(\sum_{i=1}^{n} \sigma^2 - n \cdot \frac{\sigma^2}{n} \right)$$
$$= \sum_{j=1}^{k} (n\sigma^2 - \sigma^2)$$
$$= k(n-1)\sigma^2 = (kn-k)\sigma^2 = (N-k)\sigma^2$$

mikä on haluttu tulos.

■

Varianssiestimaattorin *MSG* harhattomuus

Jos nollahypoteesi

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

pätee, niin estimaattori

$$MSG = \frac{1}{k-1} SSG$$

on *harhaton* havaintojen y_{ij} varianssille σ^2 :

$$E(MSG) = \sigma^2$$

Perustelu:

Oletetaan perustelun yksinkertaistamiseksi, että

$$n_1 = n_2 = \cdots = n_k = n$$

jolloin

$$N = kn$$

Estimaattori

$$MSG = \frac{1}{k-1} SSG = \frac{1}{k-1} \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2$$

on *harhaton* varianssille σ^2, jos

$$E(MSG) = \frac{1}{k-1} E(SSG) = \sigma^2$$

Todistamme, että estimaattori *MSG* on *harhaton* varianssille σ^2, *jos nollahypoteesi*

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

pätee.

Yksisuuntaisen varianssianalyysin tilastollinen malli voidaan esittää seuraavassa muodossa (ks. tarkemmin seuraavaa kappaletta):

$$y_{ij} = \mu + \tau_j + \varepsilon_{ij} \text{ , } i = 1, 2, \text{K} , n \text{ , } i = 1, 2, \text{K} , k$$

jossa

$$\sum_{j=1}^{k} \tau_j = 0$$

ja satunnaismuuttujat ε_{ij} ovat *riippumattomia* ja *normaalijakautuneita*:

$$\varepsilon_{ij} \quad \mathrm{N}(0, \sigma^2),\ i = 1, 2, \mathrm{K}, n,\ j = 1, 2, \mathrm{K}, k$$

Todetaan ensin, että

$$\bar{y}_j = \frac{1}{n} \sum_{i=1}^{n} y_{ij} = \mu + \tau_j + \bar{\varepsilon}_j,\ j = 1, 2, \mathrm{K}, k$$

jossa

$$\bar{\varepsilon}_j = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ij},\ j = 1, 2, \mathrm{K}, k$$

Lisäksi

$$\bar{y} = \frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{n} y_{ji} = \frac{1}{N} \sum_{j=1}^{k} n(\mu + \tau_j + \bar{\varepsilon}_j) = \mu + \frac{n}{N} \sum_{j=1}^{k} \tau_j + \bar{\varepsilon} = \alpha + \bar{\varepsilon}$$

jossa

$$\bar{\varepsilon} = \frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{n} \varepsilon_{ij} = \frac{1}{N} \sum_{j=1}^{k} n\bar{\varepsilon}_j$$

Satunnaismuuttujat

$$\bar{\varepsilon}_j = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ij},\ j = 1, 2, \mathrm{K}, k$$

ovat *riippumattomia* ja *normaalijakautuneita*:

$$\bar{\varepsilon}_j \quad \mathrm{N}\left(0, \frac{\sigma^2}{n}\right),\ j = 1, 2, \mathrm{K}, k$$

Myös satunnaismuuttuja

$$\bar{\varepsilon} = \frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{n} \varepsilon_{ij}$$

on *normaalijakautunut*:

$$\bar{\varepsilon} \quad \mathrm{N}\left(0, \frac{\sigma^2}{N}\right)$$

Lisäksi

$$\bar{y}_j - \bar{y} = \bar{\varepsilon}_j - \bar{\varepsilon} + \tau_j,\ j = 1, 2, \mathrm{K}, k$$

Edellä olevan mukaan

$$
\begin{aligned}
SSG &= \sum_{j=1}^{k} n(\bar{y}_j - \bar{y})^2 = \sum_{j=1}^{k} n(\bar{\varepsilon}_j - \bar{\varepsilon} + \tau_j)^2 \\
&= \sum_{j=1}^{k} n\left[(\bar{\varepsilon}_j - \bar{\varepsilon}) + \tau_j\right]^2 \\
&= \sum_{j=1}^{k} n(\bar{\varepsilon}_j - \bar{\varepsilon})^2 + 2\sum_{j=1}^{k} n(\bar{\varepsilon}_j - \bar{\varepsilon})\tau_i + n\sum_{j=1}^{k} \tau_j^2 \\
&= \sum_{j=1}^{k} n(\bar{\varepsilon}_j^2 - 2\bar{\varepsilon}_j\bar{\varepsilon} + \bar{\varepsilon}^2) + 2\sum_{j=1}^{k} n(\bar{\varepsilon}_j - \bar{\varepsilon})\tau_i + n\sum_{j=1}^{k} \tau_j^2 \\
&= \sum_{j=1}^{k} n\bar{\varepsilon}_j^2 - 2kn\bar{\varepsilon}^2 + kn\bar{\varepsilon}^2 + 2\sum_{j=1}^{k} n(\bar{\varepsilon}_j - \bar{\varepsilon})\tau_j + n\sum_{j=1}^{k} \tau_j^2 \\
&= \sum_{j=1}^{k} n\bar{\varepsilon}_j^2 - N\bar{\varepsilon}^2 + 2\sum_{j=1}^{k} n(\bar{\varepsilon}_j - \bar{\varepsilon})\tau_j + n\sum_{j=1}^{k} \tau_j^2
\end{aligned}
$$

Siten

$$
\begin{aligned}
\mathrm{E}(SSG) &= \mathrm{E}\left[\sum_{j=1}^{k} n(\bar{y}_j - \bar{y})^2\right] \\
&= \mathrm{E}\left[\sum_{j=1}^{k} n\bar{\varepsilon}_j^2 - N\bar{\varepsilon}^2 + 2\sum_{j=1}^{k} n(\bar{\varepsilon}_j - \bar{\varepsilon})\tau_j + n\sum_{j=1}^{k} \tau_j^2\right] \\
&= \sum_{j=1}^{k} n\,\mathrm{E}(\bar{\varepsilon}_j^2) - N\,\mathrm{E}(\bar{\varepsilon}^2) + 2\sum_{j=1}^{k} n\tau_j\,\mathrm{E}(\bar{\varepsilon}_j - \bar{\varepsilon}) + n\sum_{j=1}^{k} \tau_j^2 \\
&= kn\frac{\sigma^2}{n} - N\frac{\sigma^2}{N} + 2\sum_{j=1}^{k} n\tau_j \cdot 0 + n\sum_{j=1}^{k} \tau_j^2 \\
&= (k-1)\sigma^2 + n\sum_{j=1}^{k} \tau_j^2
\end{aligned}
$$

Siten olemme todistaneet, että

$$
\mathrm{E}(MSG) = \mathrm{E}\left(\frac{SSG}{k-1}\right) = \sigma^2 + \frac{n\sum_{j=1}^{k} \tau_j^2}{k-1}
$$

Koska nollahypoteesi

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

on *ekvivalentti* nollahypoteesin

$$\mathrm{H}_0 : \tau_1 = \tau_2 = \cdots = \tau_k = 0$$

kanssa (ks. tarkemmin seuraavaa kappaletta), niin *nollahypoteesin* H_0 *pätiessä MSG on varianssin* σ^2 *harhaton estimaattori*:

$$\mathrm{E}(MSG) = \sigma^2$$

■

Varianssianalyysitaulukko

Varianssianalyysin tulokset esitetään tavallisesti **varianssianalyysitaulukon** muodossa:

Vaihtelun lähde	**Neliö-summa** ***SS***	**Vapaus-asteet** ***df***	**Varianssi-estimaattori** ***MS***	***F*-testisuure**
Ryhmien välinen vaihtelu	SSG	$k-1$	$MSG = \frac{1}{k-1} SSG$	$F = \frac{MSG}{MSE} = \frac{N-k}{k-1} \cdot \frac{SSG}{SSE}$
Ryhmien sisäinen vaihtelu	SSE	$N-k$	$MSE = \frac{1}{N-k} SSE$	
Kokonais-vaihtelu	SST	$N-1$		

Varianssianalyysitaulukon *neliösummat* toteuttavat yhtälön

$$SST = SSG + SSE$$

Yhtälö on *varianssianalyysihajotelma*. Varianssianalyysitaulukon neliösummien *vapausasteet df* (*df* = degrees of freedom) toteuttavat yhtälön

$$N - 1 = (k - 1) + (N - k)$$

20.4. Yksisuuntaisen varianssianalyysin malli ja sen parametrointi

Yksisuuntaisen varianssianalyysin tilastollinen malli voidaan *parametroida* kahdella erilaisella tavalla. Parametroinnit ovat kuitenkin *ekvivalentteja*.

Parametrointi 1

Yksisuuntaisen varianssianalyysin tilastollinen malli voidaan **parametroida** seuraavalla tavalla:

$$(1) \qquad y_{ij} = \mu_j + \varepsilon_{ij}\ ,\ i = 1, 2, \mathrm{K}, n_j\ ,\ j = 1, 2, \mathrm{K}, k$$

jossa *jäännös-* eli *virhetermit* ε_{ij} ovat *riippumattomia* ja *normaalijakautuneita* satunnaismuuttujia:

$$\varepsilon_{ij} \quad \mathrm{N}(0, \sigma^2)\ ,\ i = 1, 2, \mathrm{K}, n_j\ ,\ j = 1, 2, \mathrm{K}, k$$

Mallissa (1)

y_{ij} = y-muuttujan i. havaintoarvo ryhmässä j

μ_j = y-muuttujan odotusarvo ryhmässä j

$j = 1, 2, \ldots, k$

Ei-satunnaiset vakiot μ_j ja jäännösvarianssi σ^2 ovat yksisuuntaisen varianssianalyysin tilastollisen mallin (1) *parametrit*.

Mallia (1) koskevista oletuksista seuraa, että

$$\mathrm{E}(y_{ij}) = \mu_j \,,\ i = 1, 2, \mathrm{K}, n_j \,,\ j = 1, 2, \mathrm{K}, k$$

ja

$$\mathrm{D}^2(y_{ij}) = \sigma^2 \,,\ i = 1, 2, \mathrm{K}, n_j \,,\ j = 1, 2, \mathrm{K}, k$$

Parametrointi 2

Yksisuuntaisen varianssianalyysin malli voidaan **parametroida** myös seuraavalla tavalla:

(2) $$y_{ij} = \alpha + \tau_j + \varepsilon_{ij} \,,\ i = 1, 2, \mathrm{K}, n_j \,,\ j = 1, 2, \mathrm{K}, k$$

jossa

(3) $$\sum_{j=1}^{k} n_j \tau_j = 0$$

ja

$$\varepsilon_{ij} \quad \mathrm{N}(0, \sigma^2) \,,\ i = 1, 2, \mathrm{K}, n_j \,,\ j = 1, 2, \mathrm{K}, k$$

Ei-satunnaiset vakiot μ ja τ_j sekä jäännösvarianssi σ^2 ovat yksisuuntaisen varianssianalyysin tilastollisen mallin (2) *parametrit*.

Mallia (2) koskevista oletuksista seuraa, että

$$\mathrm{E}(y_{ij}) = \mu + \tau_j \,,\ i = 1, 2, \mathrm{K}, n_j \,,\ j = 1, 2, \mathrm{K}, k$$

ja

$$\mathrm{D}^2(y_{ij}) = \sigma^2 \,,\ i = 1, 2, \mathrm{K}, n_j \,,\ j = 1, 2, \mathrm{K}, k$$

Parametrointien 1 ja 2 ekvivalenssi

Mallissa

(1) $$y_{ij} = \mu_j + \varepsilon_{ij} \,,\ i = 1, 2, \mathrm{K}, n_j \,,\ j = 1, 2, \mathrm{K}, k$$

y-havainnot esitetään seuraavien tekijöiden summana:

μ_j = **ryhmäkohtainen odotusarvo**, $j = 1, 2, \ldots, k$

ε_{ij} = *jäännöstermi*, $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, k$

Mallissa

(2) $$y_{ij} = \alpha + \tau_j + \varepsilon_{ij} \,,\ i = 1, 2, \mathrm{K}, n_j \,,\ j = 1, 2, \mathrm{K}, k$$

jossa

(3) $$\sum_{j=1}^{k} n_j \tau_j = 0$$

y-havainnot esitetään seuraavien tekijöiden summana:

μ = **yleisodotusarvo**

τ_j = **ryhmittelevän tekijän *A* tason** A **vaikutus**, $j = 1, 2, \ldots, k$

ε_{ij} = *jäännöstermi*, $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, k$

Kaavan (1) ja kaavojen (2) ja (3) määrittelemät mallit *ovat ekvivalentteja* – mallit on vain *parametroitu* eri tavoilla.

Perustelu:

Määritellään

$$\mu = \frac{1}{N}\sum_{j=1}^{k} n_j \mu_j$$

jossa

$$N = n_1 + n_2 + \mathrm{L} + n_k$$

Kirjoitetaan identiteetti

$$y_{ij} = \mu + (\mu_j - \mu) + (y_{ij} - \mu_j),\ i = 1, 2, \mathrm{K}, n_j,\ j = 1, 2, \mathrm{K}, k$$

ja merkitään

$$\tau_j = \mu_j - \mu,\ j = 1, 2, \mathrm{K}, k$$

ja

$$y_{ij} - \mu_j = \varepsilon_{ij},\ i = 1, 2, \mathrm{K}, n_j,\ j = 1, 2, \mathrm{K}, k$$

Siten

$$y_{ij} = \mu_j + \varepsilon_{ij}$$
$$y_{ij} = \mu + (\mu_j - \mu) + (y_{ij} - \mu_j)$$
$$y_{ij} = \mu + \tau_j + \varepsilon_{ij}$$
$$i = 1, 2, \mathrm{K}, n_j,\ j = 1, 2, \mathrm{K}, k$$

jossa

$$\mu = \frac{1}{N}\sum_{j=1}^{k} n_j \mu_j,\ N = n_1 + n_2 + \mathrm{L} + n_k$$
$$\tau_j = \mu_j - \mu,\ j = 1, 2, \mathrm{K}, k,\ \sum_{j=1}^{k} n_j \tau_j = 0$$

ovat *yksisuuntaisen varianssianalyysin tilastollisen mallin ekvivalentteja esitysmuotoja*.

■

Mallin

(1) $$y_{ij} = \mu_j + \varepsilon_{ij},\ i = 1, 2, \mathrm{K}, n_j,\ j = 1, 2, \mathrm{K}, k$$

nollahypoteesia

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

vastaa kaavojen

(2) $$y_{ij} = \alpha + \tau_j + \varepsilon_{ij},\ i = 1, 2, \mathrm{K}, n_j,\ j = 1, 2, \mathrm{K}, k$$

ja

$$(3) \qquad y_{ij} = \alpha + \tau_j + \varepsilon_{ij} \,,\, i = 1, 2, \mathrm{K}, n_j \,,\, j = 1, 2, \mathrm{K}, k$$

määrittelemässä mallissa **nollahypoteesi**

$$\mathrm{H}_0 : \tau_1 = \tau_2 = \cdots = \tau_k = 0$$

Malli (1) kiinnittää huomion suoraan *ryhmäkohtaisiin odotusarvoihin* μ_j , kun taas kaavojen (2) ja (3) määrittelemässä mallissa *ryhmäkohtaiset odotusarvot* μ_j on esitetty *yleisodotusarvon* μ ja *ryhmittelevän tekijän tasoon j liittyvän vaikutuksen* (efektin) τ_j summana.

Yksisuuntaisen varianssianalyysin malli ja yleinen lineaarinen malli

Yksisuuntaisen varianssianalyysin malli on erikoistapaus **yleisestä lineaarisesta mallista**; ks. lukua **Yleinen lineaarinen malli**.

Olkoon

y_{ij} = *i*. havainto ryhmässä *j* , $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, k$

Määritellään *ryhmäindikaattorit*

$$I_{ij} = \begin{cases} 1, \text{ jos havainto } y_{ij} \text{ kuuluu ryhmään } j \\ 0, \text{ jos havainto } y_{ij} \text{ ei kuulu ryhmään } j \end{cases}$$
$$i = 1, 2, \mathrm{K}, n_j \,,\, j = 1, 2, \mathrm{K}, k$$

Tällöin yksisuuntaisen varianssianalyysin malli

$$(1) \qquad y_{ij} = \mu_j + \varepsilon_{ij} \,,\, i = 1, 2, \mathrm{K}, n_j \,,\, j = 1, 2, \mathrm{K}, k$$

voidaan esittää muodossa

$$(4) \qquad y_{ij} = \mu_1 I_{i1} + \mu_2 I_{i2} + \mathrm{L} + \mu_k I_{ik} + \varepsilon_{ij} \,,\, i = 1, 2, \mathrm{K}, n_j \,,\, j = 1, 2, \mathrm{K}, k$$

Malli (4) on **lineaarinen regressiomalli**, jossa **selitettävänä muuttujana** on *y*, **selittävinä muuttujina** ovat *ryhmäindikaattorit*

I_j , $j = 1, 2, \ldots, k$

ja **regressiokertoimina** ovat ryhmäkohtaiset odotusarvot

$\mu_1, \mu_2, \ldots, \mu_k$

Malli (4) on **yleisen lineaarisen mallin** erikoistapaus, jossa selitettävä muuttuja *y* on *kvantitatiivinen*, mutta selittäjät I_j *kvalitatiivisia* (*kategorisia*) *muuttujia*.

Huomaa, että regressiomallissa (4) *ei saa* olla vakiotermiä, koska sen lisääminen malliin loisi selittävien muuttujan arvojen välille eksaktin *lineaarisen riippuvuuden*:

$$I_{i1} + I_{i2} + \mathrm{L} + I_{ik} = 1 \,,\, i = 1, 2, \mathrm{K}, n_j$$

sillä kaikille *i* täsmälleen yksi ryhmäindikaattoreista I_j , $j = 1, 2, \ldots, k$ saa arvon = 1 ja kaikki muut saavat arvon = 0.

20.5. Yksisuuntaisen varianssianalyysin mallin parametrien estimointi

Yksisuuntaisen varianssianalyysin malli

Olkoon

(1) $$y_{ij} = \mu_j + \varepsilon_{ij}\ ,\ i = 1, 2, \mathrm{K}, n_j\ ,\ j = 1, 2, \mathrm{K}, k$$

yksisuuntaisen varianssianalyysin tilastollinen malli, jossa *jäännös-* eli *virhetermit* ε_{ij} ovat *riippumattomia* ja *normaalijakautuneita* satunnaismuuttujia:

$$\varepsilon_{ij} \quad \mathrm{N}(0, \sigma^2)\ ,\ i = 1, 2, \mathrm{K}, n_j\ ,\ j = 1, 2, \mathrm{K}, k$$

Mallissa (1)

y_{ij} = y-muuttujan i. havaintoarvo ryhmässä j

μ_j = y-muuttujan odotusarvo ryhmässä j

Jos nollahypoteesi

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

otetaan huomioon, malli (1) saa muodon

(2) $$y_{ij} = \mu + \varepsilon_{ij}\ ,\ i = 1, 2, \mathrm{K}, n_j\ ,\ j = 1, 2, \mathrm{K}, k$$

Pienimmän neliösumman estimointi

Tarkastellaan mallin (1) parametrien μ_j , j = 1, 2, … , k *pienimmän neliösumman estimointia*.

Mallin (1) parametrien $\mu_1, \mu_2, \mathrm{K}, \mu_k$ **pienimmän neliösumman estimaattoreiksi** $\hat{\mu}_1, \hat{\mu}_2, \mathrm{K}, \hat{\mu}_k$ saadaan havaintoarvojen **ryhmäkeskiarvot** $\bar{y}_j$:

$$\hat{\mu}_j = \bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}\ ,\ j = 1, 2, \mathrm{K}, k$$

Perustelu:

Estimoidaan mallin

(1) $$y_{ij} = \mu_j + \varepsilon_{ij}\ ,\ i = 1, 2, \mathrm{K}, n_j\ ,\ j = 1, 2, \mathrm{K}, k$$

parametrit μ_j , j = 1, 2, … , k *PNS-menetelmällä*. Etsitään *neliösumman*

$$SS(\mu_1, \mu_2, \mathrm{K}, \mu_k) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \varepsilon_{ij}^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \mu_j)^2$$

minimi parametrien suhteen tavanomaiseen tapaan:

(i) *Derivoidaan* neliösumma $SS(\mu_1, \mu_2, \mathrm{K}, \mu_k)$ parametrien μ_j , j = 1, 2, … , k suhteen.

(ii) Merkitään derivaatat nolliksi.

(iii) Ratkaistaan saadut *normaaliyhtälöt* parametrien μ_j , j = 1, 2, … , k suhteen.

Normaaliyhtälöiksi saadaan:

$$\frac{\partial}{\partial \mu_j} SS(\mu_1, \mu_2, \mathsf{K}, \mu_k) = \frac{\partial}{\partial \mu_j} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \mu_j)^2$$
$$= -2 \sum_{i=1}^{n_j} (y_{ij} - \mu_j)$$
$$= -2 \left(\sum_{i=1}^{n_j} y_{ij} - n_j \mu_j \right)$$
$$= 0 \,,\; j = 1, 2, \mathsf{K}, k$$

Normaaliyhtälöiden ratkaisuiksi saadaan *ryhmäkeskiarvot*

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} = \overline{y}_j \,,\; j = 1, 2, \mathsf{K}, k$$

■

Estimoidun mallin **sovitteet** saadaan kaavoilla

$$\hat{y}_{ij} = \hat{\mu}_j = \overline{y}_j \,, i = 1, 2, \mathsf{K}, n_j \,,\; j = 1, 2, \mathsf{K}, k$$

ja **residuaalit** kaavoilla

$$e_{ij} = y_{ij} - \hat{y}_j = y_{ij} - \overline{y}_j \,, i = 1, 2, \mathsf{K}, n_j \,,\; j = 1, 2, \mathsf{K}, k$$

Jäännösneliösummaksi saadaan *ryhmien sisäistä vaihtelua kuvaava neliösumma*

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_j)^2$$

Pienimmän neliösumman estimointi yhtä suurten odotusarvojen tapauksessa

Mallin (2) parametrin μ **pienimmän neliösumman estimaattoriksi** $\hat{\mu}$ saadaan havaintoarvojen y_{ij} **yleis**- eli **kokonaiskeskiarvo** $\overline{y}$:

$$\hat{\mu} = \overline{y} = \frac{1}{N} \sum_{j=1}^{k} \sum_{j=1}^{n_j} y_{ij} = \frac{1}{N} \sum_{j=1}^{k} n_j \overline{y}_j$$

jossa

$$N = n_1 + n_2 + \mathsf{L} + n_k$$

ja

$$\overline{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \,,\; j = 1, 2, \mathsf{K}, k$$

Perustelu:

Estimoidaan mallin

(2) $$y_{ij} = \mu + \varepsilon_{ij} \,, i = 1, 2, \mathsf{K}, n_j \,,\; j = 1, 2, \mathsf{K}, k$$

parametri μ *PNS-menetelmällä*. Etsitään *neliösumman*

$$SS(\mu) = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \varepsilon_{ij}^2 = \sum_{j=1}^{k}\sum_{j=1}^{n_j} (y_{ij} - \mu)^2$$

minimi tavanomaiseen tapaan:

(i) *Derivoidaan* neliösumma $S(\mu)$ parametrin μ suhteen.

(ii) Merkitään derivaatta nollaksi.

(iii) Ratkaistaan saatu *normaaliyhtälö* parametrin μ suhteen.

Normaaliyhtälöksi saadaan:

$$\begin{aligned}\frac{\partial}{\partial\mu} SS(\mu) &= \frac{\partial}{\partial\mu}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \mu)^2 \\ &= -2\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \mu) \\ &= -2\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij} - N\mu\right) \\ &= 0\end{aligned}$$

jossa

$$N = n_1 + n_2 + \mathsf{L} + n_k$$

Normaaliyhtälön ratkaisuksi saadaan *yleiskeskiarvo*

$$\hat{\mu} = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij} = \bar{y}$$

■

Estimoidun mallin **sovitteet** saadaan kaavoilla

$$\hat{y}_{ij} = \hat{\mu} = \bar{y}\ ,\ i = 1, 2, \mathsf{K}, n_j\ ,\ j = 1, 2, \mathsf{K}, k$$

ja **residuaalit** kaavoilla

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}\ ,\ i = 1, 2, \mathsf{K}, n_j\ ,\ j = 1, 2, \mathsf{K}, k$$

Jäännösneliösummaksi saadaan havaintoarvojen y_{ji} *kokonaisvaihtelua kuvaava kokonais-neliösumma*

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2$$

Testi odotusarvojen samuudelle

Jos nollahypoteesi

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

otetaan huomioon, että regressiomallin (1) kertoimille tulee asetetuksi $(k - 1)$ **lineaarista rajoitusta** eli **side-ehtoa**:

(3) $\mu_1 - \mu_j = 0$, $j = 2, 3, \ldots, k$

Yleisen lineaarisen mallin teorian mukaan (ks. luvun **Erityiskysymyksiä yleisen lineaarisen mallin soveltamisessa** kappaletta **Rajoitettu pienimmän neliösumman menetelmä**) side-ehtojen (3) testaaminen voidaan perustaa *F*-testisuureeseen

$$F = \frac{N-k}{k-1} \cdot \frac{SST - SSE}{SSE}$$

jossa

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \overline{y})^2$$

on havaintoarvojen y_{ji} *kokonaisvaihtelua kuvaava kokonaisneliösumma* ja

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_j)^2$$

on *ryhmien sisäistä vaihtelua kuvaava jäännösneliösumma*. Ottamalla huomioon *varianssianalyysi-hajotelma*

$$SST = SSG + SSE$$

testisuure *F* voidaan kirjoittaa myös muotoon

$$F = \frac{N-k}{k-1} \cdot \frac{SSG}{SSE}$$

jossa

$$SSG = SST - SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\overline{y}_j - \overline{y})^2 = \sum_{j=1}^{k} n_j (\overline{y}_j - \overline{y})^2$$

on *ryhmien välistä* (systemaattista) *vaihtelua kuvaava ryhmäneliösumma*. Jos side-ehdot (3) *pätevät*, testisuure *F* noudattaa *F-jakaumaa* vapausastein $(k-1)$ ja $(N-k)$:

$$F \quad F(k-1, N-k)$$

Suuret testisuureen *F* arvot viittaavat siihen, että nollahypoteesi H_0 *on syytä hylätä*.

Sovitteet ja residuaalit

Kuten edellä jo todettiin, estimoidun yksisuuntaisen varianssianalyysin mallin **sovitteet** saadaan kaavoilla

$$\hat{y}_{ij} = \hat{\mu}_j = \overline{y}_j \text{ , } i = 1, 2, \mathsf{K}, n_j \text{ , } j = 1, 2, \mathsf{K}, k$$

ja estimoidun mallin **residuaalit** saadaan kaavoilla

$$e_{ij} = y_{ij} - \hat{y}_j = y_{ij} - \overline{y}_j \text{ , } i = 1, 2, \mathsf{K}, n_j \text{ , } j = 1, 2, \mathsf{K}, k$$

Mallin residuaalit on aina syytä alistaa samanlaisiin **diagnostisiin tarkistuksiin** kuin minkä tahansa regressiomallin residuaalit. Residuaalien avulla voidaan selvittää onko havaintojen joukossa **poikkeavia havaintoja** sekä pätevätkö mallin jäännöstermeistä tehdyt **homoskedastisuus-**, **korreloimattomuus-** ja **normaalisuusoletukset**; ks. yksityiskohtia luvusta **Regressiodiagnostiikka**.

20.6. Yksisuuntaisen varianssianalyysin mallin matriisiesitys

Olkoon

$$(1) \qquad y_{ij} = \mu_j + \varepsilon_{ij}\ ,\ i = 1,2,\mathsf{K}\ ,n_j\ ,\ j = 1,2,\mathsf{K}\ ,k$$

yksisuuntaisen varianssianalyysin tilastollinen malli, jossa *jäännös*- eli *virhetermit* ε_{ij} ovat *riippumattomia* ja *normaalijakautuneita* satunnaismuuttujia:

$$\varepsilon_{ij} \quad \mathrm{N}(0,\sigma^2)\ ,\ i = 1,2,\mathsf{K}\ ,n_j\ ,\ j = 1,2,\mathsf{K}\ ,k$$

Mallissa (1)

y_{ij} = y-muuttujan i. havaintoarvo ryhmässä j

μ_j = y-muuttujan odotusarvo ryhmässä j

Määritellään *ryhmäindikaattorit*

$$I_{ij} = \begin{cases} 1, \text{ jos havainto } y_{ij} \text{ kuuluu ryhmään } j \\ 0, \text{ jos havainto } y_{ij} \text{ ei kuulu ryhmään } j \end{cases}$$
$$i = 1,2,\mathsf{K}\ ,n_j\ ,\ j = 1,2,\mathsf{K}\ ,k$$

Tällöin yksisuuntaisen varianssianalyysin malli

$$(1) \qquad y_{ij} = \mu_j + \varepsilon_{ij}\ ,\ i = 1,2,\mathsf{K}\ ,n_j\ ,\ j = 1,2,\mathsf{K}\ ,k$$

voidaan esittää muodossa

$$(2) \qquad y_{ij} = \mu_1 I_{i1} + \mu_2 I_{i2} + \mathsf{L} + \mu_k I_{ik} + \varepsilon_{ij}\ ,\ i = 1,2,\mathsf{K}\ ,n_j\ ,\ j = 1,2,\mathsf{K}\ ,k$$

Matriisiesitys

Yhtälöt (2) voidaan kirjoittaa *matriisein* seuraavassa muodossa:

$$\begin{bmatrix} y_{11} \\ \mathsf{M} \\ y_{n_1 1} \\ y_{12} \\ \mathsf{M} \\ y_{n_2 2} \\ \mathsf{M} \\ \mathsf{M} \\ \mathsf{M} \\ y_{1k} \\ \mathsf{M} \\ y_{n_k k} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \mathsf{L} & 0 \\ \mathsf{M} & \mathsf{M} & \mathsf{M} & & \mathsf{M} \\ 1 & 0 & 0 & \mathsf{L} & 0 \\ 0 & 1 & 0 & \mathsf{L} & 0 \\ \mathsf{M} & \mathsf{M} & \mathsf{M} & & \mathsf{M} \\ 0 & 1 & 0 & \mathsf{L} & 0 \\ \mathsf{M} & \mathsf{M} & \mathsf{M} & \mathsf{L} & \mathsf{M} \\ \mathsf{M} & \mathsf{M} & \mathsf{M} & & \mathsf{M} \\ \mathsf{M} & \mathsf{M} & \mathsf{M} & \mathsf{L} & \mathsf{M} \\ 0 & 0 & 0 & \mathsf{L} & 1 \\ \mathsf{M} & \mathsf{M} & \mathsf{M} & & \mathsf{M} \\ 0 & 0 & 0 & \mathsf{L} & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mathsf{M} \\ \mu_k \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \mathsf{M} \\ \varepsilon_{n_1 1} \\ \varepsilon_{12} \\ \mathsf{M} \\ \varepsilon_{n_2 2} \\ \mathsf{M} \\ \mathsf{M} \\ \mathsf{M} \\ \varepsilon_{1k} \\ \mathsf{M} \\ \varepsilon_{n_k k} \end{bmatrix}$$

Esitetään tämä matriisiyhtälö muodossa

$$\mathbf{y} = \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

jossa

$\mathbf{y}$ = havaintoarvojen y_{ji} muodostama N-vektori, $N = n_1 + n_2 + \mathrm{L} + n_k$

$\mathbf{X}$ = nollien ja ykkösten muodostama *täysiasteinen* $N \times k$-matriisi

$\boldsymbol{\mu}$ = regressiokertoimien (= ryhmäodotusarvojen) muodostama k-vektori

$\boldsymbol{\varepsilon}$ = jäännöstermien ε_{oj} muodostama N-vektori

Huomaa, että matriisin **X** jokaisella rivillä on *täsmälleen yksi ykkönen* ja muut ko. rivin alkiot ovat nollia.

Pienimmän neliösumman estimointi

Regressiokertoimien vektorin $\boldsymbol{\mu}$ *PNS-estimaattori* on

$$\hat{\boldsymbol{\mu}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Matriisin **X** erikoisen rakenteen takia nähdään helposti, että matriisi $\mathbf{X}'\mathbf{X}$ on *diagonaalimatriisi*:

$$\mathbf{X}'\mathbf{X} = \mathrm{diag}(n_1, n_2, \mathrm{K}, n_k) = \begin{bmatrix} n_1 & 0 & 0 & \mathrm{L} & 0 \\ 0 & n_2 & 0 & \mathrm{L} & 0 \\ 0 & 0 & n_3 & \mathrm{L} & 0 \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ 0 & 0 & 0 & \mathrm{L} & n_k \end{bmatrix}$$

Lisäksi

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \Sigma y_{i1} \\ \Sigma y_{i2} \\ \Sigma y_{i3} \\ \mathrm{M} \\ \Sigma y_{ik} \end{bmatrix}$$

Koska $\mathbf{X}'\mathbf{X}$ on diagonaalimatriisi, niin

$$(\mathbf{X}'\mathbf{X})^{-1} = \mathrm{diag}(1/n_1, 1/n_2, \mathrm{K}, 1/n_k) = \begin{bmatrix} \frac{1}{n_1} & 0 & 0 & \mathrm{L} & 0 \\ 0 & \frac{1}{n_2} & 0 & \mathrm{L} & 0 \\ 0 & 0 & \frac{1}{n_3} & \mathrm{L} & 0 \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ 0 & 0 & 0 & \mathrm{L} & \frac{1}{n_k} \end{bmatrix}$$

Siten regressiokertoimien vektorin $\boldsymbol{\mu}$ PNS-estimaattoriksi $\hat{\boldsymbol{\mu}}$ saadaan ryhmäkeskiarvojen

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij},\ j = 1, 2, \mathrm{K}, k$$

muodostama vektori:

$$\hat{\boldsymbol{\mu}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} \frac{1}{n_1}\Sigma y_{i1} \\ \frac{1}{n_2}\Sigma y_{i2} \\ \frac{1}{n_3}\Sigma y_{i3} \\ \vdots \\ \frac{1}{n_k}\Sigma y_{ik} \end{bmatrix} = \begin{bmatrix} \overline{y}_1 \\ \overline{y}_2 \\ \overline{y}_3 \\ \vdots \\ \overline{y}_k \end{bmatrix}$$

Edellä todettiin, että jos nollahypoteesi

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

otetaan huomioon, niin malli

$$(1) \qquad y_{ij} = \mu_j + \varepsilon_{ij}\ ,\ i = 1, 2, \mathrm{K}\ , n_j\ ,\ j = 1, 2, \mathrm{K}\ , k$$

saa muodon

$$(3) \qquad y_{ij} = \mu + \varepsilon_{ij}\ ,\ i = 1, 2, \mathrm{K}\ , n_j\ ,\ j = 1, 2, \mathrm{K}\ , k$$

Yhtälöt (3) voidaan kirjoittaa *matriisein* seuraavaan muotoon:

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{n_1 1} \\ y_{12} \\ \vdots \\ y_{n_2 2} \\ \vdots \\ \vdots \\ \vdots \\ y_{1k} \\ \vdots \\ y_{n_k k} \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ \vdots \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix} \mu + \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{n_1 1} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{n_2 2} \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_{1k} \\ \vdots \\ \varepsilon_{n_k k} \end{bmatrix}$$

Esitetään tämä matriisiyhtälö muodossa

$$\mathbf{y} = \mathbf{1}\mu + \boldsymbol{\delta}$$

jossa

$\mathbf{y}$ = havaintoarvojen y_{ji} muodostama N-vektori, $N = n_1 + n_2 + \mathrm{L} + n_k$

$\mathbf{1}$ = ykkösten muodostama N-vektori

μ = regressiokerroin

$\boldsymbol{\delta}$ = jäännöstermien δ_{ij} muodostama N-vektori

Regressiokertoimen μ *PNS-estimaattori* on

$$\hat{\mu} = (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{y}$$

Helposti nähdään, että

$$\mathbf{1}'\mathbf{1} = N$$

ja

$$\mathbf{1}'\mathbf{y} = \Sigma\Sigma y_{ij}$$

Siten

$$(\mathbf{1}'\mathbf{1})^{-1} = \frac{1}{N}$$

ja regressiokertoimen μ *PNS-estimaattoriksi* $\hat{\mu}$ saadaan havaintoarvojen yleiskeskiarvo $\overline{y}$:

$$\hat{\mu} = (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{y} = \frac{1}{N}\Sigma\Sigma y_{ji} = \overline{y}$$

20.7. Laskutoimitusten suorittaminen

Olkoon

$$(1) \qquad y_{ij} = \mu_j + \varepsilon_{ij}\ ,\ i = 1, 2, \mathsf{K}, n_j\ ,\ j = 1, 2, \mathsf{K}, k$$

yksisuuntaisen varianssianalyysin tilastollinen malli, jossa *jäännös-* eli *virhetermit* ε_{ij} ovat *riippumattomia* ja *normaalijakautuneita* satunnaismuuttujia:

$$\varepsilon_{ij} \quad \mathrm{N}(0, \sigma^2)\ ,\ i = 1, 2, \mathsf{K}, n_j\ ,\ j = 1, 2, \mathsf{K}, k$$

Mallissa (1)

y_{ij} = y-muuttujan i. havaintoarvo ryhmässä j

μ_j = y-muuttujan odotusarvo ryhmässä j

Määritellään *ryhmän* $j = 1, 2, \ldots, k$ *havaintoarvojen* y_{ij} *summa* kaavalla

$$T_j = \sum_{i=1}^{n_j} y_{ij}\ ,\ j = 1, 2, \mathsf{K}, k$$

ja *kaikkien havaintoarvojen* y_{ij} *kokonaissumma* kaavalla

$$T = \sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij} = \sum_{j=1}^{k} T_j$$

Olkoon lisäksi

$$N = n_1 + n_2 + \mathsf{L} + n_k$$

havaintojen kokonaislukumäärä.

Havaintoarvojen *ryhmäkeskiarvot* saadaan kaavoilla

$$\overline{y}_j = \frac{1}{n_j} T_j \ , \ j = 1, 2, \mathrm{K}, k$$

ja havaintoarvojen *yleiskeskiarvo* saadaan kaavalla

$$\overline{y} = \frac{1}{N} T$$

Edelleen *kokonaisneliösumma SST* voidaan kirjoittaa muotoon

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \overline{y})^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} y_{ij}^2 - \frac{1}{N} T^2$$

ja *ryhmien välistä vaihtelua kuvaava neliösumma SSG* voidaan kirjoittaa muotoon

$$SSG = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\overline{y}_j - \overline{y})^2 = \sum_{j=1}^{k} n_j (\overline{y}_j - \overline{y})^2 = \sum_{j=1}^{k} \frac{1}{n_j} T_j^2 - \frac{1}{N} T^2$$

Varianssianalyysihajotelmasta seuraa, että *ryhmien sisäistä vaihtelua kuvaava neliösumma SSE* saadaan kaavalla

$$SSE = SST - SSG$$

F-testisuure nollahypoteesin

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

testaamiseksi on muotoa

$$F = \frac{N - k}{k - 1} \times \frac{SSG}{SSE}$$

Bartlettin testissä (ks. alla) tarvittavat *ryhmäkohtaiset varianssit* saadaan kaavoilla

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_j)^2 = \frac{1}{n_j - 1} \left(\sum_{i=1}^{n_j} y_{ij}^2 - \frac{1}{n} T_j^2 \right), \ j = 1, 2, \mathrm{K}, k$$

Siten yksisuuntaisen varianssianalyysin perustehtävien suorittamiseksi riittää laskea seuraavat *summat* ja *neliösummat*:

$$T_j = \sum_{i=1}^{n_j} y_{ij} \ , \ j = 1, 2, \mathrm{K}, k \qquad\qquad T = \sum_{j=1}^{k} \sum_{i=1}^{n_j} y_{ij}$$

$$\sum_{i=1}^{n_j} y_{ij}^2 \ , \ j = 1, 2, \mathrm{K}, k \qquad\qquad \sum_{j=1}^{k} \sum_{i=1}^{n_j} y_{ij}^2$$

Laskutoimitukset voidaan järjestää esimerkiksi seuraavan taulukon muotoon:

Ryhmä 1	Ryhmä 2	L	Ryhmä k	Summa
n_1	n_2	L	n_k	$N = \sum_{j=1}^{k} n_j$
y_{11}	y_{12}	L	y_{1k}	
M	M		M	
$y_{n_1 1}$	$y_{n_2 2}$	L	$y_{n_k k}$	
$T_1 = \sum_{i=1}^{n_1} y_{i1}$	$T_2 = \sum_{i=1}^{n_2} y_{i2}$	L	$T_k = \sum_{i=1}^{n_k} y_{ik}$	$T = \sum_{j=1}^{k} T_j$
T_1^2	T_2^2	L	T_k^2	
$\sum_{i=1}^{n_1} y_{i1}^2$	$\sum_{i=1}^{n_2} y_{i2}^2$	L	$\sum_{i=1}^{n_k} y_{ik}^2$	$\sum_{j=1}^{k} \sum_{i=1}^{n_j} y_{ij}^2$

20.8. Bartlettin testi

Bartlettin testi on tilastollinen testi, jolla voidaan testata yksisuuntaisen varianssianalyysin oletusta *havaintojen ryhmäkohtaisten varianssien yhtäsuuruudesta*.

Testausasetelma

Olkoon

(1) $$y_{ij} = \mu_j + \varepsilon_{ij}\ ,\ i = 1, 2, \mathrm{K}, n_j\ ,\ j = 1, 2, \mathrm{K}, k$$

yksisuuntaisen varianssianalyysin tilastollinen malli. Mallissa (1)

y_{ij} = y-muuttujan i. havaintoarvo ryhmässä j

μ_j = y-muuttujan odotusarvo ryhmässä j

Olkoon lisäksi

$$N = n_1 + n_2 + \mathrm{L} + n_k\ .$$

havaintojen kokonaislukumäärä.

Tehdään oletus, että *jäännös-* eli *virhetermit* ε_{ij} ovat *riippumattomia* ja *normaalijakautuneita* satunnaismuuttujia:

$$\varepsilon_{ij} \quad \mathrm{N}(0, \sigma_j^2)\ ,\ i = 1, 2, \mathrm{K}, n_j\ ,\ j = 1, 2, \mathrm{K}, k$$

Oletuksen mukaan *ryhmäkohtaiset varianssit*

$$\mathrm{D}^2(y_{ij}) = \sigma_j^2\ ,\ i = 1, 2, \mathrm{K}, n\ ,\ j = 1, 2, \mathrm{K}, k$$

voivat erota toisistaan. Yksisuuntaisen varianssianalyysin *F-testi* nollahypoteesille

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

nojaa oletukseen, jonka mukaan *kaikilla havainnoilla* y_{ij} *on* (ryhmästä riippumatta) *sama varianssi*:

$$\mathrm{D}^2(y_{ij}) = \sigma^2\ ,\ i = 1, 2, \mathrm{K}, n\ ,\ j = 1, 2, \mathrm{K}, k$$

Asetetaan siksi nollahypoteesi

$$\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2 = \mathrm{L} = \sigma_j^2 = \sigma^2$$

Testisuure ja sen jakauma

Määritellään *havaintoarvojen y_{ij} ryhmäkohtaiset otosvarianssit* s_j^2 kaavalla

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \ , \ j = 1, 2, \mathsf{K}, k$$

jossa

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \ , \ j = 1, 2, \mathsf{K}, k$$

on *ryhmän i havaintoarvojen aritmeettinen keskiarvo* ja ryhmäkohtaisista otosvariansseista s_j^2 *yhdistetty varianssi* s_P^2 kaavalla

$$s_P^2 = \frac{1}{N-k} \sum_{j=1}^{k} (n_j - 1) s_j^2 = \frac{1}{N-k} SSE = MSE$$

Bartlettin testisuure on

$$B = \frac{Q}{h}$$

jossa

$$Q = (N-k) \log(s_P^2) - \sum_{j=1}^{k} (n_j - 1) \log(s_j^2)$$

ja

$$h = 1 + \frac{1}{3(k-1)} \left(\sum_{j=1}^{k} \frac{1}{n_j - 1} - \frac{1}{N-k} \right)$$

Jos nollahypoteesi $\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2 = \mathsf{L} = \sigma_j^2 = \sigma^2$ pätee, Bartlettin testisuure *B* noudattaa *suurissa otoksissa approksimatiivisesti* χ^2*-jakaumaa* vapausastein (k – 1):

$$B \ \ {}_a \ \chi^2(k-1)$$

Testisuureen *B normaaliarvo* on suurissa otoksissa approksimatiivisesti (k – 1), koska

$$\mathrm{E}(\chi^2) = k - 1$$

jossa

$$\chi^2 \quad \chi^2(k-1)$$

Siten *suuret* testisuureen χ^2 arvot viittaavat siihen, että nollahypoteesi H_0 on syytä hylätä.

20.9. Odotusarvoparien vertailu

Olkoon

(1) $$y_{ij} = \mu_j + \varepsilon_{ij} \ , i = 1, 2, \mathsf{K}, n_j \ , \ j = 1, 2, \mathsf{K}, k$$

yksisuuntaisen varianssianalyysin tilastollinen malli, jossa *jäännös-* eli *virhetermit* ε_{ij} ovat *riippumattomia* ja *normaalijakautuneita* satunnaismuuttujia:

$$\varepsilon_{ij} \quad \mathrm{N}(0,\sigma^2),\ i=1,2,\mathrm{K},n_j,\ j=1,2,\mathrm{K},k$$

Mallissa (1)

y_{ij} = y-muuttujan i. havaintoarvo ryhmässä j

μ_j = y-muuttujan odotusarvo ryhmässä j

Olkoon lisäksi

$$N = n_1 + n_2 + \mathrm{L} + n_k$$

havaintojen kokonaislukumäärä.

Jos yksisuuntaisen varianssianalyysin nollahypoteesi

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$

hylätään tiedetään, että *ainakin kaksi odotusarvoista*

$$\mu_j,\ j = 1, 2, \ldots, k$$

eroaa tilastollisesti merkitsevästi toisistaan. Jos nollahypoteesi H_0 hylätään, varianssianalyysia voidaan jatkaa *ryhmittelyllä*, jossa selvitetään *missä ryhmissä odotusarvojen erot ovat tilastollisesti merkitseviä*. Vertailu voidaan tehdä käyttämällä *luottamusvälejä* tai *testejä*.

Luottamusvälit ja odotusarvojen parivertailu

Oletetaan, että haluamme verrata odotusarvoja μ_k ja μ_l. Odotusarvojen vertailu voidaan tehdä siten, että konstruoidaan odotusarvojen μ_k ja μ_l erotukselle

$$\mu_k - \mu_l$$

luottamusväli ja tutkitaan kuuluuko *nolla* konstruoituun väliin vai ei.

Käytetään **erotuksen** $\mu_k - \mu_l$ **luottamusvälinä** väliä

$$(\bar{y}_k - \bar{y}_l) \pm t_{\alpha/2}\, s_P \sqrt{\frac{1}{n_k} + \frac{1}{n_l}}$$

jossa

$$s_P^2 = \frac{1}{N-k}\sum_{j=1}^{k}(n_j - 1)s_j^2 = \frac{1}{N-k}SSE = MSE$$

on ns. *yhdistetty varianssi*, jossa

$$s_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2,\ j = 1,2,\mathrm{K},k$$

on havaintoarvojen y_{ij} varianssi ryhmässä i ja *luottamustasoa* $(1 - \alpha)$ *vastaavat luottamuskertoimet* $-t_{\alpha/2}$ ja $+t_{\alpha/2}$ on valittu siten, että

$$\Pr(-t_{\alpha/2} \le t \le +t_{\alpha/2}) = 1 - \alpha$$

jossa satunnaismuuttuja t noudattaa *t-jakaumaa* vapausastein $(N - k)$:

$$t \quad t(N-k)$$

Huomautuksia:

- Tässä käytettävä luottamusvälin kaava eroaa tavanomaisesta kaavasta siten, että yhdistetyn varianssin s_P^2 kaavassa on yhdistetty otosvarianssit *kaikista ryhmistä* eikä vain ryhmistä *k* ja *l*.
- Konstruoidut luottamusvälit *eivät ole simultaanisia*, vaan koskevat vain ryhmien *k* ja *l* odotusarvoja.
- Tehtävien *parivertailujen lukumäärä* on

$$\binom{k}{2} = \frac{k(k-1)}{2}$$

Testit ja odotusarvojen parivertailu

Oletetaan, että haluamme verrata odotusarvoja μ_k ja μ_l. Odotusarvojen vertailu voidaan tehdä siten, että *testataan* nollahypoteesia

$$\mathrm{H}_0 : \mu_k - \mu_l = 0$$

vaihtoehtoista hypoteesia

$$\mathrm{H}_1 : \mu_k - \mu_l \neq 0$$

vastaan.

Käytetään testisuureena ***t*-testisuuretta**

$$t = \frac{\bar{y}_k - \bar{y}_l}{s_P \sqrt{\frac{1}{n_k} + \frac{1}{n_l}}}$$

Kaavassa

$$s_P^2 = \frac{1}{N-k} \sum_{j=1}^{k} (n_j - 1) s_j^2 = \frac{1}{N-k} SSE = MSE$$

on ns. *yhdistetty varianssi*, jossa

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \ , \ j = 1, 2, \mathrm{K}, k$$

on havaintoarvojen varianssi ryhmässä *i*.

Jos nollahypoteesi

$$\mathrm{H}_0 : \mu_k - \mu_l = 0$$

pätee, niin testisuure *t* noudattaa *t-jakaumaa* vapausastein ($N - k$):

$$t \quad t(N-k)$$

Itseisarvoltaan suuret testisuureen *t* arvot johtavat nollahypoteesin hylkäämiseen.

Huomautuksia:

- Tässä käytettävä testisuureen kaava eroaa tavanomaisesta kahden riippumattoman otoksen *t*-testin kaavasta siten, että yhdistetyn varianssin s_P^2 kaavassa on yhdistetty otosvarianssit *kaikista ryhmistä* eikä vain ryhmistä *k* ja *l*.
- Konstruoidut testit *eivät ole simultaanisia*, vaan koskevat vain ryhmien *k* ja *l* odotusarvoja.
- Tehtävien *parivertailujen lukumäärä* on

$$\binom{k}{2} = \frac{k(k-1)}{2}$$

Luottamusvälien ja testien ekvivalenssi

Edellä esitetty *luottamusvälejä* käyttävä menettely, jossa *luottamustasoksi* valitaan luku

$$1 - \alpha$$

ja edellä esitetty *testausmenettely*, jossa *merkitsevyystasoksi* valitaan luku

$$\alpha$$

ovat *ekvivalentteja*.

Simultaaniset luottamusvälit ja testit

Simultaanisten luottamusvälien tai testien konstruoimiseen on useita erilaisia menetelmiä. Simultaaniset luottamusvälit tai testit ovat tavallisesti kuitenkin vain *approksimatiivisia*.

Bonferronin menetelmässä käytetään parivertailujen luottamusvälejä tai testejä, mutta luku α korvataan luvulla

$$\alpha/m$$

jossa *m on tehtävien parivertailujen lukumäärä*:

$$m = \binom{k}{2} = \frac{k(k-1)}{2}$$

Bonferronin epäyhtälö

Olkoot

$$A_1,\ A_2,\ \ldots,\ A_m$$

tapahtumia. Tällöin pätee **Bonferronin epäyhtälö**

$$\Pr(A_1 \cap A_2 \cap \mathsf{L} \cap A_m) \geq 1 - \left[\Pr(A_1^c) + \Pr(A_2^c) + \mathsf{L} + \Pr(A_m^c)\right]$$

Perustelu:

Todistetaan Bonferronin epäyhtälö tapauksessa $m = 2$. Yleinen tapaus voidaan todistaa samaan tapaan kuin tapaus $m = 2$ käyttäen *induktiota*.

Olkoot siis A_1 ja A_2 kaksi otosavaruuden *S* tapahtumaa.

Tällöin *yleisestä yhteenlaskusäännöstä* seuraa, että

(1) $$\Pr(A_1^c \cup A_2^c) = \Pr(A_1^c) + \Pr(A_2^c) - \Pr(A_1^c \cap A_2^c) \leq \Pr(A_1^c) + \Pr(A_2^c)$$

De Morganin lain mukaan

$$A_1 \cap A_2 = (A_1^c \cup A_2^c)^c$$

Siten *komplementtitapahtuman todennäköisyyden* säännöstä ja kaavasta (1) seuraa, että

$$\Pr(A_1 \cap A_2) = \Pr((A_1^c \cup A_2^c)^c) = 1 - \Pr(A_1^c \cup A_2^c) \geq 1 - [\Pr(A_1^c) + \Pr(A_2^c)]$$

■

Bonferronin epäyhtälö ja simultaaniset testit

Bonferronin menetelmä odotusarvojen vertailussa perustuu **Bonferronin epäyhtälöön**.

Oletetaan, että tehtävänä on suorittaa *m* tilastollista testiä. Tarkastellaan todennäköisyyttä α', että *vähintään yksi testien nollahypoteeseista hylätään virheellisesti*.

Määritellään tapahtuma

A_i = ”Nollahypoteesia *ei hylätä* virheellisesti testissä *i*”, $i = 1, 2, \ldots, m$

Tällöin

A_i^c = ”Nollahypoteesi *hylätään* virheellisesti testissä *i*”, $i = 1, 2, \ldots, m$

Jos kaikissa odotusarvojen vertailutesteissä käytetään *samaa merkitsevyystasoa* α, niin

$$\Pr(A_i) = 1 - \alpha, \; i = 1, 2, \ldots, m$$

ja

$$\Pr(A_i^c) = \alpha, \; i = 1, 2, \ldots, m$$

Jos testit $i = 1, 2, \ldots, m$ ovat *riippumattomia*, niin *todennäköisyys, että nollahypoteesia ei hylätä virheellisesti yhdessäkään testissä* on

$$\Pr(A_1 \cap A_2 \cap \ldots \cap A_m) = \Pr(A_1) \Pr(A_2) \ldots \Pr(A_m)$$

Jos jokaisessa testissä käytetään merkitsevyystasona samaa lukua α, niin tällöin

$$\Pr(A_1 \cap A_2 \cap \ldots \cap A_m) = (1 - \alpha)^m$$

ja siten

$$\alpha' = \Pr(\text{”Vähintään yksi virheellinen hylkäys”}) = 1 - (1 - \alpha)^m$$

Jos testit $i = 1, 2, \ldots, m$ *riippuvat toisistaan*, niin tämä yhtälö pätee vain *approksimatiivisesti*.

Bonferronin epäyhtälön mukaan

$$\begin{aligned} &\Pr(\text{”Ei yhtään virheellistä hylkäystä”}) \\ &\quad = \Pr(A_1 \cap A_2 \cap \ldots \cap A_m) \\ &\quad \geq 1 - [\Pr(A_1^c) + \Pr(A_2^c) + \ldots + \Pr(A_m^c)] \end{aligned}$$

Jos jokaisessa testissä käytetään merkitsevyystasona samaa lukua α, niin saamme epäyhtälön

$$\Pr(\text{”Ei yhtään virheellistä hylkäystä”}) \geq 1 - m\alpha$$

Siten vähintään yhden virheellisen hylkäyksen todennäköisyydelle α' on saatu arvio

$$\begin{aligned} \alpha' &= \text{Pr(”Vähintään yksi virheellinen hylkäys”)} \\ &= 1 - \text{Pr(”Ei yhtään virheellistä hylkäystä”)} \\ &\leq m\alpha \end{aligned}$$

Siten valinta

$$\alpha = \beta/m$$

takaa sen, että

$$\alpha' \leq \beta$$

20.10. Kontrastit

Olkoon

$$(1) \qquad y_{ij} = \mu_j + \varepsilon_{ij}\ ,\ i = 1, 2, \mathsf{K}\ , n_j\ ,\ j = 1, 2, \mathsf{K}\ , k$$

yksisuuntaisen varianssianalyysin tilastollinen malli, jossa *jäännös-* eli *virhetermit* ε_{ij} ovat *riippumattomia* ja *normaalijakautuneita* satunnaismuuttujia:

$$\varepsilon_{ij} \quad \mathrm{N}(0, \sigma^2)\ ,\ i = 1, 2, \mathsf{K}\ , n_j\ ,\ j = 1, 2, \mathsf{K}\ , k$$

Mallissa (1)

y_{ij} = y-muuttujan i. havaintoarvo ryhmässä j

μ_j = y-muuttujan odotusarvo ryhmässä j

Olkoon lisäksi

$$N = n_1 + n_2 + \mathsf{L} + n_k$$

havaintojen kokonaislukumäärä.

Kontrasti

Parametrien $\mu_1, \mu_2, \ldots, \mu_k$ *lineaarikombinaatio*

$$\Gamma = \sum_{j=1}^{k} c_j \mu_j$$

on **kontrasti**, jos

$$\sum_{j=1}^{k} c_j = 0$$

Kontrastit

$$\Gamma = \sum_{j=1}^{k} c_j \mu_j$$

ja

$$\Delta = \sum_{j=1}^{k} d_j \mu_j$$

ovat **ortogonaalisia**, jos

$$\sum_{j=1}^{k} \frac{c_j d_j}{n_j} = 0$$

Jos

$$n_1 = n_2 = \cdots = n_k = n$$

niin kontrastit Γ ja Δ ovat ortogonaalisia, jos

$$\sum_{j=1}^{k} c_j d_j = 0$$

Kontrastien estimointi

Olkoon

$$C = \sum_{j=1}^{k} c_j \bar{y}_j$$

kontrastin

$$\Gamma = \sum_{j=1}^{k} c_j \mu_j$$

estimaattori, jossa

$$\bar{y}_j = \frac{1}{n_j} \sum_{j=1}^{n_j} y_{ij} \ , \ j = 1, 2, \mathrm{K}, k$$

on havaintoarvojen aritmeettinen keskiarvo ryhmässä *j* ja

$$n_j \ , j = 1, 2, \ldots , k$$

on ryhmän *j* koko. Yksisuuntaisen varianssianalyysin mallista tehdyistä oletuksista seuraa, että estimaatori *C* noudattaa *normaalijakaumaa*:

$$C \quad \mathrm{N}(\mu_C, \sigma_C^2)$$

jossa

$$\mu_C = \mathrm{E}(C) = \sum_{j=1}^{k} c_j \mu_j = \Gamma$$

ja

$$\sigma_C^2 = \mathrm{D}^2(C) = \sigma^2 \sum_{j=1}^{k} \frac{c_j^2}{n_j}$$

Kontrasteja koskevat testit

Asetetaan *nollahypoteesi*

$$\mathrm{H}_0 : \Gamma = \sum_{j=1}^{k} c_j \mu_j = 0$$

ja sille kaksisuuntainen *vaihtoehtoinen hypoteesi*

$$\mathrm{H}_1 : \Gamma = \sum_{j=1}^{k} c_j \mu_j \neq 0$$

Määritellään ***F*-testisuure**

$$F = \frac{\left(\sum_{j=1}^{k} c_j \bar{y}_j \right)^2}{MSE \sum_{j=1}^{k} \frac{c_j^2}{n_j}}$$

jossa

$$\bar{y}_j = \frac{1}{n_j} \sum_{j=1}^{n_j} y_{ij} \ , \ j = 1, 2, \mathrm{K}, k$$

on havaintoarvojen aritmeettinen keskiarvo ryhmässä *j*,

n_j , j = 1, 2, … , k

on ryhmän *j* koko ja

$$MSE = \frac{SSE}{N-k} = \frac{1}{N-k} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 = \frac{1}{N-k} \sum_{j=1}^{k} (n_j - 1) s_j^2$$

ryhmien sisäistä vaihtelua kuvaava neliösumma, jossa

$$N = n_1 + \ n_2 + \cdots + n_k$$

havaintojen kokonaislukumäärä ja

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \ , \ j = 1, 2, \mathrm{K}, k$$

on havaintoarvojen varianssi ryhmässä *j*. Jos nollahypoteesi

$$\mathrm{H}_0 : \Gamma = \sum_{j=1}^{k} c_j \mu_j = 0$$

pätee, niin testisuure *F* noudattaa *F-jakaumaa* vapaustein 1 ja ($N - k$):

$$F \sim F(1, N - k)$$

Suuret testisuureen *F* arvot johtavat nollahypoteesin H_0 hylkäämiseen.

Perustelu:

Olkoon

$$Q_1 = \frac{C^2}{\mathrm{D}^2(C)}$$

jossa

$$C = \sum_{j=1}^{k} c_j \overline{y}_j$$

ja

$$\mathrm{D}^2(C) = \sigma^2 \sum_{j=1}^{k} \frac{c_j^2}{n_j}$$

Jos nollahypoteesi

$$\mathrm{H}_0 : \Gamma = \sum_{j=1}^{k} c_j \mu_j = 0$$

pätee, niin

$$Q_1 \quad \chi^2(1)$$

Olkoon

$$Q_2 = \frac{SSE}{\sigma^2}$$

jossa

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_j)^2 = \sum_{j=1}^{k} (n_j - 1) s_j^2$$

Voidaan osoittaa, että (ks. kappaletta **Yksisuuntainen varianssianalyysi ja sen suorittaminen**)

$$Q_2 \quad \chi^2(N-k)$$

jossa

$$N = n_1 + n_2 + \mathrm{L} + n_k$$

Määritellään *F*-testisuure

$$F = \frac{Q_1 / 1}{Q_2 / (N-k)} = (N-k) \frac{Q_1}{Q_2}$$

Koska neliösummat Q_1 ja Q_2 ovat *riippumattomia*,

$$F \sim F(1, N-k)$$

Todetaan lopuksi, että testisuure *F* voidaan kirjoittaa muotoon

$$F = \frac{\left(\sum_{j=1}^{k} c_j \overline{y}_j \right)^2}{MSE \sum_{j=1}^{k} \frac{c_j^2}{n_j}}$$

jossa

$$MSE = \frac{SSE}{N-k} = \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2 = \frac{1}{N-k}\sum_{j=1}^{k}(n_j-1)s_j^2$$

■

Ottamalla edellä esitetystä *F*-testisuureesta *neliöjuuri*, saadaan ***t*-testisuure**

$$t = \frac{\sum_{j=1}^{k} c_j \bar{y}_j}{\sqrt{MSE\sum_{j=1}^{k}\frac{c_j^2}{n_j}}}$$

Jos nollahypoteesi

$$\mathrm{H}_0 : \Gamma = \sum_{j=1}^{k} c_j \mu_j = 0$$

pätee, niin testisuure *t* noudattaa *t-jakaumaa* vapausastein ($N - k$):

$$t \sim t(N-k)$$

Itseisarvoltaan suuret testisuureen *t* arvot johtavat nollahypoteesin H_0 hylkäämiseen.

Edellä esitetyt *F*-testi ja *t*-testi ovat *ekvivalentteja*.

Kontrastien luottamusvälit

Kontrastin

$$\Gamma = \sum_{i=1}^{k} c_i \mu_i$$

luottamusväli luottamustasolla $(1 - \alpha)$ on muotoa

$$\sum_{j=1}^{k} c_j \bar{y}_j \pm t_{\alpha/2}\sqrt{MSE\left(\sum_{j=1}^{k}\frac{c_j^2}{n_j}\right)}$$

jossa *luottamuskertoimet* $-t_{\alpha/2}$ ja $+t_{\alpha/2}$ on määrätty niin, että

$$\Pr(-t_{\alpha/2} \le t \le +t_{\alpha/2}) = 1 - \alpha$$

ja

$$t \sim t(N-k)$$

Ortogonaalisten kontrastien testaaminen

Kontrastit

$$\Gamma = \sum_{j=1}^{k} c_j \mu_j$$

ja

$$\Delta = \sum_{j=1}^{k} d_j \mu_j$$

ovat **ortogonaalisia**, jos

$$\sum_{j=1}^{k} \frac{c_j d_j}{n_j} = 0$$

Toisistaan *riippumattomien* ortogonaalisia kontrastien lukumäärä on

$$k - 1$$

jossa

k = Ryhmien lukumäärä

Ortogonaaliset kontrastit *dekomponoivat ryhmien välistä* (systemaatista) *vaihtelua kuvaavan neliösumman*

$$SSG = \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2$$

($k - 1$) *komponenttiin*, joista jokaisen aste = 1. Siten *ortogonaalisiin kontrasteihin liittyvät testit ovat riippumattomia*.

Perustelu:

Olkoot

$$\Gamma_l = \sum_{j=1}^{k} c_{lj} \mu_j \ , l = 1, 2, \mathsf{K}, k-1$$

($k - 1$) *ortogonaalista kontrastia* odotusarvoille

$$\mu_1, \mu_2, \ldots, \mu_k$$

Olkoon

$$SS_l = \frac{\left(\sum_{j=1}^{k} c_{lj} \bar{y}_j \right)^2}{\sum_{j=1}^{k} \frac{c_{lj}^2}{n_j}}, l = 1, 2, \mathsf{K}, k-1$$

Tällöin pätee

$$SSG = \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2 = SS_1 + SS_2 + \mathsf{L} + SS_{k-1}$$

Kappaleessa **Yksisuuntainen varianssianalyysi ja sen suorittaminen** todettiin, että

$$\frac{SSG}{\sigma^2} \quad \chi^2(k-1)$$

Edellä esitetyn mukaan satunnaismuuttujat

$$\frac{SS_1}{\sigma^2}, \frac{SS_2}{\sigma^2}, \mathrm{K}, \frac{SS_{k-1}}{\sigma^2}$$

noudattavat χ^2*-jakaumaa* yhdellä vapausasteella:

$$SS_l \quad \chi^2(1), l = 1, 2, \mathrm{K}, k-1$$

Siten *Cochranin lauseesta* (ks. kappaletta **Yksisuuntainen varianssianalyysi ja sen suorittaminen**) seuraa, että satunnaismuuttujat

$$\frac{SS_1}{\sigma^2}, \frac{SS_2}{\sigma^2}, \mathrm{K}, \frac{SS_{k-1}}{\sigma^2}$$

ovat *riippumattomia.*

Koska *F-testisuureet* kontrasteille

$$\Gamma_1, \Gamma_2, \ldots, \Gamma_{k-1}$$

voidaan esittää muodossa

$$F_l = \frac{SS_l}{MSE}, l = 1, 2, \mathrm{K}.k-1$$

näemme, että *testit ortogonaalisille kontrasteille ovat riippumattomia*.

■

21. Kaksisuuntainen varianssianalyysi

21.1. Varianssianalyysi: Johdanto

21.2. Kaksisuuntainen varianssianalyysi ja sen suorittaminen

21.3. Kaksisuuntaisen varianssianalyysin malli ja sen parametrointi

21.4. Kaksisuuntaisen varianssianalyysin mallin parametrien estimointi

21.5 Laskutoimitusten suorittaminen

Yksisuuntaisessa varianssianalyysissa perusjoukko on jaettu ryhmiin *yhden ryhmittelevän* **tekijän** suhteen ja tavoitteena on testata hypoteesia, jonka mukaan **kiinnostuksen kohteena olevan muuttujan ryhmäkohtaiset odotusarvot ovat yhtä suuria**.

Kaksi- tai **useampisuuntaisessa varianssianalyysissa** perusjoukko on jaettu ryhmiin **kahden** tai **useamman ryhmittelevän tekijän** suhteen ja nytkin tavoitteena on testata hypoteesia, jonka mukaan **kiinnostuksen kohteena olevan muuttujan ryhmäkohtaiset odotusarvot ovat yhtä suuria**.

Tässä luvussa tarkastellaan **kaksisuuntaista varianssianalyysia**. Tarkastelun kohteena ovat mm. **kaksisuuntaisen varianssianalyysin malli** ja sen **parametrointi**, **parametrien estimointi**, **odotusarvojen yhtäsuuruuden testaaminen** ja **laskutoimitusten suorittaminen**.

Avainsanat:

Aritmeettinen keskiarvo, Cochranin lause, Estimointi, *F*-testi, Faktori, Harha, Harhattomuus, Indikaattorimuuttuja, Interaktio, Jäännösvaihtelu, Jäännösvarianssi, Kaksisuuntainen varianssianalyysi, Kokonaiskeskiarvo, Kokonaisvaihtelu, Merkitsevyystaso, Muuttuja, Normaalijakauma, Neliösumma, Otos, Otostunnusluku, Parametri, Parametrointi, Pienimmän neliösumman estimaattori, Pienimmän neliösumman menetelmä, Päävaikutus, Residuaali, Reunakeskiarvo, Riippumattomuus, Ryhmien sisäinen vaihtelu, Ryhmien välinen vaihtelu, Ryhmittely, Ryhmä, Ryhmäkeskiarvo, Satunnaisuus,
Side-ehto, Sovite, *t*-testi, Taso, Tekijä, Testi, Vapausaste, Varianssi, Varianssianalyysihajotelma, Varianssianalyysitaulukko, Yhdysvaikutus, Yksisuuntainen varianssianalyysi, Yleinen lineaarinen malli, Yleiskeskiarvo

21.1. Varianssianalyysi: Johdanto

Kahden riippumattoman otoksen *t*-testi

Suhdeasteikollisille muuttujille tarkoitettuja testejä käsitelleessä kappaleessa tarkastellaan **kahden riippumattoman otoksen *t*-testiä.** Testin testausasetelma on seuraava:

(i) Perusjoukko koostuu *kahdesta* ryhmästä.

(ii) Havainnot noudattavat kummassakin ryhmässä *normaalijakaumaa*.

(iii) Kummastakin ryhmästä on poimittu *toisistaan riippumattomat satunnaisotokset*.

(iv) Tehtävänä on *testata ryhmäkohtaisten odotusarvojen samuutta*.

Varianssianalyysin perusongelma

Varianssianalyysi voidaan ymmärtää *kahden riippumattoman otoksen t-testin yleistykseksi* tilanteisiin, jossa perusjoukko koostuu *kahdesta* tai *useammasta ryhmästä*:

(i) Perusjoukko koostuu *kahdesta* tai *useammasta* ryhmästä.

(ii) Havainnot noudattavat jokaisessa ryhmässä *normaalijakaumaa*.

(iii) Jokaisesta ryhmästä poimitaan *toisistaan riippumattomat satunnaisotokset*.

(iv) Tehtävänä on *testata ryhmäkohtaisten odotusarvojen samuutta*.

Perusjoukon *jako ryhmiin* voidaan tehdä *yhden* tai *useamman* **faktorin** eli **tekijän** (*muuttujan*) *arvojen perusteella*. Jos perusjoukon jako ryhmiin perustuu *yhteen tekijään*, puhutaan **yksisuuntaisesta varianssianalyysista**. Jos perusjoukon jako ryhmiin perustuu *m tekijään*, puhutaan ***m*-suuntaisesta varianssianalyysista**.

Huomautus:

- Tässä luvussa käsitellään **kaksisuuntaista varianssianalyysia**; **yksisuuntaista varianssianalyysia** on käsitelty edellisessä luvussa **kolmisuuntaista varianssianalyysia** käsitellään seuraavassa luvussa.

Varianssianalyysin nimi *johtaa helposti harhaan*. Varianssianalyysissa ei testata varianssien vaan **odotusarvojen** yhtäsuuruutta tilanteessa. Nimi johtuu siitä, että odotusarvojen yhtäsuuruuden testaaminen perustuu eri tavoilla määrättyjen *varianssien yhtäsuuruuden testaamiseen*.

21.2. Kaksisuuntaisen varianssianalyysin perusasetelma

Oletetaan, että tutkimuksen kohteena oleva perusjoukko voidaan *jakaa ryhmiin kahden* **faktorin** eli **tekijän** (*muuttujan*) *A* ja *B arvojen suhteen* ja oletetaan, että tekijällä *A* on *J* **tasoa** ja tekijällä *B* on *K* **tasoa**, jolloin jaossa syntyy **ryhmiä** $J \times K$ kappaletta. Oletetaan edelleen, että ryhmistä on poimittu toisistaan riippumattomat satunnaisotokset, joiden kaikkien koko on *I*. Koska ryhmäkoot on oletettu yhtä suuriksi, sanomme, että asetelma on **tasapainotettu**.

Olkoon

y_{ijk} = i. havainto tekijän A tason A_j ja tekijän B tason B_k määräämässä ryhmässä (j,k), $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$, $k = 1, 2, \ldots, K$

Käytetystä otantamenetelmästä seuraa, että havainnot y_{ijk} voidaan olettaa *riippumattomiksi* (ja siten myös *korreloimattomiksi*) satunnaismuuttujiksi.

Oletetaan, että kaikilla *samaan ryhmään* (j,k) *kuuluvilla havainnoilla* on *sama* odotusarvo:

$$\mathrm{E}(y_{ijk}) = \mu_{jk} \text{ , } i = 1, 2, \dots, I \text{ , } j = 1, 2, \dots, J \text{ , } k = 1, 2, \dots, K$$

ja kaikilla havainnoilla on *ryhmästä riippumatta sama* varianssi:

$$\mathrm{D}^2(y_{ijk}) = \sigma^2 \text{ , } i = 1, 2, \dots, I \text{ , } j = 1, 2, \dots, J \text{ , } k = 1, 2, \dots, K$$

Oletetaan lisäksi, että havainnot y_{ijk} ovat *normaalijakautuneita*:

$$y_{ijk} \sim \mathrm{N}(\mu_{jk} \text{ , } \sigma^2) \text{ , } i = 1, 2, \dots, I \text{ , } j = 1, 2, \dots, J \text{ , } k = 1, 2, \dots, K$$

Haluamme testata **nollahypoteesia** siitä, että *ryhmäkohtaiset odotusarvot* $\mathrm{E}(y_{ijk}) = \mu_{jk}$ *ovat yhtä suuria*:

$$\mathrm{H}_0 : \mu_{jk} = \mu \text{ , } j = 1, 2, \dots, J \text{ , } k = 1, 2, \dots, K$$

Jos nollahypoteesi *ryhmäkohtaisten odotusarvojen yhtäsuuruudesta pätee, ryhmät voidaan yhdistää* kaikissa havaintojen keskimääräisiä arvoja koskevissa tarkasteluissa.

Kaksisuuntaisessa varianssianalyysissa nollahypoteesi

$$\mathrm{H}_0 : \mu_{jk} = \mu \text{ , } j = 1, 2, \dots, J \text{ , } k = 1, 2, \dots, K$$

on tapana jakaa *kolmeksi* nollahypoteesiksi, jotka koskevat **tekijän *A* päävaikutusta**, **tekijän *B* päävaikutusta** sekä **tekijöiden *A* ja *B* interaktiota** eli **yhdysvaikutusta**. Tämä tekee ryhmä-kohtaisten odotusarvojen yhtäsuuruutta koskevan testausongelman olennaisesti *moni-mutkaisemmaksi* kuin yksisuuntaisessa varianssianalyysissa. Tämä johtuu siitä, että tekijöiden *A* ja *B* *päävaikutuksia ei voida tarkastella erillisinä*, jos tekijöillä *A* ja *B on yhdysvaikutusta*.

Siten kaksisuuntaisessa varianssianalyysissa *testattavia nollahypoteeseja* on *kolme kappaletta*:

(i) *Tekijöiden A* ja *B* **yhdysvaikutusta koskeva nollahypoteesi** on muotoa

$$\mathrm{H}_{AB} : \text{Ei yhdysvaikutusta}$$

Jos nollahypoteesi H_{AB} jää voimaan, tekijöiden *A* ja *B* päävaikutuksia voidaan tarkastella *erillisinä*.

(ii) *Tekijän A* **päävaikutusta koskeva nollahypoteesi** on muotoa

$$\mathrm{H}_A : \text{Ei } A\text{-vaikutusta}$$

(iii) *Tekijän B* **päävaikutusta koskeva nollahypoteesi** on muotoa

$$\mathrm{H}_B : \text{Ei } B\text{-vaikutusta}$$

Huomautus:

- Nollahypoteesit H_A ja H_B *ovat yksisuuntaisen varianssianalyysin nollahypoteeseja*.

Kaksisuuntainen varianssianalyysi tarkoittaa em. testausasetelman nollahypoteesien

$$\mathrm{H}_{AB} : \text{Ei yhdysvaikutusta}$$

$$\mathrm{H}_A : \text{Ei } A\text{-vaikutusta}$$

$$\mathrm{H}_B : \text{Ei } B\text{-vaikutusta}$$

testaamista.

Interaktio: Havainnollistus

Tarkastellaan yksinkertaisten esimerkkikuvioiden avulla ryhmittelevien tekijöiden *A* ja *B interaktion* eli *yhdysvaikutuksen* ilmenemistä ryhmä-kohtaisia odotusarvoja kuvaavissa *odotusarvo-diagrammeissa*.

Oletetaan, että molemmilla ryhmittelevällä tekijöillä *A* ja *B* on *kaksi tasoa*:

$$A : A_j \text{ , } j = 1, 2$$

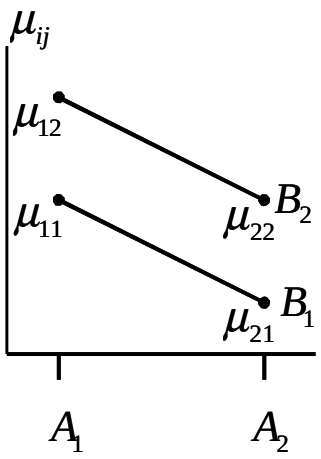
$$B : B_k \text{ , } k = 1, 2$$

Olkoot vastaavat *ryhmäodotusarvot*

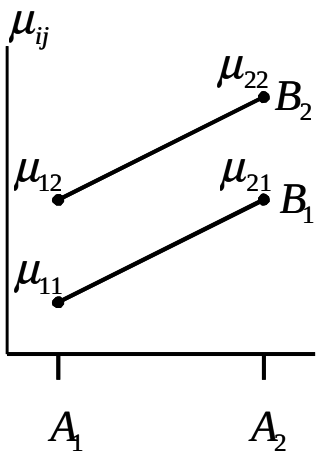
$$\mu_{jk} \text{ , } j = 1,2 \text{ , } k = 1,2$$

Kuvio 1: Ei yhdysvaikutusta.

Tapaus 1:



Tapaus 2:

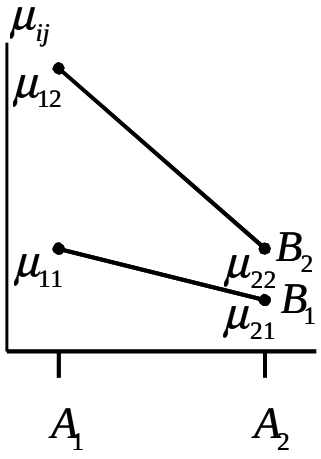


Jos tekijän *B* tasoa muutetaan, *ryhmäodotusarvoissa tapahtuva muutos ei riipu tekijän A tasosta*:

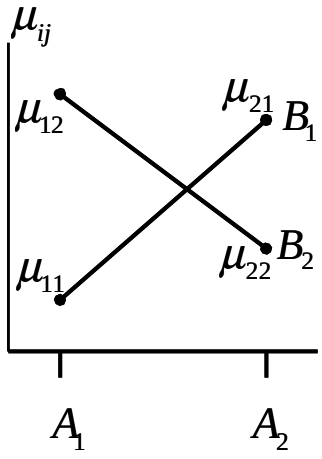
$$\mu_{11} - \mu_{12} = \mu_{21} - \mu_{22}$$

Kuvio 2: Yhdysvaikutusta esiintyy.

Tapaus 1:



Tapaus 2:



Jos tekijän *B* tasoa muutetaan, *ryhmäodotusarvoissa tapahtuva muutos riippuu tekijän A tasosta*:

$$\mu_{11} - \mu_{12} \neq \mu_{21} - \mu_{22}$$

Kaksisuuntainen varianssianalyysi ja koesuunnittelu

Kaksisuuntaista varianssianalyysiä voidaan käyttää koetulosten analyysiin seuraavassa **koeasetelmassa**:

(i) Oletetaan, että kokeen tavoitteena on verrata, miten **käsittelyiden**

$$A_1, A_2, \ldots, A_J$$

ja

$$B_1, B_2, \ldots, B_K$$

yhdistelmät (A_j , B_k) vaikuttavat kiinnostuksen kohteena olevan **vastemuuttujan** *y keskimääräisiin arvoihin.*

(ii) Valitaan jokaisen käsittelykombinaation (A_j , B_k) kohteeksi kokeen kaikkien mahdollisten kohteiden joukosta *satunnaisesti I* yksilöä, $j = 1, 2, \ldots, J$, $k = 1, 2, \ldots, K$, jolloin *havaintojen kokonaislukumääräksi* tulee

$$IJK = N$$

(iii) Mitataan **vasteet** y_{ijk} eli kiinnostuksen kohteena olevan muuttujan *y* arvot:

$$y_{ijk}\ ,\ i = 1, 2, \ldots, I, j = 1, 2, \ldots, J\ ,\ k = 1, 2, \ldots, K$$

Huomaa, että koeasetelma on **täydellisesti satunnaistettu**: *Sattuma määrää täydellisesti millaisen käsittelyn kohteeksi kokeen kohteiksi valitut yksilöt joutuvat*.

21.3. Kaksisuuntaisen varianssianalyysin suorittaminen

Havaintojen keskiarvot

Määritellään havaintoarvojen y_{ijk} **ryhmäkeskiarvot** eli *ryhmäkohtaiset aritmeettiset keskiarvot* tekijän *A* tason A_j ja tekijän *B* tason B_k määräämässä ryhmässä (*j*,*k*):

$$\bar{y}_{jk} = \frac{1}{I}\sum_{i=1}^{I} y_{ijk}\ ,\ j = 1, 2, \mathrm{K}, J\ ,\ k = 1, 2, \mathrm{K}, K$$

Jos nollahypoteesit H_{AB}, H_A ja H_B pätevät, on odotettavissa, että ryhmäkeskiarvot *eivät poikkea kovin paljon toisistaan.*

Jos ryhmäkohtaiset otokset yhdistetään yhdeksi otokseksi, yhdistetyn otoksen havaintoarvojen **yleis-** eli **kokonaiskeskiarvo** on

$$\bar{y} = \frac{1}{IJK}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ijk}$$

jossa

$$IJK = N$$

on yhdistetyn otoksen havaintojen kokonaislukumäärä.

Määritellään havaintoarvojen y_{ijk} **marginaali**- eli **reunakeskiarvot** kaavoilla:

$$\bar{y}_{j} = \frac{1}{IK}\sum_{k=1}^{K}\sum_{i=1}^{I} y_{ijk}\ ,\ j = 1, 2, \mathrm{K}, J$$

$$\bar{y}_{k} = \frac{1}{IJ}\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ijk}\ ,\ k = 1, 2, \mathrm{K}, K$$

Reunakeskiarvo $\bar{y}_{j}$ on havaintojen y_{ijk} keskiarvo tekijän A tason A_j määräämässä ryhmässä j, kun B-ryhmitystä ei oteta huomioon ja reunakeskiarvo $\bar{y}_{k}$ on havaintojen y_{ijk} keskiarvo tekijän B tason B_k määräämässä ryhmässä k, kun A-ryhmitystä ei oteta huomioon.

Huomaa, että kokonaiskeskiarvo $\bar{y}$ on ryhmäkeskiarvojen $\bar{y}_{jk}$ keskiarvo:

$$\bar{y} = \frac{1}{JK}\sum_{k=1}^{K}\sum_{j=1}^{J} \bar{y}_{jk}$$

Myös reunakeskiarvot $\bar{y}_{j}$ ja $\bar{y}_{k}$ voidaan määritellä ryhmäkeskiarvojen $\bar{y}_{jk}$ avulla:

$$\bar{y}_{j} = \frac{1}{K}\sum_{k=1}^{K} \bar{y}_{jk}\ ,\ j = 1, 2, \mathrm{K}, J$$

$$\bar{y}_{k} = \frac{1}{J}\sum_{j=1}^{J} \bar{y}_{jk}\ ,\ k = 1, 2, \mathrm{K}, K$$

Varianssianalyysihajotelma

Kirjoitetaan identiteetti

$$y_{ijk} - \bar{y} = (\bar{y}_{j} - \bar{y}) + (\bar{y}_{k} - \bar{y}) + (\bar{y}_{jk} - \bar{y}_{j} - \bar{y}_{k} + \bar{y}) + (y_{ijk} - \bar{y}_{jk})$$

2-suuntaisen varianssianalyysin testit nollahypoteeseille H_{AB}, H_{A} ja H_{B} perustuvat *poikkeamien*

$$(\bar{y}_{j} - \bar{y}), (\bar{y}_{k} - \bar{y}), (\bar{y}_{jk} - \bar{y}_{j} - \bar{y}_{k} + \bar{y}), (y_{ijk} - \bar{y}_{jk})$$

neliösummille.

Testi nollahypoteesille

H_{AB} : Ei yhdysvaikutusta

perustuu poikkeamien

$$(\bar{y}_{jk} - \bar{y}_{j} - \bar{y}_{k} + \bar{y}), (y_{ijk} - \bar{y}_{jk})$$

neliösummille. Jos nollahypoteesi H_{AB} pätee, on odotettavissa, että erotukset

$$(\bar{y}_{jk} - \bar{y}_{j} - \bar{y}_{k} + \bar{y})$$

eivät ole itseisarvoiltaan kovin suuria.

Testi nollahypoteesille

H_{A} : Ei A-vaikutusta

perustuu poikkeamien

$$(\bar{y}_{j} - \bar{y}), (y_{ijk} - \bar{y}_{jk})$$

neliösummille. Jos nollahypoteesi H_A pätee, on odotettavissa, että erotukset

$$(\bar{y}_{\,j} - \bar{y}\;)$$

eivät ole itseisarvoiltaan kovin suuria.

Testi nollahypoteesille

H_B : Ei *B*-vaikutusta

perustuu poikkeamien

$$(\bar{y}_{\,k} - \bar{y}\;), (y_{ijk} - \bar{y}_{\,jk})$$

neliösummille. Jos nollahypoteesi H_B pätee, on odotettavissa, että erotukset

$$(\bar{y}_{\,k} - \bar{y}\;)$$

eivät ole itseisarvoiltaan kovin suuria.

Määritellään **havaintoarvojen kokonaisvaihtelua kuvaava kokonaisneliösumma**:

$$SST = \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I}(y_{ijk} - \bar{y}\;)^2$$

Jos ryhmäkohtaiset otokset yhdistetään yhdeksi otokseksi, saadun yhdistetyn otoksen *varianssi* on

$$s_y^2 = \frac{1}{N-1}SST$$

jossa

$$N = IJK$$

on yhdistetyn otoksen havaintojen kokonaislukumäärä.

Määritellään **tekijän *A* päävaikutusta kuvaava neliösumma**:

$$SSA = IK\sum_{j=1}^{J}(\bar{y}_{\,j} - \bar{y}\;)^2$$

Määritellään **tekijän *B* päävaikutusta kuvaava neliösumma**:

$$SSB = IJ\sum_{k=1}^{K}(\bar{y}_{\,k} - \bar{y}\;)^2$$

Määritellään **tekijöiden *A* ja *B* yhdysvaikutusta kuvaava neliösumma**:

$$SSAB = I\sum_{k=1}^{K}\sum_{j=1}^{J}(\bar{y}_{\,jk} - \bar{y}_{\,j} - \bar{y}_{\,k} + \bar{y}\;)^2$$

Määritellään **ryhmien sisäistä vaihtelua kuvaava** (**jäännös-**) **neliösumma**:

$$SSE = \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I}(y_{ijk} - \bar{y}_{\,jk})^2$$

Havaintoarvojen y_{ijk} **ryhmävarianssit** eli **ryhmäkohtaiset varianssit** saadaan lausekkeista

$$s_{jk}^2 = \frac{1}{I-1}\sum_{i=1}^{I}(y_{ijk} - \bar{y}_{jk})^2 \text{ , } j = 1,2,\text{K},J \text{ , } k = 1,2,\text{K},K$$

Siten ryhmien sisäistä vaihtelua kuvaava neliösumman *SSE* lauseke voidaan esittää myös muodossa

$$SSE = (I-1)\sum_{k=1}^{K}\sum_{j=1}^{J} s_{jk}^2$$

Korottamalla identiteetti

$$y_{ijk} - \bar{y} = (\bar{y}_{j} - \bar{y}) + (\bar{y}_{k} - \bar{y}) + (\bar{y}_{jk} - \bar{y}_{j} - \bar{y}_{k} + \bar{y}) + (y_{ijk} - \bar{y}_{jk})$$

potenssiin kaksi ja laskemalla yhteen saadaan **varianssianalyysihajotelma**

$$\begin{aligned}\Sigma\Sigma\Sigma(y_{ijk} - \bar{y})^2 &= \Sigma\Sigma\Sigma(\bar{y}_{j} - \bar{y})^2 + \Sigma\Sigma\Sigma(\bar{y}_{k} - \bar{y})^2 \\ &\quad + \Sigma\Sigma\Sigma(\bar{y}_{jk} - \bar{y}_{j} - \bar{y}_{k} + \bar{y})^2 \\ &\quad + \Sigma\Sigma\Sigma(y_{ijk} - \bar{y}_{jk})^2 \\ &= IK\Sigma(\bar{y}_{j} - \bar{y})^2 + IJ\Sigma(\bar{y}_{k} - \bar{y})^2 \\ &\quad + I\Sigma\Sigma(\bar{y}_{jk} - \bar{y}_{j} - \bar{y}_{k} + \bar{y})^2 \\ &\quad + \Sigma\Sigma\Sigma(y_{ijk} - \bar{y}_{jk})^2\end{aligned}$$

Edellä esitettyjen neliösummien määritelmien perusteella varianssianalyysihajotelma voidaan esittää muodossa

$$SST = SSA + SSB + SSAB + SSE$$

Varianssianalyysihajotelmassa *kokonaisneliösumma SST* on hajotettu *neljän* osatekijän summaksi, jossa osatekijä

$$SSAB = I\Sigma\Sigma(\bar{y}_{jk} - \bar{y}_{j} - \bar{y}_{k} + \bar{y})^2$$

kuvaa tekijöiden *A* ja *B yhdysvaikutusta*, osatekijät

$$SSA = IK\Sigma(\bar{y}_{j} - \bar{y})^2$$

ja

$$SSB = IJ\Sigma(\bar{y}_{k} - \bar{y})^2$$

kuvaavat tekijöiden *A* ja *B päävaikutuksia* ja osatekijä

$$SSE = \Sigma\Sigma\Sigma(y_{ijk} - \bar{y}_{jk})$$

kuvaa *ryhmien sisäistä vaihtelua*.

Kaksisuuntaisen varianssianalyysin testit

Jos tekijöiden *A* ja *B yhdysvaikutusta kuvaava neliösumma*

$$SSAB = I\Sigma\Sigma(\bar{y}_{jk} - \bar{y}_{j} - \bar{y}_{k} + \bar{y})^2$$

on suuri verrattuna ryhmien sisäistä vaihtelua kuvaavaan *jäännösneliösummaan*

$$SSE = \Sigma\Sigma\Sigma(y_{ijk} - \bar{y}_{jk})$$

nollahypoteesi

$$\mathrm{H}_{AB} \text{ : Ei yhdysvaikutusta}$$

on asetettava *kyseenalaiseksi*. Määritellään ***F*-testisuure**

$$F_{AB} = \frac{JK(I-1)}{(J-1)(K-1)} \cdot \frac{SSAB}{SSE}$$

Jos havainnot noudattavat *normaalijakaumaa* ja *nollahypoteesi*

$$\mathrm{H}_{AB} \text{ : Ei yhdysvaikutusta}$$

pätee, testisuure F_{AB} noudattaa *F-jakaumaa* vapausastein $(J-1)(K-1)$ ja $JK(I-1)$:

$$F_{AB} \quad F((J-1)(K-1), JK(I-1))$$

Testisuureen F_{AB} *normaaliarvo* on

$$\mathrm{E}(F_{AB}) \underset{\mathrm{H}_{AB}}{=} \frac{IJK - JK}{IJK - JK - 2}$$

ja

$$\mathrm{E}(F_{AB}) \approx 1$$

jos $N = IJK$ on *suuri*. *Suuret* testisuureen F_{AB} arvot johtavat nollahypoteesin H_{AB} hylkäämiseen.

Jos tekijän *A päävaikutusta kuvaava neliösumma*

$$SSA = IK \sum (\bar{y}_{\,j} - \bar{y}\,)^2$$

on suuri verrattuna ryhmien sisäistä vaihtelua kuvaavaan *jäännösneliösummaan*

$$SSE = \sum \sum \sum (y_{ijk} - \bar{y}_{\,jk})$$

nollahypoteesi

$$\mathrm{H}_{A} \text{ : Ei } A\text{-vaikutusta}$$

on asetettava *kyseenalaiseksi*. Määritellään ***F*-testisuure**

$$F_{A} = \frac{JK(I-1)}{J-1} \cdot \frac{SSA}{SSE}$$

Jos havainnot noudattavat *normaalijakaumaa* ja *nollahypoteesi*

$$\mathrm{H}_{A} \text{ : Ei } A\text{-vaikutusta}$$

pätee, testisuure F_A noudattaa *F-jakaumaa* vapausastein $(J-1)$ ja $JK(I-1)$:

$$F_{A} \quad F((J-1), JK(I-1))$$

Testisuureen F_A *normaaliarvo* on

$$\mathrm{E}(F_{A}) \underset{\mathrm{H}_{A}}{=} \frac{IJK - JK}{IJK - JK - 2}$$

ja

$$\mathrm{E}(F_{A}) \approx 1$$

jos $N = IJK$ on *suuri*. *Suuret* testisuureen F_A arvot johtavat nollahypoteesin H_A hylkäämiseen.

Jos tekijän *B päävaikutusta kuvaava neliösumma*

$$SSB = IJ \sum(\bar{y}_{\ k} - \bar{y}\ )^2$$

on suuri verrattuna ryhmien sisäistä vaihtelua kuvaavaan *jäännösneliösummaan*

$$SSE = \sum\sum\sum(y_{ijk} - \bar{y}_{\ jk})$$

nollahypoteesi

H_B : Ei *B*-vaikutusta

on asetettava *kyseenalaiseksi*. Määritellään *testisuure*

$$F_B = \frac{JK(I-1)}{K-1} \cdot \frac{SSB}{SSE}$$

Jos havainnot noudattavat *normaalijakaumaa* ja *nollahypoteesi*

H_B : Ei *B*-vaikutusta

pätee, testisuure F_B noudattaa *F-jakaumaa* vapausastein $(K-1)$ ja $JK(I-1)$:

$$F_B \quad F((K-1), JK(I-1))$$

Testisuureen F_B *normaaliarvo* on

$$\mathrm{E}(F_B) \underset{H_B}{=} \frac{IJK - JK}{IJK - JK - 2}$$

ja

$$\mathrm{E}(F_B) \approx 1$$

jos $N = IJK$ on *suuri*. *Suuret* testisuureen F_B arvot johtavat nollahypoteesin H_B hylkäämiseen.

Neliösummien jakaumat

Voidaan osoittaa, että *aina pätee*

$$\frac{SSE}{\sigma^2} \quad \chi^2(JK(I-1))$$

Edelleen voidaan osoittaa, että *jos nollahypoteesit*

H_{AB} : Ei yhdysvaikutusta

H_A : Ei *A*-vaikutusta

H_B : Ei *B*-vaikutusta

pätevät, niin

$$\frac{SST}{\sigma^2} \quad \chi^2(IJK-1)$$

$$\frac{SSAB}{\sigma^2} \quad \chi^2((J-1)(K-1))$$

$$\frac{SSA}{\sigma^2} \quad \chi^2(J-1)$$

$$\frac{SSB}{\sigma^2} \quad \chi^2(K-1)$$

Varianssianalyysihajotelman mukaan

$$SST = SSA + SSB + SSAB + SSE$$

Edellä esitetyn mukaan satunnaismuuttujat

$$\frac{SST}{\sigma^2}, \frac{SSA}{\sigma^2}, \frac{SSB}{\sigma^2}, \frac{SSAB}{\sigma^2}, \frac{SSE}{\sigma^2}$$

noudattavat *nollahypoteesien* H_{AB}, H_A ja H_B *pätiessä* χ^2*-jakaumaa* vapausastein, jotka toteuttavat yhtälön

$$IJK - 1 = (J-1) + (K-1) + (J-1)(K-1) + JK(I-1)$$

Siten satunnaismuuttujat

$$\frac{SSA}{\sigma^2}, \frac{SSB}{\sigma^2}, \frac{SSAB}{\sigma^2}, \frac{SSE}{\sigma^2}$$

ovat *Cochranin lauseen mukaan riippumattomia* (ks. luvun **Yksisuuntainen varianssianalyysi** kappaletta **Yksisuuntainen varianssianalyysi ja sen suorittaminen**).

Edellä esitetyn nojalla *testisuureet* F_{AB}, F_A ja F_B *noudattavat nollahypoteesien* H_{AB}, H_A ja H_B *pätiessä F-jakaumaa* suoraan *F*-jakauman määritelmän mukaan:

$$F_{AB} = \frac{SSAB/(J-1)(K-1)}{SSE/JK(I-1)} \quad F((J-1)(K-1), JK(I-1))$$

$$F_A = \frac{SSA/(J-1)}{SSE/JK(I-1)} \quad F((J-1), JK(I-1))$$

$$F_B = \frac{SSB/(K-1)}{SSE/JK(I-1)} \quad F((K-1), JK(I-1))$$

Testisuureet F_{AB}, F_A ja F_B voidaan tulkita *varianssien vertailutestisuureiksi*, joissa *variansseja*

$$MSAB = \frac{SSAB}{(J-1)(K-1)}$$

$$MSA = \frac{SSA}{J-1}$$

$$MSB = \frac{SSB}{K-1}$$

verrataan ryhmien sisäiseen varianssiin

$$MSE = \frac{1}{JK(I-1)} SSE$$

Varianssiestimaattoreiden harhattomuus

Estimaattori

$$MSE = \frac{1}{JK(I-1)} SSE$$

on *aina harhaton* havaintojen y_{ijk} varianssille σ^2. Sen sijaan *estimaattorit*

$$MSAB = \frac{SSAB}{(J-1)(K-1)}$$

$$MSA = \frac{SSA}{J-1}$$

$$MSB = \frac{SSB}{K-1}$$

ovat *harhattomia* havaintojen y_{ijk} varianssille σ^2 *vain, jos nollahypoteesit* H_{AB}, H_A ja H_B *pätevät*.

Tarkastelemme seuraavassa lähemmin ehtoja, joiden pätiessä estimaattorit *MSE*, *MSAB*, *MSA* ja *MSB ovat harhattomia* havaintojen y_{ijk} varianssille σ^2. Käytämme hyväksi sitä, että kaksisuuntaisen varianssianalyysin tilastollinen malli voidaan esittää muodossa (ks. tarkemmin seuraavaa kappaletta):

$$y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk}\ ,\ i = 1,2,\mathrm{K}\ ,I\ ,\ J = 1,2,\mathrm{K}\ ,J\ ,\ k = 1,2,\mathrm{K}\ ,K$$

jossa

$$\sum_{j=1}^{J} \alpha_j = \sum_{k=1}^{K} \beta_k = \sum_{j=1}^{J} (\alpha\beta)_{jk} = \sum_{k=1}^{K} (\alpha\beta)_{jk} = 0$$

ja

$$\varepsilon_{ijk} \quad \mathrm{N}(0,\sigma^2)\ ,\ i = 1,2,\mathrm{K}\ ,I\ ,\ j = 1,2,\mathrm{K}\ ,J\ ,\ k = 1,2,\mathrm{K}\ ,K$$

Voidaan osoittaa, että nollahypoteesit

H_{AB} : Ei yhdysvaikutusta

H_A : Ei *A*-vaikutusta

H_B : Ei *B*-vaikutusta

ovat *ekvivalentteja* seuraavien ehtojen kanssa:

H_{AB} : $(\alpha\beta)_{jk} = 0\ ,\ j = 1,2,\mathrm{K}\ ,J\ ,\ k = 1,2,\mathrm{K}\ ,K$

H_A : $\alpha_j = 0\ ,\ j = 1,2,\mathrm{K}\ ,J$

H_B : $\beta_k = 0\ ,\ k = 1,2,\mathrm{K}\ ,K$

Voidaan osoittaa, että *aina pätee*:

$$\mathrm{E}(MSE) = \frac{1}{JK(I-1)} \mathrm{E}(SSE) = \sigma^2$$

Siten *MSE* on *aina* varianssin σ^2 *harhaton* estimaattori.

Voidaan osoittaa, että

$$\mathrm{E}(MSAB) = \frac{1}{(J-1)(K-1)}\mathrm{E}(SSAB) = \sigma^2 + \frac{I\sum_{j=1}^{K}\sum_{k=1}^{J}(\alpha\beta)_{jk}^2}{(J-1)(K-1)}$$

Siten *MSAB* on varianssin σ^2 *harhaton estimaattori, jos nollahypoteesi*

$$\mathrm{H}_{AB} \;:\; (\alpha\beta)_{jk} = 0\,,\; j = 1, 2, \mathrm{K}\,, J\,,\; k = 1, 2, \mathrm{K}\,, K$$

pätee.

Voidaan osoittaa, että

$$\mathrm{E}(MSA) = \frac{1}{J-1}\mathrm{E}(SSA) = \sigma^2 + \frac{IK\sum_{j=1}^{J}\alpha_j^2}{J-1}$$

Siten *MSA* on varianssin σ^2 *harhaton estimaattori, jos nollahypoteesi*

$$\mathrm{H}_{A} \;:\; \alpha_j = 0\,,\; j = 1, 2, \mathrm{K}\,, J$$

pätee.

Voidaan osoittaa, että

$$\mathrm{E}(MSB) = \frac{1}{K-1}\mathrm{E}(SSB) = \sigma^2 + \frac{IJ\sum_{k=1}^{K}\beta_k^2}{K-1}$$

Siten *MSB* on varianssin σ^2 *harhaton estimaattori, jos nollahypoteesi*

$$\mathrm{H}_{B} \;:\; \beta_k = 0\,,\; k = 1, 2, \mathrm{K}\,, K$$

pätee.

Varianssianalyysitaulukko

Varianssianalyysin tulokset esitetään tavallisesti alla esitetyn **varianssianalyysitaulukon** muodossa.

Varianssianalyysitaulukon *neliösummat SS* toteuttavat yhtälön

$$SST = SSA + SSB + SSAB + SSE$$

Yhtälö on *varianssianalyysihajotelma*. Varianssianalyysitaulukon neliösummien *vapausasteet df* (*df* = degrees of freedom) toteuttavat yhtälön

$$N - 1 = IJK - 1 = (J - 1) + (K - 1) + (J - 1)(K - 1) + JK(I - 1)$$

Vaihtelun lähde	**Neliö-summa *SS***	**Vapaus-asteet *df***	**Varianssi-estimaattori *MS***	***F*-testisuure**
A	SSA	$I-1$	$MSA=\frac{SSA}{I-1}$	$F=\frac{MSA}{MSE}$
B	SSB	$J-1$	$MSB=\frac{SSB}{J-1}$	$F=\frac{MSB}{MSE}$
AB	$SSAB$	$(I-1)(J-1)$	$MSAB=\frac{SSAB}{(I-1)(J-1)}$	$F=\frac{MSAB}{MSE}$
Jäännös	SSE	$IJ(K-1)$	$MSE=\frac{SSE}{IJ(K-1)}$	
Kokonais-vaihtelu	SST	$IJK-1$		

21.4. Kaksisuuntaisen varianssianalyysin malli ja sen parametrointi

Kaksisuuntaisen varianssianalyysin tilastollinen malli voidaan *parametroida* kahdella erilaisella tavalla. Parametroinnit ovat kuitenkin *ekvivalentteja*.

Parametrointi 1

Kaksisuuntaisen varianssianalyysin tilastollinen malli voidaan **parametroida** seuraavalla tavalla:

(1) $$y_{ijk}=\mu_{jk}+\varepsilon_{ijk},\ i=1,2,\mathsf{K},I,\ j=1,2,\mathsf{K},J,\ k=1,2,\mathsf{K},K$$

jossa jäännöstermit ε_{ijk} ovat *riippumattomia* ja *normaalijakautuneita*:

$$\varepsilon_{ijk}\quad \mathrm{N}(0,\sigma^2),\ i=1,2,\mathsf{K},I,\ j=1,2,\mathsf{K},J,\ k=1,2,\mathsf{K},K$$

Mallissa (1)

y_{ijk} = y-muuttujan k. havaintoarvo ryhmässä (j, k)

μ_{jk} = y-muuttujan odotusarvo ryhmässä (j, k)

Ei-satunnaiset vakiot μ_{jk}, j = 1, 2, … , J, k = 1, 2, … , K ja jäännösvarianssi σ^2 ovat kaksisuuntaisen varianssianalyysin tilastollisen mallin (1) *parametreja*.

Mallia (1) koskevista oletuksista seuraa, että

$$\mathrm{E}(y_{ijk})=\mu_{jk},\ i=1,2,\mathsf{K},I,\ j=1,2,\mathsf{K},J,\ k=1,2,\mathsf{K},K$$

ja

$$\mathrm{D}^2(y_{ijk})=\sigma^2,\ i=1,2,\mathsf{K},I,\ j=1,2,\mathsf{K},J,\ k=1,2,\mathsf{K},K$$

Parametrointi 2

Kaksisuuntaisen varianssianalyysin tilastollinen malli voidaan **parametroida** myös seuraavalla tavalla:

$$(2) \qquad y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk} \ , i = 1, 2, \mathsf{K}, I \ , \ j = 1, 2, \mathsf{K}, J \ , k = 1, 2, \mathsf{K}, K$$

jossa

$$(3) \qquad \sum_{j=1}^{J} \alpha_j = \sum_{k=1}^{K} \beta_k = \sum_{j=1}^{J} (\alpha\beta)_{jk} = \sum_{k=1}^{K} (\alpha\beta)_{jk} = 0$$

jossa jäännöstermit ε_{ijk} ovat *riippumattomia* ja *normaalijakautuneita*:

$$\varepsilon_{ijk} \quad \mathrm{N}(0, \sigma^2) \ , i = 1, 2, \mathsf{K}, I \ , \ j = 1, 2, \mathsf{K}, J \ , k = 1, 2, \mathsf{K}, K$$

Ei-satunnaiset vakiot μ, α_j, β_k, $(\alpha\beta)_{jk}$, $j = 1, 2, \ldots, J$, $k = 1, 2, \ldots, K$ ja jäännösvarianssi σ^2 ovat kaksisuuntaisen varianssianalyysin tilastollisen mallin (2) *parametreja*.

Mallia (2) koskevista oletuksista seuraa, että

$$\mathrm{E}(y_{ijk}) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} \ , i = 1, 2, \mathsf{K}, I \ , \ j = 1, 2, \mathsf{K}, J \ , k = 1, 2, \mathsf{K}, K$$

ja

$$\mathrm{D}^2(y_{ijk}) = \sigma^2 \ , i = 1, 2, \mathsf{K}, I \ , \ j = 1, 2, \mathsf{K}, J \ , k = 1, 2, \mathsf{K}, K$$

Parametrointien 1 ja 2 ekvivalenssi

Mallissa

$$(1) \qquad y_{ijk} = \mu_{jk} + \varepsilon_{ijk} \ , i = 1, 2, \mathsf{K}, I \ , \ j = 1, 2, \mathsf{K}, J \ , k = 1, 2, \mathsf{K}, K$$

y-havainnot esitetään seuraavien tekijöiden summana:

μ_{jk} = **ryhmäkohtainen odotusarvo**

ε_{ijk} = jäännöstermi

Mallissa

$$(2) \qquad y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk} \ , i = 1, 2, \mathsf{K}, I \ , \ j = 1, 2, \mathsf{K}, J \ , k = 1, 2, \mathsf{K}, K$$

y-havainnot esitetään seuraavien tekijöiden summana:

μ = **yleisodotusarvo**

α_j = **ryhmittelevän tekijän *A* tason A_j vaikutus**

β_k = **ryhmittelevän tekijän *B* tason B_k vaikutus**

$(\alpha\beta)_{jk}$ = **yhdysvaikutus**

ε_{ijk} = jäännöstermi

Mallit (1) ja (2) *ovat ekvivalentteja* – mallit on vain *parametroitu* eri tavoilla.

Määritellään keskiarvot

$$\mu_{j\cdot} = \frac{1}{K}\sum_{k=1}^{K}\mu_{jk}$$

$$\mu_{\cdot k} = \frac{1}{J}\sum_{j=1}^{J}\mu_{jk}$$

$$\mu = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}\mu_{jk} = \frac{1}{J}\sum_{j=1}^{J}\mu_{j\cdot} = \frac{1}{K}\sum_{k=1}^{K}\mu_{\cdot k}$$

Kirjoitetaan identiteetti

$$y_{ijk} = \mu + (\mu_{j\cdot} - \mu) + (\mu_{\cdot k} - \mu) + (\mu_{jk} - \mu_{j\cdot} - \mu_{\cdot k} + \mu) + (y_{ijk} - \mu_{jk})$$

ja merkitään

$$\begin{aligned}
\alpha_j &= \mu_{j\cdot} - \mu \\
\beta_k &= \mu_{\cdot k} - \mu \\
(\alpha\beta)_{jk} &= \mu_{jk} - \mu_{j\cdot} - \mu_{\cdot k} + \mu \\
\varepsilon_{ijk} &= y_{ijk} - \mu_{jk}
\end{aligned}$$

Tällöin

$$\sum_{j=1}^{J}\alpha_j = 0$$

$$\sum_{k=1}^{K}\beta_k = 0$$

$$\sum_{j=1}^{J}(\alpha\beta)_{jk} = \sum_{k=1}^{K}(\alpha\beta)_{jk} = 0$$

Perustelu:

$$\sum_{j=1}^{J}\alpha_j = \sum_{j=1}^{J}(\mu_{j\cdot} - \mu) = \sum_{j=1}^{J}\mu_{j\cdot} - J\mu = J\mu - J\mu = 0$$

$$\sum_{k=1}^{K}\beta_k = \sum_{k=1}^{K}(\mu_{\cdot k} - \mu) = \sum_{k=1}^{K}\mu_{\cdot k} - K\mu = K\mu - K\mu = 0$$

$$\begin{aligned}
\sum_{J=1}^{J}(\alpha\beta)_{jk} &= \sum_{j=1}^{J}(\mu_{jk} - \mu_{j\cdot} - \mu_{\cdot k} + \mu) \\
&= \sum_{j=1}^{J}\mu_{jk} - \sum_{j=1}^{J}\mu_{j\cdot} - J\mu_{\cdot k} + J\mu \\
&= J\mu_{\cdot k} - J\mu - J\mu_{\cdot k} + I\mu = 0
\end{aligned}$$

$$\begin{aligned}
\sum_{k=1}^{K}(\alpha\beta)_{jk} &= \sum_{k=1}^{K}(\mu_{jk} - \mu_{j\cdot} - \mu_{\cdot k} + \mu) \\
&= \sum_{k=1}^{K}\mu_{jk} - K\mu_{j\cdot} - \sum_{k=1}^{K}\mu_{\cdot k} + K\mu \\
&= K\mu_{j\cdot} - K\mu_{j\cdot} - K\mu + K\mu = 0
\end{aligned}$$

■

Siten

$$y_{ijk} = \mu_{jk} + \varepsilon_{ijk}$$
$$y_{ijk} = \mu + (\mu_j - \mu) + (\mu_k - \mu) + (\mu_{jk} - \mu_j - \mu_k + \mu) + (y_{ijk} - \mu_{jk})$$
$$y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk}$$

ovat *kaksisuuntaisen varianssianalyysin tilastollisen mallin ekvivalentteja esitysmuotoja,* jos ehdot (3) pätevät.

Edellä esitetystä seuraa, että nollahypoteesit

H_{AB} : Ei yhdysvaikutusta

H_A : Ei *A*-vaikutusta

H_B : Ei *B*-vaikutusta

ovat *ekvivalentteja* seuraavien ehtojen kanssa:

H_{AB} : $(\alpha\beta)_{jk} = 0$, $j = 1, 2, \mathsf{K}, J$, $k = 1, 2, \mathsf{K}, K$

H_A : $\alpha_j = 0$, $j = 1, 2, \mathsf{K}, J$

H_B : $\beta_k = 0$, $k = 1, 2, \mathsf{K}, K$

21.5. Kaksisuuntaisen varianssianalyysin mallin parametrien estimointi

Kaksisuuntaisen varianssianalyysin malli

Tarkastellaan *kaksisuuntaisen varianssianalyysin mallin* **parametrointia 2** (ks. edellistä kappaletta).

Olkoon siis

(2) $$y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk} \text{ , } i = 1, 2, \mathsf{K}, I \text{ , } j = 1, 2, \mathsf{K}, J \text{ , } k = 1, 2, \mathsf{K}, K$$

kaksisuuntaisen varianssianalyysin tilastollinen malli, jossa *jäännös-* eli *virhetermit* ε_{ijk} ovat riippumattomia ja normaalijakautuneita satunnaismuuttujia:

$$\varepsilon_{ijk} \quad \mathrm{N}(0, \sigma^2) \text{ , } i = 1, 2, \mathsf{K}, I \text{ , } j = 1, 2, \mathsf{K}, J \text{ , } k = 1, 2, \mathsf{K}, K$$

Mallissa (2)

y_{ijk} = *y*-muuttujan *i*. havaintoarvo ryhmässä (*j*, *k*)

μ = yleisodotusarvo

α_j = ryhmittelevän tekijän *A* tason A_j vaikutus

β_k = ryhmittelevän tekijän *B* tason B_k vaikutus

$(\alpha\beta)_{jk}$ = yhdysvaikutus

ε_{ijk} = jäännöstermi

Pienimmän neliösumman estimointi

Tarkastellaan mallin (2) parametrien *pienimmän neliösumman estimointia*.

Mallin (2) parametrit voidaan estimoida pienimmän neliösumman menetelmällä, jos *ehdot*

$$(3) \qquad \sum_{j=1}^{J}\alpha_j = \sum_{k=1}^{K}\beta_k = \sum_{j=1}^{J}(\alpha\beta)_{jk} = \sum_{k=1}^{K}(\alpha\beta)_{jk} = 0$$

pätevät. Parametrien **pienimmän neliösumman estimaattoreiksi** saadaan

$$\begin{aligned}
\hat{\mu} &= \bar{y} \\
\hat{\alpha}_j &= \bar{y}_{j\cdot} - \bar{y} \quad , j = 1,2,\mathrm{K},J \\
\hat{\beta}_k &= \bar{y}_{\cdot k} - \bar{y} \quad , k = 1,2,\mathrm{K},K \\
(\hat{\alpha}\hat{\beta})_{jk} &= \bar{y}_{jk} - \bar{y}_{j\cdot} - \bar{y}_{\cdot k} + \bar{y} \quad , j = 1,2,\mathrm{K},J \; , k = 1,2,\mathrm{K},K
\end{aligned}$$

Perustelu:

Estimoidaan mallin

$$(2) \qquad y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk} \; , i = 1,2,\mathrm{K},I \; , \; j = 1,2,\mathrm{K},J \; , k = 1,2,\mathrm{K},K$$

parametrit *PNS-menetelmällä*, kun oletamme, että ehdot

$$(3) \qquad \sum_{j=1}^{J}\alpha_j = \sum_{k=1}^{K}\beta_k = \sum_{j=1}^{J}(\alpha\beta)_{jk} = \sum_{k=1}^{K}(\alpha\beta)_{jk} = 0$$

pätevät. Etsitään *neliösumman*

$$SS = \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I}(y_{ijk} - \mu - \alpha_j - \beta_k - (\alpha\beta)_{jk})^2$$

minimi parametrien suhteen tavanomaiseen tapaan:

(i) *Derivoidaan* neliösumma *SS* parametrien suhteen.

(ii) Merkitään derivaatat nolliksi.

(iii) Ratkaistaan saadut *normaaliyhtälöt* parametrien suhteen ottamalla huomioon *ehdot* (3).

Normaaliyhtälöt voidaan esittää seuraavassa muodossa:

$$\begin{aligned}
\mu \quad &: \; IJK\mu + IK\sum_{j=1}^{J}\alpha_j + IJ\sum_{k=1}^{K}\beta_k + I\sum_{j=1}^{J}\sum_{k=1}^{K}(\alpha\beta)_{jk} = T \\
\alpha_j \quad &: \; IK\mu + IK\alpha_J + I\sum_{k=1}^{K}\beta_k + I\sum_{k=1}^{K}(\alpha\beta)_{jk} = T_{j\cdot} \\
\beta_k \quad &: \; IJ\mu + I\sum_{j=1}^{J}\alpha_j + IJ\beta_k + I\sum_{j=1}^{J}(\alpha\beta)_{jk} = T_{\cdot k} \\
(\alpha\beta)_{jk} \quad &: \; I\mu + I\alpha_j + I\beta_k + I(\alpha\beta)_{jk} = T_{jk} \\
&\qquad j = 1,2,\mathrm{K},J \; , k = 1,2,\mathrm{K},K
\end{aligned}$$

jossa

$$T_{jk} = \sum_{I=1}^{I} y_{ijk}$$

$$T_{j} = \sum_{k=1}^{K}\sum_{i=1}^{I} y_{ijk} = \sum_{k=1}^{K} T_{jk}$$

$$T_{k} = \sum_{j=1}^{J}\sum_{I=1}^{I} y_{ijk} = \sum_{j=1}^{J} T_{jk}$$

$$T = \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{I=1}^{I} y_{ijk} = \sum_{j=1}^{J}\sum_{k=1}^{K} T_{jk} = \sum_{j=1}^{J} T_{j} = \sum_{k=1}^{K} T_{k}$$

$$j = 1, 2, \mathrm{K}, J \;,\; k = 1, 2, \mathrm{K}, K$$

Huomaa, että sekä estimoitavien parametrien lukumäärä että ratkaistavien normaaliyhtälöiden yhtälöiden lukumäärä on

$$1 + J + K + JK$$

Yhtälöt ovat kuitenkin *yliparametroituja*:

(i) Yhtälöiden (2) summana saadaan yhtälö (1).

(ii) Yhtälöiden (3) summana saadaan yhtälö (1).

(iii) Yhtälöiden (4) summana kiinteälle j saadaan yhtälö (2).

(iv) Yhtälöiden (4) summana kiinteälle k saadaan yhtälö (3).

Siten yhtälösysteemin yhtälöiden välillä on

$$1 + J + K$$

lineaarista riippuvuutta ja *systeemillä ei ole yksikäsitteistä ratkaisua. Yhtälösysteemi saadaan kuitenkin ratkaistuksi ottamalla huomioon ehdot*

(3) $$\sum_{j=1}^{J} \alpha_j = \sum_{k=1}^{K} \beta_k = \sum_{j=1}^{J} (\alpha\beta)_{jk} = \sum_{k=1}^{K} (\alpha\beta)_{jk} = 0$$

Huomaa, että *riippumattomien* ehtojen lukumäärä on yhtälösysteemin ratkaisemiseksi tarvittava

$$1 + I + J$$

Kun ehdot (3) otetaan huomioon yhtälösysteemin ratkaisuksi saadaan

$$\hat{\mu} = \bar{y}$$

$$\hat{\alpha}_j = \bar{y}_{j} - \bar{y} \quad , j = 1, 2, \mathrm{K}, J$$

$$\hat{\beta}_k = \bar{y}_{k} - \bar{y} \quad , k = 1, 2, \mathrm{K}, K$$

$$(\hat{\alpha}\hat{\beta})_{jk} = \bar{y}_{jk} - \bar{y}_{j} - \bar{y}_{k} + \bar{y} \quad , j = 1, 2, \mathrm{K}, J \;,\; k = 1, 2, \mathrm{K}, K$$

■

Sovitteet ja residuaalit

Estimoidun kaksisuuntaisen varianssianalyysin mallin **sovitteet** saadaan kaavoilla

$$\begin{aligned}\hat{y}_{ijk} &= \hat{\mu} + \hat{\alpha}_j + \hat{\beta}_k + (\hat{\alpha}\hat{\beta})_{jk} \\ &= \bar{y}\ \ + (\bar{y}_{\,j} - \bar{y}\ \ ) + (\bar{y}_{\,k} - \bar{y}\ \ ) + (\bar{y}_{\,jk} - \bar{y}_{\,j} - \bar{y}_{\,k} + \bar{y}\ \ ) \\ &= \bar{y}_{\,jk} \\ i &= 1, 2, \mathrm{K}, I\ ,\ j = 1, 2, \mathrm{K}, J\ ,\ k = 1, 2, \mathrm{K}, K\end{aligned}$$

ja estimoidun mallin **residuaalit** saadaan kaavoilla

$$\begin{aligned}e_{ijk} &= y_{ijk} - \hat{y}_{ijk} = y_{ijk} - \bar{y}_{\,jk} \\ i &= 1, 2, \mathrm{K}, I\ ,\ j = 1, 2, \mathrm{K}, J\ ,\ k = 1, 2, \mathrm{K}, K\end{aligned}$$

Mallin residuaalit on aina syytä alistaa samanlaisiin **diagnostisiin tarkistuksiin** kuin minkä tahansa regressiomallin residuaalit. Residuaalien avulla voidaan selvittää onko havaintojen joukossa **poikkeavia havaintoja** sekä pätevätkö mallin jäännöstermeistä tehdyt **homoskedastisuus-**, **korreloimattomuus-** ja **normaalisuusoletukset**; ks. yksityiskohtia luvusta **Regressiodiagnostiikka**.

21.6. Laskutoimitusten suorittaminen

Jos kaksisuuntaisen varianssianalyysin laskutoimitukset joudutaan tekemään *käsin* tai *laskimella*, kannattaa laskutoimituksissa käyttää alla esitettäviä kaavoja.

Oletetaan, että ryhmittelevällä tekijällä A on J tasoa:

$$A_1\ ,\ A_2\ ,\ \ldots\ ,\ A_J$$

ja ryhmittelevällä tekijällä B on K tasoa:

$$B_1\ ,\ B_2\ ,\ \ldots\ ,\ B_K$$

Olkoon

y_{ijk} = k. havainto ryhmässä, jonka määrittelee tekijän A taso j ja tekijän B taso k,
$i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$, $k = 1, 2, \ldots, K$

Määritellään seuraavat summat:

$$\begin{aligned}T_{\,jk} &= \sum_{I=1}^{I} y_{ijk} \\ T_{\,j} &= \sum_{k=1}^{K}\sum_{i=1}^{I} y_{ijk} = \sum_{k=1}^{K} T_{\,jk} \\ T_{\,k} &= \sum_{j=1}^{J}\sum_{I=1}^{I} y_{ijk} = \sum_{j=1}^{J} T_{\,jk} \\ T\ \ &= \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{I=1}^{I} y_{ijk} = \sum_{j=1}^{J}\sum_{k=1}^{K} T_{\,jk} = \sum_{j=1}^{J} T_{\,j} = \sum_{k=1}^{K} T_{\,k} \\ & \qquad j = 1, 2, \mathrm{K}, J\ ,\ k = 1, 2, \mathrm{K}, K\end{aligned}$$

Olkoon lisäksi

$$\sum_{i=1}^{I} y_{ijk}^2\ ,\ j = 1, 2, \mathrm{K}, J\ ,\ k = 1, 2, \mathrm{K}, K$$

tekijän A tason A_j ja tekijän B tason B_k määräämän ryhmän (j, k) havaintoarvojen y_{ijk} neliöiden summa ja

$$\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ijk}^2$$

kaikkien havaintoarvojen y_{ijk} neliöiden kokonaissumma.

Havaintoarvojen y_{ijk} *ryhmäkeskiarvot* saadaan kaavoilla

$$\overline{y}_{jk} = \frac{1}{I}\sum_{i=1}^{I} y_{ijk} = \frac{1}{I}T_{jk}\ ,\ j = 1,2,\mathrm{K},J\ ,\ k = 1,2,\mathrm{K},K$$

Havaintoarvojen y_{ijk} *reunakeskiarvot* saadaan kaavoilla

$$\overline{y}_{j} = \frac{1}{IJ}\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ijk} = \frac{1}{IJ}T_{j}\ ,\ j = 1,2,\mathrm{K},J$$

$$\overline{y}_{k} = \frac{1}{IK}\sum_{k=1}^{K}\sum_{i=1}^{I} y_{ijk} = \frac{1}{IK}T_{k}\ ,\ k = 1,2,\mathrm{K},K$$

Havaintoarvojen y_{ijk} *kokonaiskeskiarvo* saadaan kaavalla

$$\overline{y} = \frac{1}{IJK}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ijk} = \frac{1}{IJK}T$$

Havaintoarvojen y_{ijk} *ryhmävarianssit* saadaan kaavoilla

$$s^2_{jk} = \frac{1}{I-1}\sum_{i=1}^{I}(y_{ijk} - \overline{y}_{jk})^2 = \frac{1}{I-1}\left(\sum_{i=1}^{I} y^2_{jk} - \frac{1}{I}T^2_{jk}\right),\ j = 1,2,\mathrm{K},J\ ,\ k = 1,2,\mathrm{K},K$$

Havaintoarvojen y_{ijk} *kokonaisvarianssi* saadaan kaavalla

$$s^2 = \frac{1}{IJK-1}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I}(y_{ijk} - \overline{y})^2 = \frac{1}{IJK-1}\left(\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} y^2_{ijk} - \frac{1}{IJK}T^2\right)$$

Havaintoarvojen *kokonaisvaihtelua kuvaava neliösumma SST* voidaan laskea kaavalla

$$SST = \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I}(y_{ijk} - \overline{y})^2 = \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} y^2_{ijk} - \frac{1}{IJK}T^2$$

Tekijän A päävaikutusta kuvaava neliösumma SSA voidaan laskea kaavalla

$$SSA = IK\sum_{j=1}^{J}(\overline{y}_{j} - \overline{y})^2 = \frac{1}{IK}\sum_{j=1}^{J}T^2_{j} - \frac{1}{IJK}T^2$$

Tekijän B päävaikutusta kuvaava neliösumma SSB voidaan laskea kaavalla

$$SSB = IJ\sum_{k=1}^{K}(\overline{y}_{k} - \overline{y})^2 = \frac{1}{IJ}\sum_{k=1}^{K}T^2_{k} - \frac{1}{IJK}T^2$$

Tekijöiden A ja *B yhdysvaikutusta kuvaava neliösumma SSAB* kannattaa laskea kahdessa vaiheessa. Lasketaan ensin ryhmäkeskiarvojen kokonaisvaihtelua kuvaava neliösumma

$$SS = I\sum_{j=1}^{J}\sum_{k=1}^{K}(\overline{y}_{jk} - \overline{y})^2 = \frac{1}{I}\sum_{j=1}^{J}\sum_{k=1}^{K}T^2_{jk} - \frac{1}{IJK}T^2$$

Tällöin

$$SSAB = I\sum_{j=1}^{J}\sum_{k=1}^{K}(\bar{y}_{jk} - \bar{y}_{j} - \bar{y}_{k} + \bar{y})^2 = SS - SSA - SSB$$

Ryhmien sisäistä vaihtelua kuvaava jäännösneliösumma SE saadaan kaavalla

$$SSE = SST - SSAB - SSA - SSB = SST - SS$$

Käsin tai laskimella laskettaessa havainnot kannattaa järjestää seuraavan taulukon muotoon:

	A_1	A_2	L	A_J
B_1	$y_{111}, y_{211}, \mathrm{K}, y_{I11}$	$y_{121}, y_{221}, \mathrm{K}, y_{I21}$	L	$y_{1J1}, y_{2J1}, \mathrm{K}, y_{IJ1}$
B_2	$y_{112}, y_{212}, \mathrm{K}, y_{I12}$	$y_{122}, y_{222}, \mathrm{K}, y_{I22}$	L	$y_{1J2}, y_{2J2}, \mathrm{K}, y_{IJ2}$
M	M	M		M
B_K	$y_{11K}, y_{21K}, \mathrm{K}, y_{I1K}$	$y_{12K}, y_{22K}, \mathrm{K}, y_{I2K}$	L	$y_{1JK}, y_{2JK}, \mathrm{K}, y_{IJK}$

Tästä taulukosta lasketaan solukohtaiset summat

$$T_{jk} = \sum_{i=1}^{I} y_{ijk}\ ,\ j = 1, 2, \mathrm{K}, J\ , k = 1, 2, \mathrm{K}, K$$

ja kaikkien havaintojen neliöiden summa

$$\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ijk}^2$$

Solusummat

$$T_{jk}\ ,\ j = 1, 2, \mathrm{K}, J\ , k = 1, 2, \mathrm{K}, K$$

järjestetään seuraavaksi taulukoksi, josta loput tarvittavista summista saadaan rivi- ja sarakesummina:

	A_1	A_2	L	A_J	Summa
B_1	T_{11}	T_{21}	L	T_{J1}	T_{1}
B_2	T_{12}	T_{22}	L	T_{J2}	T_{2}
M	M	M		M	M
B_K	T_{1K}	T_{2K}	L	T_{JK}	T_{K}
Summa	T_{1}	T_{2}	L	T_{J}	T

22. Kolmi- ja useampisuuntainen varianssianalyysi

22.1. Varianssianalyysi: Johdanto

22.2. Kolmisuuntainen varianssianalyysi ja sen malli

22.3. Kolmisuuntainen varianssianalyysi ja sen suorittaminen

22.4. Laskutoimitusten suorittaminen kolmisuuntaisessa varianssianalyysissa

Yksisuuntaisessa varianssianalyysissa perusjoukko on jaettu ryhmiin *yhden ryhmittelevän* **tekijän** suhteen ja tavoitteena on testata hypoteesia, jonka mukaan **kiinnostuksen kohteena olevan muuttujan ryhmäkohtaiset odotusarvot ovat yhtä suuria**.

Kaksi- tai **useampisuuntaisessa varianssianalyysissa** perusjoukko on jaettu ryhmiin **kahden** tai **useamman ryhmittelevän tekijän** suhteen ja nytkin tavoitteena on testata hypoteesia, jonka mukaan **kiinnostuksen kohteena olevan muuttujan ryhmäkohtaiset odotusarvot ovat yhtä suuria**.

Tässä luvussa tarkastellaan **kolmisuuntaista varianssianalyysia** esimerkkinä kolmi- ja useampisuuntaisesta varianssianalyysista. Kolmisuuntaisen varianssianalyysin tarkasteleminen *riittää* tämän esityksen tarpeisiin, koska neli- tai useampisuuntaisen varianssianalyysin asetelma ei vaadi uusia teoreettisia näkökulmia. Tarkastelun kohteena ovat mm. **kolmisuuntaisen varianssianalyysin malli** ja sen **parametrointi**, **parametrien estimointi**, **odotusarvojen yhtäsuuruuden testaaminen** ja **laskutoimitusten suorittaminen**.

Avainsanat:

Aritmeettinen keskiarvo, Cochranin lause, Estimointi, *F*-testi, Faktori, Harha, Harhattomuus, Indikaattorimuuttuja, Interaktio, Jäännösvaihtelu, Jäännösvarianssi, Kaksisuuntainen varianssianalyysi, Kokonaiskeskiarvo, Kokonaisvaihtelu, Kolmisuuntainen varianssianalyysi, Merkitsevyystaso, Muuttuja, Normaalijakauma, Neliösumma, Otos, Otostunnusluku, Parametri, Parametrointi, Pienimmän neliösumman estimaattori, Pienimmän neliösumman menetelmä, Päävaikutus, Residuaali, Reunakeskiarvo, Riippumattomuus, Ryhmien sisäinen vaihtelu, Ryhmien välinen vaihtelu, Ryhmittely, Ryhmä, Ryhmäkeskiarvo, Satunnaisuus, Side-ehto, Sovite, *t*-testi, Taso, Tekijä, Testi, Vapausaste, Varianssi, Varianssianalyysihajotelma, Varianssianalyysitaulukko, Yhdysvaikutus, Yksisuuntainen varianssianalyysi, Yleinen lineaarinen malli, Yleiskeskiarvo

22.1. Varianssianalyysi: Johdanto

Kahden riippumattoman otoksen *t*-testi

Suhdeasteikollisille muuttujille tarkoitettuja testejä käsitelleessä kappaleessa tarkastellaan **kahden riippumattoman otoksen *t*-testiä.** Testin testausasetelma on seuraava:

(i) Perusjoukko koostuu *kahdesta* ryhmästä.

(ii) Havainnot noudattavat kummassakin ryhmässä *normaalijakaumaa*.

(iii) Kummastakin ryhmästä on poimittu *toisistaan riippumattomat satunnaisotokset*.

(iv) Tehtävänä on *testata ryhmäkohtaisten odotusarvojen samuutta*.

Varianssianalyysin perusongelma

Varianssianalyysi voidaan ymmärtää *kahden riippumattoman otoksen t-testin yleistykseksi* tilanteisiin, jossa perusjoukko koostuu *kahdesta* tai *useammasta ryhmästä*:

(i) Perusjoukko koostuu *kahdesta* tai *useammasta* ryhmästä.

(ii) Havainnot noudattavat jokaisessa ryhmässä *normaalijakaumaa*.

(iii) Jokaisesta ryhmästä poimitaan *toisistaan riippumattomat satunnaisotokset*.

(iv) Tehtävänä on *testata ryhmäkohtaisten odotusarvojen samuutta*.

Perusjoukon *jako ryhmiin* voidaan tehdä *yhden* tai *useamman* **faktorin** eli **tekijän** (*muuttujan*) *arvojen perusteella*. Jos perusjoukon jako ryhmiin perustuu *yhteen tekijään*, puhutaan **yksisuuntaisesta varianssianalyysista**. Jos perusjoukon jako ryhmiin perustuu *m tekijään*, puhutaan ***m*-suuntaisesta varianssianalyysista**.

Huomautus:

- Tässä luvussa käsitellään useampisuuntaisesta varianssianalyysista esimerkkinä **kolmisuuntaista varianssianalyysia**; **yksi**- ja **kaksisuuntaista varianssianalyysia** on käsitelty edellisissä luvuissa.

Varianssianalyysin nimi *johtaa helposti harhaan*. Varianssianalyysissa ei testata varianssien vaan **odotusarvojen** yhtäsuuruutta tilanteessa. Nimi johtuu siitä, että odotusarvojen yhtäsuuruuden testaaminen perustuu eri tavoilla määrättyjen *varianssien yhtäsuuruuden testaamiseen*.

22.2. Kolmisuuntainen varianssianalyysi ja sen malli

Kolmisuuntaisen varianssianalyysin perusasetelma

Oletetaan, että tutkimuksen kohteena oleva perusjoukko voidaan *jakaa ryhmiin kolmen* **faktorin** eli **tekijän** (*muuttujan*) *A*, *B* ja *C arvojen suhteen* ja oletetaan, että tekijällä *A* on *J* **tasoa**, tekijällä *B* on *K* **tasoa** ja tekijällä *C* on *L* **tasoa**, jolloin jaossa syntyy **ryhmiä** $J \times K \times L$ kappaletta. Oletetaan edelleen, että ryhmistä on poimittu *toisistaan riippumattomat yksinkertaiset satunnaisotokset*, joiden kaikkien koko on *I*.

Olkoon

y_{ijkl} = *i*. havainto tekijän *A* tason A_j, tekijän *B* tason B_k ja tekijän *C* tason C_l määräämässä ryhmässä (j, k, l)

$$i = 1, 2, \ldots, I \, , \, j = 1, 2, \ldots, J \, , \, k = 1, 2, \ldots, K \, , \, l = 1, 2, \ldots, L$$

Käytetystä otantamenetelmästä seuraa, että havainnot y_{ijkl} voidaan olettaa *riippumattomiksi* (ja siten myös *korreloimattomiksi*) satunnaismuuttujiksi.

Oletetaan, että havainnot y_{ijkl} ovat *normaalijakautuneita*:

$$y_{ijkl} \sim \mathrm{N}(\mu_{jkl} \, , \, \sigma^2) \, , \, i = 1, 2, \ldots, I \, , \, j = 1, 2, \ldots, J \, , \, k = 1, 2, \ldots, K \, , \, l = 1, 2, \ldots, L$$

Havainnoista y_{ijkl} tehdystä oletuksesta seuraa:

(i) Kaikilla *samaan ryhmään* (*j*,*k*,*l*) *kuuluvilla havainnoilla* on *sama* odotusarvo:

$$\mathrm{E}(y_{ijkl}) = \mu_{jkl} \, , \, i = 1, 2, \ldots, I \, , \, j = 1, 2, \ldots, J \, , \, k = 1, 2, \ldots, K \, , \, l = 1, 2, \ldots, L$$

(ii) Kaikilla havainnoilla on *ryhmästä riippumatta sama* varianssi:

$$\mathrm{D}^2(y_{ijkl}) = \sigma^2 \, , \, i = 1, 2, \ldots, I \, , \, j = 1, 2, \ldots, J \, , \, k = 1, 2, \ldots, K \, , \, l = 1, 2, \ldots, L$$

Haluamme testata nollahypoteesia siitä, että *ryhmäkohtaiset odotusarvot* $\mathrm{E}(y_{ijkl}) = \mu_{jkl}$ *ovat yhtä suuria*. Asetetaan siis **nollahypoteesi**

$$\mathrm{H}_0 : \mu_{jkl} = \mu \, , \, j = 1, 2, \ldots, J \, , \, k = 1, 2, \ldots, K \, , \, l = 1, 2, \ldots, L$$

Jos nollahypoteesi *ryhmäkohtaisten odotusarvojen yhtäsuuruudesta pätee*, *ryhmät voidaan yhdistää* kaikissa havaintojen keskimääräisiä arvoja koskevissa tarkasteluissa.

Kolmisuuntaisessa varianssianalyysissa nollahypoteesi H_0 on tapana jakaa *seitsemäksi* nollahypoteesiksi, jotka koskevat **tekijöiden *A***, ***B*** ja ***C* päävaikutuksia**, **tekijöiden *A***, ***B*** ja ***C* pareittaisia interaktiota** eli **yhdysvaikutuksia** ja **tekijöiden *A***, ***B*** ja ***C* yhdysvaikutusta**.

Tekijöiden A, B ja *C yhdysvaikutusta koskeva nollahypoteesi* on muotoa

H_{ABC} : Ei yhdysvaikutusta *ABC*

Jos nollahypoteesi H_{ABC} jää voimaan, tekijöiden *A*, *B* ja *C* vaikutuksia voidaan tutkia *pareittain*.

Tekijöiden A ja *B yhdysvaikutusta koskeva nollahypoteesi* on muotoa

H_{AB} : Ei yhdysvaikutusta *AB*

Tekijöiden A ja *C yhdysvaikutusta koskeva nollahypoteesi* on muotoa

H_{AC} : Ei yhdysvaikutusta *AC*

Tekijöiden B ja *C yhdysvaikutusta koskeva nollahypoteesi* on muotoa

H_{BC} : Ei yhdysvaikutusta *BC*

Jos nollahypoteesit H_{ABC} , H_{AB} , H_{AC} , H_{BC} jäävät voimaan, tekijöiden *A*, *B* ja *C* vaikutusta voidaan tutkia *erillisinä*.

Tekijän A vaikutusta koskeva nollahypoteesi on muotoa

H_A : Ei *A*-vaikutusta

Tekijän B vaikutusta koskeva nollahypoteesi on muotoa

H_B : Ei *B*-vaikutusta

Tekijän C vaikutusta koskeva nollahypoteesi on muotoa

H_C : Ei *C*-vaikutusta

Huomautus:

- Nollahypoteesit H_A , H_B , H_C ovat yksisuuntaisen varianssianalyysin nollahypoteeseja.

Kolmisuuntainen varianssianalyysi tarkoittaa em. testausasetelman nollahypoteesien H_{ABC} , H_{AB} , H_{AC} , H_{BC} , H_A , H_B , H_C testaamista.

Kolmisuuntaisen varianssianalyysin tilastollinen malli

Kolmisuuntaisen varianssianalyysin **tilastollinen malli** voidaan *parametroida* seuraavalla tavalla:

$$y_{ijkl} = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl} + (\alpha\beta\gamma)_{jkl} + \varepsilon_{ijkl}$$
$$i = 1, 2, \mathrm{K}, I \ , \ j = 1, 2, \mathrm{K}, J \ , \ k = 1, 2, \mathrm{K}, K \ , \ l = 1, 2, \mathrm{K}, L$$

jossa jäännöstermit ε_{ijkl} ovat *riippumattomia* ja *normaalijakautuneita*:

$$\varepsilon_{ijkl} \quad \mathrm{N}(0, \sigma^2) \ , \ i = 1, 2, \mathrm{K}, I \ , \ j = 1, 2, \mathrm{K}, J \ , \ k = 1, 2, \mathrm{K}, K \ , \ l = 1, 2, \mathrm{K}, L$$

Ei-satunnaiset vakiot

$$\mu$$
$$\alpha_j \ , \ \beta_k \ , \ \gamma_l$$
$$(\alpha\beta)_{jk} \ , \ (\alpha\gamma)_{jl} \ , \ (\beta\gamma)_{kl} \ , \ (\alpha\beta\gamma)_{jkl}$$
$$j = 1, 2, \ldots, J \ , \ k = 1, 2, \ldots, K \ , \ l = 1, 2, \ldots, L$$

ja *jäännösvarianssi*

$$\sigma^2$$

ovat kolmisuuntaisen varianssianalyysin tilastollisen mallin *parametrit*.

Kolmisuuntaisen varianssianalyysin tilastollisen mallin parametrien on toteutettava seuraavat *ehdot*:

$$\sum_{j=1}^{J} \alpha_j = \sum_{k=1}^{K} \beta_k = \sum_{l=1}^{L} \gamma_l = 0$$
$$\sum_{j=1}^{J} (\alpha\beta)_{jk} = \sum_{k=1}^{K} (\alpha\beta)_{jk} = 0$$
$$\sum_{j=1}^{J} (\alpha\gamma)_{jl} = \sum_{l=1}^{L} (\alpha\gamma)_{jl} = 0$$
$$\sum_{k=1}^{K} (\beta\gamma)_{kl} = \sum_{l=1}^{L} (\beta\gamma)_{kl} = 0$$
$$\sum_{j=1}^{J} (\alpha\beta\gamma)_{jkl} = \sum_{k=1}^{K} (\alpha\beta\gamma)_{jkl} = \sum_{l=1}^{L} (\alpha\beta\gamma)_{jkl} = 0$$

Mallia koskevista oletuksista seuraa, että

$$\mathrm{E}(y_{ijkl}) = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl} + (\alpha\beta\gamma)_{jkl}$$
$$i = 1, 2, \mathrm{K}, I \ , \ j = 1, 2, \mathrm{K}, J \ , \ k = 1, 2, \mathrm{K}, K \ , \ l = 1, 2, \mathrm{K}, L$$

ja

$$\mathrm{Var}(\varepsilon_{ijkl}) = \sigma^2 \ , \ i = 1, 2, \mathrm{K}, I \ , \ j = 1, 2, \mathrm{K}, J \ , \ k = 1, 2, \mathrm{K}, K \ , \ l = 1, 2, \mathrm{K}, L$$

Kolmisuuntaisen varianssianalyysin nollahypoteesit voidaan ilmaista mallin parametrien avulla seuraavassa muodossa:

$$H_{ABC} : (\alpha\beta\gamma)_{jkl} = 0$$

$$H_{AB} : (\alpha\beta)_{jk} = 0$$

$$H_{AC} : (\alpha\gamma)_{jl} = 0$$

$$H_{BC} : (\beta\gamma)_{kl} = 0$$

$$H_{A} : \alpha_{j} = 0$$

$$H_{B} : \beta_{k} = 0$$

$$H_{C} : \gamma_{l} = 0$$

Kolmisuuntainen varianssianalyysi ja koesuunnittelu

Kolmisuuntaista varianssianalyysiä voidaan käyttää koetulosten analyysiin seuraavassa **koeasetelmassa**:

(i) Oletetaan, että kokeen tavoitteena on verrata, miten **käsittelyiden**

$$A_1 , A_2 , \ldots , A_J$$

ja

$$B_1 , B_2 , \ldots , B_K$$

ja

$$C_1 , C_2 , \ldots , C_L$$

yhdistelmät vaikuttavat kiinnostuksen kohteena olevan **vastemuuttujan** *y keskimääräisiin arvoihin*.

(ii) Valitaan käsittelykombinaation (A_j, B_k, C_l) kohteeksi kaikkien kokeen mahdollisten kohteiden joukosta *satunnaisesti I* yksilöä, $j = 1, 2, \ldots , J$, $k = 1, 2, \ldots , K$, $l = 1, 2, \ldots , L$ ja olkoon

$$IJKL = N$$

havaintojen kokonaislukumäärä.

(iii) Mitataan **vasteet** y_{ijkl} eli kiinnostuksen kohteena olevan muuttujan *y* arvot:

$$y_{ijkl} , i = 1, 2, \ldots , I , j = 1, 2, \ldots , J , k = 1, 2, \ldots , K , l = 1, 2, \ldots , L$$

Huomaa, että koeasetelma on **täydellisesti satunnaistettu**: *Sattuma määrää täydellisesti millaisen käsittelyn kohteeksi kokeen kohteiksi valitut yksilöt joutuvat*.

22.3. Kolmisuuntaisen varianssianalyysin suorittaminen

Havaintojen keskiarvot

Määritellään havaintoarvojen y_{ijkl} **ryhmäkeskiarvot** eli *ryhmäkohtaiset aritmeettiset keskiarvot* tekijän *A* tason A_j, tekijän *B* tason B_k ja tekijän *C* tason C_l määräämässä ryhmässä (*j*,*k*,*l*):

$$\bar{y}_{jkl} = \frac{1}{I}\sum_{i=1}^{I} y_{ijkl} , \; j = 1, 2, \mathrm{K} , J , \; k = 1, 2, \mathrm{K} , K , \; l = 1, 2, \mathrm{K} , L$$

Jos ryhmäkohtaiset otokset yhdistetään yhdeksi otokseksi, yhdistetyn otoksen havaintoarvojen y_{ijkl} **yleis-** eli **kokonaiskeskiarvo**

on

$$\bar{y} = \frac{1}{IJKL}\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ijkl}$$

jossa

$$IJKL = N$$

on havaintojen kokonaislukumäärä.

Määritellään havaintoarvojen y_{ijkl} **1. kertaluvun marginaali-** eli **reunakeskiarvot** kaavoilla:

$$\bar{y}_{j} = \frac{1}{IKL}\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{i=1}^{I} y_{ijkl}\ ,\ j = 1,2,\mathrm{K},J$$

$$\bar{y}_{k} = \frac{1}{IJL}\sum_{l=1}^{L}\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ijkl}\ ,\ k = 1,2,\mathrm{K},K$$

$$\bar{y}_{l} = \frac{1}{IJK}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ijkl}\ ,\ l = 1,2,\mathrm{K},L$$

Määritellään havaintoarvojen y_{ijkl} **2. kertaluvun marginaali-** eli **reunakeskiarvot** kaavoilla:

$$\bar{y}_{jk} = \frac{1}{JK}\sum_{l=1}^{L}\sum_{i=1}^{I} y_{ijkl}\ ,\ j = 1,2,\mathrm{K},J\ ,\ k = 1,2,\mathrm{K},K$$

$$\bar{y}_{jl} = \frac{1}{JL}\sum_{k=1}^{K}\sum_{i=1}^{I} y_{ijkl}\ ,\ j = 1,2,\mathrm{K},J\ ,\ l = 1,2,\mathrm{K},L$$

$$\bar{y}_{kl} = \frac{1}{KL}\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ijkl}\ ,\ k = 1,2,\mathrm{K},K\ ,\ l = 1,2,\mathrm{K},L$$

Varianssianalyysihajotelma

Kirjoitetaan identiteetti

$$\begin{aligned} y_{ijkl} - \bar{y} = {} & (\bar{y}_{j} - \bar{y}) + (\bar{y}_{k} - \bar{y}) + (\bar{y}_{l} - \bar{y}) \\ & + (\bar{y}_{jk} - \bar{y}_{j} - \bar{y}_{k} + \bar{y}) \\ & + (\bar{y}_{jl} - \bar{y}_{j} - \bar{y}_{l} + \bar{y}) \\ & + (\bar{y}_{kl} - \bar{y}_{k} - \bar{y}_{l} + \bar{y}) \\ & + (\bar{y}_{jkl} - \bar{y}_{jk} - \bar{y}_{jl} - \bar{y}_{kl} + \bar{y}_{j} + \bar{y}_{k} + \bar{y}_{l} - \bar{y}) \\ & + (y_{ijkl} - \bar{y}_{jkl}) \end{aligned}$$

3-suuntaisen varianssianalyysin testit perustuvat näiden sulkulausekkeilla esitettyjen *poikkeamien neliösummille*.

Määritellään **havaintoarvojen kokonaisvaihtelua kuvaava kokonaisneliösumma**:

$$SST = \sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} (y_{ijkl} - \bar{y})^2$$

Jos ryhmäkohtaiset otokset yhdistetään yhdeksi otokseksi, saadun yhdistetyn otoksen *varianssi* on

$$s_y^2 = \frac{1}{IJKL-1} SST$$

jossa

$$IJKL = N$$

on yhdistetyn otoksen havaintojen kokonaislukumäärä.

Määritellään **tekijöiden *A* ja *B* ja *C* päävaikutuksia kuvaavat neliösummat**:

$$SSA = IKL \sum_{j=1}^{J} (\bar{y}_{\ j} - \bar{y})^2$$

$$SSB = IJL \sum_{k=1}^{K} (\bar{y}_{\ k} - \bar{y})^2$$

$$SSC = IJK \sum_{l=1}^{L} (\bar{y}_{\ l} - \bar{y})^2$$

Määritellään **tekijöiden *A* ja *B*, *A* ja *C*, *B* ja *C* yhdysvaikutuksia kuvaavat neliösumma**:

$$SSAB = IL \sum_{k=1}^{K} \sum_{j=1}^{J} (\bar{y}_{jk} - \bar{y}_{j} - \bar{y}_{\ k} + \bar{y})^2$$

$$SSAC = IK \sum_{l=1}^{L} \sum_{j=1}^{J} (\bar{y}_{j\ l} - \bar{y}_{j} - \bar{y}_{\ l} + \bar{y})^2$$

$$SSBC = IJ \sum_{l=1}^{L} \sum_{k=1}^{K} (\bar{y}_{\ kl} - \bar{y}_{\ k} - \bar{y}_{\ l} + \bar{y})^2$$

Määritellään **tekijöiden *A* ja *B* ja *C* yhdysvaikutusta kuvaava neliösumma**:

$$SSABC = I \sum_{l=1}^{L} \sum_{k=1}^{K} \sum_{j=1}^{J} (\bar{y}_{jkl} - \bar{y}_{jk} - \bar{y}_{j\ l} - \bar{y}_{\ kl} + \bar{y}_{j} + \bar{y}_{\ k} + \bar{y}_{\ l} - \bar{y})^2$$

Määritellään **ryhmien sisäistä vaihtelua kuvaava** (**jäännös**-) **neliösumma**:

$$SSE = \sum_{l=1}^{L} \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{i=1}^{I} (y_{ijkl} - \bar{y}_{\ jkl})^2$$

Havaintoarvojen y_{ijkl} *ryhmävarianssit* eli *ryhmäkohtaiset varianssit* saadaan lausekkeista

$$s_{jkl}^2 = \frac{1}{I-1} \sum_{i=1}^{I} (y_{ijkl} - \bar{y}_{\ jkl})^2 \ , \ j = 1, 2, \mathrm{K}, J \ , k = 1, 2, \mathrm{K}, K \ , l = 1, 2, \mathrm{K}, L$$

Siten ryhmien sisäistä vaihtelua kuvaava neliösumman *SSE* lauseke voidaan esittää myös muodossa

$$SSE = (I-1) \sum_{l=1}^{L} \sum_{k=1}^{K} \sum_{j=1}^{J} s_{jkl}^2$$

Neliösummat *SST*, *SSA*, *SSB*, *SSC*, *SSAB*, *SSAC*, *SSBC*, *SSABC*, *SSE* toteuttavat **varianssianalyysi-hajotelman**

$$SST = SSA + SSB + SSC + SSAB + SSAC + SSBC + SSABC + SSE$$

ja neliösummiin liittyvät vapausasteiden lukumäärät toteuttavat yhtälön

$$\begin{aligned} IJKL - 1 = {} & (J-1) + (K-1) + (L-1) \\ & + (J-1)(K-1) + (J-1)(L-1) + (K-1)(L-1) \\ & + (J-1)(K-1)(L-1) \\ & + JKL(I-1) \end{aligned}$$

Testisuureet ja niiden jakaumat

Määritellään ***F*-testisuure**

$$F_{ABC} = \frac{JKL(I-1)}{(J-1)(K-1)(L-1)} \cdot \frac{SSABC}{SSE}$$

jossa *SSABC* on tekijöiden *A* ja *B* ja *C yhdysvaikutusta kuvaava neliösumma* ja *SSE* on *ryhmien sisäistä vaihtelua kuvaava neliösumma*. *Jos nollahypoteesi*

H_{ABC} : Ei yhdysvaikutusta *ABC*

pätee, niin

$$F_{ABC} \quad F((J-1)(K-1)(L-1), JKL(I-1))$$

Suuret testisuureen F_{ABC} arvot johtavat nollahypoteesin hylkäämiseen.

Määritellään ***F*-testisuure**

$$F_{AB} = \frac{JKL(I-1)}{(J-1)(K-1)} \cdot \frac{SSAB}{SSE}$$

jossa *SSAB* on tekijöiden *A* ja *B yhdysvaikutusta kuvaava neliösumma* ja *SSE* on *ryhmien sisäistä vaihtelua kuvaava neliösumma*. *Jos nollahypoteesi*

H_{AB} : Ei yhdysvaikutusta *AB*

pätee, niin

$$F_{AB} \quad F((J-1)(K-1), JKL(I-1))$$

Suuret testisuureen F_{AB} arvot johtavat nollahypoteesin hylkäämiseen.

Määritellään ***F*-testisuure**

$$F_{AC} = \frac{JKL(I-1)}{(J-1)(L-1)} \cdot \frac{SSAC}{SSE}$$

jossa *SSAC* on tekijöiden *A* ja *C yhdysvaikutusta kuvaava neliösumma* ja *SSE* on *ryhmien sisäistä vaihtelua kuvaava neliösumma*. *Jos nollahypoteesi*

H_{AC} : Ei yhdysvaikutusta *AC*

pätee, niin

$$F_{AC} \quad F((J-1)(L-1), JKL(I-1))$$

Suuret testisuureen F_{AC} arvot johtavat nollahypoteesin hylkäämiseen.

Määritellään ***F*-testisuure**

$$F_{BC} = \frac{JKL(I-1)}{(K-1)(L-1)} \cdot \frac{SSBC}{SSE}$$

jossa *SSBC* on tekijöiden *B* ja *C yhdysvaikutusta kuvaava neliösumma* ja *SSE* on *ryhmien sisäistä vaihtelua kuvaava neliösumma*. *Jos nollahypoteesi*

H_{BC} : Ei yhdysvaikutusta *BC*

pätee, niin

$$F_{BC} \quad F((K-1)(L-1), JKL(I-1))$$

Suuret testisuureen F_{BC} arvot johtavat nollahypoteesin hylkäämiseen.

Määritellään ***F*-testisuure**

$$F_A = \frac{JKL(I-1)}{J-1} \cdot \frac{SSA}{SSE}$$

jossa *SSA* on tekijän *A päävaikutusta kuvaava neliösumma* ja *SSE* on *ryhmien sisäistä vaihtelua kuvaava neliösumma*. *Jos nollahypoteesi*

H_A : Ei päävaikutusta *A*

pätee, niin

$$F_A \quad F((J-1), JKL(I-1))$$

Suuret testisuureen F_A arvot johtavat nollahypoteesin hylkäämiseen.

Määritellään ***F*-testisuure**

$$F_B = \frac{JKL(I-1)}{K-1} \cdot \frac{SSB}{SSE}$$

jossa *SSB* on tekijän *B päävaikutusta kuvaava neliösumma* ja *SSE* on *ryhmien sisäistä vaihtelua kuvaava neliösumma*. *Jos nollahypoteesi*

H_B : Ei päävaikutusta *B*

pätee, niin

$$F_B \quad F((K-1), JKL(I-1))$$

Suuret testisuureen F_B arvot johtavat nollahypoteesin hylkäämiseen.

Määritellään ***F*-testisuure**

$$F_C = \frac{JKL(I-1)}{L-1} \cdot \frac{SSC}{SSE}$$

jossa *SSC* on tekijän *C päävaikutusta kuvaava neliösumma* ja *SSE* on *ryhmien sisäistä vaihtelua kuvaava neliösumma*. *Jos nollahypoteesi*

H_C : Ei päävaikutusta *C*

pätee, niin

$$F_C \quad F((L-1), JKL(I-1))$$

Suuret testisuureen F_C arvot johtavat nollahypoteesin hylkäämiseen.

Varianssianalyysitaulukko

Varianssianalyysin tulokset esitetään tavallisesti **varianssianalyysitaulukon** muodossa:

Vaihtelun lähde	***SS***	***df***	***MS = SS/df***	***F = MS/MSE***
A	SSA	$I - 1$	MSA	$F_A = MSA/MSE$
B	SSB	$J - 1$	MSB	$F_B = MSB/MSE$
C	SSC	$K - 1$	MSC	$F_C = MSC/MSE$
AB	$SSAB$	$(I - 1)(J - 1)$	$MSAB$	$F_{AB} = MSAB/MSE$
AC	$SSAC$	$(I - 1)(K - 1)$	$MSAC$	$F_{AC} = MSAC/MSE$
BC	$SSBC$	$(J - 1)(K - 1)$	$MSBC$	$F_{BC} = MSBC/MSE$
ABC	$SSABC$	$(I - 1)(J - 1)(K - 1)$	$MSABC$	$F_{ABC} = MSABC/MSE$
Jäännös	SSE	$IJK(L - 1)$	MSE	
Kokonais-vaihtelu	SST	$IJKL - 1$		

Varianssianalyysitaulukon *neliösummat* toteuttavat yhtälön

$$SST = SSA + SSB + SSC + SSAB + SSAC + SSBC + SSABC + SSE$$

Yhtälö on *varianssianalyysihajotelma*. Varianssianalyysitaulukon neliösummien *vapausasteet df* (*df* = degrees of freedom) toteuttavat yhtälön

$$\begin{aligned} IJKL - 1 = {} & (J - 1) + (K - 1) + (L - 1) \\ & + (J - 1)(K - 1) + (J - 1)(L - 1) + (K - 1)(L - 1) \\ & + (J - 1)(K - 1)(L - 1) \\ & + JKL(I - 1) \end{aligned}$$

22.4. Laskutoimitusten suorittaminen

Olkoon

y_{ijkl} = i. havainto tekijän A tason A_j, tekijän B tason B_k ja tekijän C tason C_l määräämässä ryhmässä (j, k, l)

$i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$, $k = 1, 2, \ldots, K$, $l = 1, 2, \ldots, L$

Määritellään seuraavat summat:

$$T_{jkl} = \sum_{i=1}^{I} y_{ijkl}$$

$$T_{j} = \sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{i=1}^{I} y_{ijkl}$$

$$T_{k} = \sum_{l=1}^{L}\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ijkl}$$

$$T_{l} = \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ijkl}$$

$$j = 1, 2, \text{K}, J,\ k = 1, 2, \text{K}, K,\ l = 1, 2, \text{K}, L$$

Määritellään edelleen myös seuraavat summat:

$$T_{jk} = \sum_{l=1}^{L}\sum_{i=1}^{I} y_{ijkl}$$

$$T_{jl} = \sum_{k=1}^{K}\sum_{i=1}^{I} y_{ijkl}$$

$$T_{kl} = \sum_{j=1}^{J}\sum_{i=1}^{I} y_{ijkl}$$

$$T = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L} y_{lijk}$$

$$j = 1, 2, \text{K}, J,\ k = 1, 2, \text{K}, K,\ l = 1, 2, \text{K}, L$$

Määritellään *tekijän A tason* A_j*, tekijän B tason* B_k ja *tekijän C tason* C_l *määräämän ryhmän* (j, k, l) *havainto-arvojen* y_{ijkl} *neliöiden summa* kaavalla

$$\sum_{i=1}^{I} y_{ijkl}^2,\ j = 1, 2, \text{K}, J,\ k = 1, 2, \text{K}, K,\ l = 1, 2, \text{K}, L$$

ja *kaikkien havaintoarvojen* y_{ijkl} *neliöiden kokonaissumma* kaavalla

$$\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} y_{ijkl}^2$$

Havaintoarvojen y_{ijkl} *ryhmävarianssit* saadaan kaavoilla

$$s^2_{jkl} = \frac{1}{I-1}\left(\sum_{i=1}^{I} y^2_{jkl} - \frac{1}{I}T^2_{jkl}\right),\ j = 1, 2, \text{K}, J,\ k = 1, 2, \text{K}, K,\ l = 1, 2, \text{K}, L$$

Havaintoarvojen y_{ijkl} *kokonaisvarianssi* saadaan kaavalla

$$s^2 = \frac{1}{IJKL-1}\left(\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} y^2_{ijkl} - \frac{1}{IJKL}T^2\right)$$

Kokonaisneliösumma SST voidaan laskea kaavalla

$$SST = \sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I} y^2_{ijkl} - \frac{1}{IJKL}T^2$$

Tekijöiden A, B ja *C päävaikutuksia kuvaavat neliösummat SSA, SSB* ja *SSC* saadaan kaavoilla

$$SSA = \frac{1}{IKL}\sum_{j=1}^{J} T_j^2 \; - \frac{1}{IJKL} T^2$$

$$SSB = \frac{1}{IJL}\sum_{k=1}^{K} T_k^2 \; - \frac{1}{IJKL} T^2$$

$$SSC = \frac{1}{IJK}\sum_{l=1}^{L} T_l^2 - \frac{1}{IJKL} T^2$$

Tekijöiden *A* ja *B*, *A* ja *C*, *B* ja *C yhdysvaikutuksia kuvaavat neliösumma SSAB, SSAC* ja *SSBC* saadaan kaavoilla

$$SSAB = \frac{1}{IL}\sum_{k=1}^{K}\sum_{j=1}^{J} T_{jk}^2 \; - \frac{1}{IJKL} T^2 \; - SSA - SSB$$

$$SSAC = \frac{1}{IK}\sum_{l=1}^{L}\sum_{j=1}^{J} T_{jl}^2 - \frac{1}{IJKL} T^2 \; - SSA - SSC$$

$$SSBC = \frac{1}{IJ}\sum_{l=1}^{L}\sum_{k=1}^{K} T_{kl}^2 - \frac{1}{IJKL} T^2 \; - SSB - SSC$$

Tekijöiden *A* ja *B* ja *C yhdysvaikutusta kuvaava neliö-summa SSABC* saadaan kaavalla

$$SSABC = SS - SSA - SSB - SSC - SSAB - SSAC - SSBC$$

jossa

$$SS = I\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{j=1}^{J}(\bar{y}_{jkl} - \bar{y}\;)^2 = \frac{1}{I}\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{j=1}^{J} T_{jkl}^2 - \frac{1}{IJKL} T^2$$

on *ryhmäkeskiarvojen kokonaisvaihtelua kuvaava neliösumma*.

Ryhmien sisäistä vaihtelua kuvaava jäännösneliösumma SSE saadaan *varianssianalyysihajotelman* nojalla kaavalla

$$SSE = SST - SSA - SSB - SSC - SSAB - SSAC - SSBC - SSABC$$

tai kaavalla

$$SSE = SST - SS$$

jossa *SS* on ryhmäkeskiarvojen kokonaisvaihtelua kuvaava neliösumma.