Szigma, XXVI. (1995) 1–2. 1

VARIANCIAFELBONTÁS: ELŐFELTEVÉSEK ÉS KÖVETKEZTETÉSEK[1]

HAJDU OTTÓ – HUNYADI LÁSZLÓ
BKE Statisztikai Tanszék

A varianciafelbontás a statisztika elméletének egyik legáltalánosabb, legtöbb területen alkalmazható összefüggését, nevezetesen a heterogén sokaságok varianciájának az ún. külső és a belső variancia összegére való bontását eredményezi. A varianciafelbontás tulajdonságait tankönyvek, szakkönyvek és tanulmányok sora (pl. [1,2,3,4,5,6,7]) vizsgálja, néhány fontos sokasági összefüggésre Hunyadi [4] cikke hívta fel a figyelmet. Jelen tanulmány a varianciafelbontás során előálló külső, belső és teljes eltérésnégyzetösszegek *mintavételi ingadozásában* rejlő, a mintavételi következtetések alapjául szolgáló törvényszerűségek, alkalmazási előfeltevések áttekinthető rendszerbe foglalását tűzi ki céljául.

Valahányszor egy ismeretlen sokaság jellegzetességeinek a vizsgálata áll érdeklődésünk homlokterében, és a teljes sokaság megfigyelése vagy lehetetlen, vagy fölösleges, következtetéseink során kénytelenek vagyunk egy mintavétel eredményeire hagyatkozni. Praktikus megközelítésből tehát értelmetlennek tűnik sokasági és mintabeli jellemzők egymás mellett való szerepeltetése, hiszen ismert sokaság mellett szükségtelen a mintavétel, ismeretlen sokaság esetén pedig csak mintaadatok állnak rendelkezésünkre. A következtetéseink minőségét azonban az alkalmazott mintavételi mód, továbbá a mintabeli információ feldolgozásának a módja alapvetően befolyásolja. Ha ismerjük e tényezők hatásmechanizmusát, s közben a sokaságról tökéletes információval rendelkezünk, akkor szembeállítva következtetéseinket és a valóságot, megítélhetővé válik, hogy milyen hatékonyan használtuk fel a mintavétel eredményeit. A valóságot természetesen nem ismerjük, viszont föltételezhetjük több, vagy kevesebb jellegzetességének az ismeretét, s ezek birtokában megadhatjuk magának a mintavételnek, illetve a mintabeli információk feldolgozásának azon módját, amely mellett következtetéseink várhatóan a legmegbízhatóbban fogják közelíteni az ismeretlen sokasági jellemzőket.

A fenti gondolatmenetet csoportosított (rétegzett) sokaságra, s így csoportosított mintára alkalmazva arra kívánunk rámutatni, hogy milyen mintavételi tervet kell készíteni, illetve milyen statisztikákat célszerű számolni a rétegzett sokaság minél megbízhatóbb jellemzése érdekében akkor, ha a sokaságról

[1] Beérkezett 1994. november 3.

több, vagy kevesebb ismeretünk van. A sokaságról alkotott ismereteink ún. előfeltevések formájában kerülnek megfogalmazásra. Mint általában a mintavételi következtetések, a varianciafelbontáson alapulók is kétirányúak, becslési, valamint hipotézisvizsgálati célúak lehetnek. Míg becslési feladatok esetén minden, a sokaságról alkotott előfeltevésünket egy minél jobb becslés érdekében mozgósítjuk, addig a hipotézisvizsgálat során maga a hipotézis is előfeltevéseink egyike, amely mellé más feltevések is társulhatnak.

Lévén a vizsgálandó sokaság *rétegzett,* teljeskörű leírása a szóbanforgó jelenség (változó) csoporton belüli jellemzőinek, és a csoportok egymáshoz való arányának ismeretét igényli. A sokaság megadása során e kérdés a vizsgált változó csoporton belüli eloszlásainak összehasonlítására irányul. Alapvető kérdés az, hogy a változó normális eloszlású-e vagy sem, a csoporton belüli varianciák egyenlőek-e vagy sem, továbbá, hogy a várható értékek különböznek-e, vagy sem. A tanulmány a varianciafelbontás mintavételi várható értékeinek általános formulákba öntése alapján néhány alapvető, széleskörűen használt modellnek a fenti kritériumokra, továbbá a mintaelemszám elosztására, és a becslő-, illetve tesztfüggvény megválasztására való érzékenységét vizsgálja.

Ennek érdekében elsőként a rétegzett sokaság ismeretét feltételező *általános* modellt definiáljuk, majd meghatározzuk a mintából számított eltérésnégyzetösszegek várható értékét ezen általános modell keretei között. Ezt követően megmutatjuk, hogy az egyenlő rétegvarianciákra és egyenlő rétegátlagokra, továbbá a mintaelemszám elosztására vonatkozó feltételek különkülön, vagy egyidejű figyelembevétele miként eredményezi az általános modell szűkülését, a különféle eltérésnégyzetösszegek várható értékei hogyan egyszerűsödnek, nyernek statisztikai tartalmat azáltal, hogy az általános modellt fokozatosan speciális modellekké redukáljuk. Végül azt tárgyaljuk, hogy a mintavételi várható értékek fokozatos egyszerűsödése hogyan szolgálja néhány közismert, a varianciafelbontás elvén alapuló becslési, illetve hipotézisvizsgálati eljárás működését, gyakorlati alkalmazhatóságát. A gyakorlati vonatkozásokat illetően figyelmünket – messze a teljesség igénye nélkül – az arányos rétegzésből végrehajtott becslések, illetve az egyszempontú, klasszikus varianciaanalízis elméletére koncentráljuk.

Modellfeltevések

Az alábbiakban egy *normális eloszlású, rétegzett* sokaságot leíró általános modellt adunk meg, amely modellt a következő jelölésrendszer foglal egységbe. Tekintsük a sokaságot, mely $j = 1, \ldots, m$ számú rétegre tagolódik. Az egyes rétegekre, illetve a sokaságra vonatkozóan az 1. táblázatban foglalt jellemzők

ismeretét feltételezzük:

1. táblázat: A sokaság leírása

Jellemző	Rétegek			Alap-
	1	j	m	sokaság
Változó	Y_1	Y_j	Y_m	Y
Általános egyed	Y_{i1}	Y_{ij}	Y_{im}	Y_i
Rétegarány	P_1	P_j	P_m	1
Várható érték	μ_1	μ_j	μ_m	μ
Variancia	σ_1^2	σ_j^2	σ_m^2	σ^2
Réteghatás	$\tau_1 = \mu_1 - \mu$	$\tau_j = \mu_j - \mu$	$\tau_m = \mu_m - \mu$	0

A táblázattal kapcsolatban megjegyzendő, hogy a rétegarányokat kifejező P_j értékek végtelen sokaság esetén a j-edik rétegbe kerülés valószínűségét jelentik, bár ezek a valószínűségek nem szükségképpen ismertek. A j-edik réteg varianciája definíció szerint[2]

$$\sigma_j^2 = E\big((Y_j - \mu_j)^2\big)\,, \tag{1}$$

a teljes sokaságé pedig

$$\sigma^2 = E\big((Y - \mu)^2\big)\,. \tag{2}$$

Az alábbiakban sorra vesszük a csoportosított sokaságra vonatkozó, a csoportok és a teljes sokaság viszonyát leíró mindazon összefüggéseket, amelyek ismerete a tanulmány mondanivalója szempontjából elengedhetetlen.

Rögzítsük a *centrális tendenciát*[3] kifejező, rétegen belüli várható értékeket rendre az

$$E(Y_j) = \mu_j \qquad (j = 1, \ldots, m) \tag{3}$$

szinteken. Jelölje a rétegen belüli várható értékek súlyozott számtani átlagát

$$\overline{\mu}(v) = \sum_{j=1}^{m} v_j \mu_j \tag{4}$$

a

$$\sum_{j=1}^{m} v_j = 1 \tag{5}$$

általános súlyrendszer felhasználásával. Ekkor a sokaságra értelmezett

$$\mu = \overline{\mu}(P) \tag{6}$$

[2]Az Y véletlen változó várható értékét a továbbiakban $E(Y)$ jelöli.

[3]Centrális tendencia alatt azt értjük, hogy egy sokaság egyedeinek túlnyomó többsége egy rögzitett számhoz közeli értékkel bir, akörül, és ahhoz közel ingadozik.

középérték is rögzített. Ebből következően a réteghatások tényleges rétegarányokkal súlyozott átlaga

$$\overline{\tau}(P) = 0 \,, \tag{7}$$

valamely feltételezett rétegarányokkal súlyozott átlaga azonban a konkrét súlyrendszer függvénye:

$$\overline{\tau}(v) = ?$$

Ebből következően a rétegátlagok súlyozott számtani átlaga csak akkor egyenlő a sokaság átlagával, ha súlyként az egzakt P_j rétegarányok állnak rendelkezésre, egyébként az egyenlőség nem áll fenn:[4]

$$\overline{\mu}(v) = \sum_{j=1}^{m} v_j \mu_j = \sum_{j=1}^{m} v_j(\mu + \tau_j) = \mu + \overline{\tau}(v) \neq \mu \,. \tag{8}$$

A rétegek szóródását a *külső szórásnégyzettel* jellemezzük. A külső varianciát a réteghatásoknak a rétegarányokkal súlyozott szórásnégyzete állítja elő. Mivel modellünk szerint a réteghatásoknak a rétegarányokkal súlyozott számtani átlaga zérus, ezért a külső szórásnégyzet egyben a réteghatásoknak a rétegarányokkal súlyozott négyzetes átlaga négyzetével is megegyezik:

$$\sigma_\tau^2(P) = \sum_{j=1}^{m} P_j \left(\tau_j - \overline{\tau}(P)\right)^2 = \sum_{j=1}^{m} P_j \tau_j^2 = \overline{\tau}_q^2(P) = \sigma_K^2 \,. \tag{9}$$

A külső szórásnégyzet természetesen csak abban az esetben zérus, ha valamennyi rétegátlag (rétegen belüli várható érték) egyenlő egymással, s így a sokaság átlagával (várható értékével). Ekkor valamennyi réteghatás zérus:

$$\sigma_K^2 = 0 \qquad \text{ha} \qquad \mu_1 = \mu_2 = \ldots = \mu_m = \mu \,, \tag{10}$$

vagyis

$$\sigma_K^2 = 0 \qquad \text{ha} \qquad \tau_1 = \tau_2 = \ldots = \tau_m = 0 \,. \tag{11}$$

A rétegek belső szóródását összefoglalóan az átlagos rétegen belüli szórásnégyzettel jellemezzük, amely két esetben egyezik meg a *belső szórásnégyzettel.* Egyrészt akkor, ha súlyként a rétegarányok rendelkezésünkre állnak:

$$\overline{\sigma}^2(P) = \sigma_B^2 \,, \tag{12}$$

[4] A rétegarányok ismerete nem irreális feltevés, még végtelen (megszámolhatóan végtelen) sokaság esetén sem. Gondoljunk pl. a budapesti napi mozibevételek sokaságára, melyet két rétegre, a hétközi és a hétvégi napi bevételekre bonthatunk. Itt a rétegarányok 5/7 és 2/7, a rétegátlagok jelentése pedig a napi átlagos hétközi és hétvégi bevétel.

másrészt akkor, ha valamennyi rétegen belül a szórásnégyzet megegyezik. Ekkor ugyanis a súlyoknak nincs befolyásuk az átlagos értékre:

$$\overline{\sigma}^2(v) = \sigma_B^2 \qquad \text{ha} \qquad \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_m^2 = \sigma_e^2 \,. \tag{13}$$

A varianciafelbontás ismert eredményeképpen pedig a teljes variancia a külső és a belső variancia összege:

$$\sigma^2 = \sigma_K^2 + \sigma_B^2 \,. \tag{14}$$

Mintavételi várható értékek

A rétegzett sokaságra vonatkozó következtetésünket – legyen annak célja akár valamely jellemző becslése, akár valamely hipotézis vizsgálata – a mintabeli eltérésnégyzetösszeg dekompozíciójára, s a komponensek várható értékeire alapozzuk. Ezért az alábbiakban elsőként a külső, belső és teljes eltérésnégyzetösszegek várható értékeinek általános formuláit adjuk meg, amelyek a vizsgált feltevések mellett leegyszerűsödnek, ezáltal érdemi, a tanulmány központi mondanivalóját jelentő következtetések levonását teszik lehetővé.

Vegyünk rétegenként *függetlenül,* rendre $n_1, \ldots, n_m$ elemű, rétegen belül független és azonos eloszlású (FAE) mintákat, amelyekre

$$\sum_{j=1}^{m} n_j = n, \qquad \text{és} \qquad w_j = n_j/n \,. \tag{15}$$

A mintaelemszám rétegek közötti w_j megoszlását *mintaelosztásnak* nevezzük, s ez feltevésünk szerint nincs mintavételi ingadozásnak kitéve. Ekkor a mintabeli csoportátlagok rendre

$$\overline{y}_1, \ldots, \overline{y}_m \,, \tag{16}$$

az n elemű (teljes) minta átlaga

$$\overline{y}(w) = \overline{y} \,, \tag{17}$$

a mintabeli korrigált szórásnégyzetek pedig rendre

$$s_j^2 = \sum_{j=1}^{n_j} (y_{ij} - \overline{y}_j)^2/(n_j - 1) \qquad (j = 1, \ldots, m) \tag{18}$$

alakúak lesznek. Tekintsük ezután a mintabeli teljes eltérésnégyzetösszeg külső és belső összetevőkre bontását:

$$SS = SS_K + SS_B \tag{19}$$

ahol

$$SS = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij} - \overline{y})^2 , \tag{20}$$

$$SS_K = \sum_{j=1}^{m} n_j(\overline{y}_j - \overline{y})^2 , \tag{21}$$

$$SS_B = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2 . \tag{22}$$

A mintavétel következtében mind a teljes, mind a belső, mind a külső eltérésnégyzetösszeg mintavételi ingadozásnak van kitéve. Az egyes eltérésnégyzetösszegek mintavételi várható értékét az alábbi általános formulák[5] szolgáltatják:

$$E(SS) = (n-1)\overline{\sigma}^2(w) + \sum_{j=1}^{m} n_j\left(\tau_j - \overline{\tau}(w)\right)^2 = (n-1)\overline{\sigma}^2(w) + n\sigma_\tau^2(w) \tag{23}$$

$$E(SS_B) = E(\sum_{j=1}^{m}(n_j - 1)s_j^2) = \sum_{j=1}^{m}(n_j - 1)\sigma_j^2 = \sum_{j=1}^{m} n_j(\sigma_j^2 - \sigma_j^2/n_j) \tag{24}$$

$$E(SS_K) = \sum_{j=1}^{m}(1 - w_j)\sigma_j^2 + \sum_{j=1}^{m} n_j\left(\tau_j - \overline{\tau}(w)\right)^2 = \sum_{j=1}^{m}(1 - w_j)\sigma_j^2 + n\sigma_\tau^2(w) . \tag{25}$$

Látható, hogy a fenti várható értékek az alábbi tényezők függvényei:

1. mintanagyság,
2. mintaelosztás,
3. rétegvarianciák,
4. átlagos rétegvariancia,
5. réteghatások varianciája.

Vegyük észre, hogy mind az átlagos rétegvariancia, mind a réteghatások varianciája súlyozottan értendő, ahol súlyrendszerként a (w) mintaelosztás szerepel. Nevezetes esetekben e várható értékek egyszerűbb alakot öltenek. E nevezetes esetek a következők:

[5] A várható értékek alábbi formuláinak a levezetését az Olvasó az I. Függelék (F.1)-(F.8) azonosságai alapján ellenőrizheti.

Homoszkedaszticitás: valamennyi rétegvariancia egyenlő egymással, s így rétegarányoktól függetlenül a közös variancia egyben a belső varianciát is jelenti:

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_m^2 = \sigma_e^2 = \sigma_B^2 \ .$$

Korrelálatlanság:[6] valamennyi rétegátlag egyenlő egymással, s így a sokaság átlagával, amiből következően valamennyi réteghatás, s ezért a külső variancia is zérus:

$$\tau_1 = \tau_2 = \ldots = \tau_m = 0 = \sigma_K^2 \ .$$

Arányosság: a teljes n mintaelemszám rétegek közötti szétosztása a tényleges rétegarányoknak megfelelően történik:

$$w_j = P_j, \qquad \text{vagyis} \qquad \overline{\tau}(w) = 0 \ .$$

Ha a fenti három tulajdonság külön-külön, vagy valamilyen kombinációban egyidejűleg teljesül, akkor a (23), (24) és (25) eltérésnégyzetek várható értékei a következők szerint alakulnak:

Arányosság esetén:

$$E(SS) = (n-1)\sigma_B^2 + \sum_{j=1}^{m} n_j \tau_j^2 = (n-1)\sigma_B^2 + n\overline{\tau}_q^2(P) = (n-1)\sigma_B^2 + n\sigma_K^2 \quad (26)$$

$$E(SS_B) = \sum_{j=1}^{m}(n_j - 1)\sigma_j^2 = n\overline{\sigma}^2(P) - \sum_{j=1}^{m}\sigma_j^2 = n\sigma_B^2 - \sum_{j=1}^{m}\sigma_j^2 \quad (27)$$

$$E(SS_K) = \sum_{j=1}^{m}(1-P_j)\sigma_j^2 + \sum_{j=1}^{m} n_j\tau_j^2 = \sum_{j=1}^{m}(1-P_j)\sigma_j^2 + n\overline{\tau}_q^2(P) = \sum_{j=1}^{m}\sigma_j^2 - \sigma_B^2 + n\sigma_K^2 \ . \quad (28)$$

Korrelálatlanság esetén:

$$E(SS) = (n-1)\overline{\sigma}^2(w) \quad (29)$$

$$E(SS_B) = n\overline{\sigma}^2(w) - \sum_{j=1}^{m}\sigma_j^2 \quad (30)$$

[6] A tanulmányban korrelálatlanság alatt azt értjük, hogy a rétegátlagok egyenlők egymással, vagyis a csoportképző (rétegképző) változó nincs sztochasztikus kapcsolatban a vizsgált Y mennyiségi változóval. Megjegyzendő, hogy nominális skálán mért rétegképző ismérv esetén a sztochasztikus kapcsolatnak ezt a típusát a magyar statisztikai terminológia nem korrelációnak, hanem vegyes kapcsolatnak nevezi.

$$E(SS_K) = \sum_{j=1}^{m} \sigma_j^2 - \overline{\sigma}^2(w) \,. \tag{31}$$

Homoszkedaszticitás esetén:

$$E(SS) = (n-1)\sigma_e^2 + n\sigma_\tau^2(w) \tag{32}$$

$$E(SS_B) = (n-m)\sigma_e^2 \tag{33}$$

$$E(SS_K) = (m-1)\sigma_e^2 + n\sigma_\tau^2(w) \,. \tag{34}$$

Arányosság és homoszkedaszticitás esetén:

$$E(SS) = (n-1)\sigma_e^2 + n\overline{\tau}_q^2(P) = (n-1)\sigma_e^2 + n\sigma_K^2 \tag{35}$$

$$E(SS_B) = (n-m)\sigma_e^2 \tag{36}$$

$$E(SS_K) = (m-1)\sigma_e^2 + n\overline{\tau}_q^2(P) = (m-1)\sigma_e^2 + n\sigma_K^2 \,. \tag{37}$$

Arányosság és korrelálatlanság esetén:

$$E(SS) = (n-1)\sigma_B^2 \tag{38}$$

$$E(SS_B) = n\sigma_B^2 - \sum_{j=1}^{m} \sigma_j^2 \tag{39}$$

$$E(SS_K) = \sum_{j=1}^{m} \sigma_j^2 - \sigma_B^2 \,. \tag{40}$$

Korrelálatlanság és homoszkedaszticitás esetén:

$$E(SS) = (n-1)\sigma_e^2, \quad E(SS_B) = (n-m)\sigma_e^2, \quad E(SS_K) = (m-1)\sigma_e^2 \,. \tag{41}$$

A fenti várható értékeket azért kell ismernünk, mert segítségükkel tudjuk megválaszolni azt a kérdést, hogy a sokaság milyen jellegzetességei mellett mely statisztika becsüli torzítatlanul a sokaság külső, belső, illetve teljes varianciáját. A továbbiakban sorra vesszük mindazon statisztikák torzítatlansági[7] tulajdonságait, amelyek számítását a mintabeli eltérés-négyzetösszeg dekompozíciója teszi lehetővé, majd megmutatjuk, hogy ezek a tulajdonságok mit jelentenek néhány fontos gyakorlati alkalmazás szemszögéből.

[7]Egy becslőfüggvény torzitatlan, ha mintavételi várható értéke megegyezik a becsülni kívánt sokasági jellemzővel.

Torzítatlansági következmények

A teljes eltérésnégyzetösszeg tulajdonságai

A rétegenkénti független mintavétel alkalmazását feltéve, általában SS semmilyen transzformációja nem alkalmas a sokaság valamely varianciájának torzítatlan becslésére. Még az $SS/(n-1)$ statisztika[8] sem, mivel (23) alapján

$$E\Big(\frac{SS}{n-1}\Big) = \overline{\sigma}^2(w) + \frac{n}{n-1}\sigma_\tau^2(w) \neq \sigma^2 \ . \tag{42}$$

Speciálisan azonban, arányos rétegzés és korrelálatlanság egyidejű teljesülése esetén az $SS/(n-1)$ statisztika (38) szerint alkalmas a belső, s egyben a teljes variancia torzítatlan becslésére:

$$E\Big(\frac{SS}{n-1}\Big) = \sigma_B^2 = \sigma^2 \ . \tag{43}$$

Ebből is látható, hogy korrelálatlanság esetén értelmetlen a rétegek megkülönböztetése.

A belső eltérésnégyzetösszeg tulajdonságai

1. A teljes mintaelemszámmal való osztás útján nyert SS_B/n statisztika általában nem alkalmas a σ_B^2 belső variancia torzítatlan becslésére, mivel (24) figyelembevételével:

$$E\Big(\frac{SS_B}{n}\Big) = \sum_{j=1}^{m} w_j\sigma_j^2 - \sum_{j=1}^{m}\frac{\sigma_j^2}{n} = \overline{\sigma}^2(w) - \sum_{j=1}^{m}\frac{w_j\sigma_j^2}{n_j} \neq \sigma_B^2 \ . \tag{44}$$

Arányos rétegzés, vagy homoszkedaszticitás esetén azonban $\overline{\sigma}^2(w) = \sigma_B^2$, tehát a lefelé torzítás mértéke meghatározható:

$$\sum_{j=1}^{m}\frac{w_j\sigma_j^2}{n_j} \ , \tag{45}$$

amely arányosság esetén[9]

$$\sum_{j=1}^{m} P_j\sigma_{\overline{y}_j}^2 \ , \tag{46}$$

[8] Az $SS/(n-1)$ statisztika az n elemű minta egyszerű véletlen módon való kiválasztása mellett nyújtana torzítatlan becslést σ^2-re.

[9] $\sigma_{\overline{y}_j}^2$ a j-edik réteg mintaátlagának mintavételi varianciáját jelöli.

homoszkedaszticitás esetén pedig

$$\frac{m}{n}\sigma_e^2 \tag{47}$$

lesz.

2. Általánosságban az $SS_B/(n-m)$ statisztika sem becsli torzítatlanul a belső varianciát, hiszen (24) alapján:

$$E\Big(\frac{SS_B}{n-m}\Big) = E\Big(\sum_{j=1}^{m}\frac{(n_j-1)s_j^2}{n-m}\Big) = \sum_{j=1}^{m}\frac{(n_j-1)\sigma_j^2}{n-m} \neq \sigma_B^2 \,. \tag{48}$$

Homoszkedaszticitás teljesülésekor azonban torzítatlan statisztika, mivel (33)-at tekintve:

$$E\Big(\frac{SS_B}{n-m}\Big) = \sigma_e^2 = \sigma_B^2 \,. \tag{49}$$

A külső eltérésnégyzetösszeg vizsgálata

A homoszkedaszticitás és a korrelálatlanság együttes teljesülésekor az $SS_K/(m-1)$ statisztika a közös variancia, s így egyben a belső és a teljes variancia torzítatlan becslésére alkalmas, (41) alapján ugyanis:

$$E\Big(\frac{SS_K}{m-1}\Big) = \sigma_e^2 = \sigma_B^2 = \sigma^2 \,. \tag{50}$$

Ha azonban csupán a homoszkedaszticitás érvényesülését feltételezzük, akkor (34) szerint:

$$E\Big(\frac{SS_K}{m-1}\Big) = \sigma_e^2 + \frac{n}{m-1}\sigma_\tau^2(w) \,. \tag{51}$$

Az (50) és (51) várható értékek viszonylatában a

$$\sigma_e^2 \leq \sigma_e^2 + \frac{n}{m-1}\sigma_\tau^2(w) \tag{52}$$

reláció mindig teljesül, viszont w megválasztásának a függvényében $\sigma_\tau^2(w)$ is változik.

Alkalmazások

A belső variancia becslése

Mivel (49) alapján az $SS_B/(n-m)$ statisztika várható értéke csak homoszkedaszticitás esetén egyezik meg a belső varianciával, ezért valahányszor a

belső variancia torzítatlan becslése a célunk és a homoszkedaszticitás teljesülését semmi nem támasztja alá, akkor a belső variancia becslésére másik becslőfüggvényt kell keresnünk. Tekintsük az

$$\overline{s}^2(w) = \sum_{j=1}^{m} w_j s_j^2$$

statisztikát, amelynek várható értéke minden esetben

$$E\big(\overline{s}^2(w)\big) = \overline{\sigma}^2(w) \, .$$

Mivel $\overline{\sigma}^2(w) = \sigma_B^2$ mind arányos rétegzés, mind homoszkedaszticitás esetén teljesül, ezért $\overline{s}^2(w)$ mindkét esetben a belső variancia torzítatlan becslését nyújtja. A belső variancia nem homoszkedasztikus körülmények közötti torzítatlan becslésének igénye tipikusan a *klasszikus rétegzett mintavételen alapuló becslések* készítésekor merül fel.

Ezzel szemben az

$$s_p^2 = \frac{SS_B}{n-m} = \sum_{j=1}^{m} \frac{(n_j - 1)s_j^2}{n-m}$$

formulával definiált ún. „pooled" mintabeli variancia

$$E(s_p^2) = \sum_{j=1}^{m} \frac{(n_j - 1)\sigma_j^2}{n-m}$$

várható értéke *általában* nem egyezik meg a belső varianciával, ezért s_p^2 kizárólag homoszkedaszticitás esetén lesz a közös σ_e^2, s így a belső variancia torzítatlan becslőfüggvénye. Ezek a feltételek pedig tipikusan a varianciaanalízisnek, és kétmintás esetének, a kétmintás "pooled" t-tesztnek az alkalmazási feltételei.

Ugyanakkor belátható,[10] hogy a mintabeli korrigált s_j^2 varianciákat más és más v_j súlyrendszerekkel átlagolva, $\overline{s}^2(v)$ minimumát az s_p^2 pooled variancia szolgáltatja.

Becslés független részmintákból

Független részmintákból való becslés esetén az n elemű teljes mintát véletlenszerűen és függetlenül m számú egyenlő részre, mondhatni rétegre bontjuk úgy, hogy

$$n_j = n_e \qquad \text{és} \qquad n = mn_e$$

[10] Speciális esetre lásd a II. Függelék azonosságait.

teljesüljön. A rétegenkénti egyenlő mintaelemszám alkalmazását a

$$P_j = \frac{1}{m}$$

feltevés elfogadásának megfelelő arányos rétegzés indokolja, míg a rétegek független és véletlen kialakítása – a rétegképzés és a vizsgált jelenség korrelálatlan voltának biztosításával – a külső variancia zérus értékének a feltételezését teszi reálissá.

Feltételezve tehát a homoszkedasztikus korrelálatlanság és az arányosság együttes teljesülését, a sokaság μ várható értékének torzítatlan becslőfüggvénye független részminták esetén

$$\overline{y}_{FRM} = \sum_{j=1}^{m} \frac{\overline{y}_j}{m} ,$$

amelynek elméleti varianciája

$$\mathrm{Var}\,(\overline{y}_{FRM}) = \frac{\sigma_e^2}{n} .$$

Ebből következően viszont látható, hogy független részminták alkalmazása mellett az átlagbecslés varianciájának becslésére használt

$$\frac{\sum_{j=1}^{m}(\overline{y}_j - \overline{y}_{FRM})^2}{m(m-1)}$$

statisztika torzítatlanul becsli az elméleti varianciát, mivel homoszkedasztikus korrelálatlanság esetén (41) szerint

$$E\Big(\frac{SS_K}{n(m-1)}\Big) = \frac{\sigma_e^2}{n} = E\Big(\frac{n_e \sum_{j=1}^{m}(\overline{y}_j - \overline{y}_{FRM})^2}{n_e m(m-1)}\Big) = E\Big(\frac{\sum_{j=1}^{m}(\overline{y}_j - \overline{y}_{FRM})^2}{m(m-1)}\Big).$$

A varianciaanalízis teszt[11] ereje

A varianciaanalízis teszt statisztikájának egyik működési alapelve azon közismert tétel, miszerint a szabadságfokkal osztott (korrigált) külső eltérésnégyzetösszeg várható értéke homoszkedaszticitás esetén nem kisebb, mint a korrigált belső eltérésnégyzetösszeg várható értéke, az egyenlőség pedig csak a

[11] A varianciaanalizis nullhipotézise szerint $H_0 : \tau_1 = \tau_2 = \ldots \tau_m$, amely tesztelésének – normális eloszlású csoportok esetében – a homoszkedaszticitás teljesülése alkalmazási előfeltétele.

homoszkedasztikus korrelálatlanság esetén áll fenn, mikor a kérdéses várható értékek éppen a közös σ_e^2 varianciával egyenlők:

$$E\Big(\frac{SS_K}{m-1} \mid \sigma_1^2 = \ldots = \sigma_m^2, \tau_1 = \ldots = \tau_m\Big) = \sigma_e^2 = E\Big(\frac{SS_B}{n-m}\Big) \leq$$
$$E\Big(\frac{SS_K}{m-1} \mid \sigma_1^2 = \ldots = \sigma_m^2\Big) = \sigma_e^2 + \frac{n}{m-1}\sigma_\tau^2(w)\,.$$

Minél nagyobb tehát a réteghatások szóródása, s így annak $\sigma_\tau^2(w)$ mértéke, annál magasabb a korrigált külső eltérésnégyzetösszeg várható értéke a varianciaanalízis alternatív hipotézisének helyessége esetén, vagyis annál nagyobb e próba szelektivitása, azaz a próba ereje. Mivel azonban a w mintaelosztás megfelelő megválasztásával $\sigma_\tau^2(w)$ befolyásolható, így a varianciaanalízis ereje növelhető, de csökkenthető is. Ugyanakkor viszont a τ_j réteghatások a gyakorlati számítások esetén eredendően ismeretlenek, tehát nem tudjuk megmondani egzaktan azt a w súlyrendszert, amely $\sigma_\tau^2(w)$-t a sokasági réteghatások mellett maximálja. E maximálás során tehát w nevezetes eseteit tudjuk csak kezelni. A réteghatások varianciájának

$$\sigma_\tau^2(w) = \overline{\tau}_q^2(w) - \overline{\tau}^2(w)$$

formuláját tekintve, ilyen feltevés lehet a mintaelosztás tekintetében az *arányosság* esete. Ekkor ugyanis a réteghatások $\overline{\tau}(P)$ átlagának zérus az értéke, s ezért a varianciaanalízis teszt erejének viszonylagos nagysága ez esetben a

$$\sigma_\tau^2(w) \leq \overline{\tau}_q^2(P)$$

reláció teljesülésének, vagy nem teljesülésének a kérdése, amely a

$$\overline{\tau}_q^2(w) - \overline{\tau}_q^2(P) \leq \overline{\tau}^2(w) \tag{54}$$

formában is írható. Amennyiben tehát (54) fennáll, úgy a varianciaanalízis ereje *arányosság* esetén nagyobb, mint nem arányos mintaelosztás mellett.[12] A varianciaanalízis erejét a "pooled" variancia minimum tulajdonságának a szempontjából vizsgálva továbbá az is látható, hogy mivel s_p^2 a varianciaanalízis F próbafüggvényének a nevezőjében szerepel,[13] ezért a *pooled* variancia használata is az ANOVA erejének a növelését szolgálja.

[12] Az (54) alatti reláció pl. abban az esetben biztosan teljesül, ha a réteghatásoknak a mintaelosztással súlyozott négyzetes átlaga kisebb, mint a rétegarányokkal súlyozott, hiszen ez esetben (54) baloldala negatív, jobboldala viszont mindig pozitív.

[13] $F = \big(SS_K/(m-1)\big)/\big(SS_B/(n-m)\big)$.

A mintaelemszám elosztásának kérdése

Számos statisztikaelméleti irodalom (pl. [1, 5]) a varianciaanalízis modellfeltevéseinek a tárgyalásakor a

$$\overline{\tau}(w) = \sum_{j=1}^{m} w_j \mu_j - \mu = 0$$

feltétel rögzítéséből indul ki. Mivel azonban modellünk szerint a réteghatások előre rögzítettek, és $\overline{\mu}(P) = \mu$, ezért $\overline{\tau}(w) = 0$ előre történő rögzítése csak az alábbi esetekben indokolt:

- egyrészt, ha valamennyi réteghatás zérus;
- másrészt, különböző réteghatások melletti arányosan rétegzett mintavétel mellett;
- harmadrészt, ha egy végtelen sokaság rétegarányai nem ismertek, de azokat egyenlőnek feltételezve, rétegenként azonos elemszámú mintákat veszünk.

Összefoglalás, következtetések

A tanulmányban a heterogén sokaságokat jellemző általános modellt vizsgáltuk. Általános volt a modell abból a szempontból, hogy feltételeit olyan tágra szabtuk, hogy abba mind a varianciaanalízis, mind a rétegzett mintavétel, mind a részmintás becslések belefértek, ugyanakkor szűk volt a modell abban az értelemben, hogy csak egyetlen ismérv (változó) szerint csoportosítottuk a sokaságot.

Módszerünk az volt, hogy a mintából számított négyzetösszegek várható értékét tekintettük, és bemutattuk azok egyszerűsödési lehetőségét háromféle feltétel mellett.

Főbb következtetéseink az alábbiak voltak:

a) kimutattuk, hogy arányosan rétegzett mintavétel esetén a mintasúlyokkal átlagolt korrigált mintavariancia ad torzítatlan becslést a belső varianciára;

b) megmutattuk, hogy megfelelő feltételek mellett a pooled variancia torzítatlan, és minimális varianciájú becslést ad a közös (belső) varianciára;

c) beláttuk, hogy a szokásos feltételek mellett a független részminták módszere az átlagbecslés varianciájára torzítatlan becslőfüggvényt ad;

d) megvilágítottuk, hogy a varianciaanalízisnél szokásos $\overline{\tau}(w) = 0$ feltevés valójában az egyenlő sokasági rétegarányok melletti arányos mintavétel hallgatólagos feltételezésével azonos, végül

e) kimutattuk, hogy a mintaelosztás változtatásával a varianciaanalízis F-tesztjének szelektivitása (és ereje) növelhető, és általános esetben ez azt jelenti, hogy az egyenletes mintaelosztásból való következtetés javítja a teszt erejét.

Függelék

I. A mintabeli eltérésnégyzetösszegek mintavételi várható értékének a meghatározásához szükséges azonosságok

Irjuk fel a mintabeli teljes, belső és külső eltérésnégyzetösszegeket az alábbi formákban, kihasználva a tényt, hogy az eltérésnégyzetösszeg a mintaelemek konstanssal való eltolására invariáns.

A teljes eltérésnégyzetösszeg, valamennyi mintaelemnek ugyanazon μ konstanssal való eltolása (csökkentése) nyomán:

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

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij} - \mu)^2 - n(\overline{y} - \mu)^2 = \sum_{j=1}^{m}\sum_{i=1}^{n_j}\big((y_{ij} - \mu_j) + (\mu_j - \mu)\big)^2 - n(\overline{y} - \mu)^2 =$$

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij} - \mu_j)^2 + \sum_{j=1}^{m} n_j(\mu_j - \mu)^2 + 2\sum_{j=1}^{m}(\mu_j - \mu)\sum_{i=1}^{n_j}(y_{ij} - \mu_j) - n(\overline{y} - \mu)^2 \tag{F.1}$$

A belső eltérésnégyzetösszeg, a mintaelemeknek csoporton belüli, rendre μ_j $(j = 1, \ldots, m)$ konstanssal való csökkentése után:

$$SS_B = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2 = \sum_{j=1}^{m}\Big(\sum_{i=1}^{n_j}(y_{ij} - \mu_j)^2 - n_j(\overline{y}_j - \mu_j)^2\Big) = \tag{F.2}$$

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

Végül a külső eltérésnégyzetösszeg egyszerű kivonással:

$$SS_K = \sum_{j=1}^{m} n_j(\overline{y}_j - \mu_j)^2 + \sum_{j=1}^{m} n_j(\mu_j - \mu)^2 + 2\sum_{j=1}^{m}(\mu_j - \mu)\sum_{i=1}^{n_j}(y_{ij} - \mu_j) - n(\overline{y} - \mu)^2 \tag{F.3}$$

Tekintsük továbbá a következő nevezetes várható értékeket:

$$E_j(y_{ij} - \mu_j) = 0\,, \tag{F.4}$$

$$E\big((\overline{y}_j - \mu_j)^2\big) = \mathrm{Var}\,(\overline{y}_j) = \frac{\sigma_j^2}{n_j}\,, \tag{F.5}$$

$$E(\overline{y}) = E(\sum_{j=1}^{m} w_j\overline{y}_j) = \sum_{j=1}^{m} w_j\mu_j\,. \tag{F.6}$$

Az (F.4), (F.5) és (F.6) azonosságok felhasználásával

$$E\Big[\big(\overline{y} - E(\overline{y})\big)^2\Big] = \mathrm{Var}\,(\overline{y}) = \sum_{j=1}^{m} w_j^2\mathrm{Var}\,(\overline{y}_j) = \sum_{j=1}^{m}\frac{w_j^2\sigma_j^2}{n_j} = \frac{1}{n}\sum_{j=1}^{m} w_j\sigma_j^2 = \frac{\overline{\sigma}^2(w)}{n}\,, \tag{F.7}$$

majd (F.7) figyelembevételével

$$E\big((\overline{y} - \mu)^2\big) = E\big((\overline{y} - E(\overline{y}) + E(\overline{y}) - \mu)^2\big) = \mathrm{Var}\,(\overline{y}) + (E(\overline{y}) - \mu)^2 = \frac{\overline{\sigma}^2(w)}{n} + \Big(\sum_{j=1}^{m}\frac{n_j\mu_j}{n} - \mu\Big)^2 = \frac{\overline{\sigma}^2(w)}{n} + \frac{\Big(\sum_{j=1}^{m} n_j\tau_j\Big)^2}{n^2} = \frac{\overline{\sigma}^2(w)}{n} + \overline{\tau}^2(w) \tag{F.8}$$

A fenti várható értékeknek az (F.1), (F.2), (F.3) formulákba helyettesítésével nyerjük a (23), (24) és (25) alatti általános várható érték képleteket.

II. A mintabeli „pooled" variancia minimum tulajdonsága

Általában belátható, hogy ha $\hat{\theta}_1$ és $\hat{\theta}_2$ torzítatlan és független becslőfüggvények θ-ra, akkor

a) $\alpha\hat{\theta}_1 + (1-\alpha)\hat{\theta}_2 = \hat{\theta}_3$ is torzítatlan, és

b) $\mathrm{Var}\,(\hat{\theta}_3)$ akkor lesz minimális, ha

$$\alpha = \frac{\mathrm{Var}\,(\hat{\theta}_2)}{\mathrm{Var}\,(\hat{\theta}_1) + \mathrm{Var}\,(\hat{\theta}_2)}\,, \tag{F.9}$$

ami belátható, ha felírjuk a

$$\mathrm{Var}(\hat{\theta}_3) = \alpha^2 \mathrm{Var}(\hat{\theta}_1) + (1-\alpha)^2 \mathrm{Var}(\hat{\theta}_2)$$

egyenletet, és megkeressük a minimumát α függvényében.

Speciálisan két, normális eloszlású réteg, és rétegenkénti független mintavétel esetén legyen $\theta = \sigma_e^2$, $\hat{\theta}_1 = s_1^2$, $\hat{\theta}_2 = s_2^2$. Ekkor mindkét becslőfüggvény torzítatlan, varianciájuk pedig

$$\mathrm{Var}(s_1^2) = \frac{2\sigma_e^4}{n_1 - 1}$$

és

$$\mathrm{Var}(s_2^2) = \frac{2\sigma_e^4}{n_2 - 1}\,.$$

Az (F.9) eredmény figyelembevételével esetünkben

$$\alpha = \frac{n_1 - 1}{n_1 + n_2 - 2}\,,$$

amely felhasználásával az

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

"pooled" variancia valóban torzítatlan és minimális varianciájú. Ez az eredmény kettőnél több rétegre is azonnal általánosítható.

Irodalom

1. Canavos, C. G.: Applied Probability and Statistical Methods. Little Brown and Co., Boston, 1984.
2. Dunn, O. J.–Clark, V. A.: Applied Statistics: Analysis of Variance and Regression. (2nd Edition). John Wiley & Sons, Inc., New York, 1987. item Greene, W. H.: Econometric Analysis (2nd Edition). McMillan P. Co., New York, 1993.
3. Hunyadi L.: A varianciafelbontásról. Statisztikai Szemle, (70), 1992. 1037–1047. old.
4. Hunyadi L.–Vita L.: Statisztika 1. AULA Kiadó, Budapest, 1991.
5. Hunyadi L.–Mundruczó Gy.–Vita L.: Statisztika II. Aula Kiadó, Budapest, 1992.
6. Meszéna Gy.–Ziermann M.: Valószínűségelmélet és matematikai statisztika. Közgazdasági és Jogi Könyvkiadó, Budapest, 1981.

VARIANCE DECOMPOSITION: ASSUMPTIONS AND CONCLUSIONS

In the paper a general model of the variance decomposition of heterogeneous populations is investigated. The aim of the study is to show how the commonly used, classical statistical procedures can be derived from the general model by introducing different assumptions and restrictions. The detailed analysis of the expectations of the sums of squares under different restrictions revealed the common root of the given methods (ANOVA, two sample t test, estimation from stratified samples, estimation by means of the independent subsamples), helped us to understand them better. Besides, the paper promotes the correct and efficient applications of the above mentioned methods.