ELMÉLETI KIEGÉSZÍTÉS A 7. FEJEZETHEZ

Ennek a kiegészítésnek az a célja, hogy egyes, a tankönyvben csak érintőlegesen szereplő módszereket, eljárásokat bemutasson, illetve az azokhoz fűzött állítások hátterét, elméleti alapjait megmutassa. Az itt leírtak nem képezik a tananyag, illetve a vizsgaanyag részét, de aki választ akar kapni az egyes módszerek kapcsán a miért? kérdésre, aki szeretne mélyebben megismerkedni a tananyag egyes részeivel, vagy egyáltalában szeretne egy kicsit többet tudni a statisztika elméletéről annak érdemes az itt következő néhány oldalt áttanulmányozni!

Tartalom:

1. Az MSE mutató származtatása

Az MSE a variancia olyan általánosítása, ahol az eltéréseket nem a $\hat{\theta}$ becslések várható értékétől, hanem az elméleti θ -tól mérjük. Így az MSE definiciója:
$MSE\left(\hat{\theta}\right)=E\left(\hat{\theta}-\theta\right)^2$, míg a varianciáé: $Var\left(\hat{\theta}\right)=E\left(\hat{\theta}-E\left(\hat{\theta}\right)\right)^2$.
A kettő formális összefüggése könnyen megmutatható:

$$MSE\left(\hat{\theta}\right)=E\left(\hat{\theta}-\theta\right)^2=E\left[\hat{\theta}-E\left(\hat{\theta}\right)+E\left(\hat{\theta}\right)-\theta\right]^2=E\left[\left(\hat{\theta}-E\left(\hat{\theta}\right)\right)+\left(E\left(\hat{\theta}\right)-\theta\right)\right]^2=$$
$$=Var\left(\hat{\theta}\right)+2E\left[\left(\hat{\theta}-E\left(\hat{\theta}\right)\right)\left(E\left(\hat{\theta}\right)-\theta\right)\right]+E\left[E\left(\hat{\theta}\right)-\theta\right]^2$$

Mivel a második tag utolsó tényezője azonosan egyenlő 0-val, az egész második tag eltűnik, a harmadik pedig nem más, mint a torzítás négyzete, hiszen a szögletes zárójelben nincs változó, ezért a várható értéke önmaga. Innen a végeredmény:

$$\boxed{MSE\left(\hat{\theta}\right)=Var\left(\hat{\theta}\right)+Bs^2\left(\hat{\theta}\right)}.$$

E feladat kapcsán érdemes emlékeztetni a könyvben szereplő „lőtáblára”: a kis szóródás esetenként kompenzálhatja a torzítást!

[1] Egyutasnak is szokás nevezni.

2. A sokasági szórásnégyzet torzítatlan becslése FAE mintából

Legyen $s^{*2}=\dfrac{\sum(y_i-\bar{y})^2}{n}$ és első lépésben belátandó, hogy ez torzítottan becsüli σ^2-et, pontosabban $E(s^{*2})=\sigma^2-\dfrac{\sigma^2}{n}$.

Először tekintsük a számláló várható értékét! Ehhez alakítsuk át a számlálót a következő módon:

$$\sum(y_i-\bar{y})^2=\sum(y_i-\mu+\mu-\bar{y})^2=\sum((y_i-\mu)-(\bar{y}-\mu))^2=$$
$$=\sum\left[(y_i-\mu)^2-2(y_i-\mu)(\bar{y}-\mu)+(\bar{y}-\mu)^2\right]=\sum(y_i-\mu)^2-2(\bar{y}-\mu)\sum(y_i-\mu)+n(\bar{y}-\mu)^2=$$
$$=\sum(y_i-\mu)^2-n(\bar{y}-\mu)^2.$$

A várható értékre áttérve:

$E(y_i-\mu)^2=\sigma^2$ a sokasági variancia definíciójából és $E(\bar{y}-\mu)^2=\dfrac{\sigma^2}{n}$ a mintaátlag varianciájának korábban származtatott összefüggése miatt. A várható érték összegtulajdonságából adódóan $E\sum(y_i-\mu)^2=\sum E(y_i-\mu)^2=n\sigma^2$, és $\sum(y_i-\mu)^2-n(\bar{y}-\mu)^2=n\sigma^2-\sigma^2$, innen pedig

$$E\left(\frac{\sum(y_i-\bar{y})^2}{n}\right)=\sigma^2-\frac{\sigma^2}{n}=\left(\frac{n-1}{n}\right)\sigma^2.$$

A torzítás könnyen eltüntethető, hiszen a korrigált szórásnégyzet már torzítatlan lesz:

$E\left(\dfrac{n}{n-1}s^{*2}\right)=E\left(\dfrac{\sum(y_i-\bar{y})^2}{n-1}\right)=E(s^2)=\sigma^2$, ahol tehát

$$\boxed{s^2=\frac{\sum(y_i-\bar{y})}{n-1}}$$

a korrigált tapasztalati (mintából számított) variancia. Ez FAE minta esetén torzítatlanul becsüli a sokasági varianciát. Megjegyezzük, hogy a számoló- és számítógépes eljárások világosan megkülönböztetik a kétféle szóródási mutatót. A gyakran használt Excel például a függvények közt SZÓRÁS néven az s, SZÓRÁSP néven az s^* mutatót számítja!

3. Becslés a momentumok módszerével

A momentumok módszere korábban igen népszerű volt eloszlások paramétereinek becslésére, elsősorban kényelmes kezelhetősége folytán. Mára a számítástechnikai eszközök fejlődése más irányt adott a becsléseknek, így a momentumok módszere eredeti formájában már némiképp avittnak számít, ám általánosított változata egyre népszerűbbé válik. A következőkben az eredeti módszer alkalmazására mutatunk be két példát.

3.1. Exponenciális eloszlás paraméterbecslése (1 paraméter)

Az eloszlás sűrűségfüggvénye: $f(y)=\dfrac{1}{\theta}\exp\left(-\dfrac{y}{\theta}\right)$, várható értéke pedig $E(Y)=\theta$.

Egy paraméter esetén egyetlen egyenlet szükséges, ami kimondja, hogy olyan paramétert becsülünk, amely esetén a sokaság és a minta első momentumai megegyeznek. Itt az első sokasági momentum a várható érték, az első mintamomentum pedig a mintaátlag. Innen azonnal adódik a becslés:

$$\boxed{\hat{\theta}_{mom} = \bar{y}}.$$

3.2. Folytonos egyenletes eloszlás paraméterbecslése (2 paraméter)

Az eloszlás sűrűségfüggvénye az $[\alpha, \beta]$ intervallumban $f(y) = \frac{1}{\beta - \alpha}$, másutt 0.

Az eloszlás első momentuma és második centrális momentuma a következő:
$E(Y) = \frac{\alpha + \beta}{2}$, illetve $Var(Y) = \frac{(\beta - \alpha)^2}{12}$. (Ezek valószínűség számításból ismert eredmények.)

A megfelelő két mintabeli momentum a mintaátlag $(\bar{y})$ és a minta varianciája (s^{*2}). Ezért a „momentum-egyenletek”:

$$\bar{y} = \frac{\hat{\alpha} + \hat{\beta}}{2} \quad \text{és} \quad s^{*2} = \frac{(\hat{\beta} - \hat{\alpha})^2}{12}.$$

Ezt a két egyenletet megoldva a következőket kapjuk:

$$\boxed{\hat{\alpha}_{mom} = \bar{y} - s^* \sqrt{3} \quad \text{és} \quad \hat{\beta}_{mom} = \bar{y} + s^{*2} \sqrt{3}}$$

4. Normális eloszlás paramétereinek becslése FAE mintából a maximum likelihood módszerrel

A normális eloszlású Y változó sűrűségfüggvénye: $f(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2} \right)$, s mivel minden y_i mintaelem ilyen eloszlású, és függetlenek, a likelihood függvény, az n elemű minta bekövetkezését leíró sűrűségfüggvény ezek szorzatával lesz arányos, azaz

$$L(y_1, y_2, \ldots, y_n \mid \mu, \sigma^2) = c \cdot \prod_{i=1}^{n} f(y_i) = \left(\frac{1}{\sqrt{2\pi}} \right)^n \left(\frac{1}{\sigma^2} \right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n} (y_i - \mu)^2}{2\sigma^2} \right).$$

Ennek maximumát keresve célszerű logaritmálni (hiszen a monoton logaritmus transzformáció nem változtatja meg a szélsőérték helyét). Ekkor az ún. log-likelihood függvény

$$\log L = \log c - n \log \sqrt{2\pi} - \frac{n}{2} \log \sigma^2 - \frac{\sum (y_i - \mu)^2}{2\sigma^2}.$$

A maximum likelihood módszer azokat a $\hat{\mu}$ és $\hat{\sigma}^2$ paramétereket keresi, amelyek mellett ez a függvény felveszi maximumát. A μ szerinti parciális derivált:

$\frac{\partial \log L}{\partial \mu} = \frac{\sum (y_i - \mu)}{n}$, és ezt 0-ra megoldva azt kapjuk, hogy

$$\boxed{\hat{\mu} = \bar{y}}$$

mellett lehet szélsőértéke.

Hasonlóan elkészítve a σ^2 szerinti parciális deriváltat:

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum (y_i - \mu)^2}{2\sigma^4}.$$

Ezt 0-ra megoldva, majd μ-be annak becsült értékét helyettesítve azt kapjuk, hogy

$$\boxed{\hat{\sigma}^2 = \frac{\sum (y_i - \bar{y})^2}{n} = s^{*2}}.$$

Könnyen belátható, hogy ilyen paraméterértékek mellett a likelihood függvénynek maximuma van, azaz ilyen paraméterek esetén a leginkább hihető, hogy az adott minta ebből a sokaságból származik.

5. A véges korrekciós faktor származtatása

Mint arra utaltunk, az EV mintából való becsléskor a FAE minták alapján számított standard hibák szorzódnak egy véges korrekciós faktorral, ami $f = \frac{N-n}{N-1} \cong 1 - \frac{n}{N}$ alakú. Ennek származtatása a következő:

a) $Var(\bar{y}_{EV}) = Var\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) = \frac{1}{n^2} Var\left(\sum y_i\right) = \frac{1}{n^2}\left[nVar(y_i) + \sum_{i \neq j}\sum_{j \neq i} Cov(y_i, y_j)\right]$

Az utolsó tagban $Var(y_i) = \sigma^2$, a kovarianciákat pedig az alábbiak szerint számíthatjuk:

b) Általában igaz, hogy

$$Cov(y_i, y_j) = E(y_i y_j) - E(y_i)E(y_j).$$

Ebben a kifejezésben $E(y_i) = E(y_j) = \bar{Y}$, így az utolsó tag értéke $\bar{Y}^2$. Kiszámítandó az együttes várható érték:

$$E(y_i y_j) = \frac{\sum_{i \neq j}\sum_{j \neq i} Y_i Y_j}{N(N-1)}.$$

Két általános, ismert összefüggés:

- $\sigma^2 = \frac{\sum Y_i^2}{N} - \bar{Y}^2$ és ezt átrendezve $\frac{\sum Y_i^2}{N} = \bar{Y}^2 + \sigma^2$

- $\left(\sum Y_i\right)^2 = \sum Y_i^2 + \sum_{i\neq j}\sum_{j\neq i} Y_i Y_j$ és ezt átrendezve $\sum_{i\neq j}\sum_{j\neq i} Y_i Y_j = \left(\sum Y_i\right)^2 - \sum Y_i^2$ és végül mindkét oldalt $N(N-1)$-gyel osztva:

$$\frac{\sum_{i\neq j}\sum_{j\neq i} Y_i Y_j}{N(N-1)} = \frac{\left(\sum Y_i\right)^2}{N(N-1)} - \frac{\sum Y_i^2}{N(N-1)} = \frac{\overline{Y}^2 N}{N-1} - \left(\frac{\sigma^2 + \overline{Y}^2}{N-1}\right).$$

- A jobb oldalt tovább alakítva

$$\frac{\overline{Y}^2 N}{N-1} - \left(\frac{\sigma^2 + \overline{Y}^2}{N-1}\right) = \frac{1}{N-1}\left[(N-1)\overline{Y}^2 - \sigma^2\right] = \overline{Y}^2 - \frac{\sigma^2}{N-1}.$$

d) Visszaírva ezt b)-be azt kapjuk, hogy

$$Cov(y_i, y_j) = E(y_i y_j) - E(y_i)E(y_j) = \overline{Y}^2 - \frac{\sigma^2}{N-1} - \overline{Y}^2 = -\frac{\sigma^2}{N-1}.$$

A kovariancia értelmezésekor 3 momentumra hívja fel a figyelmet:

- A kovariancia a sokaság nagyságával fordítottan arányos, azaz nagy sokaságok esetén (abszolút értékben) kicsi;
- A kovariancia előjele negatív, mert véges sokaságban egy „nagy” elem kiválasztása esetén másodiknak várhatóan kis elem kerül a mintába;
- A kovariancia nem függ *i*-től se *j*-től, azaz állandó.

Ez utóbbi tulajdonságát kihasználva a kovariancia visszaírható a)-ba:

$$Var(\overline{y}_{EV}) = \frac{1}{n^2}\left[nVar(y_i) + \sum_{i\neq j}\sum_{j\neq i} Cov(y_i, y_j)\right] = \frac{1}{n^2}\left[n\sigma^2 - n(n-1)\frac{\sigma^2}{N-1}\right] = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

Ha a sokaság elég nagy, azaz N és $N-1$ közti különbség (a hányadosban) elhanyagolható, akkor

$$Var(\overline{y}_{EV}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \cong \frac{\sigma^2}{n}\left(\frac{N-n}{N}\right) = \frac{\sigma^2}{n}\left(1-\frac{n}{N}\right) = Var(\overline{y}_{FAE})\cdot f$$

6. Az átlagbecslés varianciája arányosan rétegzett EV minta esetén

Az átlagbecslés varianciája rétegzett EV minta esetén: $Var(\overline{y}_R) \cong \sum_j W_j^2 \frac{\sigma_j^2}{n}\left(1-\frac{n_j}{N_j}\right)$, ahol $W_j = \frac{N_j}{N}$. Arányos rétegzés esetén $\frac{n_j}{N_j} = \frac{n}{N}$ és $\frac{n_j}{n} = \frac{N_j}{N}$ és ezeket beírva a variancia képletébe:

$$Var(\bar{y}_{AR}) = \sum_j \frac{N_j}{N}\frac{n_j}{n}\frac{\sigma_j^2}{n_j}\left(1-\frac{n}{N}\right) = \frac{1}{n}\left(1-\frac{n}{N}\right)\sum_j \frac{N_j}{N}\sigma_j^2 = \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{\sum_j N_j\sigma_j^2}{N}$$, és mivel a heterogén sokaságokról tanultak alapján $\sigma_B^2 = \frac{\sum_j N_j\sigma_j^2}{N}$, így a keresett variancia:

$$\boxed{Var(\bar{y}_{AR}) = \frac{\sigma_B^2}{n}\left(1-\frac{n}{N}\right)}.$$

Megjegyzendő, hogy a mintavételek elméletével foglalkozó szakirodalom a sokasági szórásnégyzet $\left(\sigma^2 = \frac{\sum(Y-\bar{Y})^2}{N}\right)$ helyett inkább a korrigált sokasági szórásnégyzettel $\left(S^2 = \frac{\sum(Y-\bar{Y})^2}{N-1}\right)$ dolgozik, amely használata esetén az eredmények részben áttekinthetőbbek, részben pontosabbak lesznek. Mi ebben a tananyagban az eredeti, jobban értelmezhető szóródásmutatót használjuk következetesen, vállalva az apró, a gyakorlat számára többnyire elhanyagolható pontatlanságot.

7. A p -érték közelítő meghatározása táblázatok segítségével

Ha adott a próbafüggvény *eloszlásfüggvényének* táblázata, akkor a táblázat segítségével szinte pontosan[2] meghatározható a p-érték. Ha azonban csak a próbafüggvény eloszlásának *bizonyos kvantiliseit* tartalmazó táblázat áll rendelkezésre, mint például a t-eloszlás esetében, akkor a *p*-érték rendszerint csak a $p_a < p < p_f$ módon közelíthető. Ebben a felírásban p_a és p_f a próbafüggvény adott mintára vonatkozó konkrét értékét közrefogó két táblázatbeli kvantilis rendszáma, ami nem tévesztendő össze a *p*-értékkel. (Egyes esetekben a táblázat alapján csak $p < p_f$ alakú megállapítás tehető.) Erre mutatunk be az alábbiakban néhány, a gyakorlást megkönnyítő példát.

7.1. Egy Z próbafüggvény valamely mintából nyert értéke 2,15. Határozzuk meg a hozzá tartozó egy- és kétoldali p -értéket!

Ekkor az egyoldali p -érték 0,0158, mert $\Phi(2{,}15) = 0{,}9842 = P(Z \leq 2{,}15)$, a kétoldali pedig $2 \times 0{,}0158 = 0{,}0316$.

7.2. Egy Z próbafüggvény valamely mintából nyert értéke –3,84. Határozzuk meg a hozzá tartozó egy- és kétoldali p -értéket!

[2] A pontosság ilyenkor csak a táblázat részletezettségétől függ.

Ekkor csak annyi állítható az egyoldali p-értékről, hogy $p < 0{,}0002$, mert az F1.1. táblabeli legnagyobb z-érték esetében a függvényérték 0,9998. Ez a z-érték azonban abszolút értékben nagyobb 3,84-nél. A kétoldali p-érték biztosan kisebb 0,0004-nél.

7.3. Egy T próbafüggvény valamely mintából nyert értéke 1,5. Határozzuk meg a hozzá tartozó egy- és kétoldali p-értéket, ha $\nu = 14$!

Az F1.2. táblázat adott szabadságfokhoz tartozó sora szerint $1{,}34 < 1{,}5 < 1{,}36$, s a tartomány két határához 5, illetve 10 %-os jobb oldali szignifikanciaszint tartozik. Ezért az egyoldali p-érték alsó határa 0,05, felső határa 0,10. Ennek megfelelően a kétoldali p-értékre $0{,}10 < p < 0{,}20$ teljesül.

7.4. A χ^2-próba a *variancia* nagyságára vonatkozó előírás teljesülésének vizsgálatára irányult. A próbafüggvény valamely 15 elemű mintából nyert értéke 6,0. Határozzuk meg a hozzá tartozó egy- és kétoldali p-értéket!

Az F1.3. táblázat 14 szabadságfokhoz tartozó sora szerint a próbafüggvény mintából nyert értékét a 2,5 és 5 %-os szignifikanciaszinthez tartozó baloldali kritikus értékek fogják közre. Ezért az egyoldali p-értékre $0{,}025 < p < 0{,}05$, a kétoldalira pedig $0{,}05 < p < 0{,}10$ áll fenn.

7.5. Egy χ^2-próbával végzett függetlenségvizsgálat esetén a próbafüggvény értéke valamely 4 sorból és 3 oszlopból álló kontingencia-táblázat alapján 6,0. Mekkora a hozzá tartozó p-érték?

Ebben az esetben $\nu = 3 \times 2 = 6$. Az F1.3. táblázat e szabadságfokhoz tartozó sora szerint a próbafüggvény mintából nyert értékét a 25 és 50 %-os szignifikanciaszinthez tartozó kritikus értékek fogják közre, amelyek a hipotézisvizsgálat tárgya miatt csakis jobboldali kritikus értékeknek tekintendők. Ezért $0{,}25 < p < 0{,}50$, ami meglehetősen tág intervallum.

7.6. Egy varianciaanalízist elvégezve az $F = 5{,}8$ eredményt kaptuk. A szabadságfokok $\nu_1 = 7$ és $\nu_2 = 10$. Mekkora a megfelelő p-érték?

A jelen esetben nyilván a jobboldali p-értékre adható alsó és felső határt keressük. Ehhez a rendelkezésre álló minden F-táblázat adott szabadságfok-párhoz tartozó celláját végig kell néznünk. Az F1.4. táblázatok alapján megállapítható, hogy

Kvantilis rendszám	c_f
0,9	2,41
0,95	3,14
0,975	3,95
0,99	5,20
0,995	6,30

Ennek alapján a keresett p-érték határai $0{,}005 < p < 0{,}010$.

Amennyiben ugyanez a próbafüggvény érték egy varianciák összehasonlítására irányuló kétoldali hipotézisvizsgálat eredménye lenne, ahol 8 a nagyobb, 11 a kisebb becsült varianciát adó minta nagysága, akkor az előbb kapott határok kétszerese adna becslést a p-értékre.

8. A másodfajú hiba valószínűségének számítására szolgáló képlet a 7.22. példa esetében

A β meghatározására használható általános számítási formula

$$\beta(\mu_1) = P\left(\mu_0 - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \bar{y} < \mu_0 + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \mid \mu = \mu_1\right) =$$

$$= \Phi\left(\frac{\mu_0 - \mu_1 + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}}{\frac{\sigma}{\sqrt{n}}}\right) - \Phi\left(\frac{\mu_0 - \mu_1 - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}}{\frac{\sigma}{\sqrt{n}}}\right) = \Phi(\Delta + z_{1-\frac{\alpha}{2}}) - \Phi(\Delta - z_{1-\frac{\alpha}{2}}),$$

ahol $\Delta = \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}$. A $\beta(\mu_1)$ jelöléssel azt hangsúlyozzuk, hogy a másodfajú hiba elkövetési valószínűsége függ a H_0-lal szemben ténylegesen fennállónak gondolt egyszerű alternatívában szereplő μ_1 várható értéktől. A képlet Excellel igen egyszerűen használható. Ezt demonstrálja a 7.22. példa Excel-munkalapja.

9. A fix hatású egyszempontos[3] varianciaanalízis modellje

A tananyagban a varianciaanalízisnek csak az alkalmazás szempontjából leglényegesebb pontjait emeltük ki, holott a látszólag egyszerű eredmények mögött nagyon finoman kimunkált feltételek, meggondolások és modellek húzódnak. Ezért az itt következőkben bemutatjuk az eredmények hátterét, igaz, csak a legegyszerűbb, fix hatású egyutas variancia-analízis esetében.

A varianciaanalízis abból a *modellből* indul ki, hogy a *j*-edik sokaságból származó *i*-edik mintaelem

$$y_{ij} = (\mu + \beta_j) + \varepsilon_{ij} = \mu_j + \varepsilon_{ij} \tag{1}$$

alakú, ahol μ és β_j ismeretlen paraméterek, ε_{ij} pedig *i* és *j* bármely értéke mellett egy $N(0,\sigma^2)$ eloszlású valószínűségi változó. Az (1) modell szerint minden megfigyelés a következő három komponens összege:

- egy, a vizsgált sokaságokból összetevődő fősokaságra jellemző μ várható érték;
- egy, csak a *j-edik* sokaságra jellemző β_j konstans;
- egy véletlen összetevő.

[3] Egyutasnak is szokás nevezni.

Az első két komponens $\mu + \beta_j$ összege így valójában a j-edik sokaságra jellemző várható érték, amit μ_j-vel is szokás jelölni. Ez azt jelenti, hogy $\beta_j = \mu_j - \mu$, azaz a j-edik sokaság és a fősokaság várható értékének eltérése. A harmadik komponens az Y változónak a j-edik sokaságra jellemző μ_j várható érték körüli véletlenszerű ingadozását fejezi ki. Ettől válik sztochasztikussá Y és a vizsgált sokaságokat egymástól megkülönböztető ismérv kapcsolata, ha ez a kapcsolat egyáltalán fennáll.

Mivel ez az ingadozás mindazon tényezők együttes hatásának tulajdonítható, melyek *nem* játszottak szerepet a kezelési sokaságok (vagy részsokaságok) egymástól való elkülönítésében, az ε_{ij} valószínűségi változót (kísérleti) *hibaváltozónak*, ε_{ij} valamely konkrét értékét pedig (kísérleti) *hibának* nevezzük.[4] A varianciaanalízis (1) modelljét *fix hatású egyszempontos (egyutas)* varianciaanalízisnek szokás nevezni. A fix hatású jelző arra utal, hogy a modellünkben szereplő μ és β_j nem valószínűségi változó, hanem rögzített érték, azaz paraméter. Az egyszempontos jelző ezzel szemben azt jelzi, hogy a (kezelési) sokaságok csak egy olyan tényező – ún. faktor – szempontjából különböznek, ami befolyást gyakorol Y alakulására. A varianciaanalízisnek az (1) modellen kívül még számos más, a most tárgyalásra kerülőnél jóval bonyolultabb modellje ismeretes. Ezekkel gyakran *kísérlettervezési* modellek néven lehet találkozni a szakirodalomban.

A varianciaanalízist eredetileg kifejezetten a kontrollált kísérletekből származó mérési eredmények értékelésére dolgozták ki, s ez még ma is rányomja bélyegét a vele kapcsolatos terminológiára. Így például gyakran lehet találkozni azzal a kifejezéssel, hogy a vizsgált sokaságok egységei valamilyen tényező (faktor) különböző „szintjeinek” megfelelő „kezelésben” részesültek. Ugyancsak a kísérlettervezés tipikus terminológiája a „kezelési hatás” és „kísérleti hiba” elnevezés is. A j-edik fajta kezelés hatásának a $\beta_j = \mu_j - \mu$ paramétert szokás nevezni, kísérleti hibának pedig mindazon hatások eredőjét, melyek a kísérlet végzője által nem kontrollált tényezőknek tulajdonítható. A módszer a kontrollált kísérletek eredményeinek értékelése mellett természetesen minden olyan esetben használható, amikor előre meghatározott részsokaságokból származó, egymástól független minták alapján áll csak módunkban a részsokaságok várható értékeinek egyezésére vagy éppen eltérésére vonatkozó feltevésünk helyességét ellenőrizni. Ilyenkor a „kísérleti hiba” mindazon hatások eredője, melyek a részsokaságok megkülönböztetésére használt *ismérv(ek)en kívül* befolyással vannak még Y nagyságára.

Az (1) modell alapján a tankönyvben megfogalmazott nullhipotézis felírható a

$$\boxed{H_0\colon \beta_1 = \beta_2 = \ldots = \beta_M = 0}$$

módon is. Az ezzel szembeállítható alternatív hipotézis természetesen az, hogy a β_j paraméterek *nem mindegyike* nulla. (Természetesen ez is ekvivalens a korábbi H_0-lal szembeállított alternatív hipotézissel.)

[4] Az (1) modellben szereplő ε_{ij} hibaváltozó minden konkrét megfigyelés esetében az $N(0,\sigma^2)$ eloszlás valamely lehetséges értékét veszi fel. Mivel az egyes minták FAE minták és függetlenek is egymástól, az (1) modellben szereplő ε_{ij} $(i = 1,2,\ldots,n_j,\ j = 1,2,\ldots,M)$ valószínűségi változók függetlenek egymástól.

Az (1) modell szerint a *j*-edik sokaságból származó n_j elemű minta $\bar{y}_j$ átlagának várható értéke

$$E(\bar{y}_j) = E\left[\frac{1}{n_j}\sum_{i=1}^{n_j}(\mu+\beta_j+\varepsilon_{ij})\right] = \mu+\beta_j\,, \tag{2}$$

hiszen $E(\varepsilon_{ij}) = 0$ minden *i* és *j* értékre, μ és β_j pedig nem valószínűségi változók, hanem ismeretlen fix paraméterek.

Az (1) modell és annak feltevései alapján könnyen belátható, hogy

$$E(\bar{y}) = E\left[\frac{1}{n}\sum_{j=1}^{M} n_j\bar{y}_j\right] = \frac{1}{n}\sum_{j=1}^{M} n_j E(\bar{y}_j) = \frac{1}{n}\sum_{j=1}^{M} n_j(\mu+\beta_j) = \mu + \frac{\sum_{j=1}^{M} n_j\beta_j}{n}\,,$$

ahol felhasználtuk a (2)-t. Eszerint főminta $\bar{y}$ átlaga csak akkor lehet torzítatlan becslése μ-nek, ha

$$\sum_{j=1}^{M} n_j\beta_j = 0\,.$$

Nyilvánvaló, hogy H_0 fennállása esetén ez a feltétel automatikusan teljesül.

Az is bizonyítható, hogy

$$E(s_b^2) = \sigma^2,\ \text{akár igaz } H_0\text{, akár nem,}$$

és

$$E(s_k^2) = \begin{cases} \sigma^2, & \text{ha } H_0 \text{ igaz,} \\ \sigma^2 + \dfrac{1}{M-1}\sum_{j=1}^{M} n_j\beta_j^2, & \text{ha } H_1 \text{ igaz} \end{cases}.$$

Ha tehát H_0 igaz, s_k^2 is, s_b^2 is torzítatlanul becsli σ^2-et, aminek következtében az *F* próbafüggvény 1 körül ingadozik. Ha viszont H_0 nem igaz, s_k^2 *felülbecsli* σ^2-et, ami *F*-et 1-nél *nagyobbá* teszi. Ezért a próba *jobb oldali* kritikus tartománnyal hajtandó végre.