8. KISZÓRÓ PONTOK

8.1. A probléma felvetése

Az 1. fejezetben röviden, a 7.1. alfejezetben pedig részletesebben is beszéltünk a *kiszóró pontok*ról. A jelen fejezetben megtárgyaljuk azonosításuk módszereit, további az akkor követendő eljárást, amikor mérési adataink között kiszóró pontokat találtunk. Mindenekelőtt megadjuk definíciójukat:

8.1. DEFINÍCIÓ. Egy ξ_i mért értéket kiszóró pontnak nevezünk, ha várható értéke – valamilyen ismeretlen okból – nem egyezik meg az illesztőfüggvénnyel:

$$\mathrm{M}(\xi_i) \neq f(x_i, \mathbf{a}). \tag{8.1}$$

Itt fontos kitétel az *ismeretlen ok*ra való utalás, ugyanis a (8.1) reláció *ismert* okból való fennállását külön, a 9. fejezetben vizsgáljuk. Az ismert ok többnyire abban keresendő, hogy az illesztőfüggvényt szeretnénk egyszerűsíteni, és ennek érdekében vállaljuk, hogy (8.1) az x_i változó bizonyos tartományaiban fennálljon. Ezzel szemben a kiszóró pontok rendszertelen és izolált pontokban jelennek meg.

Komoly érdekünk fűződik a kiszóró pontok megtalálásához és kihagyásához. Ha ugyanis bennmaradnak a kiértékelt adatok között, a keresett paraméterekre torzított becslést kapunk, akármilyen módszerrel hajtjuk is végre az illesztést. Van azonban egy ezzel ellentétes érdek is, hiszen nem szabad olyan mérési adatot kihagynunk, amely nem kiszóró pont. Ha ugyanis ilyeneket nagy számban elhagyunk, az empirikus szórásokra túlságosan kis értékek jönnek ki, vagyis a mérésünk pontosabbnak fog látszani, mint amilyen a valóságban. Meg kell találnunk a két szempont között az optimális egyensúlyt.

Egy kiszóró pont megjelenésének több oka lehet: a műszerek hibás beállítása, rossz kalibráció, a mért adatok hibás regisztrációja, az illesztő programba táplált bemenő adatok hibás volta stb. A tapasztalat azt mutatja, hogy az alább ismertetett módszerrel azonosított kiszóró pontok eredetét nagyon gyakran utólag meg tudjuk találni, és az adatokat ennek megfelelően ki tudjuk javítani. Az igazi probléma akkor merül fel, amikor ez nem sikerül. Ilyenkor kerül a kísérlet kiértékelője nehéz döntés elé: mit tegyen az ismeretlen eredetű kiszóró ponttal vagy pontokkal?

Nyilván a kiszóró pontok azonosítása csak valamilyen statisztikai próbával lehetséges (vö. F3.3. alfejezet), tehát a fent említett döntéseket csak valami-

lyen konfidenciaszinten lehet meghozni. Látni fogjuk, hogy két független próbára van szükség: az elsővel azonosítjuk a kiszóró pontokat, a másikkal pedig eldöntjük, hogy ezek valóban kiszóró pontok-e vagy sem.

A (8.1) reláció fennállását úgy tudjuk ellenőrizni, hogy összehasonlítjuk a mért ξ_i értéket az

$$\tilde{y}_i = f\left(x_i, \tilde{\mathbf{a}}\right) \tag{8.2}$$

illesztett értékkel. Tulajdonképpen ezt tesszük a *4.2. ábrák*on, amelyeken a mért pontokra rárajzoltuk a lineáris regresszióval kapott egyenest. Nyilván azok a pontok lehetnek (esetleg) kiszóró pontok, amelyekre a $\left(\xi_i - \tilde{y}_i\right)$ különbségek túlságosan nagyok. A matematikai statisztikában a "nagy" vagy "kicsi" jelzőknek csak akkor van értelmük, ha ezeket a különbségeket a szórásukhoz viszonyítjuk. Ezt fogjuk tenni a következő alfejezetben.

8.2. Általánosított Student-próba

A próba definíciója

Bevezetünk néhány jelölést. A ξ_i mért érték szórásnégyzete

$$\mathrm{D}^2\left(\xi_i\right) = \sigma_{\xi i}^2 = \sigma^2\left[\mathbf{W}^{-1}\right]_{ii}. \tag{8.3a}$$

Az illesztett érték szórásnégyzete (6.19b) szerint

$$\mathrm{D}^2\left(\tilde{y}_i\right) = \sigma_i^{*2} = \sigma^2\left[\mathbf{F}\mathbf{R}^{-1}\mathbf{F}^{\mathrm{T}}\right]_{ii}. \tag{8.3b}$$

A **W**, **F** és **R** mátrixokat a (6.9) és (6.10) képletekben definiáltuk. Az illesztett és a mért érték között erős korreláció van, ezért különbségük szórásnégyzete szórásnégyzeteik különbségével egyenlő [vö. (6.23b)], ha a szóban forgó mért értéket az illesztésben felhasználtuk. Ellenkező esetben a ξ_i mért érték és illesztett értéke statisztikailag független, tehát különbségük szórásnégyzete szórásnégyzeteik összegével egyenlő. Ezt a két esetet a

$$\mathrm{D}^2\left(\xi_i - \tilde{y}_i\right) = \sigma_{\xi i}^2 \pm \sigma_i^{*2} \tag{8.3c}$$

képletben foglaljuk össze. Rövidesen érthetővé válik, miért hagyhatunk ki egy ξ_i mért értéket az illesztésből. Mindenesetre bevezetjük a következő elnevezéseket:

- *külső pont*: olyan ξ_i mért érték, amely nincs figyelembe véve az illesztésben;
- *belső pont*: olyan ξ_i mért érték, amely figyelembe van véve az illesztésben.

Ezekkel az elnevezésekkel a (8.3c) képletben a + előjel külső pontnak, a – előjel pedig belső pontnak felel meg. Ezek után definiáljuk a

$$\zeta_i = \frac{\xi_i - \tilde{y}_i}{\sqrt{\sigma_{\xi i}^2 \pm \sigma_i^{*2}}} \tag{8.4}$$

Gauss-eloszlású valószínűségi változót, amelyre nyilván fennállnak az

$$\mathrm{M}(\zeta_i) = 0 \qquad \text{és} \qquad \mathrm{D}^2(\zeta_i) = 1$$

összefüggések, ha az *i* index *nem* kiszóró pontnak felel meg. Ha σ^2-et a (6.22) képlettel becsüljük, akkor

$$t_i = \frac{\xi_i - \tilde{y}_i}{\sqrt{\frac{Q_{\min}}{n-m}\left[\mathbf{W}^{-1} \pm \mathbf{F}\mathbf{R}^{-1}\mathbf{F}^{\mathrm{T}}\right]_{ii}}} \tag{8.5}$$

($n - m$) szabadsági fokú Student-eloszlást követ a külső pontokra, ami nem áll a belső pontokra, hiszen ezekre a nevező nem független a számlálótól.

Az ($n - m$) szabadsági fokú Student-eloszlásra a következő kvantilist definiáljuk:

$$\mathrm{P}\{|t_i| < \gamma\} = 1 - \varepsilon\,. \tag{8.6}$$

Ha valamelyik külső pontra $|t_i| > \gamma$, akkor ξ_i kiszóró pontnak minősül ε konfidenciaszinten. Ezzel kitűzött célunkat elértük – legalábbis a külső pontok esetében. De mi legyen a belső pontokkal? Ezek esetében ugyanis a (8.5) szerint definiált t_i nem Student-tört.

A belső pontokra úgy írhatunk fel statisztikai próbát, hogy külső ponttá alakítjuk, vagyis

- kihagyjuk az *i*-edik belső pontot az illesztésből; így ez külső ponttá válik;
- (8.5) szerint képezzük rá vonatkozóan a Student-törtet (+ előjellel!), amelyet t_i' -vel jelölünk;
- ha $|t_i'| > \gamma$, akkor ξ_i kiszóró pont ε konfidenciaszinten.

Ez azt jelenti, hogy a kiszóró pontok keresése érdekében annyiszor kellene az illesztést megismételni, ahány belső pont van, tehát *n*-szer. A valóságban nem ilyen rossz a helyzet: elegendő az illesztést egyszer megcsinálni *minden* pont figyelembevételével. Ugyanis a belső pontként számított t_i mennyiségekből kiszámítható a fent definiált Student-tört:

8.1. TÉTEL. Ha mind a teljes, mind az *i*-edik pont kihagyásával történő illesztés végrehajtható, továbbá az illesztés linearizálható, végül az *i*-edik ponttól eltekintve más kiszóró pont nincs, a t_i' Student-tört kiszámítható a

$$t_i' = \frac{t_i}{\sqrt{1 - \frac{t_i^2 - 1}{n - m - 1}}} \tag{8.7}$$

képlettel.

A tétel bizonyítását későbbre halasztjuk, mert áttanulmányozását csak azoknak javasoljuk, akik a 6. fejezetet elolvasták.

A megadott feltételek közül kiemeljük az utolsót: (8.7) akkor igaz az *i*-edik pontra, ha *az* i-*edik ponton kívül nincs más kiszóró pont*. Ez azt is jelenti, hogy *nem igaz a többi pontra, ha az* i-*edik pont kiszóró pont*. (8.7) alapján nyerhetünk t_i -re is kvantilist:

$$\gamma' = \frac{\gamma}{\sqrt{1 + \frac{\gamma^2 - 1}{n - m}}}. \tag{8.8}$$

Eszerint a belső pont akkor kiszóró pont ε konfidenciaszinten, ha $|t_i| > \gamma'$, ami ugyanazt jelenti, hogy $|t_i'| > \gamma$.

A fentiekben definiált statisztikai tesztet *általánosított Student-próbá*nak nevezzük. Külső pontokra a szokásos Student-próbával azonos, viszont a belső pontokban attól némileg eltér. $(n - m) \to \infty$ esetén az általánosított próba átmegy a megszokott Student-próbába. A szükséges kvantilisek a 2. függelékben találhatók.

A transzformált Student-törtek tulajdonságai

Ha a (8.7) képletet megfordítjuk, a

$$t_i = \frac{t_i'}{\sqrt{1 + \frac{t_i'^2 - 1}{n - m}}} \tag{8.9}$$

képletet kapjuk, amelyben t_i' Student-tört. Ebből leolvashatjuk a t_i transzformált Student-tört tulajdonságait. Mindenekelőtt látszik, hogy korlátos. Amikor $t_i' \to \pm\infty$, t_i határértéke $\pm\sqrt{n - m}$. Mivel szigorúan monoton növekvő függvény, ebből következik, hogy

$$|t_i| < \sqrt{n - m}. \tag{8.10}$$

t_i sűrűségfüggvényét a (3.40a) sűrűségfüggvényből tudjuk levezetni. Jelöljük az n szabadsági fokú Student-tört eloszlásfüggvényét $S_n(x)$-szel:

$$S_n(x) = \int_{-\infty}^{x} s_n(x')\mathrm{d}x'.$$

Ebből kapjuk a t_i transzformált változó eloszlásfüggvényét:

$$S^*_{n-m}(x) = \mathrm{P}\{t_i < x\} = \mathrm{P}\left\{t'_i < \frac{x}{\sqrt{1-\dfrac{x^2-1}{n-m-1}}}\right\} =$$

$$= S_{n-m-1}\left(\frac{x}{\sqrt{1-\dfrac{x^2-1}{n-m-1}}}\right),$$

hiszen a t'_i Student-tört szabadsági fokainak a száma $(n-m-1)$. Ezt x szerint deriválva – elemi számítások után – kapjuk t_i sűrűségfüggvényét:

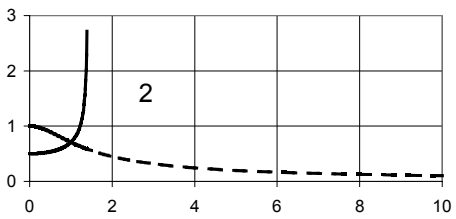
$$s^*_{n-m}(x) = \frac{1}{\sqrt{(n-m)\pi}} \frac{\Gamma\left(\dfrac{n-m}{2}\right)}{\Gamma\left(\dfrac{n-m-1}{2}\right)} \left(1-\frac{x^2}{n-m}\right)^{\frac{n-m-3}{2}}. \qquad (8.11)$$

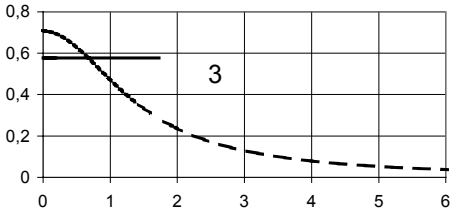
Érdekes, hogy ennek az eloszlásnak a szórásnégyzete a szabadsági fokok számától függetlenül 1:

$$\mathrm{M}(t_i) = 0 \qquad \text{és} \qquad \mathrm{D}^2(t_i) = 1, \quad n-m > 1.$$

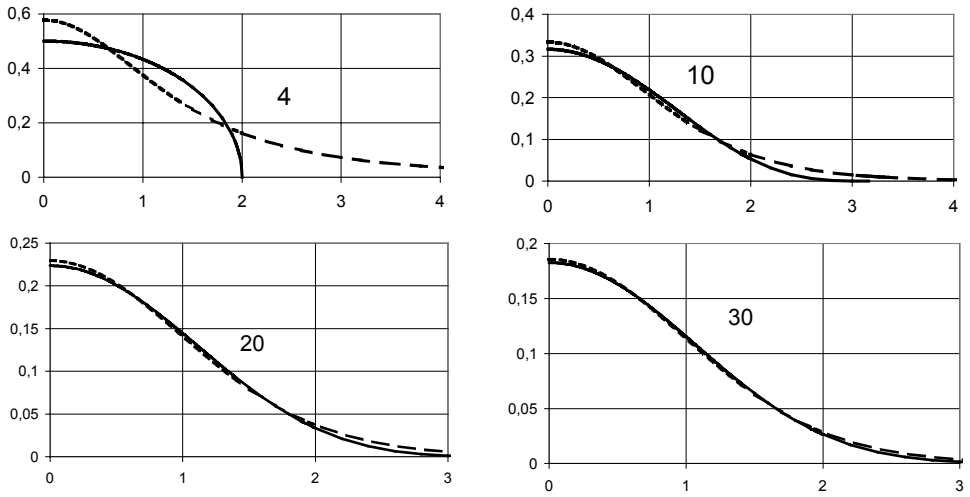
A bizonyítást az Olvasóra bízzuk.[1]

Az összetartozó Student- és módosított Student-eloszlásokat a 8.*1. ábrá*n összehasonlítjuk $(n-m)$ néhány értékére. Látható, hogy $n-m = 20$ fölött már nagyon közel van a két eloszlás egymáshoz. Természetesen az általánosított Student-próba szempontjából nem a sűrűségfüggvények alakja, hanem a γ és γ' kvantilisek a mérvadók, amelyek a 2. függelék táblázataiban találhatók $(n-m)$ és az ε konfidencia-valószínűség különböző értékeire.





[1] Útmutatás: Először lássuk be, hogy a (8.11) függvény integrálja 1. Utána használjuk ki, hogy a szórásnégyzetet megadó integrál kifejezhető az $(n-m)$-edik és az $(n-m+2)$-edik függvény integráljával.



8.1. ábra. A Student- eloszlás (folytonos görbe) és a módosított Student-eloszlás (szaggatott görbe) sűrűségfüggvénye ($n - m$) különböző értékeire

*A 8.1. TÉTEL levezetése

Jelölések

A (8.7) képlet levezetéséhez először néhány jelölésre lesz szükségünk. Az **F** mátrix *i*-edik sora az $\mathbf{f}_i^{\mathrm{T}}$ sorvektor, amivel az illesztett érték az

$$\tilde{y}_i = f\left(x_i, \tilde{\mathbf{a}}\right) = f\left(x_i, \mathbf{a}\right) + \Delta\mathbf{a}^{\mathrm{T}}\mathbf{f}_i = f\left(x_i, \mathbf{a}\right) + \mathbf{f}_i^{\mathrm{T}}\Delta\mathbf{a} \tag{8.12a}$$

alakban írható. Itt kihasználtuk a tételnek azt a feltételét, hogy az illesztési probléma linearizálható, és – ami ezzel együtt jár – a 6.3. alfejezetben tárgyalt torzítás elhanyagolható. Ezzel

$$Q_{\min} = \sum_{j=1}^{n} w_j \left(\xi_j - \tilde{y}_j\right)^2 = \sum_{j=1}^{n} w_j \left(\Delta\xi_j - \Delta\mathbf{a}^{\mathrm{T}}\mathbf{f}_j\right)^2 . \tag{8.12b}$$

A (8.3c) képletet a

$$\mathrm{D}^2\left(\xi_i - \tilde{y}_i\right) = \frac{\sigma^2}{w_i^*} \tag{8.12c}$$

alakba írjuk át, ahol (8.5) szerint

$$\frac{1}{w_i^*} = \frac{1}{w_i} - \mathbf{f}_i^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{f}_i . \tag{8.12d}$$

A (6.9b) és a (6.12b) képletek az

$$\mathbf{R} = \sum_{j=1}^{n} w_j \mathbf{f}_j \mathbf{f}_j^{\mathrm{T}} , \tag{8.12e}$$

illetve

$$\Delta\mathbf{a} = \mathbf{R}^{-1}\sum_{j=1}^{n} w_j \Delta\xi_j \mathbf{f}_j \tag{8.12f}$$

alakra hozhatók.

Hagyjuk most ki a ξ_i mért értéket az illesztésből! A (8.12) képletekkel definiált mennyiségeket úgy tudjuk kiszámítani, hogy a j-re vonatkozó összegzésekből kihagyjuk a $j = i$ indexű tagot. Az így kapott mennyiségeket a teljes illesztéshez tartozóktól az "i" indexszel különböztetjük meg:

$$\mathbf{R}_i = \sum_{j\neq i} w_j \mathbf{f}_j \mathbf{f}_j^{\mathrm{T}} = \mathbf{R} - w_i \mathbf{f}_i \mathbf{f}_i^{\mathrm{T}} \tag{8.13a}$$

[vö. (8.12e)],

$$\Delta\mathbf{a}_i = \mathbf{R}_i^{-1}\sum_{j\neq i} w_j \Delta\xi_j \mathbf{f}_j = \mathbf{R}_i^{-1}\left(\mathbf{R}\Delta\mathbf{a} - w_i \Delta\xi_i \mathbf{f}_i\right) \tag{8.13b}$$

[vö. (8.12f)], végül (8.12b) alapján

$$Q_i = \sum_{j\neq i} w_j \left(\Delta\xi_j - \Delta\mathbf{a}_i^{\mathrm{T}}\mathbf{f}_j\right)^2 = \sigma^2 \chi^2_{n-m-1}. \tag{8.13c}$$

Ennek a képletnek a második része a 6.2. TÉTELből következik, hiszen egy pontot kihagytunk, vagyis a szabadsági fokok száma 1-gyel csökkent. Fontos hangsúlyozni, hogy *ez csak akkor érvényes, ha az* i*-edik ponton kívül más kiszóró pont nincs*. Ha az i-edik pont kiszóró pont, ez nem befolyásolja (8.13c) érvényességét. Ezt a későbbiekben még ki fogjuk használni. A 6.3. TÉTELből következik, hogy ezekkel a feltételekkel Q_i és a $\left(\xi_i - \tilde{y}_i\right)$ különbség statisztikailag függetlenek.

Amikor az i-edik pontot kihagyjuk az illesztésből, a $\left(\xi_i - \tilde{y}_i\right)$ különbség így írható:

$$\xi_i - \tilde{y}_i = \Delta\xi_i - \Delta\mathbf{a}_i^{\mathrm{T}}\mathbf{f}_i. \tag{8.13d}$$

Ezt figyelembe véve (8.5) alapján kapjuk a

$$t_i' = \frac{\left(\Delta\xi_i - \Delta\mathbf{a}_i^{\mathrm{T}}\mathbf{f}_i\right)\sqrt{w_i^{**}}}{\sqrt{\dfrac{Q_i}{n-m-1}}} \tag{8.14a}$$

történt, ahol

$$\frac{1}{w_i^{**}} = \frac{1}{w_i} + \mathbf{f}_i^{\mathrm{T}}\mathbf{R}_i^{-1}\mathbf{f}_i. \tag{8.14b}$$

Azért áll itt pozitív előjel, mert az i-edik most külső pont. A jelölések befejezéseként a (8.5) egyenletet átírjuk a most bevezetett jelölésekkel:

$$t_i = \frac{(\xi_i - \tilde{y}_i)\sqrt{w_i^*}}{\sqrt{\frac{Q_{\min}}{n-m}}} = \frac{\left(\Delta\xi_i - \Delta\mathbf{a}^{\mathrm{T}}\mathbf{f}_i\right)\sqrt{w_i^*}}{\sqrt{\frac{Q_{\min}}{n-m}}}. \tag{8.15}$$

Segédtételek

A (8.7) képlet levezetése három segédtételen alapul.

8.1. LEMMA. A súlyok között fennáll a következő összefüggés:

$$w_i^* w_i^{**} = w_i^2. \tag{8.16}$$

(8.12d) és (8.14b) alapján

$$\frac{w_i^2}{w_i^* w_i^{**}} = \left(1 - w_i \mathbf{f}_i^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{f}_i\right)\left(1 + w_i \mathbf{f}_i^{\mathrm{T}}\mathbf{R}_i^{-1}\mathbf{f}_i\right) =$$

$$= 1 - w_i \mathbf{f}_i^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{f}_i + w_i \mathbf{f}_i^{\mathrm{T}}\mathbf{R}_i^{-1}\mathbf{f}_i - w_i \mathbf{f}_i^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{f}_i w_i \mathbf{f}_i^{\mathrm{T}}\mathbf{R}_i^{-1}\mathbf{f}_i =$$

$$= 1 + w_i \mathbf{f}_i^{\mathrm{T}}\left(\mathbf{R}_i^{-1} - \mathbf{R}^{-1}\right)\mathbf{f}_i - w_i \mathbf{f}_i^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{R} - \mathbf{R}_i)\mathbf{R}_i^{-1}\mathbf{f}_i =$$

$$= 1 + w_i \mathbf{f}_i^{\mathrm{T}}\left(\mathbf{R}_i^{-1} - \mathbf{R}^{-1}\right)\mathbf{f}_i - w_i \mathbf{f}_i^{\mathrm{T}}\left(\mathbf{R}_i^{-1} - \mathbf{R}^{-1}\right)\mathbf{f}_i = 1,$$

amivel (8.16)-ot igazoltuk. A második és harmadik sor között kihasználtuk a (8.13a) képletet.

8.2. LEMMA. A négyzetösszegek között fennáll a következő összefüggés:

$$Q_i = Q_{\min}\left(1 - \frac{t_i^2}{n-m}\right). \tag{8.17}$$

Ennek belátásához először kiszámítjuk a

$$\Delta Q = Q_{\min} - w_i(\xi_i - \tilde{y}_i)^2 = \sum_{j \neq i} w_j \left(\Delta\xi_j - \Delta\mathbf{a}^{\mathrm{T}}\mathbf{f}_j\right)^2$$

különbséget:

$$\Delta Q = \sum_{j \neq i} w_j \left[\left(\Delta\xi_j - \Delta\mathbf{a}_i^{\mathrm{T}}\mathbf{f}_j\right) + \left(\Delta\mathbf{a}_i^{\mathrm{T}} - \Delta\mathbf{a}^{\mathrm{T}}\right)\mathbf{f}_j\right]^2 =$$

$$= \sum_{j \neq i} w_j \left(\Delta\xi_j - \Delta\mathbf{a}_i^{\mathrm{T}}\mathbf{f}_j\right)^2 + 2\left(\Delta\mathbf{a}_i^{\mathrm{T}} - \Delta\mathbf{a}^{\mathrm{T}}\right)\sum_{j \neq i} w_j \left(\Delta\xi_j - \Delta\mathbf{a}_i^{\mathrm{T}}\mathbf{f}_j\right)\mathbf{f}_j +$$

$$+ \sum_{j \neq i} w_j \left[\left(\Delta\mathbf{a}_i^{\mathrm{T}} - \Delta\mathbf{a}^{\mathrm{T}}\right)\mathbf{f}_j\right]^2.$$

(8.13c) szerint az első tag Q_i-vel egyenlő. A második tag eltűnik, ugyanis a (8.13b) képletet balról $\mathbf{R}_i$-vel beszorozva (8.13a) alapján kapjuk:

$$\mathbf{R}_i \Delta\mathbf{a}_i = \sum_{j\neq i} w_j \mathbf{f}_j \mathbf{f}_j^{\mathrm{T}} \Delta\mathbf{a}_i = \sum_{j\neq i} w_j \Delta\mathbf{a}_i^{\mathrm{T}} \mathbf{f}_j \mathbf{f}_j = \sum_{j\neq i} w_j \Delta\xi_j \mathbf{f}_j ,$$

vagyis

$$\sum_{j\neq i} w_j \left(\Delta\xi_j - \Delta\mathbf{a}_i^{\mathrm{T}} \mathbf{f}_j\right)\mathbf{f}_j = 0 .$$

Azt kaptuk tehát, hogy

$$Q_{\min} - w_i\left(\xi_i - \tilde{y}_i\right)^2 = Q_i + \sum_{j\neq i} w_j \left[\left(\Delta\mathbf{a}_i^{\mathrm{T}} - \Delta\mathbf{a}^{\mathrm{T}}\right)\mathbf{f}_j\right]^2 =$$

$$= Q_i + \left(\Delta\mathbf{a}_i^{\mathrm{T}} - \Delta\mathbf{a}^{\mathrm{T}}\right)\mathbf{R}_i\left(\Delta\mathbf{a}_i - \Delta\mathbf{a}\right). \qquad (8.18)$$

(8.13a) és (8.13b) alapján

$$\mathbf{R}_i\left(\Delta\mathbf{a}_i - \Delta\mathbf{a}\right) = \mathbf{R}\Delta\mathbf{a} - w_i\Delta\xi_i\mathbf{f}_i - \mathbf{R}_i\Delta\mathbf{a} = \left(\mathbf{R} - \mathbf{R}_i\right)\Delta\mathbf{a} - w_i\Delta\xi_i\mathbf{f}_i =$$

$$= w_i\mathbf{f}_i\mathbf{f}_i^{\mathrm{T}}\Delta\mathbf{a} - w_i\Delta\xi_i\mathbf{f}_i = -w_i\left(\Delta\xi_i - \mathbf{f}_i^{\mathrm{T}}\Delta\mathbf{a}\right)\mathbf{f}_i = -w_i\left(\xi_i - \tilde{y}_i\right)\mathbf{f}_i .$$

Ha ezt a (8.18) egyenletbe helyettesítjük, a

$$Q_{\min} - w_i\left(\xi_i - \tilde{y}_i\right)^2 = Q_i + \left(\Delta\mathbf{a}_i^{\mathrm{T}} - \Delta\mathbf{a}^{\mathrm{T}}\right)\mathbf{R}_i\mathbf{R}_i^{-1}\mathbf{R}_i\left(\Delta\mathbf{a}_i - \Delta\mathbf{a}\right) =$$

$$= Q_i + w_i^2\left(\xi_i - \tilde{y}_i\right)^2\mathbf{f}_i^{\mathrm{T}}\mathbf{R}_i^{-1}\mathbf{f}_i$$

összefüggés adódik. (8.14b), (8.15) és (8.16) szerint ebből következik, hogy

$$Q_i = Q_{\min} - w_i\left(\xi_i - \tilde{y}_i\right)^2\left(1 + w_i\mathbf{f}_i^{\mathrm{T}}\mathbf{R}_i^{-1}\mathbf{f}_i\right) = Q_{\min} - \frac{w_i^2}{w_i^{**}}\left(\xi_i - \tilde{y}_i\right)^2 =$$

$$= Q_{\min} - w_i^*\left(\xi_i - \tilde{y}_i\right)^2 = Q_{\min} - t_i^2\,\frac{Q_{\min}}{n-m},$$

amint a lemma állítja.

8.3. LEMMA. Fennáll a következő összefüggés:

$$\left(\Delta\xi_i - \Delta\mathbf{a}_i^{\mathrm{T}}\mathbf{f}_i\right)\sqrt{w_i^{**}} = t_i\sqrt{\frac{Q_{\min}}{n-m}} . \qquad (8.19)$$

(8.13a) és (8.13b) alapján a képlet bal oldalán szereplő különbség a

$$\Delta\xi_i - \Delta\mathbf{a}_i^{\mathrm{T}}\mathbf{f}_i = \Delta\xi_i - \Delta\mathbf{a}_i^{\mathrm{T}}\mathbf{R}\mathbf{R}_i^{-1}\mathbf{f}_i + w_i\Delta\xi_i\mathbf{f}_i^{\mathrm{T}}\mathbf{R}_i^{-1}\mathbf{f}_i =$$

$$= \Delta\xi_i - \Delta\mathbf{a}_i^{\mathrm{T}}\left(\mathbf{R}_i + w_i\mathbf{f}_i\mathbf{f}_i^{\mathrm{T}}\right)\mathbf{R}_i^{-1}\mathbf{f}_i + w_i\Delta\xi_i\mathbf{f}_i^{\mathrm{T}}\mathbf{R}_i^{-1}\mathbf{f}_i =$$

$$= \left(\Delta\xi_i - \Delta\mathbf{a}_i^{\mathrm{T}}\mathbf{f}_i\right)\left(1 + w_i\mathbf{f}_i^{\mathrm{T}}\mathbf{R}_i^{-1}\mathbf{f}_i\right) = \left(\xi_i - \tilde{y}_i\right)\left(1 + w_i\mathbf{f}_i^{\mathrm{T}}\mathbf{R}_i^{-1}\mathbf{f}_i\right).$$

Ha figyelembe vesszük a (8.14b), (8.15) és (8.16) összefüggéseket, akkor ebből adódik a lemma állítása:

$$\left(\Delta\xi_i - \Delta\mathbf{a}_i^{\mathrm{T}}\mathbf{f}_i\right)\sqrt{w_i^{**}} = \left(\xi_i - \tilde{y}_i\right)\sqrt{w_i^{**}}\,\frac{w_i}{w_i^{**}} = \left(\xi_i - \tilde{y}_i\right)\sqrt{w_i^{*}} = t_i\sqrt{\frac{Q_{\min}}{n-m}}\,.$$

Végeredmény

Ha a most bizonyított két utolsó lemmában kapott képleteket (8.14a)-ba helyettesítjük, egyszerűen adódik a (8.7) képlet:

$$t_i' = \frac{t_i\sqrt{\dfrac{Q_{\min}}{n-m}}}{\sqrt{\dfrac{Q_{\min}}{n-m-1}\left(1 - \dfrac{t_i^2}{n-m}\right)}} = \frac{t_i}{\sqrt{1 - \dfrac{t_i^2 - 1}{n-m-1}}}\,.$$

A fenti levezetés tiszta algebrai eszközökkel dolgozik, tehát a (8.7) képlet meglehetősen általános feltételekkel érvényes. A 8.1. TÉTELben az illesztés végrehajthatóságára vonatkozó feltétel konkrétan azt jelenti, hogy az $\mathbf{R}$ és $\mathbf{R}_i$ mátrixok minden i-re invertálhatók.

Az általánosított Student-próba használata

A kiszóró pontok megtalálására több próba is elképzelhető – attól függően, hogyan állnak rendelkezésünkre a mérési adatok. Az alábbiakban három rokon próbát tekintünk át: Gauss-próba, Student-próba és általánosított Student-próba. Megfontolásainkat a súlyozatlan átlagolás esetére korlátozzuk, ami a fenti jelölésekkel azt jelenti, hogy $w_i \equiv 1$ és $f(x_i,\mathbf{a}) = a_1$.

Gauss-próba

Akkor alkalmazzuk a *Gauss-próbát*, amikor ismerjük az egyes mérések σ szórását. Ilyenkor nincs szükség arra, hogy az (5.7)-ben definiált s empirikus szórás segítségével becsüljük.

A Gauss-próba alapesete az $\mathrm{M}(\xi_i) = a$ hipotézis ellenőrzése ($i = 1, 2, ..., n$). Ez történhet akár a

$$\zeta = \frac{\overline{\xi} - a}{\mathrm{D}\left(\overline{\xi}\right)} = \frac{\dfrac{\sum_{i=1}^{n}\xi_i}{n} - a}{\sqrt{\dfrac{\sigma^2}{n}}}\,,$$

akár a

$$\zeta_i = \frac{\xi_i - a}{\sigma}, \qquad i = 1, 2, ..., n$$

statisztika alapján. Mindkettő Gauss-eloszlást követ, zérus várható értékkel és 1 szórással. A 2. függelék táblázataiból a választott ε konfidencia-valószínűséghez megtalálhatjuk a γkvantilist. A hipotézist elvetjük, ha

$$|\zeta| > \gamma, \quad \text{illetve} \qquad |\zeta_i| > \gamma$$

fennáll valamelyik *i*-re ($i = 1, 2, ..., n$).

A Gauss-próba használható kiszóró pontok keresésére is. Képezzük a

$$\zeta_i = \frac{\xi_i - \overline{\xi}}{\sigma'}, \qquad i = 1, 2, ..., n$$

hányadosokat, ahol a nevező a számláló szórása:

$$\mathrm{D}^2\left(\xi_i - \overline{\xi}\right) = \mathrm{D}^2\left[\xi_i\left(1 - \frac{1}{n}\right) - \frac{1}{n}\sum_{j \neq i} \xi_j\right] =$$

$$= \sigma^2\left(1 - \frac{1}{n}\right)^2 + \sigma^2 \frac{n-1}{n^2} = \sigma^2 \frac{n-1}{n},$$

tehát a

$$\sigma' = \sigma\sqrt{\frac{n-1}{n}}$$

választás megfelel. Az így kapott ζ_i hányados szintén Gauss-eloszlást követ, zérus várható értékkel és 1 szórással. Ha valamelyik *i*-re $|\zeta_i| > \gamma$, akkor a megfelelő mérési adatot kiszóró pontnak minősítjük.

Student-próba

A Student-próbát szerzője[2] eredetileg a következő célból alkotta meg. Egyetlen ξ adatot mérünk, és az $\mathrm{M}(\xi) = a$ hipotézist kívánjuk ellenőrizni. ξ szórására (σ) vonatkozóan van egy független mérésekből, *n* szabadsági fokkal becsült empirikus szórásnégyzetünk:

$$s^2 = \sigma^2 \frac{\chi_n^2}{n}.$$

Ekkor a

$$t = \frac{\xi - a}{s}$$

[2] A Student álnév, eredeti neve Gosset.

hányados Student-eloszlást követ n szabadsági fokkal. A hipotézist akkor vetjük el, amikor $|t| > \gamma$, ahol γ a választott ε konfidencia-valószínűséghez tartozó kvantilis.

A Student-próba ebben a megfogalmazásában nem igazán használható kiszóró pontok keresésére, legfeljebb a következő hipotézis vizsgálatáról lehet szó. Feltesszük, hogy n mérést végeztünk, és az a hipotézisünk, hogy egy független, $(n + 1)$-edik mérés várható értéke ugyanaz, mint a korábbiaké. Ekkor nyilván a

$$t = \frac{\xi - \bar{\xi}}{\sqrt{\frac{Q_{\min}}{n-1}}}$$

statisztikát kell használnunk. Ez akkor lenne Student-tört, ha a nevezőben a számláló szórásának a becslése állna. A jelenlegi nevező az első n mérés közös σ^2 szórásnégyzetének a becslése. A Student-próba esetében mindig feltesszük, hogy a független $(n + 1)$-edik mérés szórásnégyzete is ugyanez. A számláló szórásnégyzete ezzel a feltevéssel

$$\mathrm{D}^2\left(\xi - \bar{\xi}\right) = \mathrm{D}^2\left(\xi\right) + \mathrm{D}^2\left(\bar{\xi}\right) = \sigma^2 + \frac{\sigma^2}{n} = \sigma^2 \frac{n+1}{n},$$

vagyis a nevezőt korrigálni kell ahhoz, hogy Student-törtet kapjunk:

$$t' = t\sqrt{\frac{n}{n+1}} = \frac{\frac{\xi - \bar{\xi}}{\sigma\sqrt{\frac{n+1}{n}}}}{\sqrt{\frac{Q_{\min}}{\sigma^2(n-1)}}} = \frac{\zeta}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}.$$

Ha $|t'| > \gamma$, akkor a mondott hipotézist elvetjük.

Általánosított Student-próba

Nézzük először az $\mathrm{M}(\xi_i) = a$ hipotézist ($i = 1, 2, ..., n$) abban az esetben, amikor σ^2-et ebből az n mérésből becsüljük, és közülük az egyikre akarunk próbát felírni. Ekkor nem járhatunk el a fentiek szerint, mert a

$$\frac{\xi_i - a}{\sqrt{\frac{Q_{\min}}{n-1}}}$$

törtben a számláló nem független a nevezőtől. Független viszont az átlag, tehát az

$$\eta = \frac{\overline{\xi} - a}{\sqrt{\frac{Q_{\min}}{n-1}}}$$

hányados kapcsolatba hozható a Student-eloszlással:

$$t = \frac{\eta}{\sqrt{n}} = \frac{\frac{\overline{\xi} - a}{\sigma}}{\sqrt{\frac{Q_{\min}}{n(n-1)\sigma^2}}} = \frac{\zeta}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}},$$

amire szintén lehet kvantilist találni a 2. függelék táblázataiban.

Az általánosított Student-próba alapesete a kiszóró pontok keresése. Ha az egyes mérések szórását n mérésből becsüljük, és *közülük az egyik* mérésre akarunk próbát felírni, akkor csak a fentiekben ismertetett általánosított Student-próba marad. Eszerint tehát kiszámítjuk a

$$t_i = \frac{\xi_i - \overline{\xi}}{\sqrt{\frac{Q_{\min}}{n}}}$$

törteket, amelyekből a (8.7) transzformációval kaphatunk $(n-2)$ szabadsági fokú Student-törtet. Ha ennek a kvantilise γ, akkor a $|t_i| > \gamma'$ fennállásakor minősítjük az i-edik mérést kiszórónak, ahol γ'-t a (8.8) képlettel számítjuk ki γ-ból. Vegyük észre, hogy t_i számításához $Q_{\min}$-ot – kivételesen – nem $(n-1)$-gyel, hanem n-nel kell osztani, amint ezt a fenti képletben is tettük.

8.3. A kiszóró pontok megtalálása

A (8.7) képletet először 1935-ben vezették le a súlyozás nélküli átlagolás speciális esetében [4]. A fenti jelölésekkel ez a $w_i \equiv 1$ és $f(x_i,\mathbf{a}) = a_1$ esetnek felel meg. A kiszóró pontokkal foglalkozó irodalomban néha idézik is. Teljes általánosságban való levezetése [3]-ban található meg először (1977). A mérések kiértékelésében azonban sokáig nem játszott nagy szerepet. Először ennek okait beszéljük meg.

1935-ben írt dolgozatában Thompson javasolta a (8.7) képlet használatát, hiszen segítségével a súlyozás nélkül átlagolt mért mennyiségek közül egy Student-próba segítségével ki lehet válogatni a hibásakat, vagyis a kiszóró pontokat. Egy évvel később Pearson és ChandraSekar írtak egy cikket [4], amelyben rámutatnak a módszer gyengéire. Elismerik, hogy a javasolt módszer korrektül veszi figyelembe az *elsőfajú* hibát (vö. 4.3. alfejezet), de teljesen védtelen a *másodfajú* hibával szemben. Két oka van annak, hogy a matematikusok nem sokra becsülték a (8.7) képletet. Egyrészt nem volt kellően általános, hiszen eredeti formájában a súlyozás nélküli átlagolásra vonatkozott, viszont a kiszóró pontok főleg a függvényillesztésben izgalmasak. Másrészt

Pearson és ChandraSekar ellenvetései nagyon komolyak, és nem könnyű rájuk válaszolni. Mivel a 8.1. TÉTEL szerint a képlet nagyon általános feltételek mellett igaz, jól alkalmazható a kiszóró pontok megkeresésére, ha a másodfajú hibával kapcsolatos ellenvetésekre megtaláljuk a választ. Ezért az alábbiakban ezzel foglalkozunk először.

A másodfajú hiba

Amikor a másodfajú hiba hatásait elemezzük, az eredeti H_0 hipotézissel szemben meg kell fogalmaznunk egy ellenhipotézist. A H_0 hipotézis így hangzik:

H_0: a ξ_i mért értékek között (i = 1, 2, ..., n) nincs kiszóró pont.

Amikor az i-edik pontot vizsgáljuk, a $|t_i'| > \gamma$ kritériumot alkalmazzuk, ahol t_i'-t a (8.7) képlettel kapjuk. Az i-edik pontot akkor minősítjük kiszórónak, amikor ez az egyenlőtlenség igaz. Az elsőfajú hiba valószínűsége a (8.6) képletben szereplő ε konfidencia-valószínűség: ennyi annak a valószínűsége, hogy az i-edik pontot kiszórónak minősítjük, pedig nem az. Ezt jelenti az a kijelentés, hogy a módszer korrektül kezeli az elsőfajú hibát.

A másodfajú hiba azt jelenti, hogy a $|t_i'| > \gamma$ egyenlőtlenség hamis, pedig az i-edik pont kiszóró, vagyis fennáll rá a (8.1) reláció. A mérések kiértékelése szempontjából ennek súlyos következményei lehetnek, hiszen ez okozhatja, hogy a becsült paraméterek torzítottak lesznek. Annak a valószínűségét, hogy ez nem következik be, az alkalmazott *statisztikai próba erejé*nek nevezzük. A próba erejét meg szokás vizsgálni a $(\xi_i - y_i)$ szisztematikus hiba függvényében, és azt tekintjük a *legjobb* próbának, amelyre a próba ereje a legnagyobb. Ha egy próba a szisztematikus hiba minden értékére a legjobb, akkor azt *egyenletesen legjobb* próbának nevezzük. Az adott probléma elemzésében nem megyünk ilyen mélyre, mert célunk elsősorban a *gyakorlati* kérdésekre adandó válaszok megkeresése.[3]

A másodfajú hibát a következő ellenhipotézissel szemben vizsgáljuk:

H_1: a ξ_i mért értékek között (i = 1, 2, ..., n) 1-nél több kiszóró pont van.

Azért választjuk éppen ezt, mert ha ez igaz, akkor az alapul vett (8.7) képlet sem érvényes, tehát az alkalmazott próba az elsőfajú hibát sem fogja korrektül kezelni.

A probléma megvilágítására a 8.*1*. *táblázat*ban mutatunk két példát. Két mérést végeztek, mindkettő jól illeszthető volt egy-egy koszinusz-függvénnyel. Az egyikben a mérési pontok száma n = 20, a másikban n értéke jóval 100 fölött volt. Mindkét esetben a ξ_i mért értékek közül négyet-négyet tudatosan elrontottunk: a második számjegyet ±1-gyel megváltoztattuk. Felesleges az elrontott görbéket felrajzolni, a probléma megértéséhez elegendő az

[3] A helyzet az, hogy ez a vizsgálat nem történt meg, pedig hasznos lenne. Mindenestre ez a hiányosság nem érinti az alábbi megfontolásokat.

illesztésben kapott eredményeket megnézni. A 8.*1. táblázat* a négy-négy legnagyobb abszolút értékű Student-törtet mutatja. Kevés mérési pont esetében a $|t_i'| > \gamma$ próba a négy közül csak egyetlen kiszóró pontot talált meg, viszont sok mérési pont esetében megtalálta mind a négyet. Az előbbi esetben tehát fellépett a másodfajú hiba, a másodikban azonban nem.

8.1. táblázat. Négy kiszóró pont keresése

n kicsi		n nagy	
t_i'	Student-próba	t_i'	Student-próba
2,966	megtalálta	4,373	megtalálta
2,258	nem találta meg	–4,474	megtalálta
–1,874	nem találta meg	–5,497	megtalálta
–1,749	nem találta meg	6,538	megtalálta

Mi a probléma eredete? Nyilván szoros kapcsolatban van nem csak a kiszóró pontok számával, hanem a mérési pontokéval is. Pearson és ChandraSekar a következő matematikai jelenségre hívják fel a figyelmet. Súlyozatlan átlagolás esetében könnyű belátni, hogy a t_i törtek között fennállnak a

$$\sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} \frac{\left(\xi_i - \overline{\xi}\right)^2}{\frac{Q_{\min}}{n}} = \frac{Q_{\min}}{\frac{Q_{\min}}{n}} = n \tag{8.20a}$$

és

$$\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} \frac{\xi_i - \overline{\xi}}{\sqrt{\frac{Q_{\min}}{n}}} = 0 \tag{8.20b}$$

összefüggések. Belátásukhoz figyelembe kell venni a (8.5) képletet, amelyben $\widetilde{y}_i = \overline{\xi}$, továbbá egyszerű kiszámítani, hogy

$$\left[\mathbf{W}^{-1} \pm \mathbf{F}\mathbf{R}^{-1}\mathbf{F}^{\mathrm{T}}\right]_{ii} = \frac{n-1}{n},$$

amiből $m = 1$-re való tekintettel következik (8.20). Megjegyezzük, hogy a t_i törtek között $(m + 1)$ analóg összefüggés áll fenn az általános esetben [3].

Várhatóan a kiszóró pontokhoz tartoznak a legnagyobb abszolút értékű t_i értékek. A Student-próba akkor fogja ezeket kimutatni, ha nagyobbak, mint a (8.8) képletben szereplő γ' kvantilis. Érdemes ezért megvizsgálni, hogy a (8.20) feltételeknek eleget tevő t_i értékek abszolút értékének a maximuma egyáltalán mekkora lehet – függetlenül attól, hogy honnan származnak. A részletes vizsgálat (lásd [3] és [4]) a következő eredményt adja:

- Ha a kiszóró pontok száma k, akkor a t_i törtek abszolút értéke akkor a legnagyobb, ha abszolút értékeik egymással egyenlők, viszont a többi tört értéke $1/\sqrt{n}$ nagyságrendű vagy 0 (attól függően, hogy k páratlan, illetve páros).

- A maximális abszolút érték kiszámítható, nagysága $\sqrt{n/k}$, ha k páros, és

$$\sqrt{\frac{n}{k+\frac{1}{n-k}}},$$

ha k páratlan.

Meg lehet mutatni [3], hogy ezek a következtetések nem csak a súlyozatlan átlagolás esetében érvényesek, hanem jó közelítéssel igazak tetszőleges súlyok és illesztőfüggvény esetében is.

Mi mindebből a tanulság? Választ kapunk a 8.*1*. *táblázat*tal kapcsolatban felvetett kérdésre:

- Ha n elég nagy (100-as nagyságrendű), akkor ez a felső korlát még 4 kiszóró pont esetében is elég nagy. Például $n = 100$ és $k = 4$ esetében a felső korlát 5, ami általában nagyobb, mint a kvantilis, amely 99% konfidenciaszinten $\gamma = 2{,}5$ (vö. 2. függelék). Ekkor tehát jó esély van arra, hogy mind a 4 kiszóró pontot észrevesszük.
- Ha n nem elég nagy (mondjuk $n = 20$), akkor a felső határ $k = 4$-re 2,24, tehát 4 kiszóró pontot semmiképpen sem sikerül egyszerre kimutatni, hiszen a felső korlát kisebb, mint a $\gamma = 2{,}5$ kvantilis. $k = 3$ esetében a felső határ 2,56, ami nagyon közel van a kvantilishez, tehát 3 kiszóró pont egyszerre való kimutatására szintén kicsi az esély, bár nem kizárt.

A fenti megfontolások szerint kis számú mérési adat esetében *elemi matematikai* esély is alig van arra, hogy egynél több kiszóró pontot megtaláljunk. Kis n esetén ugyanis a t_i törtek elvi maximuma nem vagy alig haladja meg a kvantilist. Tekintsünk most el ezektől az esetektől, és tegyük fel, hogy n elég nagy ahhoz, hogy legalább az elemi matematikai esély meglegyen egynél több kiszóró pont megtalálásához. Ha például $k = 2$ kiszóró pontot keresünk, és egy $\gamma \approx 2{,}5$ kvantilissel dolgozunk, akkor n értékének a $\sqrt{n/2} \geq 2{,}5$ egyenlőtlenséget magasan teljesítenie kell, vagyis n-nek jelentősen nagyobbnak kell lennie, mint $2 \times 2{,}5^2 = 12{,}5$. Ha ez teljesül, akkor az alábbiakban *elegendően nagy mintá*ról fogunk beszélni.

A kiszóró pontok megtalálásának legkomolyabb gátja a (8.20a) megszorítás, amely mindig fennáll, amikor a σ^2 tényezőt az (5.7) szerinti empirikus szórásnégyzettel becsüljük. Ha ezt nem tesszük, mert σ^2-et valamilyen megfontolásból ismertnek tételezzük fel, akkor hasonló probléma nem merül fel, akkor nincs szükség a (8.7) képletre sem, sőt az nem is érvényes. A most talált probléma végső gyökere tehát a $\tilde{\sigma}^2 = s^2$ becslés, amitől meg tudunk szabadulni, ha σ^2-et ismertnek vesszük. Természetesen ezzel egyéb problémák nem oldódnak meg, amelyekről a későbbiekben még bőven lesz szó.

A kiszóró pontok keresésével kapcsolatban néha lehet találkozni meggondolatlan kijelentésekkel. Egy – szerencsére már visszavont – mérési útmutatóban találtuk a következő receptet: “Ha a ξ_1, ξ_2, ..., ξ_n mért adatok közül vala-

melyiknek az átlagtól való eltérése nagyobb, mint 3σ, akkor az kiszóró pont, és el kell vetni." σ-val az (5.7) szerint becsült *s* empirikus szórást jelöli szerző, továbbá a szövegkörnyezetből világos, hogy a mérések *n* száma ritkán nagyobb 10-nél. Attól a hibától is tekintsünk el, hogy (8.5) alapján *s* helyett az

$$s'^2 = \frac{\sum_{i=1}^{n}\left(\xi_i - \overline{\xi}\right)^2}{n}$$

képletet kell használni [vö. (8.20)] a próbában szereplő t_i statisztika számítására, ugyanis ez az s' a $\left(\xi_i - \overline{\xi}\right)$ *különbség* szórása. Az igazi baj máshol van. Legyen, mondjuk, $n = 10$. Ekkor a fentiek szerint $|t_i| \le \sqrt{n-1} = 3$, tehát a maximum éppen a szerző által választott $\gamma = 3$ kvantilissel egyezik meg. Ez azt jelenti, hogy a javasolt próba *sohasem fog találni* kiszóró pontot. A t_i statisztika maximumának elemzésére vonatkozó levezetésekből [3] ugyanis az következik, hogy a maximum csak a következő esetben lép fel:

$$\max(t_i) = \pm 3 \quad \text{és} \quad t_j = \mp\frac{1}{3}, \quad j \ne i\,.$$

Könnyű ellenőrizni, hogy ezek kielégítik a (8.20) egyenlőségeket. Ez következik be például a következő "mért" adatsor esetében:

$$\xi_1 = 109\,, \quad \xi_2 = \xi_3 = \ldots = \xi_{10} = 99\,.$$

Jól elvégzett mérések nem szoktak ilyen eredményre vezetni. Ha mégis ilyesmi jön ki, nem azt kell vizsgálni, hogy ξ_1 kiszóró adat-e, hanem azt, hogy mi lehet a baj a többi kilenc adattal. Egyébként, ha az egyiket egy kicsit megváltoztatjuk, a maximális t_i a kvantilisnél kisebbé válik. Legyen például $\xi_2 = 98$. Ekkor – mint egyszerűen kiszámíthatjuk – a következő adódik:

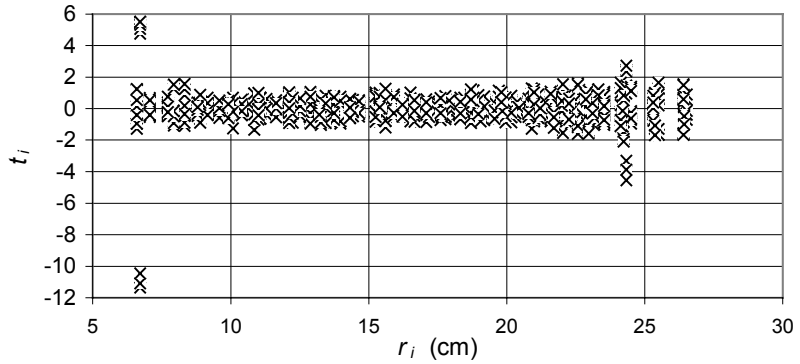
$$t_1 = 2{,}986; \quad t_2 = -0{,}623; \quad t_3 = t_4 = \ldots = t_{10} = -0{,}295\,.$$

Egyiknek az abszolút értéke sem haladja meg a $\gamma = 3$ kvantilist.

Mi legyen a kiszóró pontokkal?

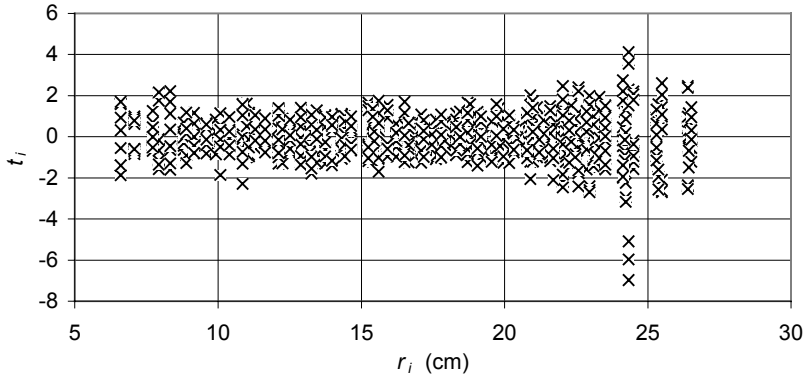
A 8.*2a. ábra* egy elegendően nagy mintára vonatkozóan mutatja a t_i történeket. Az illesztés egy, a 7.7. alfejezetben tárgyalt normálás volt, sok multiplikatív és additív korrekcióval (vö. 7.6. alfejezet): henger alakú reaktorban a különböző r_i sugarakhoz tartozó teljesítménysűrűséget mérték. Az illesztésben minden mért adatot figyelembe vettünk. Az ábráról látható, hogy a mérés tartalmaz több kiszóró pontot is. Különösen szembeszökők az $r_i = 6{,}72$ cm-hez tartozó értékek, amelyek mind a $(-\gamma', +\gamma')$ intervallumon kívül vannak, ahol a kvantilis $\varepsilon = 0{,}01$-hez tartozó értéke $\gamma' = 2{,}57$. Az ábra másik jellegzetessége, hogy a többi pont túlnyomó része a (–1, +1) intervallumba esik. Nem csak az

előbbiek, hanem ez a többség sem lehet Student-tört[4], hiszen azok 35%-nak a (–1, +1) intervallumon kívülre kell esnie.



8.2a. ábra. A t_i törtek az r_i pozíció függvényében, minden mért adat figyelembevételével

Nincs olyan gyakorlott kísérleti fizikus, aki ne ítélné az $r_i = 6{,}72$ cm-hez tartozó pontokat kiszórónak. Tekintsük magunkat ilyennek, és hagyjuk ki ezeket a pontokat az illesztésből! Az eredmény a 8.*2b. ábrá*n látszik. Most az $r_i = 24{,}33$ cm-hez tartozó pontok esnek messze kívül a $(-\gamma', +\gamma')$ intervallumon, amelyek az előbbi ábrán még alig látszottak kiszórónak. Feltűnő ugyanakkor, hogy a többi pont eloszlása kezd megfelelni a Student-eloszlásnak: kezdik kitölteni a (–2, +2) intervallumot.

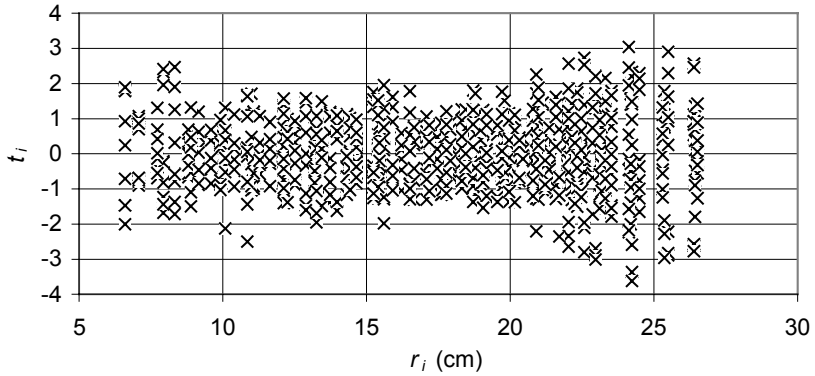


8.2b. ábra. A t_i törtek az r_i pozíció függvényében, az $r_i = 6{,}72$ cm-hez tartozó pontok kihagyásával

Nagyon úgy tűnik, hogy az $r_i = 24{,}33$ cm-hez tartozó pontok szintén kiszóró pontok, tehát hagyjuk ki ezeket is, és ismételjük meg az illesztést! Az eredmény a 8.*2c. ábrá*n látható. A kép teljesen megváltozott: jóllehet most is esnek pontok a $(-\gamma', +\gamma')$ intervallumon kívülre, de a többi pontok eloszlása már nagyon olyannak tűnik, mint amit az ember a Student-eloszlás alapján vár. A kiértékelésben most érkeztünk el az igazi dilemmához. Az eddigiekben elég magabiztosak voltunk, és gond nélkül hagytuk el a kiszórónak látszó pontokat. A

[4] A pontok száma akkora, hogy elhanyagolható a Student-eloszlás és a módosított Student-eloszlás között különbség, vö. *8.1. ábra*. Ezért a továbbiakban nyugodtan beszélhetünk egyszerűen csak Student-eloszlásról.

8.*2c. ábrá*hoz hasonló ábrák láttán még a legtapasztaltabb kísérleti fizikusok is elbizonytalanodnak.



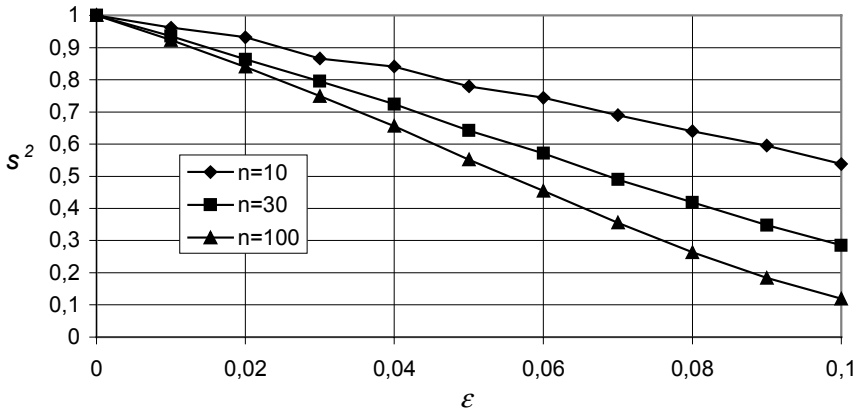
8.2c. ábra. A t_i törtek az r_i pozíció függvényében,
az r_i = 6,72 cm és 24,33 cm-hez tartozó pontok kihagyásával

Nézzük meg ezért közelebbről, mi a teendő, ha egy mérési adat kiszóró pontnak minősül. Az esetek nagy részében megtaláljuk a hiba okát: elírás, hibás kalibráció, téves adatátvitel stb. Ekkor a hiba kijavítása után rendbe szokott jönni az illesztés. Ha a hiba okát nem találjuk, *nem okos dolog kizárólag a Student-próbára hagyatkozva akár egyetlen pontot is elhagyni*. Amit a fenti mérés elemzésében egy hályogkovács magabiztosságával tettünk, nem volt helyes. Ezt az állítást akkor is fenntartjuk, ha a következő alfejezetben meg fogjuk mutatni, hogy helyes volt elhagyni, amit elhagytunk. A probléma ott van, hogy a Student-próba ehhez nem elegendő alap.

A probléma illusztrálására a következő numerikus kísérletet végeztük:

1. Generáltunk *n* darab Gauss-eloszlású véletlen számot zérus várható értékkel és 1 szórással.
2. Kiszámoltuk az átlagot és az empirikus szórásnégyzetet (s^2).
3. Elhagytuk azokat, amelyeket a Student-próba kiszórónak mutatott.
4. Ezután a 2. lépéstől folytattuk addig, amíg legalább egy kiszóró pont akadt.
5. A 8.*3. ábra* mutatja az utolsó lépésben kapott s^2-et az ε konfidencia-valószínűség és *n* függvényében.

Az ábráról látszik, hogy már 99%-os konfidenciaszinten ($\varepsilon = 0{,}01$) is jelentős a csökkenés. Tekintve, hogy a kiindulási adatokat véletlenszám-generátor állította elő, köztük kiszóró pont nem lehet. A próba mégis ilyennek minősített egyes adatokat, aminek az lett az eredménye, hogy az s^2 empirikus szórásnégyzet jelentősen lecsökkent. Emiatt *ok nélkül nem szabad kiszóró pontot elhagyni*. Fontos megjegyezni, hogy ez a következtetés arra az esetre vonatkozik, amikor σ^2-et becsüljük. Ha adottnak vesszük, akkor a kiszóró pontok elhagyása miatt nem csökken a becsült paraméterek szórása.



8.3. ábra. s^2 függése ε-tól és n-től

Akár becsüljük σ^2-et, akár nem, nem tudjuk eldönteni, miről van szó: a Student-próba elsőfajú hibájáról vagy valóságos, de ismeretlen eredetű szisztematikus hibáról. Az sem biztos továbbá, hogy nincs több kiszóró pont, mint amennyit a Student-próba annak minősít, vagyis nem lépett fel a másodfajú hiba.

Ugyanerre a konklúzióra jutnak a [6] alatt idézett szerzők is. A továbbiak szempontjából azonban a leginkább figyelemre méltó Pearson és Chandra-Sekar figyelmeztetése [4]: az a kérdés, hogy az általánosított Student-próba által kiszórónak minősített pont valóban kiszóró-e, *nem dönthető el az általánosított Student-próba keretein belül*. Ennek eldöntéséhez valamilyen független eszközre van szükség. Természetesen ugyanezt mondanák, bármilyen más próbát alkalmaznánk a kiszóró pontok keresésére.

Itt a probléma gyökere. Kell keresnünk valamilyen független eszközt. Ez lehet feljegyzéseink átnézése, műszereink beállításának, kalibrációjának ellenőrzése és ehhez hasonlók. Ha ezek nem segítenek, kell egy végső eszköz. Hogy ez mi lehet, arra a 8.*2. ábrák* elemzésekor már utaltunk. Azokon az ábrákon tűnt biztosnak, hogy kiszóró pontokkal állunk szemben, amelyeken a pontok *többsége* nem felelt meg annak, amit a Student-eloszlás alapján vártunk: a kiszóró pontok mindig együtt járnak azzal, hogy a többi ponthoz tartozó t_i törtek túlságosan kicsik. Tulajdonképpen ez a másodfajú hiba végső oka. A keresett járulékos eszköznek ezt a dolgot kell kimutatnia. Ilyen eszközök vannak, ezek az *illeszkedési próbák*, amelyek annak az ellenőrzésére szolgálnak, hogy egy adott statisztikai sokaság tekinthető-e egy adott eloszlásból vett mintának.

A mondottak alapján tehát a következő eljárást fogjuk követni:

1. Minden mérési adatra vonatkozóan elvégezzük az általánosított Student-próbát, amely vagy jelöl ki kiszóró pontokat vagy sem.
2. A kapott t_i törtek összességére vonatkozóan illeszkedési próbát végzünk annak eldöntésére, hogy ezek tekinthetők-e az általánosított Student-eloszlásból vett mintának.

3. Ha nem tekinthetők annak, akkor az 1. lépésben kiszórónak mutatkozó pontokat kiszórónak tekintjük, és elvetjük.
4. Ezt követően az elhagyott pontok nélkül megismételjük az illesztést.

Ahhoz, hogy ezt alkalmazni tudjuk, meg kell ismerkednünk az illeszkedési próbákkal. Ez lesz a 8.4. alfejezet témája.

8.4. Illeszkedési próbák

Illeszkedési próbákról általában

Legyen a η valószínűségi változó *elméleti eloszlásfüggvénye* $F(x)$. n-szer megmértük és a $\{\eta_i, i = 1, 2, \ldots, n\}$ halmazt kaptuk eredményül. Azt akarjuk ellenőrizni, tekinthető-e ez a halmaz az $F(x)$ eloszlásból vett mintának. Ehhez definiálnunk kell az *empirikus eloszlásfüggvény*t. Rendezzük a mért adatokat nagyság szerint növekvő sorba:

$$\eta_1^* \le \eta_2^* \le \ldots \le \eta_n^* .$$

Az empirikus eloszlásfüggvényt a következőképpen definiáljuk:

$$\Phi_n(x) = \frac{k}{n}, \tag{8.21}$$

ha

$$\eta_k^* < x, \qquad \text{de} \qquad \eta_{k+1}^* \ge x .$$

Azt a hipotézist kívánjuk tesztelni, hogy

$$\mathrm{M}\left[\Phi_n(x)\right] = F(x). \tag{8.22}$$

Az általánosság kedvéért az η_i jelölést használjuk, de az általános képleteket végső soron η_i helyett a t_i törtekre, $F(x)$ helyett pedig a módosított Student-eloszlás eloszlásfüggvényére fogjuk alkalmazni.

A hipotézisvizsgálathoz szükségünk van az empirikus és az elméleti eloszlásfüggvények valamilyen funkcionáljára, amelynek ismerjük az eloszlásfüggvényét. Anderson és Darling a következő funkcionált vizsgálta [7]:

$$W_n^2 = n \int_{-\infty}^{\infty} \left[\Phi_n(x) - F(x)\right]^2 \psi\left[F(x)\right] \mathrm{d}F(x), \tag{8.23}$$

ahol $\psi(t)$ valamilyen súlyfüggvény. Legyen γ_{W} a kvantilis. A (8.22) hipotézist elvetjük az adott konfidenciaszinten, ha

$$W_n^2 > \gamma_{\mathrm{W}} . \tag{8.24}$$

Két súlyfüggvényre ismertek kvantilisek (aszimptotikusan $n \to \infty$-re):
(1) Ha $\psi(t) \equiv 1$:

$$W_n^2 = n\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[F\left(\eta_i^*\right) - \frac{2i-1}{2n}\right]^2 . \qquad (8.25)$$

Kvantilis 95% konfidenciaszinten: $\gamma_W = 0{,}4614$.

(2) Ha $\psi(t) = \dfrac{1}{t(t-1)}$:

$$W_n^2 = -n + 2\sum_{i=1}^{n}\left[\ln\frac{1}{1-F\left(\eta_i^*\right)} - \frac{2i-1}{2n}\ln\frac{F\left(\eta_i^*\right)}{1-F\left(\eta_i^*\right)}\right]. \qquad (8.26)$$

Kvantilis 95% konfidenciaszinten: $\gamma_W = 2{,}4987$.

A (8.22) hipotézis ellenőrzésére több más próbát is definiáltak, de ezeket illetően az irodalomra utalunk [7]. A kiszóró pontok keresésének a céljaira tökéletesen megfelelnek a (8.25) és (8.26) funkcionálok. Érdemes megjegyezni, hogy a (8.26) funkcionál különösen érdekes a mi szempontunkból: a súlyfüggvény kiemeli a $F \to 0$ és $F \to 1$ szélsőértékeket, amelyek éppen a kiszóró pontoknak felelnek meg. Egyébként $F(1-F)$ éppen a (8.23)-ban szereplő $(\Phi_n - F)$ különbség szórásnégyzete.

A (8.26) funkcionál kiszámítása okozhat numerikus nehézségeket, amikor n nagy. Ilyenkor ugyanis előfordulhatnak nagy abszolút értékű pozitív és negatív t_i törtek, amelyekre $F(t_i)$ közel lehet 0-hoz, illetve 1-hez. Ilyenekre a (8.26)-ban levő logaritmusok argumentuma a szingularitás közelébe esik, így az összeg megfelelő tagjának a kiszámítása pontatlan. Külön gond, hogy a kiszámítandó mennyiség két nagy szám, n és az összeg kis különbsége, ami tovább rontja a számítási pontosságot. Mindezek a nehézségek gondos programozással elkerülhetők.

8.2. táblázat. Az illeszkedési próbák kvantilisei

ε	$\psi(t) \equiv 1$	$\psi(t) = 1/t(t-1)$
0,001	1,1679	7,1782
0,01	0,7435	3,9245
0,02	0,6198	3,2900
0,03	0,5489	2,9336
0,04	0,4993	2,6867
0,05	0,4614	2,4987
0,10	0,3473	1,9354
0,15	0,2841	1,6226
0,20	0,2412	1,4091
0,30	0,1843	1,1204

Az illeszkedési próbához szükséges kvantilisek a 8.*2. táblázat*ban találhatók Anderson és Darling számításai szerint [7]. Mivel az ő számításaik aszimptotikusan, $n \to \infty$ mellett érvényesek, a táblázatnak nem bemenő adata n értéke.

Grafikus módszer

Az előzőekben definiált funkcionálokon alapuló próba hasznos kiegészítője a grafikus ábrázolás. Ez a következő észrevételen alapul: az $F(\eta)$ valószínűségi változó egyenletes eloszlású, ugyanis

$$\mathrm{P}\{F(\eta) < x\} = \mathrm{P}\{\eta < F^{-1}(x)\} = F[F^{-1}(x)] = x\,.$$

Ez a gondolatmenet folytonos eloszlásokra érvényes, de az állítás igaz diszkrét eloszlásokra is. Határozzuk meg $F(\eta_i^*)$ várható értékét! Ehhez szükség van $F(\eta_i^*)$ sűrűségfüggvényére. Mi kell ahhoz, hogy

$$x < F(\eta_k^*) < x + \mathrm{d}x$$

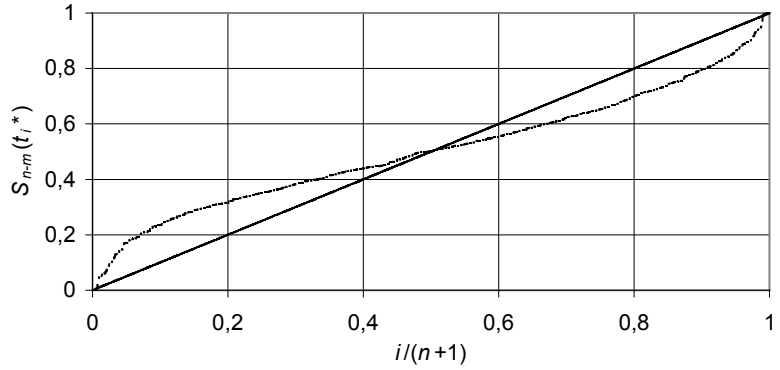
teljesüljön? $(k-1)$ darab változónak x-nél kisebbnek, $(n-k)$ darabnak pedig x-nél nagyobbnak kell lennie. (dx végtelen kicsi.) A k-adik értéket n-féleképpen választhatjuk ki, az előbbi $(k-1)$-et pedig $\binom{n-1}{k-1}$-féleképpen. Így tehát:

$$f_k(x)\mathrm{d}x = \mathrm{P}\{x < F(\eta_k^*) < x + \mathrm{d}x\} = n\binom{n-1}{k-1}x^{k-1}(1-x)^{n-k}\,\mathrm{d}x\,.$$

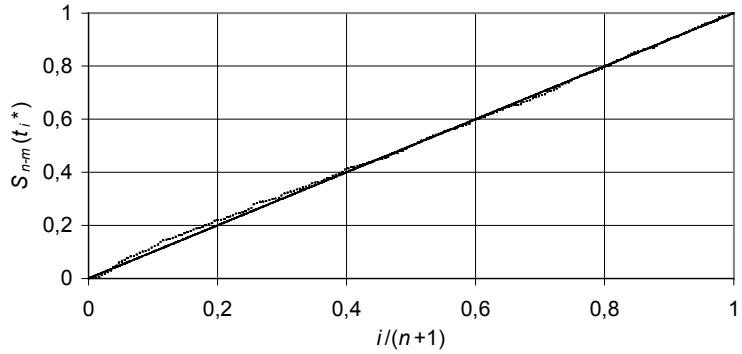
Ezzel

$$\mathrm{M}[F(\eta_k^*)] = \int_0^1 f_k(x)x\mathrm{d}x = n\binom{n-1}{k-1}\int_0^1 x^k(1-x)^{n-k}\,\mathrm{d}x = \frac{k}{n+1}\,,$$

amint ez elemi úton ellenőrizhető. Tehát ha $F(\eta_i^*)$-t ábrázoljuk $k/(n+1)$ függvényében, egy 45° alatt hajló egyenest kell kapnunk, ha igaz a hipotézis. Ilyen grafikonokat mutatnak a 8.*4a.* és *8.4b. ábrák*. Az előbbi a 8.*2a. ábrá*n, az utóbbi pedig a 8.*2c. ábrá*n látható helyzetnek felel meg. Világosan látszik, hogy az előbbi tartalmaz, de az utóbbi már nem tartalmaz kiszóró pontokat.



8.4a. ábra. A 8.*2a. ábrá*n helyzetnek megfelelő illeszkedési grafikon



8.4b. ábra. A 8.*2c. ábrá*n helyzetnek megfelelő illeszkedési grafikon

Alkalmazás a t_i törtekre

A fentieket a t_i törtekre alkalmazzuk, vagyis az elméleti eloszlásfüggvény most

$$F(x) = S^*_{n-m}(x),$$

amelynek megfelelő sűrűségfüggvényt (8.11)-ben felírtuk. Ez nagyon hasznos információt ad, de a módszer sajnos csak közelítő, mert a t_i törtek nem függetlenek, márpedig a γ_W kvantilisek számításában ezt feltételezték. Annak, hogy nem függetlenek, a (6.5) normálegyenletek az okai. Ezek ugyanis m darab egyenletet alkotnak a $(\xi_i - \tilde{y}_i)$ különbségekre vonatkozóan. Nos, (8.5) szerint velük arányosak a t_i törtek, ami azt jelenti, hogy a normálegyenletek átírhatók a t_i törtekre vonatkozó egyenletrendszerré. Az így kapott m egyenletből m darab t_i tört kifejezhető a maradék $(n - m)$ törttel. Emiatt közelítés a fenti illeszkedési próbát közvetlenül a t_i törtekre alkalmazni. A tapasztalat azt mutatja, hogy nagy n-re mégis jól alkalmazható.

Amikor azonban n nem nagyon nagy (mondjuk kisebb, mint 100), akkor célszerű ezt a lineáris függőséget megszüntetni, vagyis az n darab t_i törtet (vagy valamilyen velük arányos mennyiségeket) $(n - m)$ darab független valószínűségi változóra transzformálni. Ennek a módját tárgyaljuk a következő szakaszban.

***Transzformálás Gauss-eloszlásra**

A t_i törtek helyett egyszerűbb a $(\xi_i - \tilde{y}_i)$ különbségeket transzformálni annak érdekében, hogy megvizsgáljuk, tekinthetők-e egy Gauss-eloszlásból vett mintának.[5] Mint mondtuk, ezek nem függetlenek, hiszen (6.5) szerint kielégítenek m egyenletet:

[5] A 6.7. alfejezetben tárgyaljuk a különböző, esetleg nem Gauss-eloszlású mérési adatok kezelését. Ott megmutatjuk, hogy ezek gyakorlatilag Gauss-eloszlásúnak tekinthetők. Ezért elég csak a Gauss-eloszlásra vonatkozó hipotézist vizsgálni.

$$G_k(\tilde{\mathbf{a}}) = \sum_{i=1}^{n} w_i \frac{\partial f(x_i, \tilde{\mathbf{a}})}{\partial a_k}\left[\xi_i - f(x_i, \tilde{\mathbf{a}})\right] =$$

$$= \sum_{i=1}^{n} w_i F_{ik}(\xi_i - \tilde{y}_i) = 0\,, \tag{8.27}$$

$k = 1, 2, ..., m$, amit

$$\mathbf{F}^{\mathrm{T}}\mathbf{W}\left(\vec{\boldsymbol{\xi}} - \tilde{\mathbf{y}}\right) = 0 \tag{8.28}$$

szerint írhatunk át vektori alakba. Ez azt jelenti, hogy közülük m különbség kifejezhető a többi $(n - m)$-mel. Keresünk egy olyan $(n - m)\times n$-es $\mathbf{C}$ mátrixot, hogy a

$$\vec{\boldsymbol{\zeta}} = \mathbf{C}\mathbf{W}^{1/2}\left(\vec{\boldsymbol{\xi}} - \tilde{\mathbf{y}}\right) \tag{8.29}$$

vektor $(n - m)$ azonos szórású, független komponensből álljon, vagyis

$$\mathrm{M}\left(\vec{\boldsymbol{\zeta}}\vec{\boldsymbol{\zeta}}^{\mathrm{T}}\right) = \sigma^2 \mathbf{E}_{n-m,n-m} \tag{8.30}$$

legyen, tehát legyen arányos az $(n - m)\times(n - m)$-es egységmátrixszal. A transzformáció alapján

$$\mathrm{M}\left(\vec{\boldsymbol{\zeta}}\vec{\boldsymbol{\zeta}}^{\mathrm{T}}\right) = \mathbf{C}\mathbf{W}^{1/2}\,\mathrm{M}\left[\left(\vec{\boldsymbol{\xi}} - \tilde{\mathbf{y}}\right)\left(\vec{\boldsymbol{\xi}} - \tilde{\mathbf{y}}\right)^{\mathrm{T}}\right]\mathbf{W}^{1/2}\mathbf{C}^{\mathrm{T}} =$$

$$= \sigma^2 \mathbf{C}\mathbf{W}^{1/2}\left(\mathbf{W}^{-1} - \mathbf{F}\mathbf{R}^{-1}\mathbf{F}^{\mathrm{T}}\right)\mathbf{W}^{1/2}\mathbf{C}^{\mathrm{T}} =$$

$$= \sigma^2 \mathbf{C}\left(\mathbf{E} - \mathbf{W}^{1/2}\mathbf{F}\mathbf{R}^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{W}^{1/2}\right)\mathbf{C}^{\mathrm{T}}\,.$$

Keressük a transzformáló mátrixot a következő alakban:

$$\mathbf{C} = \left[\mathbf{E}_{n-m,n-m} \quad -\mathbf{X}_{n-m,m}\right], \tag{8.31}$$

ahol $\mathbf{X}$ egy egyelőre határozatlan $(n - m)\times m$-es mátrix. Úgy fogjuk megválasztani, hogy (8.30) teljesüljön. A deriváltak mátrixát is bontsuk ennek megfelelő blokkokra:

$$\mathbf{W}^{1/2}\mathbf{F} = \begin{bmatrix} \mathbf{F}'_{n-m,m} \\ \mathbf{F}''_{m,m} \end{bmatrix}. \tag{8.32}$$

A felső blokk $(n - m)\times m$-es, az alsó blokk pedig $m\times m$-es. Némi mátrixalgebrával az alábbi egyenletet kapjuk a keresett $\mathbf{X}$ mátrixra:

$$\sigma^2 \mathbf{E} = \sigma^2\left(\mathbf{E} + \mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathbf{F}'\mathbf{R}^{-1}\mathbf{F}'^{\mathrm{T}} + \mathbf{F}'\mathbf{R}^{-1}\mathbf{F}''^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\right) +$$

$$+ \sigma^2\left(\mathbf{X}\mathbf{F}''\mathbf{R}^{-1}\mathbf{F}'^{\mathrm{T}} - \mathbf{X}\mathbf{F}''\mathbf{R}^{-1}\mathbf{F}''^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\right).$$

Közvetlen beszorzással be lehet látni, hogy ez a mátrixegyenlet a következő alakra hozható:

$$\left(\mathbf{XF}'' - \mathbf{F}'\right)\mathbf{R}^{-1}\left(\mathbf{F}''^{\mathrm{T}}\mathbf{X}^{\mathrm{T}} - \mathbf{F}'^{\mathrm{T}}\right) = \mathbf{XX}^{\mathrm{T}}. \tag{8.33}$$

Bontsuk az **R** mátrixot két mátrix szorzatára: $\mathbf{R} = \mathbf{H}^{\mathrm{T}}\mathbf{H}$ (vö. 2.6. TÉTEL). Ezt beírva a (8.33) egyenlet kielégül, ha

$$\left(\mathbf{XF}'' - \mathbf{F}'\right)\mathbf{H}^{-1} = -\mathbf{X},$$

amiből

$$\mathbf{X} = \mathbf{F}'\mathbf{H}^{-1}\left(\mathbf{F}''\mathbf{H}^{-1} + \mathbf{E}_m\right)^{-1} = \mathbf{F}'\left(\mathbf{H} + \mathbf{F}''\right)^{-1}.$$

Azokat az *i* indexeket, amelyeket kitranszformálunk, (vagyis az $\mathbf{F}''$-nek megfelelő *i*-ket) úgy célszerű megválasztani, hogy a ζ_i és az $\left(\vec{\boldsymbol{\xi}} - \tilde{\mathbf{y}}\right)$ valószínűségi változók közötti korreláció a legerősebb legyen a megmaradó *i* indexekre. Itt nem részletezett megfontolások szerint azokat az *i*-ket célszerű az $\mathbf{F}''$ mátrix számára kiválasztani, amelyekre a $\mathrm{D}^2(\tilde{y}_i)/\mathrm{D}^2(\xi_i)$ hányados a legnagyobb.

A (8.31) ötlet Sarkadi egyik tesztjéből [7] indul ki. Sarkadi a ξ_i független valószínűségi változók normalitásának vizsgálatára javasolja a következő eljárást. Először a közös várható értéket küszöböli ki a

$$\zeta_i = \xi_i - \xi'_n, \qquad i = 1, 2, ..., n-1, \qquad \xi'_n = \frac{n\overline{\xi} + \sqrt{n}\xi_n}{n + \sqrt{n}}$$

transzformációval. Az így kapott zérus várható értékű és σ^2 szórásnégyzetű valószínűségi változók továbbra is függetlenek. Ezek a változók megfelelnek a (8.29) szerint kapott ζ_i változóknak.

Egy újabb transzformációval kiküszöböljük a σ^2 tényezőt.[6] Be lehet látni (lásd [7], 1991), hogy az

$$\eta_i = \frac{\zeta_i^2}{\sum_{j=i}^{n-m}\zeta_j^2} \tag{8.34}$$

hányadosok ($i = 1, 2, \ldots, n-m-1$) egymástól statisztikailag függetlenek, és sűrűségfüggvényük a Béta-eloszlás:

$$f_i(u) = \frac{u^{p-1}(1-u)^{q-1}}{B(p,q)}, \tag{8.35}$$

ahol

[6] Ez az ötlet szintén Sarkadi idézett dolgozatában található.

$$p = \frac{1}{2} \qquad \text{és} \qquad q = \frac{n-i}{2},$$

továbbá

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.$$

Az illeszkedési próbát az $F_i(\eta_i)$ mennyiségekre csináljuk, ahol $F_i(u)$ az előbbi sűrűségfüggvény integrálja. Ebben az illeszkedési próbában az elméleti eloszlásfüggvény az egyenletes eloszlás $F(x) = x$ eloszlásfüggvénye.

Alkalmazás a korábban tárgyalt mérésre

Ha a fenti illeszkedési próbákat alkalmazzuk a 8.*2. ábrák*on látható esetekre, a 8.*3. táblázat*ban látható eredményeket kapjuk. A táblázatban "D"-vel jelöljük a transzformálatlan ("direkt") adatokra, "G"-vel pedig a transzformált ("Gauss-eloszlású") adatokra vonatkozó próbák eredményeit. Látható a táblázatból, hogy a 8.*2c. ábrá*nak megfelelő eset lényegesen jobb funkcionálokat eredményezett, mint a két korábbi, de a "G" esetnek megfelelő W_n^2 funkcionál egy kicsivel még mindig nagyobb, mint a kvantilis. Ha további pontokat hagyunk el, ez a helyzet egyáltalán nem javul, tehát mindenképpen indokolt ennél az illesztésnél megállni.

8.3. táblázat. A (8.25) és (8.26) funkcionálok értékei a 8.*2. ábrák*on mutatott esetekre
A kvantilisek 0,4614 és 2,4987 $n\omega^2$-re, illetve W_n^2-re.

Ábra	$Q_{\min}/(n-m)$	$n\omega^2$		W_n^2	
8.2a.	11,87	D: 9,92	G: 25,5	D: 47,7	G: 126
8.2b.	4,94	D: 0,46	G: 1,06	D: 3,24	G: 6,66
8.2c.	4,07	D: 0,10	G: 0,40	D: 0,89	G: 3,20