Bevezetés a matematikai statisztikába

Dr. Ketskeméty László, Pintér Márta

Budapest, 1999. november 1.

Lektorálta: Dr. Györfi László
Szerkesztette: Győri Sándor

Tartalomjegyzék

1. fejezet

A matematikai statisztika alapfogalmai

A valószínűségszámítás elméletében az $(\Omega, \mathcal{F}, \mathbf{P})$ Kolmogorov valószínűségi mezőn fogalmaztuk meg a tételeinket, azaz a $\mathbf{P}$ valószínűségi mértéket végig adottnak tételeztük fel. A gyakorlati problémáknál azonban a valószínűség nem ismert, legfeljebb logikus előfeltételezéseink vannak róla. A matematikai statisztika alapfeladata éppen az, hogy a véletlen kísérletre, vagy a véletlen tömegjelenségre vonatkozó megfigyeléssorozat segítségével következtetni tudjunk a jelenséghez tartozó adekvát valószínűségi mértékre vagy annak egy jellemzőjére, azt megfelelő pontossággal közelíteni tudjuk. Ilyen értelemben a véletlen jelenségek matematikai modellezésénél a matematikai statisztika módszerei megelőzik a valószínűségszámítás módszereit. A matematikai statisztika fogalomköre, módszertana viszont a valószínűségszámítás fogalmain és módszerein alapul, és ilyen szempontból a matematikai statisztika követi a valószínűségszámítást.

Ugyanúgy, mint a valószínűségszámításnál, a véletlen kísérlet $(\mathcal{K})$ alapfogalmából indulunk ki. Azt is feltesszük, hogy ismert az elemi események Ω halmaza és az események $\mathcal{F}$ halmazrendszere. A $\mathbf{P}$ valószínűség pontosan nem ismert, csak azt tudjuk, hogy a $\mathcal{K}$ véletlen kísérlethez tartozó valószínűség eleme egy $\mathcal{P}$ halmaznak. Tehát $\forall \mathbf{P} \in \mathcal{P}$ esetén Kolmogorov-féle valószínűségi mezőt kapunk. A matematikai statisztika alapfeladata ezen $\mathcal{P}$ halmazból kiválasztani azt a valószínűségi mértéket, amely ténylegesen a kísérlethez tartozik. A $\mathcal{P}$ valószínűségi mértékosztályra esetenként szokásos bizonyos megkötésekkel élni. Ilyen pl. az, amikor $\mathcal{P}$-t *domináltnak* tételezzük fel valamilyen adott λ σ-véges mértékre nézve. Ezen azt értjük, hogy adott az $(\Omega, \mathcal{F})$ mérhető téren olyan λ σ-véges mérték, amelyre $\forall \mathbf{P} \in \mathcal{P}$ abszolút folytonos, azaz ha valamely $A \in \mathcal{F}$ esetén $\lambda(A) = 0$, akkor $\mathbf{P}(A) = 0$ is $\forall \mathbf{P} \in \mathcal{P}$-re.

A $\mathcal{K}$ véletlen kísérlethez megfigyeléssorozatot szervezünk, azaz adatokat gyűjtünk. Matematikailag ezt úgy fogalmazzuk meg, hogy adottnak tételezünk fel egy $X_1, \ldots, X_n$ $\mathbb{R}^d$ értékű független, azonos eloszlású valószínűségi vektorváltozó sorozatot, amelyet *statisztikai mintának* nevezünk. A $\mathbf{P} \in \mathcal{P}$ valószínűség esetén a minta közös eloszlása $\mu_X(A) = \mathbf{P}(X_1 \in A)$ lesz, ahol $A \in \mathcal{B}^d$ d-dimenziós Borel-halmaz. Tehát minden $\mathbf{P} \in \mathcal{P}$ esetén $(\mathbb{R}^d, \mathcal{B}^d, \mu_X)$ Kolmogorov-féle valószínűségi mező lesz. Jelölje $\mathcal{Q}_X$ ezen μ_X eloszlások osztályát. Az $(\mathbb{R}^{nd}, \mathcal{B}^{nd}, \mathcal{Q}_X)$ hármast *statisztikai mezőnek* nevezzük. A statisztikai vizsgálatok célja ezután az, hogy a $\mathcal{Q}_X$ eloszláscsaládból válasszuk ki az $X_1, \ldots, X_n$ mintához tartozó eloszlást.

Statisztikai modellekben általában adott egy $\boldsymbol{\vartheta} : \mathcal{Q}_X \to \mathbb{R}^k$ funkcionál, amelynek értékeit akarjuk minél pontosabban megbecsülni. Ha teljesül, hogy $\boldsymbol{\vartheta}(\mu_X^{(1)}) \neq \boldsymbol{\vartheta}(\mu_X^{(2)})$ esetén $\mu_X^{(1)} \neq \mu_X^{(2)}$, a $\boldsymbol{\vartheta}$ funkcionált paraméternek (paramétervektornak) nevezzük. Ilyenkor a $\boldsymbol{\vartheta}$-nak megfelelő eloszlást $\mu_{\boldsymbol{\vartheta}}$-val fogjuk jelölni: $\mathcal{Q}_X = \{\mu_{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta} \in \Theta\}$, ahol Θ a paramétertér, azaz a $\boldsymbol{\vartheta}$

leképezés értékkészlete. Paraméteres probléma dominált statisztikai mező esetén praktikusan azt jelenti, hogy a minta eloszlása valamilyen paramétertől függő diszkrét vagy folytonos eloszláscsaládból származhat csak. Például, ha feltesszük, hogy a mintánk eloszlása normális, akkor a $\boldsymbol{\vartheta} = (m, D)$ paramétervektor egyértelműen meghatározza a

$$\mathcal{Q}_{\mathbf{X}} = \left\{ \mu_{\boldsymbol{\vartheta}} : \mu_{\boldsymbol{\vartheta}}(B) = \int_B \mathrm{d}\Phi^n_{m,D}(\mathbf{x}) \right\}$$

kétparaméteres eloszlásosztályt, ahol

$$\Phi^n_{m,D}(\mathbf{x}) = \prod_{i=1}^{n} \int_{-\infty}^{x_i} \frac{1}{\sqrt{2\pi} D} \mathrm{e}^{-\frac{(t-m)}{2D^2}} \, \mathrm{d}t.$$

Abban az esetben viszont, ha ilyen $\boldsymbol{\vartheta}$ paraméterfüggvény nem ismert, a statisztikai mező és a rajta megfogalmazott problémák *nemparaméteresek*. Például, ha feltesszük, hogy az $\mathbf{X}$ statisztikai minta $\forall \mathbf{P} \in \mathcal{P}$ esetén véges várható értékkel rendelkezik, azaz $|\mathbf{E}_{\mathbf{P}} X_1| = \left| \int_\Omega X_1 \, \mathrm{d}\mathbf{P} \right| < \infty$, $\forall \mathbf{P} \in \mathcal{P}$-re. Ilyenkor a $\vartheta(\mathbf{P}) = \mathbf{E}_{\mathbf{P}} X_1$ funkcionál nem feltétlenül paraméter, ϑ jó becslése nem jelenti még jó valószínűségi mérték megválasztását.

Adott továbbá egy $\mathbf{t} : \mathbb{R}^n \to \mathbb{R}^k$ mérhető leképezés, melyet *statisztikai függvénynek* nevezünk. A $\mathbf{t}(X_1, X_2, \ldots, X_n)$ összetett függvény a *statisztika*. A statisztika tehát nem más, mint $\forall \mathbf{P} \in \mathcal{P}$ esetén egy valószínűségi vektorváltozó az $(\Omega, \mathcal{F}, \mathbf{P})$ Kolmogorov-féle valószínűségi mezőn.

1.1. definíció: Legyen $(\Omega, \mathcal{F})$ mérhető tér, és $\mathcal{P}$ valószínűségi mértékek egy halmaza, ahol $\forall \mathbf{P} \in \mathcal{P}$ esetén $(\Omega, \mathcal{F}, \mathbf{P})$ Kolmogorov-féle valószínűségi mező. Az $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ statisztikai megfigyelést *statisztikai mintának* nevezzük, ha X_i-k teljesen független, azonos eloszlású valószínűségi változók $\forall \mathbf{P} \in \mathcal{P}$ esetén $(\Omega, \mathcal{F}, \mathbf{P})$-n, azaz $\forall \mathbf{P} \in \mathcal{P}$-re

$$\mathbf{P}(X_i < x) = F_{\mathbf{P}}(x) \qquad (i = 1, 2, 3, \ldots, n)$$

és

$$\mathbf{P}(X_{i_1} < x_{i_1}, X_{i_2} < x_{i_2}, \ldots, X_{i_k} < x_{i_k}) = \prod_{\alpha=1}^{k} F_{\mathbf{P}}(x_{i_\alpha}) \qquad (\forall 2 \le k \le n).$$

n a *minta elemszáma*, $F_{\mathbf{P}}(x)$ a *minta eloszlásfüggvénye*, X_i az i-edik *mintaelem*, $\mu_{\mathbf{P}}(A) = \mathbf{P}(X_i \in A)$, $A \in \mathcal{B}^d$ a minta eloszlása. Egy $\omega \in \Omega$ esetén az

$$X_1(\omega) = x_1, X_2(\omega) = x_2, \ldots, X_n(\omega) = x_n$$

szám n-es a minta egy *realizációja*.

Megjegyzés:

1. Amikor egy statisztikai módszert alkalmazunk, mindig egy statisztikai minta realizáltja áll a rendelkezésünkre. Ez a szám n-es azonban a véletlentől függ, hiszen ha megismételnénk a mintavételezést, egészen biztos, hogy más adatokhoz jutnánk. A módszerek elméletének tárgyalásakor ezért a mintát független, azonos eloszlású valószínűségi változók sorozatának tekintjük.

2. Ha az $\mathbf{X}$ statisztikai minta, λ a Lebesgue-mérték, akkor a $\mathcal{P}$ eloszlásosztály domináltsága most azt jelenti, hogy a minta abszolút folytonos, azaz $\forall \mathbf{P} \in \mathcal{P}$ esetén létezik a minta sűrűségfüggvénye, amelyet $f_{\mathbf{P}}(x)$-szel jelölünk. Ha viszont λ a számláló mérték, vagyis $\lambda(B)$ azt adja meg, hogy a B halmazban mennyi elem van a minta megszámlálható értékkészletéből, a $\mathcal{P}$ domináltsága λ-ra nézve azt jelenti, hogy a statisztikai minta eloszlása diszkrét.

2. fejezet

Becsléselmélet

2.1. Torzítatlan, konzisztens becslés

Legyen $\mathcal{P} = \{\mathbf{P}\}$ egy paraméteres valószínűségi mérték-család.

Feladat olyan $\mathbf{t}_n(X_1, X_2, \ldots, X_n) \in \mathbb{R}^k \quad (n = 1, 2, \ldots)$ statisztikasorozat megadása, amely segítségével „jól" tudjuk becsülni a $\boldsymbol{\vartheta}$ paramétervektort. Ha a paramétert „pontosan" meg tudjuk becsülni, akkor ez egyben azt is jelenti, hogy az adekvát $\mu_{\boldsymbol{\vartheta}}$ eloszlást is közelítőleg megkapjuk. Az alábbiakban az elvárandó „jó", „pontos" becslési tulajdonságokat definiáljuk.

2.1.1. definíció: A $\mathbf{t}_n(X_1, X_2, \ldots, X_n) \in \mathbb{R}^k$ statisztika a $\boldsymbol{\vartheta} \in \mathbb{R}^k$ paraméter torzítatlan becslése, ha $\forall \mathbf{P} \in \mathcal{P}$ esetén a $\mathbf{t}_n$-nek mint valószínűségi vektorváltozónak létezik várhatóérték-vektora és $\mathbf{E}_{\mathbf{P}} \mathbf{t}_n = \boldsymbol{\vartheta}(\mathbf{P})$.

Megjegyzés:

1. Az $\mathbf{E}_{\mathbf{P}} \mathbf{t}_n$ azt jelöli, hogy a várhatóérték-vektor függ attól, hogy melyik $\mathbf{P}$ valószínűségi mérték alapján számoljuk az
$$F_{\mathbf{t}_n}(x_1, x_2, \ldots, x_k) = \mathbf{P}\left(t_n^{(1)} < x_1, t_n^{(2)} < x_2, \ldots, t_n^{(k)} < x_k\right)$$
eloszlásfüggvényt, majd abból a várható értéket.

2. Tudjuk, hogy egy valószínűségi változó értékei a várható értéke körül ingadoznak, tehát, hogy egy statisztika a paraméter torzítatlan becslése, azt az elvárható tulajdonságot fejezi ki, hogy a becslési statisztika realizáltjai az ismeretlen paraméter körül ingadoznak a paramétertérben.

2.1.2. definíció: A $\mathbf{t}_n(X_1, X_2, \ldots, X_n) \in \mathbb{R}^k$ statisztikasorozat a $\boldsymbol{\vartheta} \in \mathbb{R}^k$ paraméter aszimptotikusan torzítatlan becslése, ha $\forall \mathbf{P} \in \mathcal{P}$ esetén a $\mathbf{t}_n$-nek, mint valószínűségi vektorváltozónak létezik várhatóérték-vektora és $\lim\limits_{n\to\infty} \mathbf{E}_{\mathbf{P}} \mathbf{t}_n = \boldsymbol{\vartheta}(\mathbf{P})$.

A torzítatlanságból nyilvánvalóan következik az aszimptotikusan torzítatlanság, tehát ez utóbbi a gyengébb tulajdonság.

2.1.3. definíció: A $\mathbf{t}_n(X_1, X_2, \ldots, X_n) \in \mathbb{R}^k$ statisztikasorozat a $\boldsymbol{\vartheta} \in \mathbb{R}^k$ paraméter konzisztens becslése, ha $\forall \mathbf{P} \in \mathcal{P}$ és $\forall \varepsilon > 0$ esetén $\lim\limits_{n\to\infty} \mathbf{P}\left(\|\mathbf{t}_n - \boldsymbol{\vartheta}\| > \varepsilon\right) = 0$, azaz $\mathbf{t}_n \xrightarrow{st} \boldsymbol{\vartheta}$, $\mathbf{t}_n$ sztochasztikusan konvergál a $\boldsymbol{\vartheta}$ paraméterhez.

Megjegyzés:

1. A konzisztencia más követelményt fejez ki, mint a torzítatlanság. A konzisztencia tulajdonsága azt a jogos elvárást fogalmazza meg, hogy a megfigyelések számának növekedtével javuljon a becslés pontossága.

2. Mivel $\left|t_n^{(i)} - \vartheta_i\right|^2 \leq \sum_{j=1}^{k} \left|t_n^{(j)} - \vartheta_j\right|^2 = \|\mathbf{t}_n - \boldsymbol{\vartheta}\|^2 \leq k \cdot \max_{1 \leq j \leq k} \left|t_n^{(j)} - \vartheta_j\right|^2$, ezért a valószínűségi vektorváltozó sztochasztikus konvergenciája ekvivalens a koordinántánkénti sztochasztikus konvergenciával.

2.1.4. definíció: A $\mathbf{t}_n(X_1, X_2, \ldots, X_n) \in \mathbb{R}^k$ statisztikasorozat a $\boldsymbol{\vartheta} \in \mathbb{R}^k$ paraméter négyzetes középben konzisztens becslése, ha $\lim_{n\to\infty} \mathbf{E_P}||t_n - \boldsymbol{\vartheta}||^2 = 0$.

2.1.1. tétel: Ha a $\mathbf{t}_n$ $(n = 1, 2, \ldots)$ statisztikasorozat négyzetes középben konzisztens becslése $\boldsymbol{\vartheta}$-nak, akkor konzisztens becslése is.

Bizonyítás: A Markov-egyenlőtlenségből:

$$\mathbf{P}\left(\|\mathbf{t}_n - \boldsymbol{\vartheta}\|^2 > \varepsilon^2\right) \leq \frac{\mathbf{E}_\vartheta||t_n - \boldsymbol{\vartheta}||^2}{\varepsilon^2} \to 0 \qquad (n \to \infty).$$

■

2.1.2. tétel: Ha a $\mathbf{t}_n$ $(n = 1, 2, \ldots)$ statisztikasorozat aszimptotikusan torzítatlan becslése $\boldsymbol{\vartheta}$-nak és $\lim_{n\to\infty} \boldsymbol{\sigma}_\mathbf{P} t_n^{(i)} = 0$ $(i = 1, 2, \ldots, k)$, akkor konzisztens becslése is.

Bizonyítás:

$$\mathbf{E_P}(t_n^{(i)} - \vartheta_i)^2 = \mathbf{E_P}(t_n^{(i)} - \mathbf{E_P}t_n^{(i)} + \mathbf{E_P}t_n^{(i)} - \vartheta_i)^2 =$$

$$= \mathbf{E_P}(t_n^{(i)} - \mathbf{E_P}t_n^{(i)})^2 + (\mathbf{E_P}t_n^{(i)} - \vartheta_i)^2 + 2\mathbf{E_P}\left[(t_n^{(i)} - \mathbf{E_P}t_n^{(i)})(\mathbf{E_P}t_n^{(i)} - \vartheta_i)\right] =$$

$$= \boldsymbol{\sigma}^2_\mathbf{P}(t_n^{(i)}) + (\mathbf{E_P}t_n^{(i)} - \vartheta_i)^2 \to 0, \quad n \to \infty.$$

Viszont a Markov-egyenlőtlenség szerint:

$$\mathbf{P}\left(\left|t_n^{(i)} - \vartheta_i\right| > \varepsilon\right) = \mathbf{P}\left((t_n^{(i)} - \vartheta_i)^2 > \varepsilon^2\right) \leq \frac{\mathbf{E_P}(t_n^{(i)} - \vartheta_i)^2}{\varepsilon^2} \to 0,$$

amiből már következik az állítás.

■

2.1.1. példa: *(Várható érték becslése)*
Az

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

statisztikát az $X_1, X_2, \ldots, X_n$ statisztikai minta *átlag*- vagy *empirikus közép* statisztikájának nevezzük.

Legyen az X valószínűségi változó adott. Tegyük fel, hogy $\forall \mathbf{P} \in \mathcal{P}$-re $\exists \mathbf{E_P}X$. Legyen most a paraméter $\vartheta = \vartheta(\mathbf{P}) = \mathbf{E_P}X$. Legyen továbbá $X_1, X_2, \ldots, X_n, \ldots$ statisztikai minta, amelynek eloszlásfüggvénye X-ével azonos $\forall \mathbf{P} \in \mathcal{P}$-re. Akkor

(i) Az $\bar{X}_n = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ empirikus közép statisztika a ϑ várható érték torzítatlan becslése.

(ii) Ha a feltételekhez azt is hozzávesszük, hogy $\forall \mathbf{P} \in \mathcal{P}$-re $\boldsymbol{\sigma}^2_{\mathbf{P}} X < \infty$ is, úgy $\bar{X}_n$ négyzetes középben konzisztens becslés is.

Bizonyítás:

(i) $\mathbf{E}_{\mathbf{P}} \bar{X}_n = \mathbf{E}_{\mathbf{P}} \left(\frac{1}{n}\sum\limits_{i=1}^{n} X_i \right) = \frac{1}{n}\sum\limits_{i=1}^{n} \mathbf{E}_{\mathbf{P}} X_i = \frac{1}{n} n\vartheta = \vartheta.$

(ii) $\mathbf{E}_{\mathbf{P}} \left(\bar{X}_n - \vartheta\right)^2 = \boldsymbol{\sigma}^2_{\mathbf{P}} \bar{X}_n = \boldsymbol{\sigma}^2_{\mathbf{P}} \left(\frac{1}{n}\sum\limits_{i=1}^{n} X_i \right) = \frac{1}{n^2}\sum\limits_{i=1}^{n} \boldsymbol{\sigma}^2_{\mathbf{P}} X_i = \frac{1}{n^2} n \boldsymbol{\sigma}^2_{\mathbf{P}} X = \frac{\boldsymbol{\sigma}^2_{\mathbf{P}} X}{n} \to 0.$

■

2.1.2. példa: *(Szórásnégyzet becslése)*
Az
$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$
statisztikát az $X_1, X_2, \ldots, X_n$ statisztikai minta *empirikus szórásnégyzet* statisztikájának nevezzük. $s_n = +\sqrt{s_n^2}$ az *empirikus szórás* statisztika. Az
$$s_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$
statisztikát az $X_1, X_2, \ldots, X_n$ statisztikai minta *korrigált empirikus szórásnégyzet* statisztikájának nevezzük. $s_n^* = +\sqrt{s_n^{*2}}$ a *korrigált empirikus szórás* statisztika.

Legyen az X valószínűségi változó adott. Tegyük fel, hogy $\forall \mathbf{P} \in \mathcal{P}$-re $\boldsymbol{\sigma}^2_{\mathbf{P}} X < \infty$. Legyen most a paraméter $\vartheta = \vartheta(\mathbf{P}) = \boldsymbol{\sigma}^2_{\mathbf{P}} X$. Legyen továbbá $X_1, X_2, \ldots, X_n, \ldots$ statisztikai minta, amelynek eloszlásfüggvénye X-ével azonos $\forall \mathbf{P} \in \mathcal{P}$-re.

(i) Az $s_n^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \bar{X}_n)^2$ empirikus szórásnégyzet statisztika a ϑ szórásnégyzet aszimptotikusan torzítatlan becslése, az $s_n^{*2} = \frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i - \bar{X}_n)^2$ korrigált empirikus szórásnégyzet statisztika pedig a ϑ szórásnégyzet torzítatlan becslése.

(ii) Ha a feltételekhez azt is hozzávesszük, hogy $\forall \mathbf{P} \in \mathcal{P}$-re $\mathbf{E}_{\mathbf{P}} X^4$ is, úgy s_n^2 konzisztens, s_n^{*2} négyzetes középben konzisztens becslés is.

Bizonyítás:
Fel fogjuk használni a *Steiner*-tételt:
Segédtétel: (Steiner)
Tetszőleges $a, x_1, x_2, \ldots, x_n$ valós számokra
$$\frac{1}{n}\sum_{i=1}^{n}(a - x_i)^2 = (a - \bar{x}_n)^2 + \frac{1}{n}\sum_{i=1}^{n}(\bar{x}_n - x_i)^2 \geq \frac{1}{n}\sum_{i=1}^{n}(\bar{x}_n - x_i)^2.$$
Másrészt $a = 0$ választással, átrendezés után:
$$\frac{1}{n}\sum_{i=1}^{n}(\bar{x}_n - x_i)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2.$$

A segédtétel bizonyítása:

$$\frac{1}{n}\sum_{i=1}^{n}(a-x_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(a-\bar{x}_n+\bar{x}_n-x_i)^2 =$$

$$= (a-\bar{x}_n)^2 + 2(a-\bar{x}_n)\frac{1}{n}\sum_{i=1}^{n}(\bar{x}_n-x_i) + \frac{1}{n}\sum_{i=1}^{n}(\bar{x}_n-x_i)^2 .$$

A középső tag nulla, így az állítást igazoltuk.

Az állítás bizonyítása:

(i)

$$\mathbf{E}_{\mathbf{P}}s_n^2 = \mathbf{E}_{\mathbf{P}}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)^2\right) = \mathbf{E}_{\mathbf{P}}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2-(\bar{X}_n)^2\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}_{\mathbf{P}}X_i^2-\mathbf{E}_{\mathbf{P}}(\bar{X}_n)^2 =$$

$$= \frac{1}{n}\cdot n\cdot\left(\vartheta+(\mathbf{E}_{\mathbf{P}}X_1)^2\right)-\left(\frac{\vartheta}{n}+(\mathbf{E}_{\mathbf{P}}X_1)^2\right) = \frac{n-1}{n}\vartheta \to \vartheta \quad (n\to\infty)$$

Mivel $s_n^{*2} = \frac{n}{n-1}s_n^2 \implies \mathbf{E}_{\mathbf{P}}s_n^{*2} = \frac{n}{n-1}\mathbf{E}_{\mathbf{P}}s_n^2 = \vartheta$.

(ii) Belátható, hogy

$$\boldsymbol{\sigma}_{\mathbf{P}}^2 s_n^2 = \frac{n^2}{(n-1)^2}\left(\frac{\mathbf{E}_{\mathbf{P}}X^4}{n}-\frac{n-3}{n(n-1)}\vartheta^2\right) \to 0, \quad \boldsymbol{\sigma}_{\mathbf{P}}^2 s_n^{*2} \to 0.$$

Hivatkozva a 2.1.2. tételre s_n^2 konzisztenciája bizonyított.

■

2.1.3. példa: *(Kovariancia és korrelációs együttható becslése)*
Legyen most az $\left((X_1,Y_1)^T,(X_2,Y_2)^T,\dots,(X_n,Y_n)^T\right)^T$ statisztikai megfigyelés kétdimenziós statisztikai minta, ahol az $(X_i,Y_i)^T$ párok azonos eloszlású, teljesen független valószínűségi vektorváltozók. Ekkor a

$$c_n = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)\left(Y_i-\bar{Y}_i\right)$$

statisztika az $(X_1,Y_1)^T,(X_2,Y_2)^T,\dots,(X_n,Y_n)^T$ minta *empirikus kovarianciája*, $\rho_n = \frac{c_n}{s_X^* s_Y^*}$ pedig az *empirikus korrelációs együtthatója*, ahol pl.

$$s_X^* = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)^2}$$

az $X_1,X_2,\dots,X_n$ statisztikai minta korrigált empirikus szórását jelöli.

(i) A c_n empirikus kovariancia az $\mathbf{E}_{\mathbf{P}}(X-\mathbf{E}_{\mathbf{P}}X)(Y-\mathbf{E}_{\mathbf{P}}Y)$ kovariancia torzítatlan becslése. Ha még azt is feltehetjük, hogy $\exists\mathbf{E}_{\mathbf{P}}X^4$, $\mathbf{E}_{\mathbf{P}}Y^4$ is, akkor c_n négyzetes középben konzisztens becslés is.

(ii) Az ρ_n empirikus korrelációs együttható a korrelációs együttható aszimptotikusan torzítatlan becslése. Ha még azt is feltehetjük, hogy $\exists\mathbf{E}_{\mathbf{P}}X^4$, $\mathbf{E}_{\mathbf{P}}Y^4$ is, akkor ρ_n konzisztens becslés is.

Bizonyítás:

(i) Jelölje $\mathbf{cov}\,(X_i, Y_i) = c$, $\mathbf{E}X_i\mathbf{E}Y_i = m$. Ekkor

$$\mathbf{E}X_iY_i = c + m,\ \mathbf{E}X_i\bar{Y} = \mathbf{E}\bar{X}Y_i = \frac{1}{n}(c + nm) = \frac{c}{n} + m,\ \mathbf{E}\bar{X}\bar{Y} = \frac{c}{n} + m.$$

Tehát

$$(n-1)\,c_n = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = \sum_{i=1}^{n}\left(X_iY_i - X_i\bar{Y} - Y_i\bar{X} + \bar{X}\bar{Y}\right),$$

azaz

$$\mathbf{E}\left((n-1)\,c_n\right) = (nc + nm) - (c + nm) - (c + nm) + (c + nm) = (n-1)\,c.$$

Megmutatható, hogy

$$\boldsymbol{\sigma}^2 c_n = \frac{m_{22}}{n} + \frac{s_1 s_2}{n(n-1)} + \frac{c}{n(n-1)},$$

ahol

$$m_{22} = \mathbf{E}\left((X_i - \mathbf{E}X_i)^2\,(Y_i - \mathbf{E}Y_i)^2\right),\quad s_1 = \boldsymbol{\sigma}^2 X_i,\quad s_2 = \boldsymbol{\sigma}^2 Y_i.$$

Mivel $\boldsymbol{\sigma}^2 c_n \longrightarrow 0$, így a konzisztencia már következik.

(ii) Nem bizonyítjuk. Bizonyítása megtalálható Cramer: Mathematical statistics c. könyvben.

■

2.1.4. példa: *(Eloszlásfüggvény becslése)*
Tekintsük azokat az $\mathrm{ord}_k(x_1, x_2, \ldots, x_n)$ skalár–vektor függvényeket, melyek definíciója:

$$x_k^* = \underset{k}{\mathrm{ord}}(x_1, x_2, \ldots, x_n) = x_j,$$

ha x_j a k-adik legnagyobb elem $x_1, x_2, \ldots, x_n$ között. Az

$$X_k^* = \underset{k}{\mathrm{ord}}(X_1, X_2, \ldots, X_n) \quad (k = 1, 2, \ldots, n)$$

statisztikák a *rendezett mintaelem-statisztikák.*

Megjegyzés:

1. A rendezett mintaelem-statisztikák között $\forall \mathbf{P} \in \mathcal{P}$ esetén 1 valószínűséggel fennáll, hogy $X_1^* \leq X_2^* \leq \cdots \leq X_n^*$. Speciálisan $X_1^* = \min\{X_1, X_2, \ldots, X_n\}$, és $X_n^* = \max\{X_1, X_2, \ldots, X_n\}$.

2. Ha a minta eloszlásfüggvényét $F(x)$-szel jelöljük, könnyű megmutatni, hogy a rendezett mintaelemek eloszlásfüggvényeit és együttes eloszlásfüggvényeit az alábbi módon lehet számolni:

$$F_k(x) = \mathbf{P}\left(X_k^* < x\right) = \sum_{i=k}^{n}\binom{n}{i}\left[F(x)\right]^i\left[1 - F(x)\right]^{n-i},$$

$$F_{k,l}(x, y) = \mathbf{P}\left(X_k^* < x, X_l^* < y\right) =$$

$$= \sum_{i=k}^{n} \sum_{j=l}^{i} \frac{n!}{j!(i-j)!(n-i)!} \left[F(x)\right]^{j} \left[F(y)-F(x)\right]^{i-j} \left[1-F(y)\right]^{n-i},$$

$$\mathbf{P}\left(X_1^* < x_1, X_2^* < x_2, \ldots, X_n^* < x_n\right) =$$

$$= \sum_{i_n=n}^{n} \sum_{i_{n-1}=n-1}^{i_n} \cdots \sum_{i_1=1}^{i_2} \frac{n!}{i_1!(i_2-i_1)!\cdots(n-i_n)!} \left[F(x_1)\right]^{i_1} \left[F(x_2)-F(x_1)\right]^{i_2-i_1} \cdots \left[1-F(x_n)\right]^{n-i_n}.$$

Az

$$F_n(x) = \begin{cases} 0, & \text{ha } \; x \le X_1^* \\ \frac{k}{n}, & \text{ha } \; X_k^* < x \le X_{k+1}^* \quad (k = 1, 2, \ldots, n-1) \\ 1, & \text{ha } \; x > X_n^* \end{cases}$$

véletlen függvényt az $X_1, X_2, \ldots, X_n$ statisztikai minta *empirikus eloszlásfüggvényének* nevezzük.

Használatos az előzővel ekvivalens $F_n(x) = \sum_{i=1}^{n} I_{\{X_i < x\}}$ definíció is, ahol

$$I_{\{X_k < x\}} = \begin{cases} 1, & \text{ha } X_k < x \\ 0, & \text{ha } X_k \ge x \end{cases}.$$

Az empirikus eloszlásfüggvény minden rögzített $x \in \mathbb{R}$ esetén statisztika, azaz valószínűségi változó! $F_n(x)$ minden realizációja diszkrét eloszlásfüggvény, azaz olyan lépcsős függvény, melynek ugráshelyei a véletlen mintától függenek, és az ugrások magassága 1 valószínűséggel $\frac{1}{n}$.

Legyen az X valószínűségi változó adott. Legyen továbbá $X_1, X_2, \ldots, X_n, \ldots$ statisztikai minta, amelynek eloszlásfüggvénye X-ével azonos. Rögzítsük most az $x \in \mathbb{R}$ valós számot. Ekkor X eloszlásfüggvénye az x pontban a paraméter: $\vartheta = \vartheta(\mathbf{P}) = F_{\mathbf{P}}(x)$. Akkor az $F_n(x)$ empirikus eloszlásfüggvény értéke a ϑ eloszlásfüggvény-érték torzítatlan, négyzetes középben konzisztens becslése.

Bizonyítás: Az empirikus eloszlásfüggvény definíciójából nyilvánvaló, hogy

$$0 \le nF_n(x) \le n$$

és

$$\mathbf{P}(nF_n(x) = k) = \mathbf{P}(k \text{ db } i \text{ indexre } X_i < x,\ (n-k) \text{ db } j \text{ indexre } j \ne i\ \ X_j \ge x) =$$

$$= \binom{n}{k} \left[F_{\mathbf{P}}(x)\right]^{k} \left[1 - F_{\mathbf{P}}(x)\right]^{n-k} \implies nF_n(x) \in B(n, \vartheta).$$

Azaz $nF_n(x)$ binomiális eloszlású n és $F_{\mathbf{P}}(x) = \vartheta$ paraméterekkel. Viszont ekkor

$$\mathbf{E}_{\mathbf{P}}(nF_n(x)) = n\vartheta$$

és

$$\boldsymbol{\sigma}_{\mathbf{P}}^2(nF_n(x)) = n\vartheta(1-\vartheta).$$

Innét pedig

$$\mathbf{E}_{\mathbf{P}}(F_n(x)) = \vartheta$$

és

$$\boldsymbol{\sigma}_{\mathbf{P}}^2(F_n(x)) = \frac{\vartheta \cdot (1-\vartheta)}{n} \le \frac{1}{4n} \to 0 \quad (n \to \infty)$$

következik, ami az állítást igazolja. Felhasználtuk, hogy $\vartheta(1-\vartheta) \le \frac{1}{4}$.

■

Mivel a négyzetes középben való konzisztenciából következik a konzisztencia, ezért $\forall\varepsilon > 0$, $\forall x \in \mathbb{R}$, $\forall \mathbf{P} \in \mathcal{P}$-re $\mathbf{P}(|F_n(x) - F_{\mathbf{P}}(x)| > \varepsilon) \to 0 \ \ (n \to \infty)$. Ennél az állításnál lényegesen erősebbet fogalmaz meg a következő tétel: az empirikus eloszlásfüggvény 1 valószínűséggel, egyenletesen konvergál az eloszlásfüggvényhez. Elméleti jelentősége miatt a tételt a matematikai statisztika alaptételének is hívják.

2.1.3. tétel: *(A matematikai statisztika alaptétele, Glivenko–Cantelli)*
Legyen $X_1, X_2, \ldots, X_n, \ldots$ a statisztikai minta. Jelölje $F(x)$ a minta eloszlásfüggvényét, és $F_n(x)$ az empirikus eloszlásfüggvényt.

Akkor $\mathbf{P}\left(\lim\limits_{n\to\infty} \sup\limits_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0\right) = 1.$

Bizonyítás: Legyen $\varepsilon > 0$, $x \in \mathbb{R}$ tetszőleges! Megmutatjuk, hogy $\exists\ N > 0$ és $C \in \mathcal{F}$: $\mathbf{P}(C) = 1$, hogy $\forall\ \omega \in C$ esetén, ha $n > N$, úgy $|F_n(x) - F(x)| < \varepsilon$. Legyen m olyan pozitív egész szám, hogy $\frac{1}{m} < \frac{\varepsilon}{2}$, és legyenek $\mathbb{R}$ egy m intervallumból álló rendszerének osztópontjai $x_0^{(m)} = -\infty$, $x_m^{(m)} = +\infty$, $x_k^{(m)} = \sup\limits_{x\in\mathbb{R}} \left\{x : F(x) \le \frac{k}{m}\right\}$. Jelölje az intervallumokat: $J_k = \left(x_k^{(m)}, x_{k+1}^{(m)}\right]$, $k = 0, 1, \ldots, m-1$. Tegyük fel, hogy a szóban forgó x-re éppen $x \in J_{k-1} \implies x_{k-1}^{(m)} < x \le x_k^{(m)}$ teljesül most. Az eloszlásfüggvény tulajdonságai miatt:

$$\left.\begin{array}{l} F(x_k^{(m)}) \le \frac{k}{m} \le F(x_k^{(m)} + 0) \\ F(x_{k-1}^{(m)}) \le \frac{k-1}{m} \le F(x_{k-1}^{(m)} + 0) \end{array}\right\} \implies (*)\ \ F(x_k^{(m)}) \le \frac{k}{m} \le F(x_{k-1}^{(m)} + 0) + \frac{1}{m}.$$

A nagy számok erős törvénye értelmében a relatív gyakoriság 1 valószínűséggel közelíti az elméleti valószínűséget:

$$\exists A_k \in \mathcal{F}:\ \mathbf{P}(A_k) = 1 \ \text{ és } \ \forall\, \omega \in A_k :\ \lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} I_{\{X_i < x_k^{(m)}\}}(\omega)\right) = F(x_k^{(m)}).$$

$$\exists B_k \in \mathcal{F}:\ \mathbf{P}(B_k) = 1 \ \text{ és } \ \forall\, \omega \in B_k :\ \lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \le x_{k-1}^{(m)}\}}(\omega)\right) = F(x_{k-1}^{(m)} + 0).$$

Legyen $C = \prod\limits_{k=1}^{m} A_k B_{k-1}$. Akkor $\mathbf{P}(C) = \mathbf{P}\left(\prod\limits_{k=1}^{m} A_k B_{k-1}\right) = 1 \implies \mathbf{P}(C) = 1.$
Tehát $\forall\omega \in C$ esetén $\exists N :\ n > N$, akkor

$$\left|\frac{1}{n}\sum_{i=1}^{n} I_{\{X_i < x_k^{(m)}\}} - F(x_k^{(m)})\right| < \frac{\varepsilon}{2}, \quad \text{és} \quad \left|\frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \le x_{k-1}^{(m)}\}} - F(x_{k-1}^{(m)} + 0)\right| < \frac{\varepsilon}{2}.$$

Így $x \in J_{k-1}$-re

$$F(x) - F_n(x) \le F(x_k^{(m)}) - F_n(x_{k-1}^{(m)}) \le F(x_{k-1}^{(m)} + 0) + \frac{1}{m} - F_n(x_{k-1}^{(m)} + 0) \le \frac{1}{m} + \frac{\varepsilon}{2}.$$

Másrészt

$$F(x) - F_n(x) \ge F(x_{k-1}^{(m)} + 0) - F_n(x_k^{(m)}) \ge F(x_k^{(m)}) - \frac{1}{m} - F_n(x_k^{(m)}) \ge -\frac{1}{m} - \frac{\varepsilon}{2}.$$

Azaz $|F(x) - F_n(x)| < \frac{\varepsilon}{2} + \frac{1}{m} < \varepsilon \implies$ állítás.

■

2.2. Hatásos becslések

2.2.1. definíció: Legyenek $\hat{t}$ és $\tilde{t}$ a $\vartheta \in \mathbb{R}$ paraméter torzítatlan becslései, ahol $\exists\, \boldsymbol{\sigma}^2_{\mathbf{P}}\hat{t}$ és $\boldsymbol{\sigma}^2_{\mathbf{P}}\tilde{t}$ $(\forall \mathbf{P} \in \mathcal{P})$. Azt mondjuk, hogy $\hat{t}$ *hatásosabb* becslése ϑ-nak mint $\tilde{t}$, ha $\boldsymbol{\sigma}^2_{\mathbf{P}}\hat{t} \le \boldsymbol{\sigma}^2_{\mathbf{P}}\tilde{t}$ $\forall \mathbf{P} \in \mathcal{P}$-re és $\exists \mathbf{P}_0 \in \mathcal{P} : \boldsymbol{\sigma}^2_{\mathbf{P}_0}\hat{t} < \boldsymbol{\sigma}^2_{\mathbf{P}_0}\tilde{t}$.

2.2.1. példa: Legyen az X valószínűségi változó adott. Tegyük fel, hogy X egyenletes eloszlású valószínűségi változó a $[0, \vartheta]$ intervallumon, ahol $\vartheta > 0$ ismeretlen paraméter.

Most $\forall \mathbf{P} \in \mathcal{P}$-re $F_{X,\vartheta}(x) = \frac{x}{\vartheta}$, $\frac{\mathrm{d}F_{X,\vartheta}(x)}{\mathrm{d}x} = f_{X,\vartheta}(x) = \frac{1}{\vartheta}$, $x \in (0, \vartheta)$, $\mathbf{E}_\vartheta X = \frac{\vartheta}{2}$, $\boldsymbol{\sigma}^2_\vartheta X = \frac{\vartheta^2}{12}$. Legyen továbbá $X_1, X_2, \ldots, X_n, \ldots$ statisztikai minta, amelynek eloszlásfüggvénye X-ével azonos.

Tekintsük a

$$T_1 = \frac{n+1}{n} X_n^*,$$

$$T_2 = X_n^* + X_1^*,$$

$$T_3 = \frac{n+1}{n-1}\left(X_n^* - X_1^*\right),$$

$$T_4 = 2\bar{X}_n$$

statisztikákat! Megmutatjuk, hogy mindegyikük torzítatlan, de különböző szórású becslés, tehát eltér a hatásosságuk.

$$\mathbf{E}_\vartheta T_4 = \mathbf{E}_\vartheta 2\,\bar{X}_n = 2\mathbf{E}_\vartheta \bar{X}_n = 2\mathbf{E}_\vartheta X = 2\frac{\vartheta}{2} = \vartheta \implies T_4 \text{ torzítatlan.}$$

$$\boldsymbol{\sigma}^2_\vartheta T_4 = 4\boldsymbol{\sigma}^2_\vartheta \bar{X}_n = 4\frac{\boldsymbol{\sigma}^2_\vartheta X}{n} = 4\,\frac{\vartheta^2}{12n} = \frac{\vartheta^2}{3n} \to 0 \implies T_4 \text{ négyzetes középben konzisztens.}$$

Az X_n^* eloszlásfüggvénye:

$$\mathbf{P}\left(X_n^* < x\right) = \left[F_{X,\vartheta}(x)\right]^n = \left(\frac{x}{\vartheta}\right)^n, \quad x \in [0, \vartheta]$$

$\implies$ sűrűségfüggvénye

$$f_{n,\vartheta}(x) = n\,\frac{x^{n-1}}{\vartheta^n}, \quad x \in (0, \vartheta)\,.$$

$$\mathbf{E}_\vartheta X_n^* = \int_0^\vartheta x\, f_{n,\vartheta}(x)\,\mathrm{d}x = \int_0^\vartheta n\,\frac{x^n}{\vartheta^n}\,\mathrm{d}x = n\,\frac{1}{\vartheta^n}\left[\frac{x^{n+1}}{n+1}\right]_0^\vartheta = \frac{n}{n+1}\vartheta$$

$\implies \mathbf{E}_\vartheta T_1 = \vartheta$, torzítatlan.

$$\boldsymbol{\sigma}^2_\vartheta T_1 = \mathbf{E}_\vartheta T_1^2 - \left(\mathbf{E}_\vartheta T_1\right)^2 = \left(\frac{n+1}{n}\right)^2 \int_0^\vartheta x^2 n \frac{x^{n-1}}{\vartheta^n}\,\mathrm{d}x - \vartheta^2 =$$

$$= \frac{(n+1)^2}{n}\,\frac{1}{\vartheta^n}\left[\frac{x^{n+2}}{n+2}\right]_0^\vartheta - \vartheta^2 = \frac{(n^2 + 2\,n + 1 - n^2 - 2\,n)\vartheta^2}{n(n+2)} = \frac{\vartheta^2}{n(n+2)} \to 0.$$

$\implies T_1$ is négyzetes középben konzisztens.

Az X_1^* eloszlásfüggvénye:

$$\mathbf{P}\left(X_1^* < x\right) = 1 - \left[1 - F_{X,\vartheta}(x)\right]^n = 1 - \left(\frac{\vartheta - x}{\vartheta}\right)^n, \ x \in [0, \vartheta]$$

$$\Longrightarrow \ f_{1,\vartheta}(x) = n\frac{(\vartheta - x)^{n-1}}{\vartheta^n}, \ \ x \in (0, \vartheta)\,.$$

$$\mathbf{E}_\vartheta T_2 = \mathbf{E}_\vartheta X_n^* + \mathbf{E}_\vartheta X_1^* = \frac{n}{n+1}\vartheta + \int\limits_0^\vartheta x\, f_{1,\vartheta}(x)\,\mathrm{d}x = \frac{n}{n+1}\vartheta + \frac{n}{\vartheta^n}\int\limits_0^\vartheta x(\vartheta - x)^{n-1}\,\mathrm{d}x \overset{*}{=}$$

Végrehajtva a $\vartheta - x = y \implies \frac{dx}{dy} = -1$ változócserét,

$$\overset{*}{=} \frac{n}{n+1}\vartheta - \frac{n}{\vartheta^n}\int\limits_\vartheta^0 (\vartheta - y)\, y^{n-1}\,\mathrm{d}y = \frac{n}{n+1}\vartheta + \frac{n}{\vartheta^{n-1}}\,\left[\frac{y^n}{n}\right]_0^\vartheta - \frac{n}{\vartheta^n}\left[\frac{y^{n+1}}{n+1}\right]_0^\vartheta = \vartheta,$$

azaz T_2 is torzítatlan.

$$\boldsymbol{\sigma}^2_\vartheta T_2 = \boldsymbol{\sigma}^2_\vartheta X_n^* + \boldsymbol{\sigma}^2_\vartheta X_1^* + 2\,\mathbf{cov}_\vartheta(X_n^*, X_1^*).$$

X_1^* és X_n^* nem függetlenek, így ki kell számolnunk a kovarianciájukat:

$$\mathbf{P}(X_1^* < x, X_n^* < y) = \mathbf{P}(X_n^* < y) - \mathbf{P}(X_1^* \geq x, X_n^* < y) =$$

$$= [F_{X,\vartheta}(y)]^n - \mathbf{P}(x \leq X_1 < y, x \leq X_2 < y, \ldots, x \leq X_n < y) =$$

$$= [F_{X,\vartheta}(y)]^n - \prod_{i=1}^n \mathbf{P}(x \leq X_i < y) =$$

$$= [F_{X,\vartheta}(y)]^n - [F_{X,\vartheta}(y) - F_{X,\vartheta}(x)]^n\,, \quad x, y \in [0, \vartheta]\,, \ x < y.$$

X_1^* és X_n^* együttes sűrűségfüggvénye így:

$$\frac{\partial^2 \mathbf{P}(X_1^* < x, X_n^* < y)}{\partial x\,\partial y} =$$

$$= n\,(n-1)\ [F_{X,\vartheta}(y) - F_{X,\vartheta}(x)]^{n-2}\, f_{X,\vartheta}(y) f_{X,\vartheta}(x) = n\,(n-1)\,\frac{(y-x)^{n-2}}{\vartheta^n}.$$

$$\mathbf{cov}_\vartheta(X_n^*, X_1^*) = \int\limits_0^\vartheta \int\limits_0^y xyn(n-1)\left(\frac{y-x}{\vartheta}\right)^{n-2}\,\frac{1}{\vartheta^2}\,\mathrm{d}x\mathrm{d}y - \mathbf{E}_\vartheta X_n^* \mathbf{E}_\vartheta X_1^* =$$

$u = y - x$ helyettesítéssel

$$= \int\limits_0^\vartheta \left(\int\limits_0^y (y^2 - y\,u)\, n\,(n-1)\left(\frac{u}{\vartheta}\right)^{n-2}\,\frac{1}{\vartheta^2}\,\mathrm{d}u \right)\,\mathrm{d}y - \frac{n}{(n+1)^2}\,\vartheta^2 =$$

$$= \int\limits_0^\vartheta \frac{y^{n+1}}{\vartheta^n}\,\mathrm{d}y - \frac{n}{(n+1)^2}\vartheta^2 =$$

$$= \frac{1}{n+2}\vartheta^2 - \frac{n}{(n+1)^2}\vartheta^2 = \frac{1}{(n+2)\,(n+1)^2}\vartheta^2.$$

Mivel

$$\mathbf{E}\,(X_1^*)^2 = \vartheta^2 - \frac{2\,n}{n+1}\vartheta^2 + \frac{n}{n+2}\vartheta^2 =$$

$$= \frac{(n^2+3\,n+2-2\,n^2-4\,n+n^2+n)}{(n+1)\,(n+2)}\vartheta^2 = \frac{2}{(n+1)\,(n+2)}\vartheta^2,$$

így:

$$\boldsymbol{\sigma}_\vartheta^2 X_1^* = \frac{2}{(n+1)\,(n+2)}\vartheta^2 - \frac{1}{(n+1)^2}\vartheta^2 = \frac{n}{(n+1)^2\,(n+2)}\vartheta^2.$$

Hasonlóan:

$$\boldsymbol{\sigma}^2 X_n^* = \int_0^\vartheta x^2 \cdot n \cdot \frac{x^{n-1}}{\vartheta^n}\,\mathrm{d}x - \left(\frac{n\cdot\vartheta}{n+1}\right)^2 = \frac{n}{n+2}\vartheta^2 - \frac{n^2}{(n+1)^2}\vartheta^2 = \frac{n\vartheta^2}{(n+1)^2(n+2)}.$$

Tehát

$$\boldsymbol{\sigma}_\vartheta^2 T_2 = \boldsymbol{\sigma}_\vartheta^2 X_n^* + \boldsymbol{\sigma}_\vartheta^2 X_1^* + 2\,\mathbf{cov}_\vartheta(X_n^*, X_1^*) =$$

$$= \frac{n}{(n+1)^2(n+2)}\vartheta^2 + \frac{n}{(n+1)^2(n+2)}\vartheta^2 + \frac{2}{(n+1)^2(n+2)}\vartheta^2 = \frac{2}{(n+1)(n+2)}\vartheta^2 \to 0,$$

T_2 is négyzetes középben konzisztens.

$$\boldsymbol{\sigma}_\vartheta^2 T_3 = \left(\frac{n+1}{n-1}\right)^2 \left(\boldsymbol{\sigma}_\vartheta^2 X_n^* + \boldsymbol{\sigma}_\vartheta^2 X_1^* - 2\,\mathbf{cov}_\vartheta(X_n^*, X_1^*)\right) = \frac{n}{(n-1)(n+2)}\vartheta^2 \to 0,$$

T_3 is négyzetes középben konzisztens. Végül:

$$\mathbf{E}_\vartheta T_3 = \frac{n+1}{n-1}\left(\mathbf{E}_\vartheta X_n^* - \mathbf{E}_\vartheta X_1^*\right) = \frac{n+1}{n-1}\left(\frac{n}{n+1}\vartheta - \frac{1}{n+1}\vartheta\right) = \vartheta \implies \text{torzítatlan.}$$

Az eredményt összegezve: $\boldsymbol{\sigma}_\vartheta^2 T_1 < \boldsymbol{\sigma}_\vartheta^2 T_2 < \boldsymbol{\sigma}_\vartheta^2 T_3 < \boldsymbol{\sigma}_\vartheta^2 T_4$.

2.2.2. példa: A lineáris statisztikák között a $\bar{X}_n$ statisztika a leghatásosabb, azaz, ha tetszőleges $c_1, c_2, \ldots, c_n$, $\sum\limits_{i=1}^n c_i = 1$ valós súlyokkal tekintjük a $t_n = \sum\limits_{i=}^n c_i \cdot X_i$ lineáris becslést, akkor t_n torzítatlan, és $\boldsymbol{\sigma}_\mathbf{P}^2 \bar{X}_n \le \boldsymbol{\sigma}_\mathbf{P}^2 t_n$.

Bizonyítás: Először is megjegyezzük, hogy a $c_i = \frac{1}{n}$ $(i = 1, 2, \ldots, n)$ súlyválasztással az átlagstatisztikát kapjuk, tehát az átlagstatisztika is lineáris becslés. Legyen $\varepsilon_i = c_i - \frac{1}{n}$ $(i = 1, 2, \ldots, n)$. Ekkor

$$\sum_{i=1}^n \varepsilon_i = \sum_{i=1}^n c_i - 1 = 0.$$

Így

$$\boldsymbol{\sigma}_\mathbf{P}^2\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i^2 \boldsymbol{\sigma}_\mathbf{P}^2 X_i = \boldsymbol{\sigma}_\mathbf{P}^2 X \cdot \sum_{i=1}^n c_i^2 = \boldsymbol{\sigma}_\mathbf{P}^2 X \cdot \sum_{i=1}^n \left(\varepsilon_i + \frac{1}{n}\right)^2 =$$

$$= \boldsymbol{\sigma}_\mathbf{P}^2 X \left(\sum_{i=1}^n \varepsilon_i^2 + \frac{2}{n}\sum_{i=1}^n \varepsilon_i + \frac{1}{n}\right) \ge \frac{\boldsymbol{\sigma}_\mathbf{P}^2 X}{n} = \boldsymbol{\sigma}_\mathbf{P}^2 \bar{X}_n$$

■

2.2.2. definíció: Ha a t^* torzítatlan statisztikára igaz, hogy

$$\boldsymbol{\sigma}^2_{\mathbf{P}} t^* = \min_{\substack{\mathbf{E}t=\vartheta \\ \boldsymbol{\sigma}^2_{\mathbf{P}} t<\infty}} \boldsymbol{\sigma}^2_{\mathbf{P}} t \quad (\forall \mathbf{P} \in \mathcal{P}),$$

akkor t^*-ot hatásos becslésnek nevezzük.

A Csebisev-egyenlőtlenségből tudjuk, hogy egy valószínűségi változó annál kisebb mértékben ingadozik a várható értéke körül, minél kisebb a szórása. Ez az oka, hogy a torzítatlan becslések között a hatásos becslés megkeresése a cél, hisz várhatóan ez pontosabb, mint bármely más torzítatlan becslés. A következő tétel azt mondja ki, hogy ha van hatásos becslés, akkor lényegében csak egy van.

2.2.1. tétel: Ha t^* és t^{**} a paraméter hatásos becslései, akkor $\mathbf{P}(t^* = t^{**}) = 1 \ \ (\forall \mathbf{P} \in \mathcal{P})$.

Bizonyítás: Legyen t egy tetszőleges torzítatlan becslés.

$$\mathbf{E}_{\mathbf{P}} t^* = \mathbf{E}_{\mathbf{P}} t^{**} = \mathbf{E}_{\mathbf{P}} t = \vartheta, \ \boldsymbol{\sigma}^2_{\mathbf{P}} t^* = \boldsymbol{\sigma}^2_{\mathbf{P}} t^{**} \le \boldsymbol{\sigma}^2_{\mathbf{P}} t.$$

Ez akkor is igaz, ha $t = \frac{t^*+t^{**}}{2}$. Így

$$\boldsymbol{\sigma}^2_{\mathbf{P}} t^* \le \boldsymbol{\sigma}^2_{\mathbf{P}} \left(\frac{t^*+t^{**}}{2}\right) = \frac{1}{4}\left[\boldsymbol{\sigma}^2_{\mathbf{P}} t^* + \boldsymbol{\sigma}^2_{\mathbf{P}} t^{**} + 2\mathbf{E}_{\mathbf{P}}(t^*-\vartheta)(t^{**}-\vartheta)\right].$$

Innen átrendezés után

$$0 \le \boldsymbol{\sigma}^2_{\mathbf{P}} t^* = \boldsymbol{\sigma}_{\mathbf{P}} t^* \boldsymbol{\sigma}_{\mathbf{P}} t^{**} \le \mathbf{E}_{\mathbf{P}}(t^*-\vartheta)(t^{**}-\vartheta) = \mathbf{cov}(t^*, t^{**}).$$

Viszont tudjuk a Cauchy–Bunyakovszkij–Schwartz-féle egyenlőtlenségből, hogy

$$\mathbf{cov}(t^*, t^{**}) \le \boldsymbol{\sigma}_{\mathbf{P}} t^* \boldsymbol{\sigma}_{\mathbf{P}} t^{**}.$$

Ez csak úgy lehet, ha $\mathbf{cov}(t^*, t^{**}) = \boldsymbol{\sigma}_{\mathbf{P}} t^* \boldsymbol{\sigma}_{\mathbf{P}} t^{**}$, vagyis t^* és t^{**} között 1 valószínűséggel lineáris kapcsolat áll fenn: $\mathbf{P}(t^* = ct^{**}) = 1$ $(\forall \mathbf{P} \in \mathcal{P})$.

Viszont $\boldsymbol{\sigma}^2_{\mathbf{P}} t^* = \boldsymbol{\sigma}^2_{\mathbf{P}}(ct^{**}) = \boldsymbol{\sigma}^2_{\mathbf{P}} t^{**} \implies c^2 = 1, \ \mathbf{cov}(t^*, t^{**}) \ge 0 \implies c = +1$. Ahonnan már következik az állítás.

■

2.2.2. tétel: *(Cramer–Rao-egyenlőtlenség)*
Tegyük fel, hogy az $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ statisztikai minta egyparaméteres $F_{\mathbf{P}}(x) = F_\vartheta(x)$ eloszlásfüggvénye abszolút folytonos: $\exists \ \frac{\mathrm{d}F_\vartheta(x)}{\mathrm{d}x} = f_\vartheta(x), \ \ \vartheta \in (a,b)$. Jelölje

$$L_\vartheta(\mathbf{x}) = L_\vartheta(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_\vartheta(x_i)$$

a minta együttes sűrűségfüggvényét!

Feltételek:

a) $I_n(\vartheta) = \int\limits_{\mathbb{R}^n} \left(\frac{\partial L_\vartheta(\mathbf{x})}{\partial \vartheta}\right)^2 \cdot \frac{1}{L_\vartheta(\mathbf{x})} \,\mathrm{d}\mathbf{x} < \infty$ (Fisher-féle információs mennyiség.)

b) Legyen $g : (a,b) \to \mathbb{R}$ tetszőleges differenciálható függvény.

c) Legyen a $t(\mathbf{X})$ statisztika a $g(\vartheta)$ torzítatlan becslése, azaz $\mathbf{E}_\vartheta(t) = g(\vartheta)$ $(\forall\, \vartheta \in (a,b)\ )$.

d) $\exists\ \boldsymbol{\sigma}^2_\vartheta t = \int\limits_{\mathbb{R}^n} (t(\mathbf{x}) - g(\vartheta))^2 L_\vartheta(\mathbf{x})\,\mathrm{d}\mathbf{x}$.

e) $\frac{\partial}{\partial\vartheta} \int\limits_{\mathbb{R}^n} t^i(\mathbf{x}) L_\vartheta(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int\limits_{\mathbb{R}^n} t^i(\mathbf{x}) \frac{\partial L_\vartheta(\mathbf{x})}{\partial\vartheta}\,\mathrm{d}\mathbf{x},\ (i = 0,1)$.

Ekkor

$$\boldsymbol{\sigma}^2_\vartheta t \geq \frac{[g\prime(\vartheta)]^2}{I_n(\vartheta)}.$$

Bizonyítás: A c) tulajdonságból, mindkét oldalt deriválva ϑ szerint:

$$\frac{\partial}{\partial\vartheta} \int\limits_{\mathbb{R}^n} t(\mathbf{x}) L_\vartheta(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int\limits_{\mathbb{R}^n} t(\mathbf{x}) \frac{\partial L_\vartheta(\mathbf{x})}{\partial\vartheta}\,\mathrm{d}\mathbf{x} = \frac{\mathrm{d}g(\vartheta)}{\mathrm{d}\vartheta}. \tag{*}$$

Másrészt, mivel $L_\vartheta(\mathbf{x})$ együttes sűrűségfüggvény:

$$\int\limits_{\mathbb{R}^n} L_\vartheta(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1.$$

Ezt is deriválva ϑ szerint:

$$\frac{\partial}{\partial\vartheta} \int\limits_{\mathbb{R}^n} L_\vartheta(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int\limits_{\mathbb{R}^n} \frac{\partial L_\vartheta(\mathbf{x})}{\partial\vartheta}\,\mathrm{d}\mathbf{x} = \frac{\partial 1}{\partial\vartheta} = 0.$$

Mindkét oldalt beszorozva $g(\vartheta)$-val:

$$\int\limits_{\mathbb{R}^n} g(\vartheta) \frac{\partial L_\vartheta(\mathbf{x})}{\partial\vartheta}\,\mathrm{d}\mathbf{x} = 0. \tag{**}$$

$(*)$ és $(**)$ különbségét véve:

$$\int\limits_{\mathbb{R}^n} (t(\mathbf{x}) - g(\vartheta)\ ) \frac{\partial L_\vartheta(\mathbf{x})}{\partial\vartheta}\,\mathrm{d}\mathbf{x} = \frac{\mathrm{d}g(\vartheta)}{\mathrm{d}\vartheta}.$$

Most a Cauchy–Bunyakovszkij–Schwarz-féle egyenlőtlenséget alkalmazva:

$$\left(g'(\vartheta)\right)^2 = \left(\int\limits_{\mathbb{R}^n} (t(\mathbf{x}) - g(\vartheta)) \frac{\partial L_\vartheta(\mathbf{x})}{\partial\vartheta}\,\mathrm{d}\mathbf{x} \right)^2 =$$

$$= \left(\int\limits_{\mathbb{R}^n} \left((t(\mathbf{x}) - g(\vartheta)) \sqrt{L_\vartheta(\mathbf{x})} \right) \left(\frac{1}{L_\vartheta(\mathbf{x})} \cdot \frac{\partial L_\vartheta(\mathbf{x})}{\partial\vartheta} \cdot \sqrt{L_\vartheta(\mathbf{x})} \right) \mathrm{d}x \right)^2 \leq$$

$$\leq \int\limits_{\mathbb{R}^n} (t(\mathbf{x}) - g(\vartheta)\ )^2 L_\vartheta(\mathbf{x})\,\mathrm{d}\mathbf{x} \int\limits_{\mathbb{R}^n} \left(\frac{1}{L_\vartheta(\mathbf{x})} \frac{\partial L_\vartheta(\mathbf{x})}{\partial\vartheta} \right)^2 L_\vartheta(\mathbf{x}) d\mathbf{x} = \boldsymbol{\sigma}^2_\vartheta t\, I_n(\vartheta).$$

Innen átosztással, már következik az állítás.

■

Megjegyzés:

1. A Cramer–Rao-egyenlőtlenség elvi alsó korlátot ad a torzítatlan becslések szórásnégyzeteire. Ha tehát egy statisztikára belátjuk, hogy szórásnégyzete éppen az alsó korláttal egyenlő, akkor az biztosan hatásos, sőt a 2.2.1. tétel szerint az egyetlen hatásos becslés.

2. A bizonyítás során felhasznált Cauchy–Bunyakovszkij–Schwarz-egyenlőtlenségben akkor és csak akkor van egyenlőség, ha $\exists\, v(\vartheta) : \frac{1}{L_\vartheta(\mathbf{x})}\frac{\partial L_\vartheta(\mathbf{x})}{\partial\vartheta}\left(= \frac{\partial \ln\, L_\vartheta(\mathbf{x})}{\partial\vartheta}\right) = v(\vartheta)\cdot(t(\mathbf{x}) - g(\vartheta))$ majdnem minden $\mathbf{x}$-re fennáll.

3. Ha speciálisan $g(\vartheta) = \vartheta$, akkor $\boldsymbol{\sigma}^2_\vartheta t \geq \frac{1}{I_n(\vartheta)}$.

4. Mivel
$$L_\vartheta(\mathbf{x}) = L_\vartheta(x_1, x_2, \ldots, x_n) = \prod_{i=1}^n f_\vartheta(x_i) \implies \ln\, L_\vartheta(\mathbf{x}) = \sum_{i=1}^n \ln\, f_\vartheta(x_i).$$
Ebből
$$I_n(\vartheta) = \boldsymbol{\sigma}^2_\vartheta\left(\frac{\partial \ln\, L_\vartheta(\mathbf{X})}{\partial\vartheta}\right) = \boldsymbol{\sigma}^2_\vartheta\left(\sum_{i=1}^n \frac{\partial \ln\, f_\vartheta(X_i)}{\partial\vartheta}\right) = \sum_{i=1}^n \boldsymbol{\sigma}^2_\vartheta\left(\frac{\partial \ln\, f_\vartheta(X_i)}{\partial\vartheta}\right) =$$
$$= n\boldsymbol{\sigma}^2_\vartheta\left(\frac{\partial \ln\, f_\vartheta(X_i)}{\partial\vartheta}\right) = nI_1(\vartheta).$$
A levezetésben a szumma kiemelését a mintaelemek teljes függetlensége miatt tehetjük meg.

5. A Cramer–Rao-egyenlőtlenség diszkrét valószínűségeloszlások esetén is érvényben marad, ha $L_\vartheta(\mathbf{x}) = L_\vartheta(x_1, x_2, \ldots, x_n)$-t mint a minta együttes eloszlását értelmezzük:
$$L_\vartheta(\mathbf{x}) = L_\vartheta(x_1, x_2, \ldots, x_n) = \mathbf{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$
A feltételekben a többes integrálok helyett többszörös szummákat kell venni, az e) regularitási tulajdonságok a deriválás és az összegzés sorrendjének felcserélhetőségét követelik meg.

6. A Cramer–Rao-egyenlőtlenség az elemi $(\mathbf{cov}(X,Y))^2 \leq \boldsymbol{\sigma}^2 X \cdot \boldsymbol{\sigma}^2 Y$ egyenlőtlenségnek felel meg, amikor $X = t$, $Y = \frac{\partial \ln L_\vartheta(\mathbf{X})}{\partial\vartheta}$. Ugyanis
$$\mathbf{cov}\left(t, \frac{\partial \ln L_\vartheta(\mathbf{X})}{\partial\vartheta}\right) = \mathbf{E}_\vartheta\left(t \cdot \frac{\partial \ln L_\vartheta(\mathbf{X})}{\partial\vartheta}\right),$$
mert
$$\mathbf{E}_\vartheta\left(\frac{\partial \ln L_\vartheta(\mathbf{X})}{\partial\vartheta}\right) = \int\limits_{\mathbb{R}^n} \frac{\partial L_\vartheta(\mathbf{x})}{\partial\vartheta}\,\mathrm{d}\mathbf{x} = 0.$$
Így
$$\mathbf{E}_\vartheta\left(t \cdot \frac{\partial \ln L_\vartheta(\mathbf{X})}{\partial\vartheta}\right) = \int\limits_{\mathbb{R}^n} t(\mathbf{x}) \cdot \frac{1}{L_\vartheta(\mathbf{x})} \cdot \frac{\partial L_\vartheta(\mathbf{x})}{\partial\vartheta} \cdot L_\vartheta(\mathbf{x})\,\mathrm{d}\mathbf{x} =$$
$$= \frac{\partial}{\partial\vartheta}\int\limits_{\mathbb{R}^n} t(\mathbf{x}) \cdot L_\vartheta(\mathbf{x})\,\mathrm{d}\mathbf{x} = g'(\vartheta).$$

7. Belátható, hogy $I_n(\vartheta) = \boldsymbol{\sigma}^2_\vartheta\left(\frac{\partial \ln L_\vartheta(\mathbf{X})}{\partial\vartheta}\right)$, hiszen

$$\boldsymbol{\sigma}^2_\vartheta\left(\frac{1}{L_\vartheta(\mathbf{X})}\frac{\partial L_\vartheta(\mathbf{X})}{\partial\vartheta}\right) = \mathbf{E}_\vartheta\left(\frac{1}{L_\vartheta(\mathbf{X})}\frac{\partial L_\vartheta(\mathbf{X})}{\partial\vartheta}\right)^2 - \left(\mathbf{E}_\vartheta\left(\frac{1}{L_\vartheta(\mathbf{X})}\frac{\partial L_\vartheta(\mathbf{X})}{\partial\vartheta}\right)\right)^2,$$

de $\int\limits_{\mathbb{R}^n}\frac{\partial L_\vartheta(\mathbf{x})}{\partial\vartheta}\,\mathrm{d}\mathbf{x} = 0$ miatt

$$\mathbf{E}_\vartheta\left(\frac{1}{L_\vartheta(\mathbf{X})}\frac{\partial L_\vartheta(\mathbf{X})}{\partial\vartheta}\right) = \int\limits_{\mathbb{R}^n}\frac{1}{L_\vartheta(\mathbf{x})}\frac{\partial L_\vartheta(\mathbf{x})}{\partial\vartheta}L_\vartheta(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int\limits_{\mathbb{R}^n}\frac{\partial L_\vartheta(\mathbf{x})}{\partial\vartheta}\,\mathrm{d}\mathbf{x} = 0$$

és így

$$\boldsymbol{\sigma}^2_\vartheta\left(\frac{1}{L_\vartheta(\mathbf{X})}\frac{\partial L_\vartheta(\mathbf{X})}{\partial\vartheta}\right) = \mathbf{E}_\vartheta\left(\frac{1}{L_\vartheta(\mathbf{X})}\frac{\partial L_\vartheta(\mathbf{X})}{\partial\vartheta}\right)^2 =$$
$$= \int\limits_{\mathbb{R}^n}\left(\frac{1}{L_\vartheta(\mathbf{x})}\frac{\partial L_\vartheta(\mathbf{x})}{\partial\vartheta}\right)^2 L_\vartheta(\mathbf{x})\,\mathrm{d}\mathbf{x} = I_n(\vartheta).$$

2.2.3. példa: *(Az átlagstatisztika hatásossága normális esetben)*
Legyen az X valószínűségi változó adott. Legyen továbbá $X_1, X_2, \ldots, X_n, \ldots$ statisztikai minta, amelynek eloszlásfüggvénye X-ével azonos valamilyen m, D_0 paraméterű normális eloszláshoz tartozik, ahol $D_0 > 0$ ismert, m ismeretlen. Ennél a feladatnál az ismeretlen paraméter tehát a normális eloszlás várható értéke: $\vartheta = m = \mathbf{E}_\mathbf{P}X$.

A 2.1.1. példában bizonyítottuk, hogy általában az $\bar{X}_n$ átlagstatisztika az m torzítatlan becslése. A normális eloszlásnak valamennyi momentuma létezik, tehát $\bar{X}_n$ négyzetes középben konzisztens becslés is. A Cramer–Rao-egyenlőtlenség segítségével most megmutatjuk, hogy hatásos is. A minta együttes sűrűségfüggvénye most:

$$L_m(\mathbf{x}) = \prod_{i=1}^n \varphi_{m,D_0}(x_i) = \left(\frac{1}{\sqrt{2\pi}D_0}\right)^n \mathrm{e}^{-\frac{1}{2D_0^2}\sum\limits_{i=1}^n(x_i-m)^2}.$$

A Cramer–Rao-tétel utáni 2. megjegyzést figyelembe véve:

$$\ln L_m(\mathbf{x}) = \sum_{i=1}^n \ln\ \varphi_{m,D_0}(x_i) = -n\ln(\sqrt{2\pi}D_0) - \frac{1}{2D_0^2}\sum_{i=1}^n(x_i-m)^2,$$

$$\frac{\partial \ln L_m(\mathbf{x})}{\partial m} = \sum_{i=1}^n\frac{\partial}{\partial m}\ln\ \varphi_{m,D_0}(x_i) = \frac{1}{D_0^2}\sum_{i=1}^n(x_i-m) = \frac{n}{D_0^2}(\bar{x}_n - m) \Longrightarrow \bar{x}_n \text{ hatásos.}$$

2.2.4. példa: *(Az átlagstatisztika hatásossága exponenciális esetben)*
Legyen X egy valószínűségi változó. Legyen továbbá $X_1, X_2, \ldots, X_n$ statisztikai minta, amelynek eloszlásfüggvénye X-ével azonos valamilyen ismeretlen $\lambda > 0$ paraméterű exponenciális eloszláshoz tartozik. $\vartheta = \frac{1}{\lambda} = \mathbf{E}_\mathbf{P}X$. A minta együttes sűrűségfüggvénye most:

$$L_\vartheta(\mathbf{x}) = \prod_{i=1}^n \lambda\mathrm{e}^{-\lambda x_i} = \lambda^n\mathrm{e}^{-\lambda\sum\limits_{i=1}^n x_i} = \frac{1}{\vartheta^n}\mathrm{e}^{-\frac{1}{\vartheta}\sum\limits_{i=1}^n x_i}.$$

$$\frac{\partial \ln L_\vartheta(\mathbf{x})}{\partial\vartheta} = \frac{\partial}{\partial\vartheta}\left(\ln\ \frac{1}{\vartheta^n}\mathrm{e}^{-\frac{1}{\vartheta}\sum\limits_{i=1}^n x_i}\right) = \frac{\partial}{\partial\vartheta}\left(-n\ln\vartheta - \frac{1}{\vartheta}\sum_{i=1}^n x_i\right) = -\frac{n}{\vartheta} + \frac{1}{\vartheta^2}\sum_{i=1}^n x_i =$$
$$= \frac{n}{\vartheta^2}(\overline{x}_n - \vartheta) \Longrightarrow \overline{x}_n \text{ hatásos becslés.}$$

2.2.5. példa: *(Az átlagstatisztika hatásossága a Poisson-eloszlás esetében)*
Legyen X diszkrét valószínűségi változó. Legyen továbbá $X_1, X_2, \ldots, X_n$ statisztikai minta, amelynek eloszlásfüggvénye X-ével azonos valamilyen ismeretlen $\lambda > 0$ paraméterű Poisson-eloszláshoz tartozik. Ennél a példánál legyen az ismeretlen paraméter a Poisson-eloszlás elméleti várható értéke: $\vartheta = \lambda = \mathbf{E_P}X$. $L_\vartheta(\mathbf{x})$ most a minta együttes eloszlása lesz:

$$L_\vartheta(\mathbf{x}) = \mathbf{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} \mathbf{P}(X_i = x_i) = \prod_{i=1}^{n} \frac{\vartheta^{x_i}}{(x_i)!} \mathrm{e}^{-\vartheta} =$$

$$= \frac{\vartheta^{\sum\limits_{i=1}^{n} x_i}}{\prod\limits_{i=1}^{n} (x_i)!} \mathrm{e}^{-n\vartheta} \implies \ln L_\vartheta(\mathbf{X}) = (\ln \vartheta) \sum_{i=1}^{n} x_i - \ln \left(\prod_{i=1}^{n} (x_i)! \right) - n \cdot \vartheta.$$

ϑ szerinti deriválás után:

$$\frac{\partial \ln L_\vartheta(\mathbf{x})}{\partial \vartheta} = \frac{1}{\vartheta} \sum_{i=1}^{n} x_i - n = \frac{n}{\vartheta}(\overline{x}_n - \vartheta) \implies \overline{x}_n \text{ hatásos becslése } \vartheta\text{-nak.}$$

2.2.6. példa: *(Az egyenletes eloszlás esete)*
Legyen most az $X_1, X_2, \ldots, X_n$ minta eloszlása $U(0, \vartheta)$, ahol $\vartheta > 0$ ismeretlen paraméter. Láttuk a 2.2.1. példában, hogy a $T_1 = \frac{n+1}{n} X_n^*$ statisztika torzítatlan becslés volt $g(\vartheta) = \vartheta$-ra, ahol $\boldsymbol{\sigma}^2 T_1 = \frac{\vartheta^2}{n(n+2)}$. Számoljuk ki ebben az esetben az $\frac{1}{I_n(\vartheta)}$ információs alsó határt!

$$I_1(\vartheta) = \int_0^{\vartheta} \frac{\left(\frac{\partial}{\partial \vartheta} \frac{1}{\vartheta}\right)^2}{\frac{1}{\vartheta}} \, \mathrm{d}x = \frac{1}{\vartheta^2},$$

azaz

$$\frac{1}{I_n(\vartheta)} = \frac{1}{n I_1(\vartheta)} = \frac{\vartheta^2}{n}.$$

Az a meglepő eredményt kaptuk, hogy a T_1 torzítatlan becslés szórásnégyzete kisebb, mint a Cramer–Rao-tételben az információs alsó határ!

Az ellentmondás abból adódik, hogy az egyenletes eloszlás esetén nem teljesülnek a Cramer–Rao-tétel e) regularitási feltételei. Most

$$L_\vartheta(\mathbf{x}) = \left(\frac{1}{\vartheta}\right)^n, \quad \forall x_i \in (0, \vartheta),$$

és

$$\frac{\partial}{\partial \vartheta} \int_{\mathbb{R}^n} L_\vartheta(\mathbf{x}) \, \mathrm{d}\mathbf{x} = 0,$$

míg

$$\int_{\mathbb{R}^n} \frac{\partial}{\partial \vartheta} L_\vartheta(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \int_0^{\vartheta} \int_0^{\vartheta} \cdots \int_0^{\vartheta} -n \cdot \frac{1}{\vartheta^{n+1}} \, \mathrm{d}\mathbf{x} = -\frac{n}{\vartheta}.$$

2.3. Elégségesség

A statisztikák elvárt, jó tulajdonságai között alapvető fontosságú az elégségesség. Ezen azt fogjuk érteni, hogy a statisztika a minta eloszlásának paraméterére vonatkozóan minden információt magába sűrít, egymaga képes helyettesíteni a mintát. A paraméterek becsléseihez a megfelelő statisztikákat „elégséges" az elégséges statisztika függvényei között keresni.

2.3.1/a. definíció: Legyen adott a $\mathcal{P}$ paraméteres eloszláscsalád, és az $X_1, X_2, \ldots, X_n$ statisztikai minta, amelyek eloszlásfüggvénye abszolút folytonos $\forall \mathbf{P}_\vartheta \in \mathcal{P}$-re:

$$F_\vartheta(x) = \int\limits_{-\infty}^{x} f_\vartheta(t)\,\mathrm{d}t, \quad x \in \mathbb{R}.$$

$f_\vartheta(x)$ a minta sűrűségfüggvénye. Jelölje a $t_n(X_1, X_2, \ldots, X_n)$ statisztika sűrűségfüggvényét $g_{n,\vartheta}(y)$, az $X_1, X_2, \ldots, X_n$ és t_n együttes sűrűségfüggvényét pedig $h_\vartheta(x_1, x_2, \ldots, x_n, y)$. Ha az $X_1, X_2, \ldots, X_n$ mintának a t_n-re vonatkozó együttes feltételes sűrűségfüggvénye nem tartalmazza a ϑ paramétert, vagyis

$$f_{X_1,X_1,\ldots,X_1|t_n}(x_1, x_2, \ldots, x_n \mid y) = \frac{h_\vartheta(x_1, x_2, \ldots, x_n, y)}{g_{n,\vartheta}(y)},$$

nem függ ϑ-tól, akkor , a t_n statisztika a ϑ paraméter elégséges becslése.

2.3.1/b. definíció: Legyen adott a $\mathcal{P} = \{\mathbf{P}_\vartheta,\ \vartheta \in \Theta\}$, valószínűségi mértékek egy tere és az $X_1, X_2, \ldots, X_n$ statisztikai minta, amelyek eloszlása diszkrét $\forall \mathbf{P}_\vartheta \in \mathcal{P}$-re. Legyen $t_n(X_1, X_2, \ldots, X_n)$ statisztika. Ha az $X_1, X_2, \ldots, X_n$ mintának a t_n-re vonatkozó együttes feltételes eloszlása nem tartalmazza a ϑ paramétert, vagyis

$$\mathbf{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid t_n = y) = \frac{\mathbf{P}_\vartheta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, t_n = y)}{\mathbf{P}_\vartheta(t_n = y)},$$

nem függ ϑ-tól, akkor a t_n statisztika a ϑ paraméter elégséges becslése.

2.3.1. példa: *(Az átlagstatisztika elégségessége normális esetben)*
Legyen X valószínűségi változó. Legyen továbbá $X_1, X_2, \ldots, X_n$ statisztikai minta, amelynek eloszlásfüggvénye X-ével azonos valamilyen m, D_0 paraméterű normális eloszláshoz tartozik $\forall \mathbf{P} \in \mathcal{P}$-re, ahol $D_0 > 0$ ismert, m ismeretlen. Az ismeretlen paraméter a normális eloszlás elméleti várható értéke: $\vartheta = m = \mathbf{E_P}X$.

Az átlagstatisztika teljesen független, $N(\frac{\vartheta}{n}, \frac{D_0}{n})$ eloszlású valószínűségi változók konvolúciója, tehát maga is normális eloszlású, ϑ és $\frac{D_0}{\sqrt{n}}$ paraméterekkel. Így az $X_1, X_2, \ldots, X_n$ minta együttes $\bar{X}_n = y$-ra vett feltételes sűrűségfüggvénye:

$$f_{X_1,X_2,\ldots,X_n|\bar{X}_n}(x_1, x_2, \ldots, x_n \mid y) = \begin{cases} \frac{f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)}{f_{\bar{X}_n}(y)}, & \text{ha } n\,y = \sum\limits_{i=1}^{n} x_i \\ 0 & \text{egyébként} \end{cases}.$$

Mivel

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \left(\sqrt{2\pi}\right)^{-n} D_0^{-n} \mathrm{e}^{-\frac{1}{2D_0^2}\sum\limits_{i=1}^{n}(x_i-\vartheta)^2}$$

és

$$f_{\bar{X}_n}(y) = \frac{\sqrt{n}}{\sqrt{2\pi}D_0}\mathrm{e}^{-\frac{n}{2D_0^2}(y-\vartheta)^2},$$

ezért

$$f_{X_1,X_2,\ldots,X_n|\bar{X}_n}(x_1,x_2,\ldots,x_n \mid y) =$$

$$= \begin{cases} \frac{1}{\sqrt{n}\left(\sqrt{2\pi}D_0\right)^{n-1}}\mathrm{e}^{-\frac{1}{2D_0^2}\left[\sum\limits_{i=1}^{n}(x_i-\vartheta)^2-n(y-\vartheta)^2\right]}, & \text{ha } n\,y=\sum\limits_{i=1}^{n}x_i \\ 0 & \text{egyébként} \end{cases}.$$

Mivel $\sum\limits_{i=1}^{n}(x_i-\vartheta)^2-n(y-\vartheta)^2=\sum\limits_{i=1}^{n}x_i^2-n\,y^2$, ha $\sum\limits_{i=1}^{n}x_i=n\,y \implies$ a feltételes sűrűségfüggvény nem függ a paramétertől, amiből már következik az állítás.

2.3.2. példa: *(Az átlagstatisztika elégségessége exponenciális esetben)*
Legyen az X valószínűségi változó adott. Legyen továbbá $X_1X_2,\ldots,X_n$ statisztikai minta, amely eloszlásfüggvénye X-ével azonos valamilyen $\vartheta=\frac{1}{\lambda}$ paraméterű exponenciális eloszláshoz tartozik. Az ismeretlen paraméter tehát az exponenciális eloszlás várható értéke: $\mathbf{E}_{\vartheta}X=\vartheta$. Az átlagstatisztika teljesen független, $E\left(\frac{1}{\vartheta}\right)$ eloszlású valószínűségi változók konvolúciója, eloszlása $n,\frac{n}{\vartheta}$ paraméterű gamma eloszlás, melynek sűrűségfüggvénye:

$$f_{\bar{X}_n}(x)=\left(\frac{n}{\vartheta}\right)^n\frac{x^{n-1}\mathrm{e}^{-\frac{nx}{\vartheta}}}{(n-1)!}\quad x>0.$$

A minta együttes sűrűségfüggvénye most

$$f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)=\prod_{i=1}^{n}f_{X_i}(x_i)=\frac{1}{\vartheta^n}\mathrm{e}^{-\frac{\sum\limits_{i=1}^{n}x_i}{\vartheta}}\quad \forall x_i>0.$$

Az

$$f_{X_1,X_2,\ldots,X_n|\bar{X}_n}(x_1,x_2,\ldots,x_n \mid y)=\begin{cases}\frac{f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)}{f_{\bar{X}_n}(y)}, & \text{ha } n\,y=\sum\limits_{i=1}^{n}x_i \\ 0 & \text{egyébként}\end{cases}$$

képletbe behelyettesítve:

$$f_{X_1,X_2,\ldots,X_n|\bar{X}_n}(x_1,x_2,\ldots,x_n \mid y)=\begin{cases}\frac{\frac{1}{\vartheta^n}\mathrm{e}^{-\frac{\sum\limits_{i=1}^{n}x_i}{\vartheta}}}{\left(\frac{n}{\vartheta}\right)^n\frac{y^{n-1}\mathrm{e}^{-\frac{ny}{\vartheta}}}{(n-1)!}}, & \text{ha } n\,y=\sum\limits_{i=1}^{n}x_i \\ 0 & \text{egyébként}\end{cases}.$$

Egyszerűsítések után:

$$f_{X_1,X_2,\ldots,X_n|\bar{X}_n}(x_1,x_2,\ldots,x_n \mid y)=\begin{cases}\frac{(n-1)!\mathrm{e}^{-\frac{\sum\limits_{i=1}^{n}x_i}{\vartheta}+\frac{ny}{\vartheta}}}{n^n y^{n-1}}, & \text{ha } n\,y=\sum\limits_{i=1}^{n}x_i \\ 0 & \text{egyébként}\end{cases} =$$

$$=\begin{cases}\frac{(n-1)!}{n^n\ y^{n-1}}, & \text{ha } n\,y=\sum\limits_{i=1}^{n}x_i \\ 0 & \text{egyébként}\end{cases}.$$

Látható, hogy a függvény nem függ a paramétertől, azaz az átlagstatisztika ebben az esetben is elégséges becslést ad.

2.3.3. példa: *(Az átlagstatisztika elégségessége a Poisson-eloszlás esetében)*
Legyen az X diszkrét valószínűségi változó adott. Legyen továbbá $X_1, X_2, \dots, X_n$ statisztikai minta, amelynek eloszlásfüggvénye X-ével azonos valamilyen $\vartheta > 0$ paraméterű Poisson-eloszláshoz tartozik. Az ismeretlen paraméter tehát a Poisson-eloszlás várható értéke: $\vartheta = \mathbf{E}_P X$. Az átlagstatisztika eloszlása most:

$$\mathbf{P}(n\,\bar{X}_n = y) = \mathbf{P}\left(\sum_{i=1}^{n} X_i = y\right) = \frac{(n\vartheta)^y}{y!}\mathrm{e}^{-n\vartheta} \qquad y = 0, 1, 2, \dots\ .$$

A minta együttes eloszlása:

$$\mathbf{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^{n} \mathbf{P}(X_i = x_i) = \frac{\vartheta^{\sum\limits_{i=1}^{n} x_i}}{\prod\limits_{i=1}^{n} x_i!}\mathrm{e}^{-n\vartheta}.$$

Így a mintának az átlagra vonatkozó feltételes eloszlása:

$$\mathbf{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n \mid n\bar{X}_n = y) = \frac{\prod\limits_{i=1}^{n} \mathbf{P}(X_i = x_i)}{\mathbf{P}\left(n\bar{X}_n = y\right)} = \frac{y!}{n^y \prod\limits_{i=1}^{n} x_i!},$$

ha $y = \sum\limits_{i=1}^{n} x_i$, ami nem függ a paramétertől, azaz az átlagstatisztika a Poisson-eloszlás esetén is elégséges.

2.3.4. példa: *(Példa nemelégséges statisztikára)*
Vizsgáljuk meg a $t = X_1$ „statisztikát"! Most

$$\mathbf{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n \,|X_1 = y\,) = \begin{cases} \frac{\prod\limits_{i=1}^{n} \mathbf{P}_\vartheta(X_i = x_i)}{\mathbf{P}_\vartheta(X_1 = y)}, & \text{ha } x_1 = y \\ 0, & \text{ha } x_1 \neq y \end{cases} =$$

$$= \begin{cases} \prod\limits_{i=2}^{n} \mathbf{P}_\vartheta(X_i = x_i), & \text{ha } x_1 = y \\ 0, & \text{ha } x_1 \neq y \end{cases}$$

ami látható, hogy tartalmazza a paramétert.

2.3.1. tétel: *(Rao–Blackwell–Kolmogorov)*
Legyen adott $\mathcal{P}$, valószínűségi mértékek egy ϑ-paraméteres tere, és az $X_1, X_2, \dots, X_n$ statisztikai minta, amelyek eloszlásfüggvénye abszolút folytonos $\forall \mathbf{P} \in \mathcal{P}$-re. Jelölje $T_n(X_1, X_2, \dots, X_n)$ a ϑ paraméter egy elégséges statisztikáját, $t_n(X_1, X_2, \dots, X_n)$ pedig a paraméter g függvényének tetszőleges torzítatlan becslését: $\mathbf{E}_\vartheta t_n = g(\vartheta)$. Akkor létezik olyan h függvény, hogy $\mathbf{E}_\vartheta(h(T_n)) = g(\vartheta)$ és $\boldsymbol{\sigma}^2_\vartheta(h(T_n)) \leq \boldsymbol{\sigma}^2_\vartheta t_n$. Továbbá $h(T_n) = \mathbf{E}_\vartheta(t_n \,|T_n\,)$.

Bizonyítás: A $h(T_n)$ nem függ ϑ-tól, csak a mintától, hiszen T_n elégséges statisztika volt. Tehát $h(T_n)$ tényleg statisztika. A feltételes várható érték tulajdonságait felhasználva:

$$\mathbf{E}_\vartheta(h(T_n)) = \mathbf{E}_\vartheta(\mathbf{E}_\vartheta(t_n \,|T_n\,)) = \mathbf{E}_\vartheta t_n = g(\vartheta), \quad h(T_n) \text{ torzítatlan.}$$

Másrészt:

$$\boldsymbol{\sigma}_\vartheta^2 t_n = \mathbf{E}_\vartheta (t_n - g(\vartheta))^2 = \mathbf{E}_\vartheta \left[t_n - h(T_n) + h(T_n) - g(\vartheta)\right]^2 =$$
$$= \mathbf{E}_\vartheta (t_n - h(T_n))^2 + \boldsymbol{\sigma}_\vartheta^2 (h(T_n)) + 2\,\mathbf{E}_\vartheta \left[(t_n - h(T_n))(h(T_n) - g(\vartheta))\right].$$

De

$$\mathbf{E}_\vartheta \left[(t_n - h(T_n))(h(T_n) - g(\vartheta))\right] = \mathbf{E}_\vartheta \left[\mathbf{E}_\vartheta \left[(t_n - h(T_n))(h(T_n) - g(\vartheta))\,|T_n\right]\right] =$$
$$= \mathbf{E}_\vartheta \left[(h(T_n) - g(\vartheta))\mathbf{E}_\vartheta \left[(t_n - h(T_n))\,|T_n\right]\right] = 0,$$

mert

$$\mathbf{E}_\vartheta \left[(t_n - h(T_n))\,|T_n\right] = \mathbf{E}_\vartheta \left[t_n\,|T_n\right] - h(T_n) = 0.$$

Innen már $\boldsymbol{\sigma}_\vartheta^2 t_n \geq \boldsymbol{\sigma}_\vartheta^2 (h(T_n))$ adódik.

■

Ha létezik hatásos becslés, akkor az az elégséges becslés függvényeként áll elő. A tétel azt nem állítja, hogy a $h(T_n)$ már hatásos lenne, csak azt, hogy egy tetszőlegesen adott t_n torzítatlan becslésnél az elégséges statisztika segítségével lehet hatásosabbat előállítani, de az nem biztos, hogy egyben hatásos is!

2.3.2. tétel: *(Neymann–Fisher faktorizációs tétel)*
Legyen adott $\mathcal{P}$, valószínűségi mértékek egy ϑ-paraméteres tere, amelyhez adott az $X_1, X_2, \ldots, X_n$ statisztikai minta, amelyek eloszlásfüggvénye abszolút folytonos $\forall \mathbf{P} \in \mathcal{P}$-re.

A T_n statisztika a ϑ paraméter elégséges becslése $\iff$ $\exists\ k : \mathbb{R}^n \to \mathbb{R}$ és $g : \mathbb{R}^2 \to \mathbb{R}$ függvények, hogy $\forall \mathbf{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$ és $\forall\vartheta$-ra

$$L_\vartheta(x_1, x_2, \ldots, x_n) = k(x_1, x_2, \ldots, x_n) g(T_n(x_1, x_2, \ldots, x_n), \vartheta).$$

Bizonyítás: Nem bizonyítjuk. A bizonyítás megtalálható Lehman: Testing Statistical Hipotheses, 49. old.

2.3.5. példa: *(A faktorizációs tétel alkalmazása egyenletes eloszlásra)*
Legyen az $X_1, \ldots, X_n$ statisztikai minta egyenletes eloszlású a $(0, \vartheta)$ intervallumon. Ekkor a minta együttes sűrűségfüggvénye

$$L_\vartheta(\mathbf{x}) = \frac{1}{\vartheta^n} \cdot \prod_{i=1}^{n} u(0, x_i) u(x_i, \vartheta)$$

alakban írható, ahol

$$u(a, b) = \begin{cases} 1, & \text{ha } a < b \\ 0 & \text{egyébként} \end{cases}.$$

Mivel az $X_1 < \vartheta, X_2 < \vartheta, \ldots, X_n < \vartheta \iff X_n^* = \max\{X_i\} < \vartheta$, ezért

$$\prod_{i=1}^{n} u(x_i, \vartheta) = u(x_n^*, \vartheta).$$

Így

$$L_\vartheta(\mathbf{x}) = \left(\frac{1}{\vartheta^n} \cdot u(x_n^*, \vartheta)\right) \cdot \prod_{i=1}^{n} u(0, x_i),$$

azaz teljesül a faktorizációs tétel az n-edik rendezett mintaelem statisztikára. Beláttuk tehát, hogy az $X_n^* = \max\{X_1, \ldots, X_n\}$ statisztika elégséges a ϑ paraméterre.

A maradék $n-1$ elemű mintának az $X_n^* = t$ feltételre vonatkoztatott sűrűségfüggvénye nem függ a ϑ paramétertől. Megmutatható, hogy ez a sűrűségfüggvény $\prod_{i=1}^{n} \frac{f_\vartheta(x_i)}{F_\vartheta(t)}$ alakú, most speciálisan $\left(\frac{1}{t}\right)^{n-1}$. Vagyis a maradék minta egyenletes eloszlású a $(0,t)$ intervallumban. Ha szimulálunk $n-1$ véletlen számot a $(0,t)$-n, az t-vel együtt statisztikailag ekvivalens mintát fog alkotni, mint az eredeti $X_1, \dots, X_n$, amelynek eloszlása még függött ϑ-tól. Az X_n^* teljes statisztika „képviseli" a ϑ paramétert, jobban mondva magába tömöríti a ϑ-ra vonatkozó információkat.

2.4. Maximum-likelihood becslés

Eddig csak arról volt szó, hogy milyen jó tulajdonságai lehetnek egy statisztikának, de még nem tudjuk, milyen módszerekkel lehet egy adott becslési problémához alkalmas statisztikát előállítani. A következőkben két általános becslési módszert fogunk ismertetni.

2.4.1/a. definíció: Legyen adott $\mathcal{P}$, valószínűségi mértékek egy tere és az $X_1, X_2, \dots, X_n$ statisztikai minta, amelyek eloszlásfüggvénye abszolút folytonos $\forall \mathbf{P}_{\boldsymbol{\vartheta}} \in \mathcal{P}$-re. Jelölje most

$$L(\mathbf{x}, \boldsymbol{\vartheta}) = \prod_{i=1}^{n} f_{\boldsymbol{\vartheta}}(x_i)$$

a minta együttes sűrűségfüggvényét. A $\boldsymbol{\vartheta}$ paraméter maximum-likelihood becslésén azt a $\boldsymbol{\tau}_n(X_1, X_2, \dots, X_n)$ statisztikát értjük, melyre

$$L(\mathbf{x}, \boldsymbol{\tau}_n(\mathbf{x})) = \max_{\boldsymbol{\vartheta} \in \mathbb{R}^k} L(\mathbf{x}, \boldsymbol{\vartheta})$$

teljesül ($\forall \mathbf{x} \in \mathbb{R}^n$).

2.4.1/b. definíció: Legyen adott $\mathcal{P}$, valószínűségi mértékek egy tere és az $X_1, X_2, \dots, X_n$ diszkrét eloszlású statisztikai minta $E \subseteq \mathbb{R}$ értékkészlettel $\forall \mathbf{P}_{\boldsymbol{\vartheta}} \in \mathcal{P}$-re.

Jelölje most

$$L(\mathbf{x}, \boldsymbol{\vartheta}) = \mathbf{P}_{\boldsymbol{\vartheta}}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^{n} \mathbf{P}_{\boldsymbol{\vartheta}}(X_i = x_i)$$

a minta együttes eloszlását. A $\boldsymbol{\vartheta}$ paraméter maximum-likelihood becslésén azt a $\boldsymbol{\tau}_n(X_1, X_2, \dots, X_n)$ statisztikát értjük, melyre

$$L(\mathbf{x}, \boldsymbol{\tau}_n(\mathbf{x})) = \max_{\boldsymbol{\vartheta} \in \mathbb{R}^k} L(\mathbf{x}, \boldsymbol{\vartheta})$$

teljesül ($\forall \mathbf{x} \in E^n$).

Megjegyzés:

1. $L(\mathbf{x}, \boldsymbol{\vartheta})$-t likelihood függvénynek is nevezik. Az elnevezés jogos, mert most az együttes sűrűségfüggvényben nem $\mathbf{x}$-et, hanem $\boldsymbol{\vartheta}$-t tekintjük változónak.

2. A módszer alapgondolata a következő: mintavételezés során az $\mathbf{x}$ realizációt kaptuk. Feltételezzük, hogy azért éppen ezt a realizációt kaptuk, és nem mást, mert az összes realizációk közül ennek a legnagyobb a bekövetkezési valószínűsége. Vegyük tehát, az összes $\boldsymbol{\vartheta}$ paramétervektor közül azt, amelynél éppen az $\mathbf{x}$ realizáció bekövetkezése a maximális. A választ mind a folytonos, mind a diszkrét esetben a $L(\mathbf{x}, \boldsymbol{\vartheta}) \to \max_{\boldsymbol{\vartheta} \in \mathbb{R}^k}$ szélsőérték-feladat megoldásából kapjuk meg.

3. Mivel a természetes alapú logaritmusfüggvény szigorúan monoton növekvő, az $L(\mathbf{x}, \boldsymbol{\vartheta}) \to \max_{\boldsymbol{\vartheta} \in \mathbb{R}^k}$ feladat helyett sokszor célszerű az ln $L(\mathbf{x}, \boldsymbol{\vartheta}) \to \max_{\boldsymbol{\vartheta} \in \mathbb{R}^k}$ szélsőérték-feladatot megoldani, ugyanis ugyanott lépnek fel a maximumhelyek. Az $l(\mathbf{x}, \boldsymbol{\vartheta}) = \ln L(\mathbf{x}, \boldsymbol{\vartheta})$ függvényt log-likelihood függvénynek nevezzük.
$$l(\mathbf{x}, \boldsymbol{\vartheta}) = \sum_{c=1}^{n} \ln f_{\vartheta}(x_i).$$

4. A maximumhelyet az $\frac{\partial l(\mathbf{x}, \boldsymbol{\vartheta})}{\partial \vartheta_i} = 0, \quad i = 1, 2, \ldots, k$ egyenletrendszer megoldásai között kereshetjük.

2.4.1. példa: *(A várható érték maximum-likelihood becslése normális esetben, amikor ismert a szórás.)*

Legyen
$$f_{\vartheta}(x) = \frac{1}{\sqrt{2\,\pi}\,D_0}\,\mathrm{e}^{-\frac{1}{2D_0^2}(x-\vartheta)^2},$$
ahol $D_0 > 0$ ismert, $\vartheta \in \mathbb{R}$ az ismeretlen paraméter.

Most a likelihood függvény:
$$L(\mathbf{x}, \vartheta) = \left(\frac{1}{\sqrt{2\,\pi}\,D_0}\right)^n \mathrm{e}^{-\frac{1}{2D_0^2}\sum\limits_{i=1}^{n}(x_i-\vartheta)^2},$$
a log-likelihood függvény pedig
$$l(\mathbf{x}, \vartheta) = n \ln\left(\frac{1}{\sqrt{2\,\pi}\,D_0}\right) - \frac{1}{2\,D_0^2}\sum_{i=1}^{n}(x_i - \vartheta)^2.$$
A maximumhely keresése:
$$\frac{\mathrm{d}l(\mathbf{x}, \vartheta)}{\mathrm{d}\vartheta} = \frac{1}{D_0^2}\sum_{i=1}^{n}(x_i - \vartheta) = 0 \implies \vartheta = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}_n.$$
Mivel
$$\frac{\mathrm{d}^2 l(\mathbf{x}, \vartheta)}{\mathrm{d}\vartheta^2} = -\frac{1}{D_0^2} < 0,$$
a kapott stacionárius hely maximumhely. Tehát az átlagstatisztika normális esetben a várható érték maximum-likelihood becslése.

2.4.2. példa: *(A várható érték és a szórásnégyzet maximum-likelihood becslései normális esetben.)*

Legyen
$$f_{\vartheta_1,\vartheta_2}(x) = \frac{1}{\sqrt{2\,\pi\,\vartheta_2}}\,\mathrm{e}^{-\frac{1}{2\vartheta_2}(x-\vartheta_1)^2},$$
ahol $\vartheta_2 > 0$ és $\vartheta_1 \in \mathbb{R}$ az ismeretlen paraméterek.

Most a likelihood függvény:
$$L(\mathbf{x}, \vartheta_1, \vartheta_2) = \left(\frac{1}{\sqrt{2\,\pi\,\vartheta_2}}\right)^n \mathrm{e}^{-\frac{1}{2\vartheta_2}\sum\limits_{i=1}^{n}(x_i-\vartheta_1)^2},$$

a log-likelihood függvény pedig

$$l(\mathbf{x},\vartheta_1,\vartheta_2) = -n\ln\sqrt{2\pi} - \frac{n}{2}\ \ln\ \vartheta_2 - \frac{1}{2\,\vartheta_2}\sum_{i=1}^{n}(x_i-\vartheta_1)^2.$$

A maximumhely keresése:

$$\frac{\partial l(\mathbf{x},\vartheta_1,\vartheta_2)}{\partial\vartheta_1} = \frac{1}{\vartheta_2}\sum_{i=1}^{n}(x_i-\vartheta_1) = 0 \Longrightarrow \vartheta_1 = \frac{1}{n}\sum_{i=1}^{n}x_i = \bar{x}_n$$

$$\frac{\partial l(\mathbf{x},\vartheta_1,\vartheta_2)}{\partial\vartheta_2} = -\frac{n}{2\vartheta_2} + \frac{1}{2\vartheta_2^2}\sum_{i=1}^{n}(x_i-\vartheta_1)^2 = 0 \Longrightarrow \vartheta_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\vartheta_1)^2 = s_n^2$$

Mivel

$$\frac{\partial^2 l(\mathbf{x},\vartheta_1,\vartheta_2)}{\partial\vartheta_1^2} = -\frac{n}{\vartheta_2},$$

$$\frac{\partial^2 l(\mathbf{x},\vartheta_1,\vartheta_2)}{\partial\vartheta_2^2} = \frac{n}{2\vartheta_2^2} - \frac{1}{\vartheta_2^3}\sum_{i=1}^{n}(x_i-\vartheta_1)^2,$$

$$\frac{\partial^2 l(\mathbf{x},\vartheta_1,\vartheta_2)}{\partial\vartheta_2\partial\vartheta_1} = -\frac{1}{\vartheta_2^2}\sum_{i=1}^{n}(x_i-\vartheta_1),$$

a kapott stacionárius hely Hesse-mátrixa:

$$\begin{pmatrix} \frac{-n}{s_n^2} & 0 \\ 0 & \frac{-n}{2\,(s_n^2)^2} \end{pmatrix},$$

amiből látszik, hogy a hely maximumhely, tehát az átlagstatisztika és az empirikus szórásnégyzet statisztikák normális esetben az elméleti várható érték és szórásnégyzet maximum-likelihood becslései.

2.4.3. példa: *(A várható érték maximum-likelihood becslése Poisson-eloszlás esetében.)* Most a minta eloszlása:

$$p_{\vartheta,i} = \frac{\vartheta^i}{i!}\mathrm{e}^{-\vartheta} \quad i = 0,1,2,\ldots.$$

A likelihood függvény, a minta együttes eloszlásából számolható:

$$L(\mathbf{x},\vartheta) = \prod_{i=1}^{n}\frac{\vartheta^{x_i}}{x_i!} = \frac{\vartheta^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}\mathrm{e}^{-n\vartheta},$$

a log-likelihood függvény pedig:

$$l(\mathbf{x},\vartheta) = \ln\ \vartheta\sum_{i=1}^{n}x_i - n\vartheta - \ln\left(\prod_{i=1}^{n}x_i!\right).$$

A stacionárius helyek megkeresése:

$$\frac{\partial l(\mathbf{x},\vartheta)}{\partial\vartheta} = \frac{1}{\vartheta}\sum_{i=1}^{n}x_i - n = 0 \Longrightarrow \vartheta = \frac{1}{n}\sum_{i=1}^{n}x_i = \bar{x}_n.$$

Mivel

$$\frac{\partial^2 l(\mathbf{x},\vartheta)}{\partial\vartheta^2} = -\frac{1}{\vartheta^2}\sum_{i=1}^{n} x_i < 0,$$

a kapott stacionáriushely maximum. Tehát a Poisson-eloszlás esetén is a paraméternek maximum-likelihood becslése az átlagstatisztika.

2.4.4. példa: *(Maximum-likelihood becslés egyenletes eloszlás esetén)*
Legyen az $X_1, \dots, X_n$ statisztikai minta eloszlása $U(0,\vartheta)$, ahol $\vartheta > 0$ a becsülendő paraméter. A likelihood függvény most

$$L(\mathbf{x},\vartheta) = \frac{1}{\vartheta^n}\prod_{i=1}^{n} u(0,x_i)u(x_i,\vartheta),$$

ahol

$$U(a,b) = \begin{cases} 1, \text{ha } a \le b \\ 0, \text{ha } a > b \end{cases}.$$

Nyilvánvaló, hogy

$$\max_{\vartheta}\prod_{i=1}^{n} u(x_i,\vartheta) = 1,$$

és ez a maximum eléretik minden $\vartheta \ge \max\{x_1, \dots, x_n\} = x_n^*$ esetén.

Másrészt $\frac{1}{\vartheta^n} \le \left(\frac{1}{x_n^*}\right)^n$, ha $\vartheta \ge x_n^*$. Ezért $L_\vartheta(x_1, \dots, x_n)$ a maximumát éppen a $\tau_n(x_1, \dots, x_n) = x_n^*$ helyen fogja felvenni, tehát ϑ maximum-likelihood becslése az $X_n^* = \max\{X_1, \dots, X_n\}$ maximumstatisztika lesz.

A maximum-likelihood becslés rendelkezik néhány nagyon jó tulajdonsággal, amelyeket a következő két tételben fogalmazunk meg.

2.4.1. tétel: Legyen adott $\mathcal{P}$, valószínűségi mértékek egy tere és az $X_1, X_2, \dots, X_n$ statisztikai minta. Jelölje most $L(\mathbf{x},\vartheta)$ a likelihood függvényt és τ_n a maximum-likelihood statisztikát!

(i) Ha létezik hatásos becslés a ϑ paraméterre, akkor τ_n maga a hatásos becslés.

(ii) Ha létezik T_n elégséges becslés a ϑ paraméterre, akkor megadható olyan $h(x)$ függvény, mellyel $h(\tau_n) = T_n$, azaz az elégséges becslés a maximum-likelihood statisztika függvényeként áll elő.

Bizonyítás:

(i) A Cramer–Rao-tétel után tett 2. megjegyzés szerint t_n^* hatásos becslés, ha

$$\frac{\partial l(\mathbf{x},\vartheta)}{\partial\vartheta} = k(\vartheta)(t_n^*(\mathbf{x}) - \vartheta)$$

teljesül majdnem minden $\mathbf{x} \in \mathbb{R}^n$ vektorra. De a maximum-likelihood statisztikát éppen az $\frac{\partial l(\mathbf{x},\vartheta)}{\partial\vartheta} = 0$ egyenlet megoldásából kapjuk, azaz

$$k(\vartheta)(t_n^*(\mathbf{x}) - \vartheta) = 0 \implies t_n^*(\mathbf{x}) = \tau_n(\mathbf{x}) = \vartheta \implies \text{állítás.}$$

(ii) A Neymann–Fisher faktorizációs tételből: $\exists\, g, k$ függvények:

$$L(\mathbf{x},\vartheta) = g(T_n(\mathbf{x}),\vartheta)\cdot k(\mathbf{x}).$$

Innen $\frac{\partial \ln L(\mathbf{x},\vartheta)}{\partial\vartheta} = \frac{\partial g(T_n(\mathbf{x}),\vartheta)}{\partial\vartheta} = 0 \implies \exists\, h$ függvény: $h(T_n(\mathbf{X})) = \tau_n(\mathbf{x})$.

2.4.2. tétel: *(Cramer–Dugue)*
Legyen adott $\mathcal{P}$, valószínűségi mértékek egy tere és az $X_1, X_2, \ldots, X_n, \ldots$ statisztikai minta, amelyek eloszlásfüggvénye abszolút folytonos $\forall \mathbf{P}\in\mathcal{P}$-re. Tegyük fel, hogy a minta sűrűségfüggvénye $f_\vartheta(x)$, $\vartheta\in(a,b)$ kielégíti az alábbi a), b), c) feltételeket:

a) $\exists\ \frac{\partial^i \ln f_\vartheta(x)}{\partial\vartheta^i}\quad i=1,2,3\ \ \forall\vartheta\in(a,b)$.

b) $\exists H_1(x),\ H_2(x),\ H_3(x)$ függvények, melyekre:

$$\left|\frac{\partial f_\vartheta(x)}{\partial\vartheta}\right| < H_1(x),\ \left|\frac{\partial^2 f_\vartheta(x)}{\partial\vartheta^2}\right| < H_2(x),\ \left|\frac{\partial^3 f_\vartheta(x)}{\partial\vartheta^3}\right| < H_3(x).$$

$$\int_{-\infty}^{+\infty} H_1(x)\,\mathrm{d}x < \infty,\ \int_{-\infty}^{+\infty} H_2(x)\,\mathrm{d}x < \infty,$$

$$\exists K:\ \int_{-\infty}^{+\infty} H_3(x)\cdot f_\vartheta(x)\,\mathrm{d}x < K\ \ \forall\vartheta\in(a,b).$$

c) $0 < I_1(\vartheta) = \int_{-\infty}^{+\infty}\left(\frac{\partial\ \ln\ f_\vartheta(x)}{\partial\vartheta}\right)^2 f_\vartheta(x)\,\mathrm{d}x < \infty$.

Legyen továbbá τ_n a ϑ paraméter maximum-likelihood statisztikája.
Ekkor

(i) τ_n az ϑ paraméter konzisztens becslése,

(ii) τ_n aszimptotikusan normális eloszlású, azaz $\sqrt{nI_1(\vartheta)}\cdot(\tau_n-\vartheta)\xrightarrow{e} N(0,1)$.

Bizonyítás: Az (i) bizonyítása. A b) feltételből következik, hogy a deriválás és az integrálás sorrendje felcserélhető. Így mivel

$$\int_{-\infty}^{+\infty} f_\vartheta(x)\,\mathrm{d}x = 1 \implies \int_{-\infty}^{+\infty}\frac{\partial f_\vartheta(x)}{\partial\vartheta}\,\mathrm{d}x = 0\ ,\ \int_{-\infty}^{+\infty}\frac{\partial^2 f_\vartheta(x)}{\partial\vartheta^2}\,\mathrm{d}x = 0.$$

Legyen $\vartheta_0\in(a,b)$ a tényleges paraméter. A Taylor-formulából kapjuk, hogy:

$$\frac{\partial \ln f_\vartheta(x)}{\partial\vartheta} = \left.\frac{\partial \ln f_\vartheta(x)}{\partial\vartheta}\right|_{\vartheta=\vartheta_0} + \left.\frac{\partial^2 \ln f_\vartheta(x)}{\partial\vartheta^2}\right|_{\vartheta=\vartheta_0}(\vartheta-\vartheta_0) + \frac{1}{2}\,\delta\, H_3(x)\cdot(\vartheta-\vartheta_0)^2,$$

ahol $|\delta| < 1$ (δ esetleg függhet x-től és ϑ-tól is.) Mivel

$$L(\mathbf{x},\vartheta) = \prod_{i=1}^{n} f_\vartheta(x_i),$$

így
$$\frac{1}{n}\frac{\partial \ln L(\mathbf{X},\vartheta)}{\partial\vartheta}=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln f_\vartheta(X_i)}{\partial\vartheta}=B_1+B_2(\vartheta-\vartheta_0)+\frac{1}{2}\Delta\, B_3(\vartheta-\vartheta_0)^2,$$
ahol
$$B_1=\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial \ln f_\vartheta\,(X_i)}{\partial\vartheta}\right|_{\vartheta=\vartheta_0},$$
$$B_2=\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^2 \ln f_\vartheta\,(X_i)}{\partial\vartheta^2}\right|_{\vartheta=\vartheta_0}$$
és
$$B_3=\frac{1}{n}\sum_{i=1}^{n}H_3\,(X_i)\,.$$
Δ a minta és ϑ függvénye, de $|\Delta|<1$. Figyeljük meg, hogy B_1,B_2,B_3 független, azonos eloszlású valószínűségi változók átlagai!

A maximum-likelihood becslés az $\frac{\partial \ln\, L(\mathbf{X},\vartheta)}{\partial\vartheta}=0$ egyenlet megoldásából áll elő, azaz
$$B_1+B_2(\vartheta-\vartheta_0)+\frac{1}{2}\Delta B_3(\vartheta-\vartheta_0)^2=0.$$
Felhasználjuk, hogy
$$\mathbf{E}_{\vartheta_0}\left(\left.\frac{\partial\, \ln\, f_\vartheta(X_i)}{\partial\vartheta}\right|_{\vartheta=\vartheta_0}\right)=0,$$
$$\mathbf{E}_{\vartheta_0}\left(\left.\frac{\partial^2\, \ln\, f_\vartheta(X_i)}{\partial\vartheta^2}\right|_{\vartheta=\vartheta_0}\right)=-I_1,$$
hiszen
$$\mathbf{E}_{\vartheta_0}\left(\left.\frac{\partial \ln f_\vartheta(X_i)}{\partial\vartheta}\right|_{\vartheta=\vartheta_0}\right)=\int_{-\infty}^{\infty}\left.\frac{\partial f_\vartheta(x)}{\partial\vartheta}\right|_{\vartheta=\vartheta_0}\mathrm{d}x=\frac{\partial}{\partial\vartheta}\int_{-\infty}^{\infty}f_\vartheta(x)\,\mathrm{d}x\Bigg|_{\vartheta=\vartheta_0}=0$$
és
$$\mathbf{E}_{\vartheta_0}\left(\frac{1}{f_\vartheta(X_i)}\cdot\left.\frac{\partial^2 f_\vartheta(X_i)}{\partial\vartheta^2}\right|_{\vartheta=\vartheta_0}\right)=\int_{-\infty}^{\infty}\left.\frac{\partial^2 f_\vartheta(x)}{\partial\vartheta}\right|_{\vartheta=\vartheta_0}\mathrm{d}x=\frac{\partial^2}{\partial\vartheta^2}\int_{-\infty}^{\infty}f_\vartheta(x)\,\mathrm{d}x\Bigg|_{\vartheta=\vartheta_0}=0$$
miatt
$$\mathbf{E}_{\vartheta_0}\left(\left.\frac{\partial^2 \ln f_\vartheta(X_i)}{\partial\vartheta^2}\right|_{\vartheta=\vartheta_0}\right)=\mathbf{E}_{\vartheta_0}\left(-\left.\left(\frac{\partial \ln f_\vartheta(X_i)}{\partial\vartheta}\right)^2\right|_{\vartheta=\vartheta_0}\right)=$$
$$=-\boldsymbol{\sigma}^2\vartheta_0\left(\left.\frac{\partial \ln f_\vartheta(X_i)}{\partial\vartheta}\right|_{\vartheta=\vartheta_0}\right)=-I_1.$$
A nagy számok gyenge törvényéből következik, hogy
$$B_1\stackrel{st}{\to}0,\quad B_2\stackrel{st}{\to}-I_1,\quad B_3\stackrel{st}{\to}\mathbf{E}_{\vartheta_0}H_3(X)<K.$$

Ezért $\forall 0<\varepsilon<1$ és $0<\alpha<\frac{I_1}{2(K+1)}$-hez $\exists\, n_0(\varepsilon,\alpha)$ küszöbszám, hogy $n>n_0$ esetén

$$\mathbf{P}(|B_1|\geq\alpha^2)<\frac{\varepsilon}{3},$$

$$\mathbf{P}(B_2\geq-\frac{1}{2}I_1)<\frac{\varepsilon}{3},$$

$$\mathbf{P}(|B_3|\geq 2\,K)<\frac{\varepsilon}{3}.$$

A Boole-egyenlőtlenséget $\big(\mathbf{P}(\bar{A}_1\bar{A}_2\bar{A}_3)\geq 1-\mathbf{P}(A_1)-\mathbf{P}(A_2)-\mathbf{P}(A_3)\ \big)$ felhasználva:

$$\mathbf{P}(|B_1|<\alpha^2, B_2<-\frac{1}{2}I_1, |B_3|<2\,K)\geq 1-\varepsilon.$$

Megmutatjuk, hogy a $\vartheta=\vartheta_0+\alpha$ pontban a $B_1+B_2(\vartheta-\vartheta_0)+\frac{1}{2}\Delta B_3(\vartheta-\vartheta_0)^2$ kifejezés negatív értéket vesz fel:

$$\frac{1}{n}\frac{\partial \ln L(\mathbf{x},\vartheta)}{\partial\vartheta}\bigg|_{\vartheta=\vartheta_0+\alpha}=B_1+B_2\alpha+\frac{1}{2}\Delta B_3\alpha^2<\alpha^2+\alpha\left(\frac{-I_1}{2}\right)+\frac{1}{\cancel{2}}\Delta\alpha^2\cancel{2}K<$$

$$<\alpha\frac{I_1}{2\;(K+1)}(K+1)-\frac{1}{2}I_1\alpha=0.$$

Tehát $\frac{\partial \ln L(\mathbf{x},\vartheta)}{\partial\vartheta}<0$, ha $\vartheta=\vartheta_0+\alpha$ és $\mathbf{x}$ kielégíti a $|B_1|<\alpha^2, B_2<-\frac{1}{2}I_1, |B_3|<2\,K$ feltételrendszert. Másrészt $\vartheta=\vartheta_0-\alpha$-val ugyanarra az eseményre:

$$\frac{1}{n}\frac{\partial \ln L(\mathbf{x},\vartheta)}{\partial\vartheta}|_{\vartheta=\vartheta_0-\alpha}=B_1-B_2\alpha+\frac{1}{2}\Delta\,B_3\alpha^2>-\alpha^2+\alpha\left(\frac{I_1}{2}\right)-\frac{1}{\cancel{2}}\Delta\alpha^2\cancel{2}K>$$

$$>-\alpha^2(K+1)+\frac{1}{2}I_1\alpha>-\alpha\left(\frac{I_1}{2\;(K+1)}\right)(K+1)+\frac{1}{2}I_1\alpha=0.$$

Mivel az $\frac{1}{n}\frac{\partial \ln L(\mathbf{x},\vartheta)}{\partial\vartheta}$ függvény differenciálható, így folytonos, ezért a $(\vartheta_0-\alpha,\vartheta_0+\alpha)$ intervallumban kell, hogy legyen gyöke. Másképpen fogalmazva, $\forall 0<\varepsilon<1$ és $0<\alpha<\frac{I_1}{2(K+1)}$-hez $\exists\, n_0(\varepsilon,\alpha)$ küszöbszám, hogy $n>n_0$ esetén több mint $1-\varepsilon$ valószínűséggel a $\frac{\partial \ln L(\mathbf{X},\vartheta)}{\partial\vartheta}=0$ likelihood egyenletnek van gyöke a $(\vartheta_0-\alpha,\vartheta_0+\alpha)$ intervallumban, azaz

$$\mathbf{P}(|\tau_n(\mathbf{X})-\vartheta_0|<\alpha)\geq 1-\varepsilon,$$

vagyis a maximum-likelihood becslés konzisztens.

A (ii) bizonyítása. A

$$B_1+B_2(\tau_n(\mathbf{X})-\vartheta_0)\;+\frac{1}{2}\,\Delta\,B_3(\tau_n(\mathbf{X})-\vartheta_0)^2=0$$

egyenletből:

$$\tau_n(\mathbf{X})-\vartheta_0=\frac{-B_1}{B_2+\frac{1}{2}\Delta\,B_3\,(\tau_n(\mathbf{X})-\vartheta_0)},$$

majd mindkét oldalt $\sqrt{nI_1(\vartheta_0)}$-lal megszorozva:

$$\sqrt{nI_1(\vartheta_0)}(\tau_n(\mathbf{X})-\vartheta_0)=\frac{\frac{\sqrt{n}}{\sqrt{I_1(\vartheta_0)}}B_1}{-\frac{B_2}{(I_1(\vartheta_0))}-\frac{1}{2}\Delta B_3\frac{(\tau_n(\mathbf{X})-\vartheta_0)}{(I_1(\vartheta_0))}}=\frac{\frac{1}{\sqrt{I_1(\vartheta_0)}\cdot\sqrt{n}}\sum\limits_{i=1}^{n}\frac{\partial\ln f_\vartheta(X_i)}{\partial\vartheta}\Big|_{\vartheta=\vartheta_0}}{-\frac{B_2}{(I_1(\vartheta_0))}-\frac{1}{2}\Delta B_3\frac{(\tau_n(\mathbf{X})-\vartheta_0)}{(I_1(\vartheta_0))}}$$

Az $Y_i = \left.\frac{\partial \ln f_\vartheta(X_i)}{\partial\vartheta}\right|_{\vartheta=\vartheta_0}$ jelöléssel, az Y_i valószínűségi változók teljesen függetlenek és azonos eloszlásúak. Továbbá:

$$\mathbf{E}_{\vartheta_0} Y_i = \mathbf{E}_{\vartheta_0}\left(\frac{\partial \ln f_\vartheta(X_i)}{\partial\vartheta}|_{\vartheta=\vartheta_0}\right) = 0,$$

$$\boldsymbol{\sigma}^2_{\vartheta_0} Y_i = \boldsymbol{\sigma}^2_{\vartheta_0}\left(\frac{\partial \ln f_\vartheta(X_i)}{\partial\vartheta}|_{\vartheta=\vartheta_0}\right) = I_1(\vartheta_0).$$

A centrális határeloszlás tételt alkalmazva:

$$U_n = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{\sqrt{I_1(\vartheta_0)}}\sqrt{n} = \frac{1}{\sqrt{I_1(\vartheta_0)}\cdot\sqrt{n}}\sum_{i=1}^{n}\left.\frac{\partial \ln f_\vartheta(X_i)}{\partial\vartheta}\right|_{\vartheta=\vartheta_0} \xrightarrow{e} N(0,1).$$

Felhasználva a Csebisev-féle nagy számok törvényét:

$$\tau_n \xrightarrow{st} \vartheta_0,\ B_2 \xrightarrow{st} -(I_1(\vartheta_0)),\ B_3 \xrightarrow{st} \mathbf{E}_{\vartheta_0} H_3(X_i) < K,$$

amiből

$$Z_n = -\frac{B_2}{(I_1(\vartheta_0))} - \frac{1}{2}\Delta\, B_3\,\frac{(\tau_n - \vartheta_0)}{(I_1(\vartheta_0))} \xrightarrow{st} 1$$

következik.

Mivel $U_n \xrightarrow{e} N(0,1)$, $Z_n \xrightarrow{st} 1$, így $\frac{U_n}{Z_n} \xrightarrow{e} N(0,1)$, azaz $\sqrt{n\, I_1(\vartheta_0)}\,(\tau_n - \vartheta_0) \xrightarrow{e} N(0,1)$.

■

A maximum-likelihood módszer az előző tételek miatt alapvető fontosságú a becsléselméletben. Ahol lehet, célszerű alkalmazni. Vannak azonban esetek, amikor a likelihood egyenlet a paraméterre transzcendens egyenletet ad, azaz a paraméter kifejtése lehetetlen. Ilyen esetekben sokszor hasznos a momentumok módszere. A módszer lényege az, hogy a minta momentumai függvénykapcsolatban vannak az eloszlás paramétereivel, és ebbe az ismert függvénybe a mintából becsült empirikus momentumokat beírva kapjuk a becslési statisztikákat.

2.4.2. definíció: Legyen adott $\mathcal{P}$, valószínűségi mértékek egy tere és az $X_1, X_2, \ldots, X_n$ statisztikai minta. Tegyük fel, hogy léteznek az

$$m_j = \mathbf{E}_{\boldsymbol{\vartheta}} X_i^j = g_j(\boldsymbol{\vartheta}) \quad (j = 1, 2, \ldots, k)$$

momentumok, és

$$\exists g_j^{-1}(m_1, m_2, \ldots, m_k) = \vartheta_j \quad (j = 1, 2, \ldots, k).$$

Tekintsük az

$$\hat{m}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j \quad (j = 1, 2, \ldots, k)$$

empirikus momentum statisztikákat. Akkor az

$$m_j = g_j^{-1}(\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_k) \quad (j = 1, 2, \ldots, k)$$

statisztikák a ϑ_j paraméterek *momentumos becslései.*

A momentumok módszere nem rendelkezik olyan optimális tulajdonságokkal, mint a maximum-likelihood módszer, de azért az általános feltételek mellett belátható, hogy a becslései konzisztensek. A konzisztencia azon múlik, hogy az empirikus momentumok is konzisztens becslései az elméleti momentumoknak.

2.4.5. példa: *(A normális eloszlás paramétereinek becslése a momentumok módszerével)*

A minta sűrűségfüggvénye

$$f_{m,D}(x) = \frac{1}{\sqrt{2\pi D}} \mathrm{e}^{-\frac{(x-m)^2}{2D}}.$$

A normális eloszlás esetén tudjuk, hogy $m = g_1(m_1, m_2) = m_1$, $D = g_2(m_1, m_2) = m_2 - m_1^2$.

Az empirikus momentumok: $\hat{m}_1 = \frac{1}{n}\sum\limits_{i=1}^{n} X_i = \bar{X}_n$, $\hat{m}_2 = \frac{1}{n}\sum\limits_{i=1}^{n} X_i^2$. Így a momentumbecslések egyből adódnak:

$$m \approx g_1(\hat{m}_1, \hat{m}_2) = \bar{X}_n,$$

$$D \approx g_2(\hat{m}_1, \hat{m}_2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = s_n^2.$$

Látható, hogy ugyanazok a statisztikák adódtak, mint a maximum-likelihood módszernél.

2.4.6. példa: *(A Poisson-eloszlás paraméterének becslése a momentumok módszerével)*
A minta eloszlása most

$$\mathbf{P}_\vartheta(X_i = k) = \frac{\vartheta^k}{k!}\mathrm{e}^{-\vartheta} \quad (k = 0, 1, 2, \ldots).$$

A $\vartheta > 0$ paraméter éppen a várható érték, az első momentum, így a momentumbecslés egyből adódik: $\vartheta \approx \hat{m}_1 = \bar{X}_n$. Ezúttal is ugyanazt a statisztikát kaptuk, mint a maximum-likelihood módszernél.

2.5. Intervallumbecslések

A korábbi szakaszokban az ismeretlen paramétervektort a minta egy függvényével, azaz egyetlen statisztikával próbáltuk meg közelíteni. Konkrét realizációnál tehát, a paramétertér egy pontját egy másik ponttal becsüljük. Ezért beszélünk pontbecslésről. De tudjuk azt is, hogy folytonos eloszlásoknál, annak valószínűsége, hogy a valószínűségi változó az értékkészletének éppen egy teszőlegesen kiválasztott pontját fogja felvenni, nulla. Tehát folytonos esetben nulla annak valószínűsége, hogy éppen a paramétert találtuk el a becsléssel. Az intervallumbecsléseknél a mintából készített tartományokat definiálunk, amely tartományok nagy valószínűséggel lefedik a kérdéses paraméterpontot. A témakört egydimenziós paraméter esetén tárgyaljuk.

2.5.1. definíció: Legyen adott $\mathcal{P}$ valószínűségi mértékek egy tere és az $X_1, X_2, \ldots, X_n$ statisztikai minta. Legyen $0 < \varepsilon < 1$ rögzített. Azt mondjuk, hogy a ϑ paraméterhez megadtunk egy legalább $1 - \varepsilon$ szignifikanciaszintű konfidenciaintervallumot, ha $t_1(X_1, X_2, \ldots, X_n)$ és $t_2(X_1, X_2, \ldots, X_n)$ olyan statisztikák, hogy

$$\mathbf{P}_\vartheta(t_1(X_1, X_2, \ldots, X_n) \le \vartheta \le t_2(X_1, X_2, \ldots, X_n)) \ge 1 - \varepsilon$$

fennáll minden $P_\vartheta \in \mathcal{P}$-re.

Ahhoz, hogy példákat mutassunk konfidencia intervallumra, be kell bizonyítanunk a Lukács-tételt, és definiálni kell a χ^2- és a Student-eloszlásokat.

2.5.2. definíció: Legyenek $Y, X_1, X_2, \ldots, X_n$ standard normális eloszlású, teljesen független valószínűségi változók. Ekkor $\sum\limits_{i=1}^{n} X_i^2$ χ^2-eloszlást követ n szabadságfokkal, melynek sűrűségfüggvénye

$$f(x) = \frac{1}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)} \cdot \mathrm{e}^{-\frac{x}{2}} \cdot x^{\frac{n}{2}-1}, \quad x > 0,$$

ahol $\Gamma(s) = \int\limits_0^\infty \mathrm{e}^{-t} \cdot t^{s-1}\, \mathrm{d}t$ a gamma-függvény. Másrészt $\frac{Y}{\sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n}}}$ n szabadságfokú t- (Student-) eloszlást fog követni, melynek sűrűségfüggvénye

$$g(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{1}{\sqrt{n}} \cdot \left(\frac{1}{1+\frac{x^2}{n}}\right)^{\frac{n+1}{2}}, \quad x \in \mathbb{R}.$$

2.5.1. tétel: *(Lukács)*

Legyen $X_1, X_2, \ldots, X_n \in N(m, D)$ eloszlásból származó statisztikai minta.
Ekkor

(i) $\bar{X}_n \in N(m, \frac{D}{\sqrt{n}})$, azaz $m, \frac{D}{\sqrt{n}}$ paraméterű normális eloszlás,

(ii) $\frac{ns_n^2}{D^2} \in \chi^2_{n-1}$, azaz $n-1$ szabadságfokú χ^2-eloszlás,

(iii) $\bar{X}_n$ és s_n^2 függetlenek ($\bar{X}_n$ és s_n^{*2} is függetlenek).

Bizonyítás:

(i) $\bar{X}_n$ karakterisztikus függvénye:

$$\varphi_{\bar{X}_n}(t) = \mathbf{E}\exp\left(i\sum_{j=1}^{n} X_j \frac{t}{n}\right) = \prod_{j=1}^{n} \mathbf{E}\exp\left(iX_j\frac{t}{n}\right) = \prod_{j=1}^{n} \varphi_{X_j}\left(\frac{t}{n}\right) =$$

$$= \left(\exp\left(im\frac{t}{n} - \frac{D^2t^2}{2n^2}\right)\right)^n,$$

amiből leolvasható, hogy $\bar{X}_n \in N(m, \frac{D}{\sqrt{n}})$.

(ii) *Segédtétel:* Tekintsük a

$$\underline{\underline{H}}_n = \underline{\underline{E}}_n - \frac{1}{n}\mathbf{1}\,\mathbf{1}^T = \begin{pmatrix} \frac{n-1}{n} & -\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & \frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \vdots & \frac{n-1}{n} \end{pmatrix}$$

centráló mátrixot. A képletben az $\mathbf{1}$ olyan vektort jelöl, melynek mindegyik komponense 1-es, $\underline{\underline{E}}_n$ pedig az egységmátrix.
Ekkor

a) $\underline{\underline{H}}_n \underline{\underline{H}}_n = \underline{\underline{H}}_n$ (idempotens),

b) $\underline{\underline{H}}_n$ szimmetrikus, pozitív szemidefinit,

c) $\det\left(\underline{\underline{H}}_n\right) = 0$, $\operatorname{rank}\left(\underline{\underline{H}}_n\right) = n - 1$,

d) $\underline{\underline{H}}_n$ sajátértékei az 1 $(n-1)$- szeres multiplicitással, és a 0.

A segédtétel bizonyítása:

a) $\underline{\underline{H}}_n^2 = \underline{\underline{E}}_n^2 - \underline{\underline{E}}_n\left(\frac{1}{n}\mathbf{1}\mathbf{1}^T\right) - \left(\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\underline{\underline{E}}_n + \frac{1}{n^2}\mathbf{1}(\mathbf{1}^T\mathbf{1})\mathbf{1}^T = \underline{\underline{E}}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T = \underline{\underline{H}}_n$.

b) $\underline{\underline{H}}_n$ szimmetrikus triviálisan. Legyen $\mathbf{x} \in \mathbb{R}^n$ tetszőleges:

$\mathbf{x}^T\underline{\underline{H}}_n\mathbf{x} = \mathbf{x}^T\underline{\underline{H}}_n \cdot \underline{\underline{H}}_n\mathbf{x} = \left\|\underline{\underline{H}}_n\mathbf{x}\right\|^2 \geq 0$ pozitív szemidefinit, sajátértékei nemnegatívak.

c) $\det\left(\underline{\underline{H}}_n - \lambda\underline{\underline{E}}_n\right) = \det\left(\underline{\underline{H}}_n^2 - \lambda\underline{\underline{E}}_n^2\right) = \det\left(\underline{\underline{H}}_n - \sqrt{\lambda}\underline{\underline{E}}_n\right) \cdot \det\left(\underline{\underline{H}}_n + \sqrt{\lambda}\underline{\underline{E}}_n\right) = 0.$
Tehát, ha λ sajátérték, akkor $+\sqrt{\lambda}$ is az. Így csak 1 és 0 lehet sajátérték! Másrészt,

$$\operatorname{trace}\underline{\underline{H}}_n = n\frac{n-1}{n} = n - 1 = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

csak úgy lehet, ha $\lambda_1 = \lambda_2 = \cdots = \lambda_{n-1} = 1$ és $\lambda_n = 0$.

d) $\det\left(\underline{\underline{H}}_n\right) = \prod_{j=1}^{n}\lambda_j = 0$, $\operatorname{rank}\left(\underline{\underline{H}}_n\right) = \sum_{j=1}^{n}\lambda_j = n-1$, mert $n-1$ darab 1 sajátértéke van.

A segédtételt használva bizonyíthatjuk a tétel 2. állítását.

$\mathbf{X}^T\underline{\underline{H}}_n\mathbf{X} = \mathbf{X}^T\mathbf{X} - \frac{1}{n}\mathbf{X}^T\mathbf{1}\mathbf{1}^T\mathbf{X} = \sum_{i=1}^{n} X_i^2 - n\left(\bar{X}_n\right)^2 = ns_n^2$.

Legyenek $Z_i = X_i - m \in N(0, D)$, teljesen függetlenek.

$$\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i = \frac{1}{n}\sum_{i=1}^{n}(X_i - m) = \bar{X}_n - m,$$

$$\frac{1}{n}\mathbf{Z}\underline{\underline{H}}_n\mathbf{Z} = d_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}((X_i - m) - (\bar{X}_n - m))^2 = s_n^2.$$

Felhasználjuk $\underline{\underline{H}}_n$ spektrálfelbontását:

$\underline{\underline{H}}_n = \underline{\underline{G}}\,\underline{\underline{L}}\,\underline{\underline{G}}^T$, ahol $\underline{\underline{G}}\,\underline{\underline{G}}^T = \underline{\underline{G}}^T\underline{\underline{G}} = \underline{\underline{E}}_n$ és $\underline{\underline{L}} = \operatorname{diag}(1, 1, \ldots, 1, 0)$. Így

$$ns_n^2 = \mathbf{X}^T\underline{\underline{H}}_n\mathbf{X} = \mathbf{Z}^T\underline{\underline{H}}_n\mathbf{Z} = \mathbf{Z}^T\underline{\underline{G}}\,\underline{\underline{L}}\,\underline{\underline{G}}^T\mathbf{Z} = \mathbf{Y}^T\underline{\underline{L}}\mathbf{Y} = \sum_{i=1}^{n-1} Y_i^2.$$

$\mathbf{Y} = \underline{\underline{G}}^T\mathbf{Z} \in N_n(\underline{\underline{G}}\mathbf{0}, D\underline{\underline{E}}_n)$, azaz

$\frac{Y_i}{D} \in N(0,1)$ teljesen függetlenek $\frac{ns_n^2}{D^2} = \sum_{i=1}^{n-1}\frac{Y_i^2}{D^2} \in \chi_{n-1}^2$.

(iii) A $\lambda_n = 0$ sajátértékhez tartozó sajátvektor: $\mathbf{g}_n = \frac{1}{\sqrt{n}}\mathbf{1} = (\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})^T$, mert $\underline{\underline{H}}_n\mathbf{g}_n = \underline{\underline{E}}_n\frac{1}{\sqrt{n}}\mathbf{1} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\frac{1}{\sqrt{n}}\mathbf{1} = \mathbf{g}_n - \mathbf{g}_n = \mathbf{0}$. Így $Y_n = \mathbf{g}_n^T\mathbf{Z} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i = \sqrt{n}\bar{Z}_n$. Mivel $ns_n^2 = nd_n^2 = \sum_{i=1}^{n-1} Y_i^2$, $\bar{X}_n = \bar{Z}_n + m = \frac{1}{\sqrt{n}}Y_n + m$, és Y_i-k teljesen függetlenek voltak, így $\bar{X}_n$ és s_n^2 is függetlenek.

■

Felhasználva a Lukács-tételt belátható, hogy ha $X_1, X_2, \ldots, X_n$ $N(m, D)$ eloszlásból származó statisztikai minta, akkor az

$$\frac{\bar{X}_n - m}{D}\sqrt{n} \in N(0,1), \quad \text{és az} \quad \frac{(n-1)\, s_n^{*\,2}}{D^2} \in \chi^2_{n-1}$$

statisztikák függetlenek, így

$$\frac{\frac{\bar{X}_n - m}{D}\sqrt{n}}{\sqrt{\frac{\frac{(n-1)\, s_n^{*\,2}}{D^2}}{n-1}}} = \frac{\bar{X}_n - m}{s_n^*}\sqrt{n} \in t_{n-1}$$

($n-1$ szabadságfokú Student-eloszlású).

2.5.1. példa: *(Konfidenciaintervallum szerkesztése az ismeretlen várható értékre ismert szórású normális eloszlás esetében)*

Legyen $X_1, X_2, \ldots, X_n$ $N(m, D_0)$ eloszlásból származó statisztikai minta, ahol $D_0 > 0$ ismert, $m \in \mathbb{R}$ ismeretlen. Szerkesszünk m-re adott $0 < \varepsilon < 1$ mellett $(1-\varepsilon)$-szintű konfidenciaintervallumot! A Lukács-tételből tudjuk, hogy $u = \frac{\bar{X}_n - m}{D_0}\sqrt{n} \in N(0,1)$, azaz a statisztika sűrűségfüggvénye: $\varphi(x) = \frac{1}{\sqrt{2\,\pi}} \mathrm{e}^{-\frac{x^2}{2}}$. $\varphi(x)$ segítségével megadható olyan $u_\varepsilon > 0$ szám, hogy

$$\int\limits_{-u_\varepsilon}^{+u_\varepsilon} \varphi(t)\,\mathrm{d}t = \mathbf{P}(-u_\varepsilon < u < u_\varepsilon) = \Phi(u_\varepsilon) - \Phi(-u_\varepsilon) = 2\,\Phi(u_\varepsilon) - 1 = 1 - \varepsilon$$

teljesüljön. Az $u_\varepsilon > 0$ szám meghatározását a $\Phi(u_\varepsilon) = 1 - \frac{\varepsilon}{2}$ egyenletből, standard normális eloszlás táblázata segítségével határozhatjuk meg. Mivel a $\{-u_\varepsilon < u < u_\varepsilon\}$ esemény ekvivalens az $\left\{\bar{X}_n - \frac{u_\varepsilon\, D_0}{\sqrt{n}} < m < \bar{X}_n + \frac{u_\varepsilon\, D_0}{\sqrt{n}}\right\}$ eseménnyel, ezért

$$\mathbf{P}\left(\bar{X}_n - \frac{u_\varepsilon\, D_0}{\sqrt{n}} < m < \bar{X}_n + \frac{u_\varepsilon\, D_0}{\sqrt{n}}\right) = 1 - \varepsilon,$$

azaz a

$$T_1 = \bar{X}_n - \frac{u_\varepsilon\, D_0}{\sqrt{n}},$$

$$T_2 = \bar{X}_n + \frac{u_\varepsilon\, D_0}{\sqrt{n}}$$

$(1-\varepsilon)$-szintű konfidenciaintervallum m-re.

2.5.2. példa: *(Konfidenciaintervallum szerkesztése az ismeretlen várható értékre ismeretlen szórású normális eloszlás esetében)*

Legyen $X_1, X_2, \ldots, X_n$ $N(m, D)$ eloszlásból származó statisztikai minta, ahol $D > 0$ is és, $m \in \mathbb{R}$ is ismeretlen. Szerkesszünk m-re adott $0 < \varepsilon < 1$ mellett $(1-\varepsilon)$-szintű konfidenciaintervallumot! A Lukács-tétel után láttuk, hogy $\frac{\bar{X}_n - m}{s_n^*}\sqrt{n} \in t_{n-1}$, azaz az $n-1$ szabadságfokú Student-eloszláshoz tartozó táblázatból kiolvasható olyan $t_\varepsilon > 0$ szám, amellyel

$$1 - \varepsilon = \mathbf{P}(-t_\varepsilon < \frac{\bar{X}_n - \mu}{s_n^*}\sqrt{n} < t_\varepsilon) = \mathbf{P}\left(\bar{X}_n - \frac{t_\varepsilon\, s_n^*}{\sqrt{n}} < m < \bar{X}_n + \frac{t_\varepsilon\, s_n^*}{\sqrt{n}}\right)$$

azaz most a $T_1 = \bar{X}_n - \frac{t_\varepsilon\, s_n^*}{\sqrt{n}}$, $T_2 = \bar{X}_n + \frac{t_\varepsilon\, s_n^*}{\sqrt{n}}$ statisztikapár lesz $(1-\varepsilon)$-szintű konfidenciaintervallum m-re.

2.5.3. példa: *(Konfidencia intervallum szerkesztése az ismeretlen szórásra normális eloszlás esetében)*

Legyen $X_1, X_2, \ldots, X_n$ $N(m, D)$ eloszlásból származó statisztikai minta, ahol $D > 0$ is és $m \in \mathbb{R}$ is ismeretlen. Szerkesszünk D-re adott $0 < \varepsilon < 1$ mellett $(1-\varepsilon)$-szintű konfidenciaintervallumot! A Lukács-tételre hivatkozva megint: $\frac{(n-1)s_n^{*\,2}}{D^2} \in \chi^2_{n-1}$. Az $n-1$ szabadságfokú χ^2-eloszlás táblázatból megadhatók olyan $0 < c_1 < c_2$ számok, hogy

$$1 - \varepsilon = \mathbf{P}\left(c_1 < \frac{(n-1)\,s_n^{*\,2}}{D^2} < c_2\right)$$

teljesüljön. (A c_1, c_2 értékek nyilván kielégítik a $\mathbf{P}(\chi^2_{n-1} > c_1) = 1 - \frac{\varepsilon}{2}$ és $\mathbf{P}(\chi^2_{n-1} > c_2) = \frac{\varepsilon}{2}$ feltételeket.) Egyszerű átrendezéssel kapjuk, hogy

$$1 - \varepsilon = \mathbf{P}\left(\sqrt{\frac{(n-1)}{c_2}}s_n^* < D < \sqrt{\frac{(n-1)}{c_1}}s_n^*\right),$$

azaz a $T_1 = \sqrt{\frac{(n-1)}{c_2}}s_n^*$, $T_2 = \sqrt{\frac{(n-1)}{c_1}}s_n^*$ statisztikapár $(1-\varepsilon)$-szintű konfidenciaintervallum lesz D-re.

2.5.4. példa: *(Konfidenciaintervallum szerkesztése az ismeretlen paraméterre exponenciális eloszlás esetében)*

Legyen $X_1, X_2, \ldots, X_n$ $E(\lambda)$ eloszlásból származó statisztikai minta, ahol $\lambda > 0$ ismeretlen. Szerkesszünk λ-ra adott $0 < \varepsilon < 1$ mellett $(1-\varepsilon)$-szintű konfidenciaintervallumot!

A probléma megoldásához felhasználjuk az alábbi segédtételt:

Segédtétel: Legyen $X_1, X_2, \ldots, X_n$ $E(\lambda)$ eloszlásból származó statisztikai minta. Ekkor

a) $\lambda X_i \in E(1)$,

b) $\sum\limits_{j=1}^{n} \lambda X_j = \lambda n \bar{X}_n \in \Gamma(n, 1)$, azaz $n, 1$ paraméterű gamma eloszlású,

$$f_\Gamma(x) = \frac{x^{n-1}}{(n-1)!}\mathrm{e}^{-x} \quad (x > 0)$$

sűrűségfüggvénnyel.

A segédtétel bizonyítása:

a) $\mathbf{P}(\lambda X_j < x) = \mathbf{P}(X_j < \frac{x}{\lambda}) = 1 - \mathrm{e}^{-\lambda\frac{x}{\lambda}} = 1 - \mathrm{e}^{-x} \implies \lambda X_j \in E(1)$.

b) $\varphi_{\lambda X_j}(t) = \mathbf{E}\mathrm{e}^{i\lambda X_j t} = \int\limits_0^\infty \mathrm{e}^{ixt}\mathrm{e}^{-x}\,\mathrm{d}x = \left[\frac{1}{it-1}\mathrm{e}^{x(it-1)}\right]_0^\infty = \frac{1}{1-it}$.

$\varphi_{\lambda n\bar{X}_n}(t) = \prod\limits_{j=1}^{n} \varphi_{\lambda \bar{X}_\alpha}(t) = \left(\frac{1}{1-it}\right)^n \implies f_{\lambda n \bar{X}_n}(x) = \frac{x^{n-1}\mathrm{e}^{-x}}{(n-1)!}$, mert a karakterisztikus függvénye:

$$\int\limits_0^\infty \mathrm{e}^{ixt}\frac{x^{n-1}}{(n-1)!}\mathrm{e}^{-x}\,\mathrm{d}x = \left[\frac{x^{n-1}}{(n-1)!}\frac{1}{it-1}\mathrm{e}^{x(it-1)}\right]_0^\infty - \frac{1}{(n-2)!}\frac{1}{it-1}\int\limits_0^\infty x^{n-2}\mathrm{e}^{x(it-1)}\,\mathrm{d}x =$$

$$= 0 - \frac{1}{(n-2)!}\frac{1}{it-1}\left[x^{n-2}\frac{1}{it-1}\mathrm{e}^{x(it-1)}\right]_0^{\infty} + \frac{1}{(n-3)!}\left(\frac{1}{it-1}\right)^2 \int_0^{\infty} x^{n-3}\mathrm{e}^{x(it-1)}\,\mathrm{d}x =$$

$$= \frac{1}{(n-3)!}\left(\frac{1}{it-1}\right)^2 \int_0^{\infty} x^{n-3}\mathrm{e}^{x(it-1)}\,\mathrm{d}x = \cdots = (-1)^n\left(\frac{1}{it-1}\right)^n = \left(\frac{1}{1-it}\right)^n.$$

Az $n, 1$ paraméterű gamma-eloszláshoz tartozó táblázatból kiolvashatók olyan $0 < \gamma_1 < \gamma_2$ számok, amelyekkel

$$1 - \varepsilon = \mathbf{P}(\gamma_1 < \lambda\, n\, \bar{X}_n < \gamma_2) = \mathbf{P}(\frac{\gamma_1}{n\,\bar{X}_n} < \lambda < \frac{\gamma_2}{n\,\bar{X}_n}),$$

azaz a $T_1 = \frac{\gamma_1}{n\,\bar{X}_n}$, $T_2 = \frac{\gamma_2}{n\,\bar{X}_n}$ statisztika pár lesz $(1-\varepsilon)$-szintű konfidenciaintervallum λ-ra.

A γ_1, γ_2 számokat úgy kell meghatározni, hogy $\mathbf{P}(0 < \Gamma(n,1) < \gamma_1) = \mathbf{P}(\Gamma(n,1) > \gamma_2) = \frac{\varepsilon}{2}$ legyen.

3. fejezet

Hipotézisvizsgálat

3.1. Alapfogalmak

Tekintsük a $\mathcal{K}$ véletlen kísérletet és a hozzátartozó $(\Omega, \mathcal{F})$ mérhető teret, és a $\mathcal{P}$ valószínűségi mértékek osztályát, ahol $(\Omega, \mathcal{P}, \mathbf{P})$ Kolmogorov-féle valószínűségi mező $\forall \mathbf{P} \in \mathcal{P}$-re. Tegyük fel, hogy $\mathcal{P}$ két diszjunkt részhalmazra bontható: $\mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1$, és $\mathcal{P}_0 \cap \mathcal{P}_1 = \emptyset$. Statisztikai módszert (ún. próbát vagy tesztet) akarunk konstruálni annak eldöntésére, hogy a véletlen kísérlethez tartozó tényleges $\mathbf{P}$ valószínűségi mérték melyik halmazhoz tartozik $\mathcal{P}_0$ és $\mathcal{P}_1$ közül. Ehhez felállítunk egy $H_0 : \mathbf{P} \in \mathcal{P}_0$ nullhipotézist, és egy $H_1 : \mathbf{P} \in \mathcal{P}_1$ alternatív hipotézist. A nullhipotézis azt a feltevésünket fogalmazza meg, hogy az elméleti $\mathbf{P}$ valószínűség a $\mathcal{P}_0$ részhez tartozik, az alternatív hipotézisünk pedig azt, hogy ellenkezőleg, pont a $\mathcal{P}_1$ részhez. A kettő feltevés közül az eljárás végén egyértelműen kiválasztjuk és elfogadjuk majd az egyiket. A döntést az $X_1, X_2, \ldots, X_n, \ldots$ statisztikai minta segítségével fogjuk meghozni. Először is, el fogjuk készíteni a $t_n(X_1, X_2, \ldots, X_n)$ ún. *próbastatisztikát*, amely rendelkezni fog az alábbi tulajdonsággal: adott $0 < \varepsilon < 1$ számhoz megadhatók olyan $K_1(\varepsilon) < K_2(\varepsilon)$ számok, hogy $\mathbf{P}(K_1(\varepsilon) \leq t_n \leq K_2(\varepsilon)) \geq 1 - \varepsilon, \quad \forall \mathbf{P} \in \mathcal{P}_0$.

A $K_1(\varepsilon)$, $K_2(\varepsilon)$ értékeket kritikus értékeknek, a segítségükkel definiált $\mathcal{X}_e = \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^n \ , \ K_1(\varepsilon) \leq t_n(\mathbf{x}) \leq K_2(\varepsilon)\}$ n-dimenziós vektorhalmazt *elfogadási tartomány*nak, a komplemens halmazát, $\mathcal{X}_k = \mathbb{R}^n \setminus \mathcal{X}_e$-t, pedig *kritikus tartomány*nak nevezzük. Az ε szám a próba *terjedelme*, az $1 - \varepsilon$ érték pedig a próba *szignifikancia szintje*. A döntést úgy hajtjuk végre, hogy ellenőrizzük, hogy az $X_1, X_2, \ldots, X_n$ minta beleseik-e az $\mathcal{X}_e$ elfogadási tartományba. Ha beleesik, akkor a H_0 hipotézist, ellenkező esetben a H_1 alternatív hipotézist fogjuk elfogadni. A hipotézis eldöntése másképpen alakulhat az egyes ε terjedelmeken, ezért mindig jelezni kell, hogy milyen $1-\varepsilon$ szint mellett fogadjuk el (vagy vetjük el) a nullhipotézist. Természetesen számolunk azzal is, hogy a döntésünk hibás. Azt mondjuk, hogy *elsőfajú hibát* követünk el, ha elvetjük a nullhipotézist, holott valójában az igaz. *Másodfajú hibát* akkor követünk el, ha elfogadjuk a nullhipotézist, holott az nem igaz. Minden más esetben helyesen döntünk. A döntési hibafajtákat az alábbi táblázatban mutatjuk:

Döntés \ Valóság	H_0 igaz	H_1 igaz
H_0 mellett	jó döntés	másodfajú hiba
H_1 mellett	első fajú hiba	jó döntés

3.1.1. definíció: A

$$p_1(\varepsilon, n, \mathbf{P}) = \mathbf{P}((X_1, X_2, \ldots, X_n)^T \in \mathcal{X}_k), \quad \mathbf{P} \in \mathcal{P}_0, \quad 0 < \varepsilon < 1, \quad n \in \mathbb{N}$$

függvényt *elsőfajú hibavalószínűség*nek nevezzük. A

$$\sup_{\forall \mathbf{P} \in \mathcal{P}_0} \mathbf{P}((X_1, X_2, \ldots, X_n)^T \in \mathcal{X}_k) \leq \varepsilon$$

reláció teljesülése esetén *legfeljebb ε terjedelmű* próbáról beszélünk.

3.1.2. definíció: A

$$p_2(\varepsilon, n, \mathbf{P}) = \mathbf{P}((X_1, X_2, \ldots, X_n)^T \in \mathcal{X}_e), \quad \mathbf{P} \in \mathcal{P}_1, \quad 0 < \varepsilon < 1, \quad n \in \mathbb{N}$$

függvényt *másodfajú hibavalószínűség*nek nevezzük.

3.1.3. definíció: Az

$$E(\varepsilon, n, \mathbf{P}) = 1 - p_2(\varepsilon, n, \mathbf{P}) = \mathbf{P}((X_1, X_2, \ldots, X_n)^T \in \mathcal{X}_k), \quad \mathbf{P} \in \mathcal{P}_1 \quad 0 < \varepsilon < 1, \quad n \in \mathbb{N}$$

függvényt a próba *erőfüggvény*ének nevezzük. A

$$\sup_{\forall \mathbf{P} \in \mathcal{P}_1} \mathbf{P}((X_1, X_2, \ldots, X_n)^T \in \mathcal{X}_k)$$

érték a próba *ereje.*

3.1.4. definíció: Egy próba *torzítatlan*, ha

$$\mathbf{P}\left((X_1, X_2, \ldots, X_n)^T \in \mathcal{X}_k\right) \leq \varepsilon, \quad \forall \mathbf{P} \in \mathcal{P}_0\text{-ból}$$

következik, hogy

$$\mathbf{P}((X_1, X_2, \ldots, X_n)^T \in \mathcal{X}_k) \geq \varepsilon, \quad \forall\, \mathbf{P} \in \mathcal{P}_1, \quad \forall 0 < \varepsilon < 1.$$

Vagyis, ha H_0 nem áll fenn, nagyobb valószínűséggel utasítjuk el, mint amikor fennáll.

3.1.5. definíció: Egy próba *konzisztens*, ha $\lim_{n \to \infty} E(\varepsilon, n, \mathbf{P}) = 1, \forall \mathbf{P} \in \mathcal{P}_1$ és $0 < \varepsilon < 1$.

3.1.6. definíció: Egy próba *egyenletesen legjobb próba*, ha adott elsőfajú hibával rendelkező próbák között a legkisebb a másodfajú hibája.

3.2. Neyman–Pearson- és Stein-lemma

3.2.1. definíció: Egy *véletlenített próba* döntésfüggvényén azt a $\Phi : \mathbb{R}^n \to [0,1]$ függvényt értjük, amely megadja, hogy ha a minta realizáltja éppen $\mathbf{x}$, akkor $\Phi(\mathbf{x})$ valószínűséggel fogjuk a H_0 hipotézist elutasítani.

Megjegyzés:

1. Egy (nem véletlenített) statisztikai próba döntésfüggvénye $\Phi(\mathbf{x}) = I(\mathbf{x} \in \mathcal{X}_k)$, tehát a véletlenített próbák a statisztikai próbák kiterjesztését adják.

2. Véletlenített próba esetén a döntés két lépésből áll. Először az $\mathbf{X}$ minta alapján kiszámoljuk a $p = \Phi(\mathbf{X})$ valószínűséget, majd generálunk egy Y véletlen számot a $[0,1]$-en egyenletes eloszlásból. Ha $p \leq Y$, akkor elfogadjuk H_0-t, különben elvetjük.

3. Nyilván $\mathbf{E_P}\Phi(\mathbf{X})$ jelenti az elsőfajú hiba valószínűségét, ha $\mathbf{P} \in \mathcal{P}_0$ és az erőfüggvényt, ha $\mathbf{P} \in \mathcal{P}_1$.

4. Egy véletlenített próba terjedelmét a
$$\sup_{\forall \mathbf{P} \in \mathcal{P}_0} \mathbf{E_P}\Phi\left(\mathbf{X}\right),$$
az erejét pedig a
$$\sup_{\forall \mathbf{P} \in \mathcal{P}_1} \mathbf{E_P}\Phi\left(\mathbf{X}\right)$$
értékek adják.

3.2.1. tétel: *(Neyman–Pearson fundamentális lemma)*
Legyen a vizsgált $\mathcal{P}$ valószínűségi mértékosztály kételemű: $\mathcal{P} = \{\mathbf{P}_0, \mathbf{P}_1\}$. Létezzék az $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ statisztikai minta sűrűségfüggvénye mindkét valószínűségi mértékre nézve. Jelölje ezeket rendre $f_0\left(x\right)$ és $f_1\left(x\right)$. $\mathcal{P}$ nyilván dominált a λ Lebesgue-mértékre nézve. A minta együttes sűrűségfüggvényei így $L_0\left(\mathbf{x}\right) = \prod_{i=1}^{n} f_0\left(x_i\right)$ illetve $L_1\left(\mathbf{x}\right) = \prod_{i=1}^{n} f_1\left(x_i\right)$. Dönteni szeretnénk a $H_0:\ \mathbf{P} = \mathbf{P}_0$ hipotézisről a $H_1:\ \mathbf{P} = \mathbf{P}_1$ alternatív hipotézissel szemben. Ekkor

(i) tetszőleges $0 < \varepsilon < 1$ számhoz létezik olyan $0 < c_0$ és $0 < \tau < 1$ szám, amivel a
$$\Phi\left(\mathbf{x}\right) = \begin{cases} 1, & \text{ha} \quad L_1\left(\mathbf{x}\right) > c_0 L_0\left(\mathbf{x}\right) \\ \tau, & \text{ha} \quad L_1\left(\mathbf{x}\right) = c_0 L_0\left(\mathbf{x}\right) \\ 0, & \text{ha} \quad L_1\left(\mathbf{x}\right) < c_0 L_0\left(\mathbf{x}\right) \end{cases}$$
döntésfüggvény olyan véletlenített próbához tartozik, aminek ε a terjedelme,

(ii) az (i)-ben definiált próba egyenletesen legjobb próba,

(iii) ha Φ^* egy ε terjedelmű legjobb próba, akkor
$$\mathbf{P}_0\left(\Phi\left(\mathbf{X}\right) = \Phi^*\left(\mathbf{X}\right)\right) = \mathbf{P}_1\left(\Phi\left(\mathbf{X}\right) = \Phi^*\left(\mathbf{X}\right)\right) = 1.$$

Bizonyítás:

(i) Legyen $0 < \varepsilon < 1$ tetszőleges. Tekintsük a $G\left(c\right) = \mathbf{P}_0\left(L_1\left(\mathbf{X}\right) > cL_0\left(\mathbf{X}\right)\right),\ c \in \mathbb{R}$ függvényt. Mivel $L_0\left(\mathbf{x}\right)$ az $\mathbf{X}$ minta sűrűségfüggvénye H_0 mellett, ezért $\mathbf{P}_0\left(L_0\left(\mathbf{X}\right) > 0\right) = 1$, azaz $G\left(c\right) = \mathbf{P}_0\left(\frac{L_1(\mathbf{X})}{L_0(\mathbf{X})} > c\right)$. Mivel $1 - G\left(c\right)$ jobbról folytonos eloszlásfüggvénye az $Y = \frac{L_1(\mathbf{X})}{L_0(\mathbf{X})}$ valószínűségi változónak, $G\left(c\right)$ egy monoton nem növekvő, jobbról folytonos függvény, melyre $\lim_{c \to -\infty} G\left(c\right) = 1$, $\lim_{c \to \infty} G\left(c\right) = 0$. Ezért létezik olyan c_0 szám, melyre $G\left(c_0\right) \leq \varepsilon \leq G\left(c_0 - 0\right)$. Nyilván $G\left(c_0 - 0\right) - G\left(c_0\right) = \mathbf{P}_0\left(L_1\left(\mathbf{X}\right) = c_0 L_0\left(\mathbf{X}\right)\right)$. Ha G folytonos c_0-ban, akkor τ meghatározása érdektelen, hiszen úgyis egy 0-mértékű halmazon veszi csak fel Φ ezt az értéket. Ilyenkor
$$\mathbf{E}_{\mathbf{P}_0}\Phi\left(\mathbf{X}\right) = \mathbf{P}_0\left(L_1\left(\mathbf{X}\right) > c_0 L_0\left(\mathbf{X}\right)\right) = G\left(c_0\right) = \varepsilon,$$
vagyis a próba terjedelme ε. Ha viszont G nem folytonos c_0-ban, és
$$\tau = \frac{\varepsilon - G\left(c_0\right)}{G\left(c_0 - 0\right) - G\left(c_0\right)}$$

akkor

$$\mathbf{E}_{\mathbf{P}_0}\Phi(\mathbf{X}) = \mathbf{P}_0(L_1(\mathbf{X}) > c_0 L_0(\mathbf{X})) + \tau \mathbf{P}_0(L_1(\mathbf{X}) = cL_0(\mathbf{X})) =$$
$$= G(c_0) + \tau(G(c_0 - 0) - G(c_0)) = \varepsilon.$$

c_0 megválasztása lényegében egyértelmű. Tegyük fel ugyanis, hogy $G(c) = \varepsilon$, $c \in (c', c'')$. Tekintsük a

$$T = \left\{\mathbf{x} : p_0(\mathbf{x}) > 0 \,\cap\, c' < \frac{L_1(\mathbf{x})}{L_0(\mathbf{x})} < c''\right\}$$

tartományt.

$$\mathbf{P}_0(T) = G(c') - G(c'' - 0) = \varepsilon - \varepsilon = 0.$$

$\mathbf{x} \in T$ esetén $c' L_0(\mathbf{x}) < L_1(\mathbf{x}) < c'' L_0(\mathbf{x})$ miatt

$$0 = c' \int_T L_0(\mathbf{x})\, d\lambda(\mathbf{x}) < \int_T L_1(\mathbf{x})\, d\lambda(\mathbf{x}) < c'' \int_T L_0(\mathbf{x})\, d\lambda(\mathbf{x}) = 0,$$

azaz $\mathbf{P}_1(T) = 0$ is fennáll, azaz akár H_0, akár H_1 az igaz, csak 0 valószínűséggel fordulhat elő, hogy c_0 megválasztása nem egyértelmű.

(ii) Legyen most Φ^* egy tetszőleges legfeljebb ε terjedelmű véletlenített próba döntésfüggvénye: $\mathbf{E}_{\mathbf{P}_0}\Phi^*(\mathbf{X}) \le \varepsilon$. Legyenek $S^+ = \{\mathbf{x} : \Phi(\mathbf{x}) > \Phi^*(\mathbf{x})\}$ és $S^- = \{\mathbf{x} : \Phi(\mathbf{x}) < \Phi^*(\mathbf{x})\}$. Könnyen látható, hogy
$\forall \mathbf{x} \in S = S^+ \cup S^-$ esetén $(\Phi(\mathbf{x}) - \Phi^*(\mathbf{x}))(L_1(\mathbf{x}) - c_0 L_0(\mathbf{x})) \ge 0$ és
$\forall \mathbf{x} \in \bar{S}$ esetén $\Phi(\mathbf{x}) = \Phi^*(\mathbf{x})$.
Ezért

$$\int_{\mathbb{R}^n} (\Phi(\mathbf{x}) - \Phi^*(\mathbf{x}))(L_1(\mathbf{x}) - c_0 L_0(\mathbf{x}))\, d\lambda(\mathbf{x}) =$$
$$= \int_S (\Phi(\mathbf{x}) - \Phi^*(\mathbf{x}))(L_1(\mathbf{x}) - c_0 L_0(\mathbf{x}))\, d\lambda(\mathbf{x}) \ge 0,$$

azaz

$$\int_{\mathbb{R}^n} (\Phi(\mathbf{x}) - \Phi^*(\mathbf{x})) L_1(\mathbf{x})\, d\lambda(\mathbf{x}) \ge c_0 \int_{\mathbb{R}^n} (\Phi(\mathbf{x}) - \Phi^*(\mathbf{x})) L_0(\mathbf{x})\, d\lambda(\mathbf{x}) =$$
$$= c_0(\varepsilon - \mathbf{E}_{\mathbf{P}_0}\Phi^*(\mathbf{X})) \ge 0,$$

azaz $\mathbf{E}_{\mathbf{P}_1}\Phi(\mathbf{X}) \ge \mathbf{E}_{\mathbf{P}_1}\Phi^*(\mathbf{X})$, vagyis Φ erősebb, mint Φ^*.

(iii) Legyen most Φ^* egy tetszőleges legfeljebb ε terjedelmű egyenletesen legjobb próba döntésfüggvénye. Legyen $S = \{\mathbf{x} : \Phi(\mathbf{x}) \ne \Phi^*(\mathbf{x}) \,\cap\, L_1(\mathbf{x}) \ne c_0 L_0(\mathbf{x})\}$. Ha $\mathbf{x} \in S$, akkor

$$(\Phi(\mathbf{x}) - \Phi^*(\mathbf{x}))(L_1(\mathbf{x}) - c_0 L_0(\mathbf{x})) > 0$$

lesz. Ezért, ha S nem nullmértékű, vagy $\mathbf{P}_0$ vagy $\mathbf{P}_1$ szerint, akkor

$$0 < \int_S (\Phi(\mathbf{x}) - \Phi^*(\mathbf{x}))(L_1(\mathbf{x}) - c_0 L_0(\mathbf{x}))\, d\lambda(\mathbf{x}) =$$
$$= \int_{\mathbb{R}^n} (\Phi(\mathbf{x}) - \Phi^*(\mathbf{x}))(L_1(\mathbf{x}) - c_0 L_0(\mathbf{x}))\, d\lambda(\mathbf{x}).$$

Ebből következik, hogy

$$\int_{\mathbb{R}^n} (\Phi(\mathbf{x}) - \Phi^*(\mathbf{x})) L_1(\mathbf{x})\, \mathrm{d}\lambda(\mathbf{x}) > c_0 \int_{\mathbb{R}^n} (\Phi(\mathbf{x}) - \Phi^*(\mathbf{x})) L_0(\mathbf{x})\, \mathrm{d}\lambda(\mathbf{x}) = c_0(\varepsilon - \varepsilon) = 0,$$

azaz $\mathbf{E}_{\mathbf{P}_1}\Phi(\mathbf{X}) > \mathbf{E}_{\mathbf{P}_1}\Phi^*(\mathbf{X})$, ami ellentmondás azzal, hogy Φ^* egyenletesen legjobb próba döntésfüggvénye volt. Az ellentmondás abból fakadt, hogy feltettük, hogy S valamelyik valószínűségi mérték szerint nem nullmértékű. Tehát, Φ és Φ^* mindkét mérték szerint 1 valószínűséggel egybeesik.

■

3.2.2. tétel: *(Stein-lemma)*

Legyen a vizsgált $\mathcal{P}$ valószínűségi mértékosztály kételemű: $\mathcal{P} = \{\mathbf{P}_0, \mathbf{P}_1\}$. Létezzék az X sűrűségfüggvénye mindkét valószínűségi mértékre nézve. Jelölje ezeket rendre $f_0(x)$ és $f_1(x)$. Az $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ statisztikai minta együttes sűrűségfüggvényei így $L_0(\mathbf{x}) = \prod_{i=1}^{n} f_0(x_i)$ illetve $L_1(\mathbf{x}) = \prod_{i=1}^{n} f_1(x_i)$. Tegyük fel, hogy

$$|\mathbf{D}(f_0 \parallel f_1)| = \left| \mathbf{E}_{\mathbf{P}_0} \log_2 \frac{f_0(X_1)}{f_1(X_1)} \right| < \infty,$$

vagyis, véges a két eloszlás ún. *relatív entrópiája.* Dönteni szeretnénk a $H_0 : \mathbf{P} = \mathbf{P}_0$ hipotézisről a $H_1 : \mathbf{P} = \mathbf{P}_1$ alternatív hipotézissel szemben. Jelölje $\mathcal{X}_e^{(n)} \subseteq \mathbb{R}^n$ egy statisztikai próba elfogadási tartományát, $\alpha_n = \mathbf{P}_0\left(\mathbf{X} \in \mathcal{X}_k^{(n)}\right)$ az elsőfajú hibavalószínűséget, $\beta_n = \mathbf{P}_1\left(\mathbf{X} \in \mathcal{X}_e^{(n)}\right)$ pedig a másodfajú hibavalószínűséget. Legyen $0 < \varepsilon < \frac{1}{2}$ tetszőleges terjedelem, amellyel $\beta_{n,\varepsilon} = \min\limits_{\substack{\mathcal{X}_e^{(n)} \subseteq \mathbb{R}^n \\ \alpha_n \leq \varepsilon}} \beta_n$, azaz $\beta_{n,\varepsilon}$ jelöli a legfeljebb ε terjedelmű próbák esetén a minimális másodfajú hibavalószínűséget.

Akkor $\lim\limits_{n \to \infty} \frac{1}{n} \log_2 \beta_{n,\varepsilon} = -\mathbf{D}(f_0 \parallel f_1)$.

Bizonyítás: Először megkonstruálunk egy olyan $\mathcal{X}_e^{(n)} \subseteq \mathbb{R}^n$ elfogadási tartománysorozatot, amelyre

$$\alpha_n = \mathbf{P}_0\left(\mathbf{X} \in \mathcal{X}_k^{(n)}\right) < \varepsilon$$

és

$$\frac{\frac{1}{n}\log_2 \beta_n}{-\mathbf{D}(f_0 \parallel f_1)} \to 1$$

teljesül. Legyen tehát

$$\mathcal{X}_e^{(n)} = \left\{ \mathbf{x} :\ 2^{n(\mathbf{D}(f_0 \parallel f_1) - \delta)} \leq \frac{L_0(\mathbf{x})}{L_1(\mathbf{x})} \leq 2^{n(\mathbf{D}(f_0 \parallel f_1) + \delta)} \right\},$$

ahol $\delta > 0$ tetszőleges.

$$1 - \alpha_n = \mathbf{P}_0\left(\mathbf{X} \in \mathcal{X}_e^{(n)}\right) = \mathbf{P}_0\left(\frac{1}{n} \sum_{i=1}^{n} \log_2 \frac{f_0(\mathbf{X})}{f_1(\mathbf{X})} \in \Big(\mathbf{D}(f_0 \parallel f_1) - \delta, \mathbf{D}(f_0 \parallel f_1) + \delta\Big) \right).$$

A nagy számok erős törvénye miatt

$$\frac{1}{n}\sum_{i=1}^{n}\log_2\frac{f_0(\mathbf{X})}{f_1(\mathbf{X})}\to\mathbf{D}(f_0\parallel f_1)$$

1-valószínűséggel, így $\forall\delta>0$-hoz elég nagy n-re $\alpha_n<\varepsilon$ teljesül. Másrészt,

$$\beta_n=\mathbf{P}_1\left(\mathbf{X}\in\mathcal{X}_e^{(n)}\right)=\int\limits_{\mathcal{X}_e^{(n)}}L_1(\mathbf{x})\,\mathrm{d}\lambda(\mathbf{x})\le\int\limits_{\mathcal{X}_e^{(n)}}L_0(\mathbf{x})\,2^{-n(\mathbf{D}(f_0\|f_1)-\delta)}\,\mathrm{d}\lambda(\mathbf{x})=$$

$$=2^{-n(\mathbf{D}(f_0\|f_1)-\delta)}\int\limits_{\mathcal{X}_e^{(n)}}L_0(\mathbf{x})\,\mathrm{d}\lambda(\mathbf{x})=2^{-n(\mathbf{D}(f_0\|f_1)-\delta)}(1-\alpha_n).$$

Hasonlóan,

$$\beta_n\ge2^{-n(\mathbf{D}(f_0\|f_1)+\delta)}(1-\alpha_n).$$

Ebből

$$-\mathbf{D}(f_0\parallel f_1)-\delta+\frac{\log_2(1-\alpha_n)}{n}\le\frac{1}{n}\log_2\beta_n\le-\mathbf{D}(f_0\parallel f_1)+\delta+\frac{\log_2(1-\alpha_n)}{n}$$

és $n\to\infty$ határátmenettel

$$\lim_{n\to\infty}\frac{1}{n}\log_2\beta_n=-\mathbf{D}(f_0\parallel f_1)$$

következik, mert $\delta>0$ tetszőleges volt.

Megmutatjuk, hogy nincsen a fenti $\mathcal{X}_e^{(n)}$-nél jobb elfogadási tartománysorozat. Legyen $\mathcal{Y}^{(n)}$ egy másik elfogadási tartománysorozat, melyhez az $\alpha_{n,y},\beta_{n,y}$ elsőfajú- illetve másodfajú hibavalószínűség-sorozat tartozik.

$$\beta_{n,y}=\mathbf{P}_1\left(\mathbf{X}\in\mathcal{Y}^{(n)}\right)\ge\mathbf{P}_1\left(\mathbf{X}\in\mathcal{X}_e^{(n)}\cap\mathcal{Y}^{(n)}\right)=\int\limits_{\mathcal{X}_e^{(n)}\cap\mathcal{Y}^{(n)}}L_1(\mathbf{x})\,\mathrm{d}\lambda(\mathbf{x})\ge$$

$$\ge\int\limits_{\mathcal{X}_e^{(n)}\cap\mathcal{Y}^{(n)}}L_0(\mathbf{x})\,2^{-n(\mathbf{D}(f_0\|f_1)+\delta)}\,\mathrm{d}\lambda(\mathbf{x})\ge2^{-n(\mathbf{D}(f_0\|f_1)+\delta)}\int\limits_{\mathcal{X}_e^{(n)}\cap\mathcal{Y}^{(n)}}L_0(\mathbf{x})\,\mathrm{d}\lambda(\mathbf{x}).$$

A De Morgan azonosságot, majd a Boole-egyenlőtlenséget használva

$$\int\limits_{\mathcal{X}_e^{(n)}\cap\mathcal{Y}^{(n)}}L_0(\mathbf{x})\,\mathrm{d}\lambda(\mathbf{x})=\mathbf{P}_0\left(\mathbf{X}\in\mathcal{X}_e^{(n)}\cap\mathcal{Y}^{(n)}\right)=1-\mathbf{P}_0\left(\mathbf{X}\in\overline{\mathcal{X}}_e^{(n)}\cup\overline{\mathcal{Y}}^{(n)}\right)\ge$$

$$\ge1-\mathbf{P}_0\left(\mathbf{X}\in\overline{\mathcal{X}}_e^{(n)}\right)-\mathbf{P}_0\left(\mathbf{X}\in\overline{\mathcal{Y}}^{(n)}\right)=1-\alpha_{n,\varepsilon}-\alpha_{n,y}$$

adódik, azaz

$$\frac{1}{n}\log_2\beta_{n,y}\ge-\mathbf{D}(f_0\parallel f_1)+\delta+\frac{\log_2(1-\alpha_{n,\varepsilon}-\alpha_{n,y})}{n}.$$

Mivel $\delta>0$ tetszőleges volt, $\lim\limits_{n\to\infty}\frac{1}{n}\log_2\beta_{n,y}\ge-\mathbf{D}(f_0\parallel f_1)$. Tehát, $\mathcal{Y}^{(n)}$ nem jobb $\mathcal{X}_e^{(n)}$-nél, ahol elértük az alsó határt. Ebből az is következik, hogy $\beta_n=\beta_{n,\varepsilon}$.

■

3.3. Paraméteres próbák

Ha adott egy $\mathcal{P} = \{\mathbf{P}_\vartheta,\ \vartheta \in \Theta\}$ paraméteres eloszláscsalád, akkor a

$$\mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1, \quad \text{és} \quad \mathcal{P}_0 \cap \mathcal{P}_1 = \emptyset$$

felbontás helyett a Θ paramétertér

$$\Theta = \Theta_0 \cup \Theta_1, \quad \Theta_0 \cap \Theta_1 = \emptyset$$

diszjunkt felbontása segítségével is megfogalmazhatjuk a hipotéziseinket:

$$H_0 : \vartheta \in \Theta_0, \quad H_1 : \vartheta \in \Theta_1.$$

3.3.1. Egymintás u-próba

Most csak olyan $\mathbf{P}$ valószínűségi mértékeket tekintünk, ahol az $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ minta adott $D_0 > 0$ szórású, ismeretlen m várható értékű normális eloszlású lesz, a ϑ paraméter a várható érték ($\vartheta = m$). $\Theta_0 = \{m_0\}$, $\Theta_1 = \{m \neq m_0\}$, azaz most a nullhipotézis $H_0 : \mathbf{E_P}X = m_0$, az alternatív hipotézis pedig $H_1 : \mathbf{E_P}X \neq m_0$. Azt akarjuk tehát eldönteni, hogy lehet-e a minta várható értéke egy adott m_0 érték, vagy attól szignifikánsan különböző lesz. Ha a H_0 hipotézis igaz, akkor a mintaelemek $N(m_0, D_0)$ eloszlásúak, amiből következik, hogy a mintaátlag statisztika szintén normális eloszlású: $\bar{X}_n \in N(m_0, \frac{D_0}{\sqrt{n}})$. Standardizálás után: $u(\mathbf{X}) = \frac{\bar{X}_n - m_0}{D_0}\sqrt{n} \in N(0,1)$.

A standard normális eloszláshoz a $\Phi(u_\varepsilon) = 1 - \frac{\varepsilon}{2}$ összefüggés alapján megadhatók olyan $K_1(\varepsilon) = -u_\varepsilon$, $K_2(\varepsilon) = u_\varepsilon$ kritikus értékek, melyekre, ha a H_0 hipotézis igaz, akkor fenn kell állnia, hogy $\mathbf{P}(-u_\varepsilon < u(\mathbf{X}) < u_\varepsilon) = 1 - \varepsilon$. Adjuk meg tehát az u-próba kritikus tartományát az $\mathcal{X}_k = \left\{\mathbf{x} : \left|\frac{\bar{x}_n - m_0}{D_0}\sqrt{n}\right| \geq u_\varepsilon\right\}$ definícióval.

A nullhipotézist az adott mintarealizáció felhasználásával az $|u(\mathbf{x})| = \left|\frac{\bar{x}_n - m_0}{D_0}\sqrt{n}\right| < u_\varepsilon$ reláció ellenőrzése alapján döntjük el. Ha az előbbi egyenlőtlenség fennáll, akkor az adott terjedelmen elfogadjuk a nullhipotézist. Ellenkező esetben azt mondjuk, hogy a minta várható értéke szignifikánsan különbözik a hipotetikus m_0 értéktől.

A nullhipotézis annál megbízhatóbban fogadható el, minél nagyobb az ε értéke. A gyakorlatban, ha közel van 1-hez a nullhipotézis erősen igaznak mutatkozik, $\varepsilon \leq 0.01$ esetben viszont csak nagyon nagy elemszámú minta esetén célszerű elfogadni azt.
Az elsőfajú hiba valószínűségére:

$$p_1(\varepsilon, n, m_0) = \mathbf{P}_{m_0}(|u(\mathbf{X})| \geq u_\varepsilon) = 1 - \mathbf{P}_{m_0}(-u_\varepsilon \leq u(\mathbf{X}) \leq u_\varepsilon) =$$

$$= 1 - (\Phi(u_\varepsilon) - \Phi(-u_\varepsilon)) = 2 - 2\Phi(u_\varepsilon) = \varepsilon.$$

A másodfajú hiba valószínűsége pedig:

$$p_2(\varepsilon, n, m) = \mathbf{P}_m(-u_\varepsilon < u(\mathbf{X}) < u_\varepsilon) = \mathbf{P}_m(-u_\varepsilon < \frac{\bar{X}_n - m_0}{D_0}\sqrt{n} < u_\varepsilon) =$$

$$= \mathbf{P}_m\left(-u_\varepsilon - \frac{(m - m_0)\sqrt{n}}{D_0} < \frac{\bar{X}_n - m}{D_0}\sqrt{n} < u_\varepsilon - \frac{(m - m_0)\sqrt{n}}{D_0}\right) =$$

$$= \Phi\left(u_\varepsilon - \frac{(m - m_0)\sqrt{n}}{D_0}\right) - \Phi\left(-u_\varepsilon - \frac{(m - m_0)\sqrt{n}}{D_0}\right),$$

ugyanis az alternatív hipotézis fennállása esetén lesz $\frac{\bar{X}_n - m}{D_0}\sqrt{n} \in N(0,1)$.
Az egymintás u-próba erőfüggvénye:

$$E(\varepsilon, n, m) = 1 - p_2(\varepsilon, n, m) = 1 - \Phi\left(u_\varepsilon - \frac{(m - m_0)\sqrt{n}}{D_0}\right) + \Phi\left(-u_\varepsilon - \frac{(m - m_0)\sqrt{n}}{D_0}\right).$$

3.3.1. tétel: *(Az u-próba tulajdonságai)*
Az u-próba konzisztens és torzítatlan.

Bizonyítás: $a_n = -u_\varepsilon - \frac{(m-m_0)\sqrt{n}}{D_0}$, $b_n = u_\varepsilon - \frac{(m-m_0)\sqrt{n}}{D_0}$ jelöléssel rögzített ε és m mellett:

$$\lim_{n\to\infty} E(\varepsilon, n, m) = \lim_{n\to\infty}(1 + \Phi(a_n) - \Phi(b_n)) = 1,$$

mert $a_n, b_n \to +\infty$, ha $m < m_0$ és $a_n, b_n \to -\infty$, ha $m > m_0$. Így $\Phi(a_n) - \Phi(b_n) \to (1-1)$ vagy $(0-0)$. A fenti határátmenetből következik a próba konzisztenciája.
Rögzített ε és n mellett:

$$\lim_{m\to\infty} E(\varepsilon, n, m) = \lim_{m\to\infty}(1 + \Phi(a_n) - \Phi(b_n)) = 1 + \lim_{m\to\infty}\Phi(a_n) - \lim_{m\to\infty}\Phi(b_n) = 1 + 0 - 0 = 1,$$

$$\lim_{m\to-\infty} E(\varepsilon, n, m) = \lim_{m\to-\infty}(1 + \Phi(a_n) - \Phi(b_n)) = 1 + \lim_{m\to-\infty}\Phi(a_n) - \lim_{m\to-\infty}\Phi(b_n) = 1 + 1 - 1 = 1.$$

Határozzuk meg az $E(\varepsilon, n, m)$ erőfüggvény minimumát az m változónál!

$$\frac{\partial E(\varepsilon, n, m)}{\partial m} = \frac{\sqrt{n}}{D_0}\left[\varphi\left(u_\varepsilon - \frac{(m - m_0)}{D_0}\sqrt{n}\right) - \varphi\left(-u_\varepsilon - \frac{(m - m_0)}{D_0}\sqrt{n}\right)\right] = 0.$$

Mivel $\varphi(x)$ páros függvény, ezért ez csak úgy lehet, ha

$$u_\varepsilon - \frac{(m - m_0)}{D_0}\sqrt{n} = +u_\varepsilon + \frac{(m - m_0)}{D_0}\sqrt{n} \implies m = m_0.$$

$$\frac{\partial^2 E(\varepsilon, n, m)}{\partial m^2} = \frac{n}{D_0^2}\left[\varphi'\left(u_\varepsilon - \frac{(m - m_0)}{D_0}\sqrt{n}\right) - \varphi'\left(-u_\varepsilon - \frac{(m - m_0)}{D_0}\sqrt{n}\right)\right],$$

$$\frac{\partial^2 E(\varepsilon, n, m_0)}{\partial m^2} = \frac{n}{D_0^2}\left[\varphi'(u_\varepsilon) - \varphi'(-u_\varepsilon)\right].$$

Felhasználva, hogy $\varphi'(x) = -x\,\varphi(x)$ kapjuk, hogy

$$\frac{\partial^2 E(\varepsilon, n, m_0)}{\partial m^2} = 2\frac{n \cdot u_\varepsilon}{D_0^2}\varphi(u_\varepsilon) > 0$$

$\implies$ $m = m_0$ minimumhely, és $\min_{m\in\mathbb{R}} E(\varepsilon, n, m) = E(\varepsilon, n, m_0) = \varepsilon \implies$ az u-próba torzítatlan.

■

Megjegyzés: A gyakorlatban akkor is alkalmazzák az u-próbát, amikor a minta nem normális eloszlású. Az alkalmazás jogosságát a centrális határeloszlástétellel lehet indokolni.

3.3.2. A kétmintás u-próba

Adottak az $X_1, X_2, \ldots, X_n$ és az $Y_1, Y_2, \ldots, Y_k$ egymástól független statisztikai minták. Most csak olyan $\mathbf{P}$ valószínűségi mértékeket tekintünk, ahol a minták peremeloszlásai $D_1 > 0$ illetve $D_2 > 0$ ismert szórású, de ismeretlen m_1 illetve m_2 várható értékű normális eloszlásúak, azaz a két mintához tartozó együttes sűrűségfüggvény:

$$f_{m_1,m_2}(x,y) = \frac{1}{2\pi D_1 D_2} \exp\left(-\frac{(x-m_1)^2}{2D_1^2} - \frac{(y-m_2)^2}{2D_2^2}\right).$$

Hipotéziseink: $H_0 : m_1 = m_2$, $H_1 : m_1 \neq m_2$. A feltételek miatt a két minta átlagstatisztikájára: $\bar{X}_n \in N(m_1, \frac{D_1}{\sqrt{n}})$, $\bar{Y}_k \in N(m_2, \frac{D_2}{\sqrt{k}})$. Mivel a két minta független volt, így a különbségükre: $\bar{X}_n - \bar{Y}_k \in N\left(m_1 - m_2, \sqrt{\frac{D_1^2}{n} + \frac{D_2^2}{k}}\right)$. Ha feltesszük, hogy a nullhipotézis igaz, akkor $\bar{X}_n - \bar{Y}_k \in N(0, \sqrt{\frac{D_1^2}{n} + \frac{D_2^2}{k}})$ is fennáll. Standardizálás után: $\frac{\bar{X}_n - \bar{Y}_k}{\sqrt{\frac{D_1^2}{n} + \frac{D_2^2}{k}}} \in N(0,1)$.

Adott $0 < \varepsilon < 1$ esetén, tehát most az elfogadási tartomány:

$$\mathcal{X}_e = \left\{ (\mathbf{x}^T, \mathbf{y}^T)^T : \left| \frac{\bar{x}_n - \bar{y}_k}{\sqrt{\frac{D_1^2}{n} + \frac{D_2^2}{k}}} \right| < u_\varepsilon \right\},$$

ahol az $u_\varepsilon > 0$ kritikus értékre: $\Phi(u_\varepsilon) = 1 - \frac{\varepsilon}{2}$.

A hipotézis eldöntése tehát úgy történik, hogy ha az adott mintarealizációknál teljesül az $\left| \frac{\bar{x}_n - \bar{y}_k}{\sqrt{\frac{D_1^2}{n} + \frac{D_2^2}{k}}} \right| < u_\varepsilon$ reláció, akkor a nullhipotézist az adott ε terjedelmen elfogadjuk, ellenkező esetben pedig elvetjük. Ha a H_0 hipotézist fogadjuk el, úgy is fogalmazhatunk, hogy a két minta várható értékei között „nincsen szignifikáns különbség".

A kétmintás u-próba elsőfajú hibájának valószínűsége is ε.

3.3.3. Az egymintás t-próba

Most csak olyan $\mathbf{P}$ valószínűségi mértékeket tekintünk, ahol az $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ minta ismeretlen $D > 0$ szórású és ismeretlen m várható értékű normális eloszlású lesz, a ϑ paraméter a várható érték ($\vartheta = m$).

$\Theta_0 = \{m_0\}$, $\Theta_1 = \{m \neq m_0\}$. Azaz most a nullhipotézis $H_0 : \mathbf{E_P}X = m_0$, az alternatív hipotézis pedig $H_1 : \mathbf{E_P}X \neq m_0$. Azt akarjuk tehát eldönteni, hogy lehet-e a minta elméleti várható értéke egy adott m_0 érték, vagy attól szignifikánsan különböző. Ha a H_0 hipotézis igaz, akkor a mintaelemek $N(m_0, D)$ eloszlásúak, amiből következik, hogy a mintaátlag-statisztika szintén normális eloszlású: $\bar{X}_n \in N(m_0, \frac{D}{\sqrt{n}})$. Standardizálás után: $\frac{\bar{X}_n - m_0}{D}\sqrt{n} \in N(0,1)$. Az ismeretlen D szórás kiküszöbölését a Lukács-tétel segítségével végezzük. Tudjuk, hogy $\frac{(n-1)s_n^{*2}}{D^2} \in \chi^2_{n-1}$, akár igaz a nullhipotézis, akár nem. Felhasználva a Lukács-tétel utáni megjegyzést: $t(\mathbf{X}) = \frac{\bar{X}_n - m_0}{s_n^*}\sqrt{n} \in t_{n-1}$.

Az $n-1$ szabadságfokú Student-eloszlás táblázatából adott $0 < \varepsilon < 1$-hoz kiolvasható olyan $t_\varepsilon > 0$ kritikus érték, amellyel H_0 fennállása esetén $\mathbf{P}(|t(\mathbf{X})| < t_\varepsilon) = 1 - \varepsilon$ kell, hogy teljesüljön. Így a nullhipotézist aszerint fogadjuk vagy vetjük el, hogy $\left|\frac{\bar{X}_n - m_0}{s_n^*}\sqrt{n}\right| < t_\varepsilon$ fennáll-e vagy sem az adott mintarealizációnál. Mivel $p_1(\varepsilon, n, m_0) = \mathbf{P}(|t(\mathbf{X})| \geq t_\varepsilon) = \varepsilon$, így a t-próba esetében is ε az elsőfajú hiba valószínűsége.

3.3.4. A kétmintás t-próba

Adottak az $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ és az $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_k)^T$ egymástól független statisztikai minták. Most csak olyan $\mathbf{P}$ valószínűségi mértékeket tekintünk, ahol a minták peremeloszlásai $D > 0$ ismeretlen, de egyenlő szórású és ismeretlen m_1 illetve m_2 várható értékű normális eloszlásúak. A két mintához tartozó együttes sűrűségfüggvény:

$$f_{m_1,m_2}(x,y) = \frac{1}{2\pi D^2} \exp\left(-\frac{(x-m_1)^2}{2D^2} - \frac{(y-m_2)^2}{2D^2}\right).$$

Hipotéziseink: $H_0 : m_1 = m_2$, $H_1 : m_1 \neq m_2$. A feltételek miatt a két minta átlagstatisztikájára: $\bar{X}_n \in N(m_1, \frac{D}{\sqrt{n}})$, $\bar{Y}_k \in N(m_2, \frac{D}{\sqrt{k}})$. Mivel a két minta független volt, így a különbségükre: $\bar{X}_n - \bar{Y}_k \in N(m_1 - m_2, D\sqrt{\frac{1}{n} + \frac{1}{k}})$. Ha feltesszük, hogy a nullhipotézis igaz, akkor $\bar{X}_n - \bar{Y}_k \in N(0, D\sqrt{\frac{1}{n} + \frac{1}{k}})$ is fennáll. Standardizálás után: $\frac{\bar{X}_n - \bar{Y}_k}{D\sqrt{\frac{1}{n}+\frac{1}{k}}} \in N(0,1)$.

Ahhoz, hogy az ismeretlen D értéket kiküszöbölhessük, felhasználjuk, hogy $\frac{(n-1){s^*_{X,n}}^2}{D^2} \in \chi^2_{n-1}$, $\frac{(k-1){s^*_{Y,k}}^2}{D^2} \in \chi^2_{k-1}$, valamint azt, hogy az ${s^*_{X,n}}^2$, ${s^*_{Y,k}}^2$, $\bar{X}_n$, $\bar{Y}_k$ statisztikák a feltételek és a Lukács-tétel miatt függetlenek egymástól. Először is $\frac{(n-1){s^*_{X,n}}^2}{D^2} + \frac{(k-1){s^*_{Y,k}}^2}{D^2} \in \chi^2_{n+k-2}$, akár igaz a nullhipotézis, akár nem. Másrészt, a Lukács-tétel után tett megjegyzés értelmében, ha a H_0 hipotézis igaz, akkor

$$t_2(\mathbf{X},\mathbf{Y}) = \frac{\frac{\bar{X}_n - \bar{Y}_k}{D\sqrt{\frac{1}{n}+\frac{1}{k}}}}{\sqrt{\frac{(n-1){s^*_{X,n}}^2}{D^2} + \frac{(k-1){s^*_{Y,k}}^2}{D^2}}} =$$

$$= \frac{\bar{X}_n - \bar{Y}_k}{\sqrt{(n-1){s^*_{X,n}}^2 + (k-1){s^*_{Y,k}}^2}} \sqrt{\frac{nk\,(n+k-2)}{n+k}} \in t_{n+k-2}.$$

A fentiek alapján, az $n+k-2$ szabadságfokú Student-eloszlás táblázatból adott $0 < \varepsilon < 1$ terjedelemhez kiolvasható olyan $t_\varepsilon > 0$ kritikus érték, amellyel H_0 fennállása esetén $\mathbf{P}(|t_2(\mathbf{X},\mathbf{Y})| < t_\varepsilon) = 1 - \varepsilon$ kell, hogy teljesüljön. Így a nullhipotézist aszerint fogadjuk vagy vetjük el, hogy $|t_2(\mathbf{X},\mathbf{Y})| < t_\varepsilon$ fennáll-e vagy sem az adott mintarealizációnál. Mivel $p_1(\varepsilon, n, m_0) = \mathbf{P}(|t_2(\mathbf{X},\mathbf{Y})| \geq t_\varepsilon) = \varepsilon$, így a kétmintás t-próba esetében is ε az elsőfajú hiba valószínűsége.

Megjegyzés: Hangsúlyozzuk, hogy a kétmintás t-próba csak akkor alkalmazható, ha a két minta ismeretlen szórásait egyenlőnek tételezzük fel. (Különben nem tudtuk volna kiküszöbölni a $t_2(\mathbf{X},\mathbf{Y})$ próbastatisztikából D-t!) A minták szórásainak egyezését az F-próbával ellenőrizhetjük, tehát ennek meg kell előznie a kétmintás t-próbát.

Megmutatható, hogy ha X n szabadságfokú χ^2-eloszlású és Y tőle független k szabadságfokú χ^2-eloszlású, akkor a $Z = \frac{\frac{X}{n}}{\frac{Y}{k}}$ valószínűségi változó sűrűségfüggvénye

$$f_Z(x) = \frac{\Gamma\left(\frac{n+k}{2}\right)}{\Gamma(n)\Gamma(k)} x^{\frac{k}{2}-1}(k+nx)^{-\frac{k+n}{2}}, \qquad x > 0$$

lesz. Z eloszlását n, k paraméterű F- (Fisher-) eloszlásnak nevezzük, és $F_{n,k}$-val jelöljük.

3.3.5. Az F-próba

Adottak az $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ és az $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_k)^T$ egymástól független statisztikai minták. Most csak olyan $\mathbf{P}$ valószínűségi mértékeket tekintünk, ahol a minták peremeloszlásai $D_1 > 0$ illetve $D_2 > 0$ ismeretlen szórású és ismeretlen m_1 illetve m_2 várható értékű normális eloszlásúak. A két mintához tartozó együttes sűrűségfüggvény:

$$f_{m_1,m_2}(x,y) = \frac{1}{2\pi D_1 D_2} \exp\left(-\frac{(x-m_1)^2}{2D_1^2} - \frac{(y-m_2)^2}{2D_2^2}\right).$$

Felállított hipotézisek most a szórások egyezésére, illetve szignifikáns különbségére vonatkoznak: $H_0 : D_1 = D_2$, $H_1 : D_1 \neq D_2$. Ha feltesszük, hogy a nullhipotézis igaz, akkor a Lukács-tétel szerint igaz lesz, hogy $\frac{(n-1){s^*_{X,n}}^2}{D^2} \in \chi^2_{n-1}$, $\frac{(k-1){s^*_{Y,k}}^2}{D^2} \in \chi^2_{k-1}$, ahol $D_1 = D_2 = D$. A minták függetlensége miatt a két statisztika is független lesz.

Mivel független χ^2 eloszlású valószínűségi változók hányadosa F-eloszlású:

$$\frac{\frac{\frac{(n-1){s^*_{X,n}}^2}{D^2}}{n-1}}{\frac{\frac{(k-1){s^*_{Y,k}}^2}{D^2}}{k-1}} = \frac{{s^*_{X,n}}^2}{{s^*_{Y,k}}^2} \in F_{n-1,k-1},$$

azaz a minták korrigált empirikus szórásnégyzeteinek hányadosa $n-1, k-1$ szabadságfokú Fisher-eloszlást fog követni, ha a nullhipotézis igaz. Ezek alapján a nullhipotézis eldöntésére a kritikus tartományt úgy szerkeszthetjük meg, hogy adott $0 < \varepsilon < 1$ terjedelemhez az $n-1, k-1$ szabadságfokú F-eloszlás táblázatából kiolvasunk olyan $0 < K_1 < K_2$ kritikus értékeket, amelyekre $\mathbf{P}(K_1 < F_{n-1,k-1}) = 1 - \frac{\varepsilon}{2}$, $\mathbf{P}(K_2 < F_{n-1,k-1}) = \frac{\varepsilon}{2}$. Ha az adott mintarealizációnál $K_1 < \frac{{s^*_{X,n}}^2}{{s^*_{Y,k}}^2} < K_2$ reláció teljesül, akkor a nullhipotézist elfogadjuk, ellenkező esetben pedig elvetjük. A próba elsőfajú hibájának a valószínűsége most is ε, a másodfajú hiba valószínűsége az n és k mintaelemszámoktól, ε-tól és a $D_1 - D_2$ különbségtől függ.

Megjegyzés:

1. Ha $\varepsilon < 0.33$, n és k kettőnél nagyobb mintaelemszámok (ez gyakorlatilag mindig fennáll), akkor a $0 < K_1 < K_2$ kritikus értékekre mindig teljesül a $K_1 < 1 < K_2$ reláció. Így, ha ${s^*_{X,n}}^2$, ${s^*_{Y,k}}^2$ közül a nagyobbikat írjuk a számlálóba, a próba eldöntéséhez elég a próbastatisztika értékét csupán K_2 -vel összehasonlítani. Ha a számított érték kisebb, mint K_2, a nullhipotézist elfogadjuk. Ilyenkor az F-eloszlás táblázatából egyetlen kritikus érték meghatározása elégséges, de ügyeljünk arra, hogy az első szabadságfok mindig abból a mintaelemszámból képződik, amelyhez tartozó korrigált empirikus szórásnégyzet statisztika a számlálóban van!

2. Statisztikai elemzéseket napjainkban valamilyen statisztikai programrendszer segítségével szokás elvégezni. A programok egy próba esetén mindig azt a $0 < \varepsilon < 1$ elsőfajú hibavalószínűséget adják meg eredményül, amelynél már elfogadható a nullhipotézis. Ha tehát túl közel van 0-hoz, akkor az azt jelenti, hogy a nullhipotézist el kell vetni. 0.01 nél kisebb elsőfajú hibavalószínűség mellett „nem illik" elfogadni H_0-t, míg 0.1 felett a nullhipotézis fennállása erősnek mutatkozik. A két szélső érték között a felhasználó felelőssége, hogy elfogadja, vagy elveti H_0-t, vagy esetleg újabb mintavételezéssel bővíti a mintát (mintákat), majd megismétli a próbát. A mintaelemszám növelésével nő a próba ereje, tehát nagy n esetén kisebb ε terjedelem mellett is elfogadható a nullhipotézis.

3.3.6. A Welch-próba

Ha az F-próbát el kell vetnünk, nem alkalmazható a kétmintás t-próba arra, hogy ellen[rizz-k a két minta várható értékeinek egyezését. Erre az esetre dolgozta ki Welch az alábbi próbát. Adottak az $X_1, X_2, \ldots, X_n$ és az $Y_1, Y_2, \ldots, Y_k$ egymástól független statisztikai minták. Most is csak olyan $\mathbf{P}$ valószínűségi mértékeket tekintünk, ahol a minták peremeloszlásai $D_1 > 0$ illetve $D_2 > 0$ ismeretlen szórású és ismeretlen m_1 illetve m_2 várható értékű normális eloszlásúak. A két mintához tartozó együttes sűrűségfüggvény:

$$f_{m_1,m_2}(x,y) = \frac{1}{2\pi D_1 D_2} \exp\left(-\frac{(x-m_1)^2}{2D_1^2} - \frac{(y-m_2)^2}{2D_2^2}\right).$$

A hipotézisek ugyanazok mint a kétmintás t-próbánál voltak: $H_0 : m_1 = m_2$, $H_1 : m_1 \neq m_2$. Megmutatható, hogy a nullhipotézis fennállása esetén a $W_{n,k} = \frac{\bar{X}_n - \bar{Y}_k}{\sqrt{\frac{s^2_{X,n}}{n} + \frac{s^2_{Y,k}}{k}}}$ próbastatisztika közelítőleg Student-eloszlású $[f]$ (egészrész f) szabadságfokkal, ahol $\frac{1}{f} = \frac{c^2}{k-1} + \frac{(1-c^2)}{n-1}$, $c = \frac{\frac{s^2_{Y,k}}{k}}{\frac{s^2_{Y,k}}{k} + \frac{s^2_{X,n}}{n}}$. A kritikus értéket a Student-eloszlás táblázatából kiolvasva dönthetünk a szokásos módon a nullhipotézisről: elfogadjuk, ha az adott realizációknál a $|W_{n,k}|$ számított érték kisebb lesz, mint a kritikus érték. Ha $n, k \geq 40$, akkor a centrális határeloszlás-tétel alapján $W_{n,k} \approx N(0, \frac{[f]}{[f]-2})$, azaz akkor a normális eloszlás táblázatából is kiolvashatjuk a kritikus értéket.

3.4. Nemparaméteres próbák

Ha az alapsokaság (a statisztikai minta) eloszlását nem tekintjük eleve ismertnek, azaz nem tudjuk, hogy az egy adott paraméteres eloszláscsalád eleme, akkor nemparaméteres próbákról beszélünk. Ilyenkor tehát az előzetes feltevéseink nagyon általánosak, de természetesek; pl. feltesszük, hogy a minta eloszlása folytonos, vagy feltesszük, hogy a szórás véges, stb. Nyilvánvaló, mivel kevesebb feltételt követelünk meg kiinduláskor (a priori feltevések), a következtetéseink levonásához nagyobb elemszámú mintákra lesz szükségünk, mint a paraméteres próbák esetén.

3.4.1. χ^2-próbák

Az ismertetendő próbák mindegyike az alábbi alaptételen alapszik. Ehhez előkészítésképpen hivatkoznunk kell a polinomiális eloszlás definíciójára és a valószínűségi vektor karakterisztikus függvényének definíciójára. Ezek alapján a $\mathbf{V} \in Pol(n, p_1, p_2, \ldots, p_r)$ valószínűségi vektorváltozó karakterisztikus függvénye:

$$\varphi_{\mathbf{V}}(t_1, t_2, \ldots, t_r) = \mathbf{E}\mathrm{e}^{i\mathbf{V}^T\mathbf{t}} = \sum_{\substack{\forall k_1,k_2,\ldots,k_r \\ k_1+k_2+\ldots+k_r=n}} \frac{n!}{k_1!k_2!\ldots k_r!} p_1^{k_1} p_2^{k_2} \cdots p_r^{k_r} \mathrm{e}^{i\sum_{j=1}^{r} k_j t_j} =$$

$$= \left(p_1\mathrm{e}^{it_1} + p_2\mathrm{e}^{it_2} + \cdots + p_r\mathrm{e}^{it_r}\right)^n.$$

3.4.1. tétel: Ha $\mathbf{V} = (V_1, V_2, \ldots, V_r)^T$ egy $n, p_1, p_2, \ldots, p_r$ paraméterű polinomiális eloszlású valószínűségi vektorváltozó, akkor $\sum_{i=1}^{r} \frac{(V_i - np_i)^2}{np_i} \xrightarrow{e} \chi^2_{r-1}$ $(n \to \infty)$.

Bizonyítás: A bizonyítás a Helly-tételen alapul. Azt fogjuk megmutatni, hogy $\sum\limits_{i=1}^{r}\frac{(V_i-np_i)^2}{np_i}$ karakterisztikus függvényeinek sorozata egyenletesen konvergál χ^2_{r-1} karakterisztikus függvényéhez, vagyis $r-1$ teljesen független standard normális eloszlás négyzetösszegének karakterisztikus függvényéhez.

Először kiszámítjuk a $\tilde{V}_i=\frac{(V_i-np_i)}{\sqrt{np_i}}$ standardizáltak karakterisztikus függvényét.

$$\varphi_{\tilde{\mathbf{V}}}(t_1,t_2,\dots,t_r)=\mathrm{e}^{-i\sum\limits_{j=1}^{r}\sqrt{np_j}t_j}\varphi_{\mathbf{V}}\left(\frac{t_1}{\sqrt{np_1}},\frac{t_2}{\sqrt{np_2}},\dots,\frac{t_r}{\sqrt{np_r}}\right)=$$

$$=\mathrm{e}^{-i\sum\limits_{j=1}^{r}\sqrt{np_j}t_j}\left(1+\sum_{j=1}^{r}p_j\left(\mathrm{e}^{i\frac{t_j}{\sqrt{np_j}}}-1\right)\right)^n$$

Felhasználva az $\mathrm{e}^x=1+x+\frac{x^2}{2}+O(x^3)$, $\ln(1+x)=x-\frac{x^2}{2}+O(x^3)$ $(x\in[-1,1])$ McLaurin-sorfejtéseket: $\mathrm{e}^{i\frac{t_j}{\sqrt{np_j}}}-1=\frac{it_j}{\sqrt{np_j}}-\frac{t_j^2}{2np_j}+O(n^{-\frac{3}{2}})$, és így

$$\ln\,\varphi_{\tilde{\mathbf{V}}}(\mathbf{t})=-i\sqrt{n}\sum_{j=1}^{r}\sqrt{p_j}t_j+n\ln\left(1+\frac{i}{\sqrt{n}}\sum_{j=1}^{r}\sqrt{p_j}t_j-\frac{1}{2n}\sum_{j=1}^{r}t_j^2+O(n^{-\frac{3}{2}})\right)=$$

$$=-i\sqrt{n}\sum_{j=1}^{r}\sqrt{p_j}\,t_j+n\frac{i}{\sqrt{n}}\sum_{j=1}^{r}\sqrt{p_j}t_j-\frac{1}{2}\sum_{j=1}^{r}t_j^2+\frac{1}{2}\left(\sum_{j=1}^{r}\sqrt{p_j}t_j\right)^2+O(n^{-\frac{1}{2}})=$$

$$=-\frac{1}{2}\sum_{j=1}^{r}t_j^2+\frac{1}{2}\left(\sum_{j=1}^{r}\sqrt{p_j}t_j\right)^2+O(n^{-\frac{1}{2}}).$$

A fentiek alapján $\lim\limits_{n\to\infty}\ln\,\varphi_{\tilde{\mathbf{V}}}(\mathbf{t})=-\frac{1}{2}\sum\limits_{j=1}^{r}t_j^2+\frac{1}{2}\left(\sum\limits_{j=1}^{r}\sqrt{p_j}t_j\right)^2$.

A Schmidt-féle ortogonalizálási eljárással megadható olyan r-edrendű ortonormált mátrix, melynek utolsó sora a $\sqrt{p_1},\sqrt{p_2},\dots,\sqrt{p_r}$ elemekből áll:

$$\underline{\underline{\Gamma}}=\begin{pmatrix}\gamma_{11}&\gamma_{12}&\cdots&\gamma_{1r}\\\gamma_{21}&\gamma_{22}&\cdots&\gamma_{2r}\\\vdots&\vdots&\ddots&\vdots\\\sqrt{p_1}&\sqrt{p_2}&\cdots&\sqrt{p_r}\end{pmatrix},\ \underline{\underline{\Gamma}}^T\underline{\underline{\Gamma}}=\underline{\underline{\Gamma\Gamma}}^T=\underline{\underline{E}}_r.$$

Tekintsük ezek után a $\mathbf{Z}=\underline{\underline{\Gamma}}\tilde{\mathbf{V}}$ transzformáltat. Nyilván:

$$\mathbf{Z}^T\mathbf{Z}=\tilde{\mathbf{V}}^T\underline{\underline{\Gamma}}^T\underline{\underline{\Gamma}}\tilde{\mathbf{V}}=\tilde{\mathbf{V}}^T\underline{\underline{E}}\tilde{\mathbf{V}}=\tilde{\mathbf{V}}^T\tilde{\mathbf{V}}$$

és

$$Z_r=\sum_{j=1}^{r}\sqrt{p_j}\,\tilde{V}_j.$$

Továbbá, ha

$$\mathbf{u}=\underline{\underline{\Gamma}}\mathbf{t}\Longrightarrow\sum_{j=1}^{r}u_j^2=\sum_{j=1}^{r}t_j^2,u_r=\sum_{j=1}^{r}\sqrt{p_j}t_j\Longrightarrow\sum_{j=1}^{r-1}u_j^2=\sum_{j=1}^{r}t_j^2-\left(\sum_{j=1}^{r}\sqrt{p_j}t_j\right)^2.$$

Tehát $\ln \varphi_{\tilde{\mathbf{V}}}(\mathbf{t}) = \ln \varphi_{\mathbf{Z}}(\underline{\underline{\Gamma}}\,\mathbf{t}) = \ln \varphi_{\mathbf{Z}}(\mathbf{u})$. Ezért

$$\lim_{n\to\infty} \ln \varphi_{\tilde{\mathbf{V}}}(\mathbf{t}) = -\frac{1}{2}\left(\sum_{j=1}^{r} t_j^2 + \left(\sum_{j=1}^{r} \sqrt{p_j}t_j\right)^2\right) = -\frac{1}{2}\sum_{j=1}^{r-1} u_j^2 = \lim_{n\to\infty} \ln \varphi_{\mathbf{Z}}(\mathbf{u}).$$

Tehát

$$\lim_{n\to\infty} \varphi_{\mathbf{Z}}(\mathbf{u}) = \mathrm{e}^{-\frac{1}{2}\sum\limits_{j=1}^{r-1} u_j},$$

vagyis $\mathbf{Z} \in \mathbb{R}^r$ karakterisztikus függvénye $r-1$ darab teljesen független standard normális eloszlású valószínűségi változó karakterisztikus függvényéhez konvergál egyenletesen. Ebből már egyenesen következik, hogy akkor $\tilde{\mathbf{V}}^T\tilde{\mathbf{V}} = \mathbf{Z}^T\mathbf{Z} = \sum\limits_{j=1}^{r} Z_j^2$ karakterisztikus függvénye $r-1$ darab teljesen független standard normális eloszlású valószínűségi változó négyzetösszegének karakterisztikus függvényéhez konvergál egyenletesen, ami pedig az $r-1$ szabadságfokú χ^2 eloszlás!

■

Tiszta illeszkedésvizsgálat

Adott az $X_1, X_2, \ldots, X_n$ statisztikai minta. Ellenőrizni akarjuk azt a feltevést, hogy a minta eloszlásfüggvénye éppen az $F_0(x)$, az összes szóbajöhető eloszlásfüggvény között. $F_0(x)$-nek nincsenek ismeretlen paraméterei, egy bizonyos, konkrét eloszlásfüggvény. A nullhipotézisünk most $H_0 : \mathbf{P}(X < x) \equiv F_0(x)$, míg az alternatív hipotézis $H_1 : \mathbf{P}(X < x) \not\equiv F_0(x)$. Vagyük a számegyenesnek egy tetszőleges r diszjunkt intervallumból álló felosztását. Legyen $-\infty < x_1 < x_2 < \cdots < x_{r-1} < \infty$, $I_k = [x_{k-1}, x_k)$, $(k = 1, 2, \ldots, r)$, $x_0 = -\infty$, $x_r = +\infty$. Ha H_0 igaz, akkor $p_k = \mathbf{P}(X \in I_k) = F_0(x_k) - F_0(x_{k-1})$.

Jelölje V_k azt a gyakoriságot, ahány mintaelemre teljesült az $X_j \in I_k$ reláció, azaz $V_k = \sum\limits_{i=1}^{n} I(X_i \in I_k)$. Ha összevetjük ezt a polinomiális eloszlás definíciójával láthatjuk, hogy $\mathbf{V} = (V_1, V_2, \ldots, V_r)^T$ egy $n, p_1, p_2, \ldots, p_r$ paraméterű polinomiális eloszlású valószínűségi vektorváltozó lesz! De ekkor a 3.4.1. tételt alkalmazva, $\sum\limits_{i=1}^{r} \frac{(V_i - np_i)^2}{np_i} \xrightarrow{e} \chi^2_{r-1}$ $(n \to \infty)$. Vagyis, ha nagy a mintaelemszám, a $T_n = \sum\limits_{i=1}^{r} \frac{(V_i - np_i)^2}{np_i} = \sum\limits_{i=1}^{r} \frac{V_i^2}{np_i} - n$ statisztika a nullhipotézis fennállása esetén közelítőleg $r-1$ szabadságfokú χ^2-eloszlást követ. Erre alapozhatjuk a döntési eljárásunkat. Adott $0 < \varepsilon < 1$ terjedelemhez meghatározunk olyan K_ε kritikus értéket, amellyel $\mathbf{P}(\chi^2_{r-1} < K_\varepsilon) = 1 - \varepsilon$. Ezek után, ha az adott statisztikai minta realizációjánál teljesül a $T_n < K_\varepsilon$ reláció, a nullhipotézist elfogadjuk, ellenkező esetben pedig elvetjük. Az elsőfajú hibavalószínűség most csak aszimptotikusan lesz ε.

Megjegyzés:

1. Alkalmazásokban az $x_1 < x_2 < \cdots < x_{r-1}$ osztópontokat úgy célszerű megválasztani, hogy a realizálódott mintánál $V_i \geq 10$ és $p_i \approx \frac{1}{r}$ legyen minden i-re.

2. Ha $r \geq 30$, akkor a χ^2-eloszlás táblázat helyett a normális eloszlás táblázatát is használhatjuk, mert ilyenkor már $T_n \approx \chi^2_{r-1} \approx N(r-1, \sqrt{2r-2})$.

3. Ha a statisztikai minta diszkrét eloszlású, akkor az intervallumok helyett a minta értékkészletének diszjunkt felbontását vesszük. Például, ha a k-adik partíciót az $I_k = \{z_1, z_2, \ldots, z_{n_k}\}$ számhalmaz jelenti, akkor $p_k = \sum\limits_{i=1}^{n_k} \mathbf{P}(X = z_i)$.

Becsléses illeszkedésvizsgálat

Adott az $X_1, X_2, \ldots, X_n$ statisztikai minta. Ellenőrizni akarjuk azt a feltevést, hogy a minta eloszlásfüggvénye $F_{\boldsymbol{\vartheta}}(x)$ alakú, az összes szóbajöhető eloszlásfüggvény között. $F_{\boldsymbol{\vartheta}}(x)$ egy k-paraméteres eloszláscsalád eleme. A nullhipotézisünk most

$$H_0 : \exists \boldsymbol{\vartheta} \in \mathbb{R}^k : \mathbf{P}(X < x) \equiv F_{\boldsymbol{\vartheta}}(x),$$

míg az alternatív hipotézis

$$H_1 : \nexists \boldsymbol{\vartheta} \in \mathbb{R}^k : \mathbf{P}(X < x) \equiv F_{\boldsymbol{\vartheta}}(x).$$

A próba végrehajtása nagyon hasonlít az előző esetre, csak először venni kell a $\boldsymbol{\vartheta}$ paramétervektor $\mathbf{t}_n$ konzisztens becslését, majd az adott mintarealizációnál kapott $\boldsymbol{\vartheta} = \mathbf{t}_n$ becsléssel képezzük az $F_0(x) = F_{\boldsymbol{\vartheta}}(x)$ eloszlásfüggvényt, ami már konkrét, hiszen ismeretlen paramétereket már nem tartalmaz. Ezután végrehajtva mindazt, amit a tiszta illeszkedésvizsgálatnál leírtunk, kiszámoljuk a T_n próbastatisztikát. A különbség csak ott jelentkezik, hogy most az mutatható meg, hogy $T_n \xrightarrow{e} \chi^2_{r-1-k}$, ahol k a becsült paraméterek száma. Ezek alapján a döntési algoritmus az előzőekhez hasonlóan történik.

Függetlenségvizsgálat

Legyen $(X_1, Y_1)^T, (X_2, Y_2)^T, \ldots, (X_n, Y_n)^T$ n elemszámú kétdimenziós statisztikai minta. Ellenőrizni akarjuk, hogy a minta komponensei függetlenek-e egymástól, vagy pedig szignifikáns sztochasztikus összefüggés tapasztalható-e közöttük:

$$H_0 : \mathbf{P}(X_i < x, Y_i < y) = \mathbf{P}(X_i < x)\mathbf{P}(Y_i < y) \ \forall x, y;$$

$$H_1 : \mathbf{P}(X_i < x, Y_i < y) \not\equiv \mathbf{P}(X_i < x)\mathbf{P}(Y_i < y).$$

Legyen $-\infty < x_1 < x_2 < \cdots < x_{r-1} < \infty$, $I_k = [x_{k-1}, x_k)$, $(k = 1, 2, \ldots, r)$, $x_0 = -\infty$, $x_r = +\infty$ és $-\infty < y_1 < y_2 < \cdots < y_{s-1} < \infty$, $J_k = [y_{k-1}, y_k)$, $(k = 1, 2, \ldots, s)$, $y_0 = -\infty$, $y_s = +\infty$ két különböző partícióra bontása $\mathbb{R}$-nek. Azért kell két különböző felosztást tekintenünk, mert a két minta értékei másképpen oszolhatnak el a számegyenesen; az első felosztás az első komponens értékkészletét, a második partíció a második komponens értékkészletét fedi le.

Jelölje: V_{ij} azon mintaelemek számát, ahol $(X_k, Y_k)^T \in I_i \times J_j$ teljesül,

$$V_{i\cdot} = \sum_{j=1}^{s} V_{ij} = \sum_{k=1}^{n} I(X_k \in I_i), \ V_{\cdot j} = \sum_{i=1}^{r} V_{ij} = \sum_{k=1}^{n} I(Y_k \in J_j).$$

A $p_{ij} = \mathbf{P}(X_k \in I_i, Y_k \in J_j)$, $p_{i\cdot} = \mathbf{P}(X_k \in I_i) = \sum_{j=1}^{s} p_{ij}$, $p_{\cdot j} = \mathbf{P}(Y_k \in J_j) = \sum_{i=1}^{r} p_{ij}$ valószínűségek most nem ismertek, de azokat a relatív gyakoriságok segítségével becsülni lehet:

$$p_{i\cdot} \approx \hat{p}_{i\cdot} = \frac{1}{n} V_{i\cdot} = \frac{1}{n} \sum_{j=1}^{s} V_{ij}, \quad p_{\cdot j} \approx \hat{p}_{\cdot j} = \frac{1}{n} V_{\cdot j} = \frac{1}{n} \sum_{i=1}^{r} V_{ij}.$$

A becslések száma $r-1$ illetve $s-1$, mivel eloszlásokról van szó, és így $\sum_{i=1}^{r} p_{i\cdot} = \sum_{j=1}^{s} p_{\cdot j} = 1$, vagyis $p_{r\cdot} = 1 - \sum_{i=1}^{r-1} p_{i\cdot}$ és $p_{\cdot s} = 1 - \sum_{j=1}^{s-1} p_{\cdot j}$, azaz az eloszlás utolsó elemei már a többi becslésből számolhatók.

Most tehát becsléses illeszkedésvizsgálatot kell végrehajtani, ahol a becsült paraméterek száma: $r-1+s-1=r+s-2$. A $T_n=\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{s}\frac{(V_{ij}-n\,\hat{p}_{i.}\hat{p}_{.j})^2}{n\,\hat{p}_{i.}\hat{p}_{.j}}=n\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{s}\frac{V_{ij}^2}{V_{i.}V_{.j}}-n$ próbastatisztika eloszlása aszimptotikusan $rs-1-(r+s-2)=(r-1)(s-1)$ szabadságfokú χ^2 -eloszlású lesz.

A nullhipotézis eldöntéséhez táblázatból meg kell határoznunk olyan K_ε kritikus értéket, amelyre $\mathbf{P}(\chi^2_{(r-1)(s-1)}<K_\varepsilon)=1-\varepsilon$ teljesül. Ha a T_n számított értéke kisebb mint a K_ε kritikus érték, a nullhipotézist az 1- ε szignifikancia szinten elfogadjuk, ellenkező esetben az alternatív hipotézist tartjuk igaznak, azaz a komponensek között szignifikáns összefüggést regisztrálunk.

Homogenitásvizsgálat

A homogenitásvizsgálat annak a kérdésnek az eldöntésére szolgál, hogy két valószínűségi változó azonos eloszlású-e, azaz ugyanaz a függvény-e az eloszlásfüggvényük, vagy sem. Adottak az $X_1,X_2,\ldots,X_n$ és az $Y_1,Y_2,\ldots,Y_m$ statisztikai minták, amelyek egymástól is függetlenek. Eldöntendő, hogy:

$$H_0:\ \mathbf{P}(X<x)\equiv\mathbf{P}(Y<x)\quad\text{vagy}\quad H_1:\ \mathbf{P}(X<x)\not\equiv\mathbf{P}(Y<x).$$

Tekintsük most a

$$-\infty<x_1<x_2<\cdots<x_{r-1}<\infty,\ I_k=[x_{k-1},x_k)\,,\ (k=1,2,\ldots,r)\,,\ x_0=-\infty\,,\ x_r=+\infty$$

felosztást. A két minta ellenére elég most egyetlen intervallumrendszer, hiszen a homogenitás fennállása esetén ugyanaz a két változó értékkészlete. A minták és a felosztás segítségével definiáljuk a $V_k=\sum\limits_{i=1}^{n}I(X_i\in I_k)$, $U_k=\sum\limits_{i=1}^{n}I(Y_i\in I_k)$ $(k=1,2,\ldots,r)$ gyakoriságokat. A nullhipotézis fennállása esetén a két minta egyesítése is statisztikai minta.

Nyilvánvalóan: $\sum\limits_{i=1}^{r}V_i=n$, $\sum\limits_{i=1}^{r}U_i=m$.

H_0 átfogalmazható úgy, hogy az egy az $1,2,\ldots,r$ értékeket p_i hibavalószínűséggel felvevő valószínűségi változó illeszkedésére vonatkozzék, amelyhez $n+m$ elemszámú megfigyeléssorozat tartozik. A p_i értékeket nem ismerjük, de a mintákból a relatív gyakoriságokkal becsülni tudjuk: $p_i\approx\hat{p}_i=\frac{V_i+U_i}{n+m}$. Összesen $r-1$ becslést alkalmazunk, mivel az r-edik eloszláselem a többiből számolható. Tehát megint becsléses illeszkedésvizsgálatról van szó. A tiszta illeszkedésvizsgálatnál elmondottak szerint a $T_n^*=\sum\limits_{i=1}^{r}\frac{(V_i-np_i)^2}{np_i}$ és a $T_m^{**}=\sum\limits_{i=1}^{r}\frac{(U_i-mp_i)^2}{mp_i}$ statisztikák aszimptotikusan $r-1$ szabadságfokú χ^2-eloszlást követnek, ha H_0 igaz. Az összegük viszont akkor $2r-2$ szabadságfokú χ^2 -eloszlású lesz: $T_n^*+T_m^{**}\xrightarrow{e}\chi^2_{2r-1}$. Az összesen $r-1$ db paraméterbecslés miatt azonban, ahogy arra a becsléses illeszkedésvizsgálatnál utaltunk, a szabadságfokot $r-1$-gyel csökkenteni kell:

$$\sum_{i=1}^{r}\frac{(V_i-n\hat{p}_i)^2}{n\hat{p}_i}+\sum_{i=1}^{r}\frac{(U_i-m\hat{p}_i)^2}{m\hat{p}_i}=\sum_{i=1}^{r}\frac{\left(V_i-n\frac{V_i+U_i}{n+m}\right)^2}{n\frac{V_i+U_i}{n+m}}+\sum_{i=1}^{r}\frac{\left(U_i-m\frac{V_i+U_i}{n+m}\right)^2}{m\frac{V_i+U_i}{n+m}}=$$

$$=nm\sum_{i=1}^{r}\frac{\left(\frac{V_i}{n}-\frac{U_i}{m}\right)^2}{V_i+U_i}\xrightarrow{e}\chi^2_{r-1}.$$

A H_0 hipotézis eldöntéséhez, tehát az $r-1$ szabadságfokú χ^2-eloszlás táblázatból meghatározzuk azt a K_ε kritikus értéket, amelyre $1-\varepsilon=\mathbf{P}\left(\chi^2_{r-1}<K_\varepsilon\right)$ teljesül. Ezek után a H_0-t elfogadjuk, ha az adott realizálódott mintánál $nm\sum\limits_{i=1}^{r}\frac{\left(\frac{V_i}{n}-\frac{U_i}{m}\right)^2}{V_i+U_i}<K_\varepsilon$ teljesül.

3.4.2. Kolmogorov–Szmirnov-próbák

A χ^2-próbáknak az a hátránya, hogy csak nagy elemszámú minták esetén használhatók, ami a mintavételezés költségeit növeli. Másrészt nincs egyértelmű szabály a csoportok kialakítására, így a számítógépes megvalósítás is nehézkesebb. A rendezett mintákon alapuló Kolmogorov–Szmirnov-próbák kiküszöbölik az említett hátrányokat. Miután itt a konvergencia sebessége nagyobb, kisebb mintaelemszám is elégséges a próba sikeres végrehajtásához. (A minták rendezése ugyanis plusz információt jelent).

Az egymintás Kolmogorov–Szmirnov-próba illeszkedésvizsgálatra az alábbi tételen alapszik:

3.4.2. tétel: Legyen $X_1, X_2, \ldots, X_n, \ldots$ statisztikai minta, melynek eloszlásfüggvénye $F_0(x)$ abszolút folytonos. Jelölje

$$F_n(x) = \begin{cases} 0, & x \le X_1^* \\ \frac{k}{n}, & X_k^* < x \le X_{k+1}^* \quad (k = 1, 2, \ldots, n-1) \\ 1, & x > X_n^* \end{cases}$$

az empirikus eloszlásfüggvényt, ahol $X_1^* \le X_2^* \le \ldots \le X_n^*$ a rendezett minta, és legyen $D_n = \sqrt{n} \sup\limits_{x \in \mathbb{R}} |F_n(x) - F_0(x)|$. Ekkor

$$\lim_{n \to \infty} \mathbf{P}(D_n < y) = \begin{cases} K(y), & y > 0 \\ 0, & y \le 0 \end{cases},$$

ahol $K(y) = \sum\limits_{i=-\infty}^{\infty} (-1)^i \, \mathrm{e}^{-2i^2y^2}$, $y > 0$ a Kolmogorov-eloszlásfüggvény, azaz a D_n statisztika eloszlása $n \to \infty$ esetben az ún. Kolmogorov-eloszlást adja.

Bizonyítás: A tételt nem bizonyítjuk.
Megjegyzés:

1. Figyeljük meg, hogy $K(y)$ nem függ az $F_0(x)$ eloszlásfüggvénytől.

2. Mivel $F_n(x)$ mindig lépcsős függvény, ezért elég csak az ugráshelyeken vett különbségek maximumát venni:

$$D_n = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| = \sqrt{n} \max_{i=1,2,\ldots,n} |F_n(X_i^*) - F_0(X_i^*)| .$$

3. A Kolmogorov-eloszlásfüggvényre vonatkozó táblázat:

$K(x_\varepsilon)$	x_ε
0.9	1.23
0.95	1.36
0.99	1.63
0.999	1.96

A tétel segítségével próba szerkeszthető egy adott mintának a hipotetikus $F_0(x)$ eloszlásfüggvényhez való illeszkedésére.

$H_0: \ \mathbf{P}(X_i < x) \equiv F_0(x)$ és $H_1: \ \mathbf{P}(X_i < x) \not\equiv F_0(x)$.

Legyen most $0 < \varepsilon < 1$. A nullhipotézist akkor fogadjuk el $1 - \varepsilon$ szignifikancia szinten, ha $D_n < x_\varepsilon$ teljesül, ahol $K(x_\varepsilon) = 1 - \varepsilon$.

A kétmintás Kolmogorov–Szmirnov-féle próba homogenitásvizsgálatra pedig az alábbi tételen alapszik.

3.4.3. tétel: *(Kolmogorov)*
Legyen az $X_1, X_2, \ldots, X_n, \ldots$ statisztikai minta, melynek eloszlásfüggvénye $F(x)$, és $Y_1, Y_2, \ldots, Y_m, \ldots$ az előzőtől független másik statisztikai minta, melynek eloszlásfüggvénye $G(x)$. F és G abszolút folytonosak. Jelölje $F_n(x)$ és $G_m(x)$ a két mintához tartozó empirikus eloszlásfüggvényet.

Ha $F(x) \equiv G(x)$, akkor a $D_{n,m} = \sqrt{\frac{nm}{n+m}} \sup_{x\in\mathbb{R}} |F_n(x) - G_m(x)|$ statisztika eloszlásban a Kolmogorov-eloszláshoz tart, azaz

$$\lim_{n,m\to\infty} \mathbf{P}(D_{n,m} < y) = \begin{cases} K(y), & y > 0 \\ 0, & y \leq 0 \end{cases} .$$

Bizonyítás: A tételt nem bizonyítjuk.

Megjegyzés: $D_{n,m} = \sqrt{\frac{nm}{n+m}} \sup_{x\in\mathbb{R}} |F_n(x) - G_m(x)| = \sqrt{\frac{nm}{n+m}} \max_{i=1,2,\ldots,n+m} |F_n(Z_i^*) - G_m(Z_i^*)|$, ahol $Z_1^* \leq Z_2^* \leq \cdots \leq Z_{n+m}^*$ a két minta egyesítésével kapott minta rendezettje. A szuprémum meghatározását, most is visszavezettük maximum meghatározására.

Ezt a próbát homogenitásvizsgálatra használhatjuk, azaz annak eldöntésére, hogy a két változó azonos eloszlású-e. A nullhipotézisünk az, hogy a két minta eloszlásfüggvénye azonos, az alternatív hipotézis ennek a tagadása:

$H_0 :\ F(x) \equiv G(x)$ és $H_1 :\ F(x) \not\equiv G(x)$.

A hipotézis eldöntése: tetszőleges $0 < \varepsilon < 1$ -hez adható olyan x_ε kritikus érték, hogy $K(x_\varepsilon) = 1 - \varepsilon$ legyen. Ha a $D_{n,m} < x_\varepsilon$, akkor a nullhipotézist az adott szignifikancia szinten elfogadjuk.

Az alábbi tétel segítségével még tovább lehet a mintaelemszámot csökkenteni.

3.4.4. tétel: *(Gnyegyenko–Koroljuk)*
Legyen a $X_1, X_2, \ldots, X_n, \ldots$ statisztikai minta, melynek eloszlásfüggvénye $F(x)$, és $Y_1, Y_2, \ldots, Y_n, \ldots$ az előzőtől független másik statisztikai minta, melynek eloszlásfüggvénye $G(x)$. F és G abszolút folytonosak. Jelölje $F_n(x)$ és $G_n(x)$ a két mintához tartozó empirikus eloszlásfüggvényet. Tegyük fel, hogy $F(x) \equiv G(x)$.

Ekkor

$$\mathbf{P}\left(\sqrt{\frac{n}{2}} \sup_{x\in\mathbb{R}} |F_n(x) - G_n(x)| < y\right) = \begin{cases} 0, & y \leq \frac{1}{\sqrt{2n}} \\ L(y), & \frac{1}{\sqrt{2n}} < y \leq \sqrt{\frac{n}{2}} \\ 1, & \sqrt{\frac{n}{2}} < y \end{cases} ,$$

ahol

$$L(y) = \frac{1}{\binom{2n}{n}} \sum_{k=-\left[\frac{n}{2}\right]}^{\left[\frac{n}{2}\right]} (-1)^k \binom{2n}{n-kc}, \quad c = \left[y\sqrt{2n}\right] + 1.$$

Bizonyítás: A tételt nem bizonyítjuk.

Megjegyzés: A tétel nem határeloszlástétel ezúttal, hanem pontos eloszlást számol ki. Ezért lehet kis minta esetén is alkalmazni. Az $L(y)$ eloszlásfüggvény segítségével a $H_0 :\ F(x) \equiv G(x)$ nullhipotézisre a szokásos módon próba szerkeszthető.

4. fejezet

Regresszióanalízis

4.1. Véletlen megfigyelés

A feladat két, erős sztochasztikus összefüggést mutató X és Y valószínűségi változó közötti függvénykapcsolat jellegének, és paramétereinek feltárása. Y fogja jelölni a célváltozót, és X a megfigyelést, a független változót. Feladat olyan f függvény megadása, ahol $Y \approx f(X)$. Elméletileg a feladat megoldott, hiszen ha a két változó együttes eloszlása ismert, akkor meghatározható a feltételes várható érték (regresszió), amely a legjobb kapcsolatot adja meg abban az értelemben, hogy minimalizálja a négyzetes eltérés várható értékét:

$$\mathbf{E}(Y - \mathbf{E}(Y\,|X\,))^2 = \min_{\forall f} \mathbf{E}(Y - f(X))^2.$$

Gyakorlati problémáknál azonban az együttes eloszlás általában nem ismert, tehát a feltételes várható érték számítása sem lehetséges. A függvénykapcsolatot a két változóra vonatkozó $(X_1, Y_1)^T, (X_2, Y_2)^T, \ldots, (X_n, Y_n)^T$ statisztikai minta alapján kell meghatározni. A regresszióanalízis végrehajtásának csak akkor van értelme, ha kimutatható X és Y között a sztochasztikus összefüggés (pl. el kellett vetni a nullhipotézist függetlenségvizsgálatnál, vagy a minta empirikus korrelációs együtthatója közel van 1-hez). A regresszióanalízis tipikus módszere az, hogy egy jól körülírt többparaméteres függvényhalmazból határozunk meg egy bizonyos függvényt úgy, hogy annak paramétereit a minta segítségével megbecsüljük. Legyen adott tehát az $F = \{f\}$ függvényosztály. Meghatározandó az az $f^* \in F$ függvény, ahol

$$\mathbf{E}(Y - f^*(X))^2 = \min_{\forall f \in F} \mathbf{E}(Y - f(X))^2.$$

F-et legtöbbször a mintarealizációnak a koordinátarendszerben való ábrázolásával kapott szóródásgrafikon alapján lehet megválasztani, de az a változók fizikai tartalmából fakadó „elvárt" típusú függvények halmaza is lehet.

Ismeretes, hogy a két változó együttes normális eloszlása esetén, az elméleti regresszió, az $\mathbf{E}(Y \mid X = x)$ lineáris. Mivel az együttes normális eloszlás gyakran jelentkezik, alapvető fontosságú a regressziószámításnak az a speciális esete, amikor F a lineáris függvények halmaza. A lineáris összefüggés megadása azért is fontos, mert a kapott összefüggést könnyű magyarázni, interpretálni.

4.1.1. Lineáris regresszió két változó között

4.1.1. definíció: Legyen X és Y két adott valószínűségi változó. Az $a^*X + b^*$ valószínűségi változó az Y-nak az X-re vonatkozó lineáris regressziója, ha

$$\mathbf{E}(Y - a^*X - b^*)^2 = \min_{\forall a,b \in \mathbb{R}} \mathbf{E}(Y - aX - b)^2.$$

a^* a regressziós meredekség, b^* a regressziós konstans.

4.1.1. tétel: $a^* = \mathbf{R}(X,Y)\frac{\sigma Y}{\sigma X}$, $b^* = \mathbf{E}Y - \mathbf{R}(X,Y)\frac{\sigma Y}{\sigma X}\mathbf{E}X$, ahol $\mathbf{R}(X,Y)$ a két változó korrelációs együtthatóját jelöli.

Bizonyítás: Legyen $h(a,b) = \mathbf{E}(Y - aX - b)^2$. A lineáris regresszió meghatározásához ezt a kétváltozós függvényt kell minimalizálni. A minimumhely létezésének szükséges feltétele, hogy: $\frac{\partial h}{\partial a} = -2\mathbf{E}\left[(Y - aX - b)X\right] = 0$, $\frac{\partial h}{\partial b} = -2\mathbf{E}\left[Y - aX - b\right] = 0$. Innen: $a\mathbf{E}X^2 + b\mathbf{E}X = \mathbf{E}XY$, $a\mathbf{E}X + b = \mathbf{E}Y \implies b = \mathbf{E}Y - a\mathbf{E}X \implies a\mathbf{E}X^2 + (\mathbf{E}Y - a\mathbf{E}X)\mathbf{E}X = \mathbf{E}XY$
$\implies a = \mathbf{R}(X,Y)\frac{\sigma Y}{\sigma X}$, $b = \mathbf{E}Y - \mathbf{R}(X,Y)\frac{\sigma Y}{\sigma X}\mathbf{E}X$, $\begin{pmatrix} \mathbf{E}X^2 & \mathbf{E}X \\ \mathbf{E}X & 1 \end{pmatrix}$ pozitív definit, tehát a^*, b^* valóban minimumhely, és ez volt az állítás.

■

Megjegyzés: Normális esetben a lineáris regressziós és a regressziós összefüggések egybeesnek.

A gyakorlatban általában nem ismertek az X és Y változók momentumai, ezért az elméleti lineáris regressziós összefüggés nem határozható meg. Az $(X_1,Y_1)^T, (X_2,Y_2)^T, \ldots, (X_n,Y_n)^T$ statisztikai minta alapján a legkisebb négyzetek módszerével lehet az egyenes paramétereit megbecsülni.

4.1.2. definíció: Adott az $(X_1,Y_1)^T, (X_2,Y_2)^T, \ldots, (X_n,Y_n)^T$ statisztikai minta és az $F = \{f\,(x; a_1, a_2, \ldots, a_k)\}$ k-paraméteres függvényosztály. A

$$\min_{\forall a_1, a_2, \ldots, a_k} \sum_{i=1}^{n} (Y_i - f(X_i; a_1, a_2, \ldots, a_k))^2$$

szélsőérték-feladat megoldásából kapott $a_i^* = a_i^*(\mathbf{X}, \mathbf{Y})$ $(i = 1, 2, \ldots, n)$ statisztikákat, a $\min\limits_{\forall f \in F} \mathbf{E}(Y - f^*(X))^2$ regressziós probléma paramétereinek *legkisebb négyzetek* módszerével kapott *becslései*nek nevezzük.

4.1.2. tétel: Lineáris regresszió esetén a legkisebb négyzetek módszerével az egyenes paramétereinek becslései:

$$a^* = \hat{R}_n \frac{s_Y}{s_X}, \qquad b^* = \bar{Y}_n - \hat{R}_n \frac{s_Y}{s_X} \bar{X}_n,$$

ahol

$$\hat{R}_n = \frac{\sum\limits_{i=1}^{n} (Y_i - \bar{Y}_n)(X_i - \bar{X}_n)}{\sqrt{\sum\limits_{i=1}^{n} (Y_i - \bar{Y}_n)^2 \sum\limits_{i=1}^{n} (X_i - \bar{X}_n)^2}}$$

az empirikus korrelációs együttható,

$$s_Y = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2},$$

$$s_X = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2}$$

az empirikus szórások, és $\bar{X}_n$, $\bar{Y}_n$ az átlagstatisztikák.

Bizonyítás: A tétel állítása könnyen belátható, ha a 4.1.1. tétel bizonyítását megismételjük a $h(a,b)=\sum_{i=1}^{n}(Y_i-aX_i-b)^2$ kétváltozós függvénnyel.

■

Megjegyzés:

1. Látható, hogy az empirikus lineáris regresszió együtthatói az elméleti regressziós egyenes együtthatóitól annyiban különböznek, hogy a képletekben az elméleti momentumok helyett a mintából számolt megfelelő empirikus momentumok állnak.

2. Ha X és Y együttes eloszlása normális, akkor az elméleti regressziós egyenes a meredekségére konfidenciaintervallum szerkeszthető, mivel ilyenkor az $\frac{a^*-a}{\frac{s_Y}{s_X}\sqrt{1-\hat{R}_n^2}}\sqrt{n-2}$ statisztika $n-2$ szabadságfokú Student-eloszlást követ.

3. A normális esetben a korrellálatlanság és a függetlenség azonos tulajdonságok. Tehát, ha X és Y korrelációs együtthatója 0, akkor $a=0$, azaz

$$\frac{a^*}{\frac{s_Y}{s_X}\sqrt{1-\hat{R}_n^2}}\sqrt{n-2}=\frac{\hat{R}_n}{\sqrt{1-\hat{R}_n^2}}\sqrt{n-2}\in t_{n-2}.$$

A függetlenséget megfogalmazó nullhipotézisről tehát ilyenkor t-próbával dönthetünk.

4.1.2. Polinomiális regresszió

4.1.3. definíció: Amikor az $F=\{p_n(x)=a_0+a_1x+\cdots+a_mx^m\}$ függvényosztály a legfeljebb m-edrendű polinomosztály, a $\min\limits_{\forall f\in F}\mathbf{E}(Y-f(X))^2$ minimumfeladat megoldását *polinomiális regresszió*s illesztésnek nevezzük.

4.1.3. tétel: Az elméleti polinomiális regressziós görbe együtthatóit az

$$\begin{pmatrix} 1 & \mathbf{E}X & \cdots & \mathbf{E}X^m \\ \mathbf{E}X & \mathbf{E}X^2 & \cdots & \mathbf{E}X^{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{E}X^i & \mathbf{E}X^{i+1} & \cdots & \mathbf{E}X^{m+i} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{E}X^m & \mathbf{E}X^{m+1} & \cdots & \mathbf{E}X^{2m} \end{pmatrix}\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_i \\ \vdots \\ a_m \end{pmatrix}=\begin{pmatrix} \mathbf{E}Y \\ \mathbf{E}YX^2 \\ \vdots \\ \mathbf{E}YX^i \\ \vdots \\ \mathbf{E}YX^m \end{pmatrix}$$

lineáris egyenletrendszer megoldásával kaphatjuk meg. Ennek mindig van megoldása, hiszen az együtthatómátrix szimmetrikus és pozitív szemidefinit.

Bizonyítás: A feladatot a

$$h(a_0,a_1,\ldots,a_m)=\mathbf{E}(Y-(a_0+a_1X+\cdots+a_mX^m))^2$$

$m+1$ változós függvény minimumhelyének megkeresésével oldhatjuk meg:

$$\frac{\partial h(a_0,a_1,\ldots,a_m)}{\partial a_i}=-2\mathbf{E}([Y-(a_0+a_1X+\cdots+a_mX^m)]\,X^i)=0\quad(i=0,1,2,\ldots,m)\Longrightarrow$$

$\Longrightarrow\sum\limits_{j=0}^{m}a_j\mathbf{E}X^{i+j}=\mathbf{E}YX^i\Longrightarrow$ következik az állítás.

■

A tapasztalati polinomiális görbe együtthatóinak meghatározását az $(X_1, Y_1)^T, (X_2, Y_2)^T, \dots, (X_n, Y_n)^T$ statisztikai minta segítségével az

$$\begin{pmatrix} 1 & \frac{1}{n}\sum\limits_{j=1}^{n} X_j & \cdots & \frac{1}{n}\sum\limits_{j=1}^{n} X_j^m \\ \frac{1}{n}\sum\limits_{j=1}^{n} X_j & \frac{1}{n}\sum\limits_{j=1}^{n} X_j^2 & \cdots & \frac{1}{n}\sum\limits_{j=1}^{n} X_j^{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n}\sum\limits_{j=1}^{n} X_j^i & \frac{1}{n}\sum\limits_{j=1}^{n} X_j^{i+1} & \cdots & \frac{1}{n}\sum\limits_{j=1}^{n} X_j^{i+m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n}\sum\limits_{j=1}^{n} X_j^m & \frac{1}{n}\sum\limits_{j=1}^{n} X_j^{m+1} & \cdots & \frac{1}{n}\sum\limits_{j=1}^{n} X_j^{2m} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_i \\ \vdots \\ a_m \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum\limits_{j=1}^{n} Y_j \\ \frac{1}{n}\sum\limits_{j=1}^{n} Y_j X_j \\ \vdots \\ \frac{1}{n}\sum\limits_{j=1}^{n} Y_j X_j^i \\ \vdots \\ \frac{1}{n}\sum\limits_{j=1}^{n} Y_j X_j^m \end{pmatrix}$$

lineáris egyenletrendszer megoldásából kapjuk. Ehhez úgy jutunk el, hogy a

$$\hat{h}(a_0, a_1, \dots, a_m) = \frac{1}{n}\sum_{j=1}^{n}(Y_j - (a_0 + a_1 X_j + \cdots + a_m X_j^m))^2$$

függvény minimumhelyét meghatározzuk, hasonlóan, mint ahogy azt az 4.1.3 tételben tettük.

Megjegyzés: Nyilvánvalóan az n mintaelemszámnak jóval nagyobbnak kell lennie, mint az m-nek, az illesztendő polinom fokának.

4.1.3. Lineárisra visszvezethető kétparaméteres regressziós összefüggések keresése

Ha a lineáris regresszió feltételei valahol sérülnek, vagy rossz illesztést kapunk, a függő és a független változók transzformációjával kell megpróbálkozni. A transzformált input adatokon azután már lineáris regressziós elemzést hajtunk végre, de ez az eredeti adatoknál már nem lineáris összefüggést fog magyarázni. Az inverz leképezés és a regressziós együtthatók segítségével képezhetők azok a paraméterek, amelyekkel a kapcsolatot leíró függvény felírható. Tehát, ha az $F = \{f(x; a, b)\}$ függvényosztály kétparaméteres, és találhatók olyan g, h, k_1, k_2 függvények, hogy $y = f(x; a, b) \iff g(y) = k_1(a, b)h(x) + k_2(a, b)$ teljesül.

Ezután a $\min\limits_{\forall f \in F} \mathbf{E}(Y - f(X; a, b))^2$ feladat helyett a

$$\mathbf{E}(g(Y) - k_1^* h(X) - k_2^*)^2 = \min_{\forall k_1, k_2} \mathbf{E}(g(Y) - k_1 h(X) - k_2)^2$$

lineáris regressziós feladatot oldjuk meg. Végül $a^* \approx k_1^{-1}(k_1^*, k_2^*)$, $b^* \approx k_2^{-1}(k_1^*, k_2^*)$. Általában más eredményeket kapunk, mintha az eredeti függvényen hajtottuk volna végre a legkisebb négyzetek módszerével a paraméterbecslést. Viszont az eredeti problémánál, nem biztos, hogy a kapott (sokszor transzcendens) egyenletet meg tudnánk oldani. A továbbiakban megadunk néhány példát nemlineáris kapcsolatnak a lineáris regresszió segítségével való megadására.

$\mathbf{y} = f(x; a, b) = a\mathrm{e}^{bx}$ exponenciális függvénykapcsolat:

Az egyenlet két oldalát logaritmizálva már lineáris összefüggést kapunk $\ln y$ és x között: $y^* = \ln y = bx + \ln a = k_1 x + k_2$. Ilyenkor az $((X_1, \ln Y_1), (X_2, \ln Y_2), \dots, (X_n, \ln Y_n))$ transzformált mintára illesztünk egyenest. A kapott k_1^* és k_2^* együtthatókból az $a = \mathrm{e}^{k_2^*}$ és $b = k_1^*$ transzformációval kapjuk meg az eredeti összefüggés paramétereit.

$\mathbf{y} = f(x;a,b) = ax^b$ hatványfüggvénykapcsolat:

A lineáris kapcsolatot a logaritmizálás után most $\ln y$ és $\ln x$ között kell megadni: $y^* = \ln y = b \cdot \ln x + \ln a = k_1 x^* + k_2 \Longrightarrow b = k_1, a = \mathrm{e}^{k_2}$.

$\mathbf{y} = f(x;a,b) = a\mathrm{e}^{\frac{-b}{x}}$ Arrhenius függvénykapcsolat:

Logaritmizálás után: $y^* = \ln y = -b\frac{1}{x} + \ln a = k_1 x^* + k_2$ az $\ln y$ és x reciproka között lép fel a lineáris kapcsolat ($b = -k_1, a = \mathrm{e}^{k_2}$).

$\mathbf{y} = f(x;a,b) = \frac{1}{a+bx}$ reciprok függvénykapcsolat:

Itt most y reciproka és x között kell a lineáris regressziót kiszámolni.

$\mathbf{y} = f(x;a,b) = \frac{ax}{1+bx}$ racionális törtfüggvénykapcsolat:

Most az egyenlet két oldalának reciprokát képezzük: $y^* = \frac{1}{y} = \frac{1}{a}\frac{1}{x} + \frac{b}{a} = k_1 x^* + k_2 \Longrightarrow a = \frac{1}{k_1},\ b = \frac{k_2}{k_1}$, és a reciprokértékek között keresünk lineáris regressziót.

$\mathbf{y} = f(x;a,b) = ax^2 + bx$ kvadratikus függvénykapcsolat:

Ekkor ha x-szel átosztunk máris lineáris az összefüggés $\frac{y}{x}$ és x között: $y^* = \frac{y}{x} = ax + b$.

$\mathbf{y} = f(x;a,b) = a + \frac{b}{x}$ hiperbolikus függvénykapcsolat:

Ez eleve lineáris összefüggés y és $\frac{1}{x}$ között.

$\mathbf{y} = a\ln(bx) = a\ln b + a\ln x$ a logaritmikus függvénykapcsolat:

Ez lineáris kapcsolat y és $\ln x$ között.

4.1.4. A regressziós illeszkedés jóságának mérése

4.1.4. definíció: Tekintsük a X és Y valószínűségi változókat, és tegyük fel, hogy Y-t $f(X)$-szel közelítjük. A közelítés jóságának mérésére az

$$R_f^2 = 1 - \frac{\mathbf{E}(Y - f(X))^2}{\boldsymbol{\sigma}^2 Y}$$

meghatározottsági együtthatót használjuk. Ha $f(\mathbf{x}) = \mathbf{E}(Y \mid X = x)$ a regressziós függvény, akkor az R_r^2 jelölést használjuk. Adott $(X_1,Y_1)^T, (X_2,Y_2)^T, \dots, (X_n,Y_n)^T$ statisztikai minta esetén a meghatározottsági együtthatót az $1 - \frac{\sum_{i=1}^n (Y_i - f(X_i))^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$ ($\approx R_f^2$) statisztikával közelítjük, ami konzisztens becslés.

Megjegyzés: R^2 minél közelebb van az 1-hez, annál jobb a regressziós közelítés. Ha R^2 0 közeli érték, vagy negatív, a regressziós illesztés elfogadhatatlan. A meghatározottsági együttható tulajdonságait foglalja össze az alábbi tétel:

4.1.4. tétel:

(i) $R_f^2 \leq R_r^2 \leq 1$,

(ii) Ha X és Y függetlenek, akkor $R_r^2 = 0$, azaz $R_f^2 \leq 0$,

(iii) Ha $f(x) = a^*x + b^*$ lineáris regressziós függvény, akkor $R_f^2 = (\mathbf{R}(X,Y))^2$.

Bizonyítás:

(i) A feltételes várható érték tulajdonságaiból következik, hogy

$$\mathbf{E}(Y - \mathbf{E}(Y \mid X))^2 \leq \mathbf{E}(Y - f(X))^2$$

és

$$\mathbf{E}(Y - \mathbf{E}(Y \mid X))^2 \leq \boldsymbol{\sigma}^2 Y,$$

ami már igazolja az állítást. Az elvileg legjobb illesztés esetén sem biztos, hogy R_r^2 eléri az 1-et.

(ii) $\mathbf{E}(Y \mid X) = \mathbf{E}Y$, így $\mathbf{E}(Y - \mathbf{E}(Y \mid X))^2 = \boldsymbol{\sigma}^2 Y$, azaz $R_r^2 = 0$. „Rossz” f esetén az R_f^2 szám akár negatív is lehet.

(iii) Ha $f(x) = a^*X + b^*$ lineáris regressziós függvény, ahol

$$a^* = \mathbf{R}(X,Y) \cdot \frac{\boldsymbol{\sigma} Y}{\boldsymbol{\sigma} X}, \qquad b^* = \mathbf{E}Y - a^*\mathbf{E}X,$$

akkor

$$\mathbf{E}(Y - a^*X - b^*) = \mathbf{E}Y - a^*\mathbf{E}X - b^* = 0,$$

vagyis

$$\mathbf{E}(Y - a^*X - b^*)^2 = \boldsymbol{\sigma}^2(Y - a^*X - b^*) = \boldsymbol{\sigma}^2(Y - a^*X) =$$
$$= \boldsymbol{\sigma}^2 Y + (a^*)^2 \cdot \boldsymbol{\sigma}^2 X - 2a^* \cdot \mathbf{cov}(Y,X) =$$
$$\boldsymbol{\sigma}^2 Y + (\mathbf{R}(X,Y))^2 \cdot \boldsymbol{\sigma}^2 Y - 2\mathbf{R}(X,Y)\frac{\boldsymbol{\sigma} Y}{\boldsymbol{\sigma} X} \cdot \mathbf{R}(X,Y) \cdot \boldsymbol{\sigma} Y \cdot \boldsymbol{\sigma} X =$$
$$= \boldsymbol{\sigma}^2 Y(1 - \mathbf{R}^2(X,Y)).$$

Tehát

$$\mathbf{R}^2(X,Y) = R_f^2$$

■

4.2. Tervezett (determinisztikus) megfigyelés

Főleg műszaki alkalmazásokban gyakori, hogy a méréseket Y-ra előírt x beállításoknál végzik el, és így keresik az ismeretlen $Y \approx f(x)$ függvénykapcsolatot. A modell ilyenkor az, hogy $Y = f(x) + \varepsilon$, ahol ε a mérési hibát jelentő valószínűségi változó, melyre $\mathbf{E}\varepsilon = 0$ és $\boldsymbol{\sigma}^2\varepsilon < \infty$.

Tegyük fel, hogy adottak az $x_1, x_2, \ldots, x_n \in \mathbb{R}$ beállítások mellett elvégzett $Y_1, Y_2, \ldots, Y_n$ mérési eredmények. Mivel a mérések a véletlentől is függtek, feltesszük, hogy $Y_i = ax_i + b + \varepsilon_i \quad (i = 1, \ldots, n)$, ahol ε_i teljesen függetlenek és $\mathbf{E}\varepsilon_i = 0$, $\boldsymbol{\sigma}^2\varepsilon_i = D^2 < \infty$. A keresett a, b regressziós együtthatókat a legkisebb négyzetek módszerével a

$$h(a,b) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - (ax_i + b))^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$$

négyzetes eltérés átlagának minimalizálásával oldjuk meg.

4.2.1. tétel: *(Gauss–Markov-tétel)*
Ha $Y_i = ax_i + b + \varepsilon_i$ $(i = 1, 2, \ldots, n)$, ahol az ε_i teljesen független valószínűségi változók, és $\mathbf{E}\varepsilon_i = 0$, $\boldsymbol{\sigma}^2\varepsilon_i = \sigma^2$, akkor az a, b együtthatók legkisebb négyzetek módszerével kapott becslései torzítatlanok, és az összes lineáris becslés közül minimális szórással rendelkeznek.

Megjegyzés: A legkisebb négyzetek módszere a legjobb torzítatlan becslést adja, ami angolul: *best linear unbaised estimation* = BLUE.

Bizonyítás:

$$h(a,b) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - (ax_i + b))^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2.$$

$$\frac{\partial h(a,b)}{\partial a} = -\frac{2}{n}\sum_{i=1}^{n}(Y_i - (ax_i + b))\,x_i = 0 \implies \sum_{i=1}^{n}Y_i x_i - b\sum_{i=1}^{n}x_i - a\sum_{i=1}^{n}x_i^2 = 0.$$

$$\frac{\partial h(a,b)}{\partial b} = -\frac{2}{n}\sum_{i=1}^{n}(Y_i - (ax_i + b)) = 0 \implies \sum_{i=1}^{n}Y_i - bn - a\sum_{i=1}^{n}x_i = 0.$$

A fenti egyenletrendszerből:

$$a^* = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}_n)}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n}\frac{x_i - \bar{x}}{\sum\limits_{j=1}^{n}(x_j - \bar{x})^2}\cdot Y_i = \sum_{i=1}^{n}k_i Y_i,$$

$$b^* = \bar{Y}_n - a^*\bar{x} = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{x}k_i\right)Y_i = \sum_{i=1}^{n}l_i Y_i,$$

tehát lineáris becsléseket kapunk.

Fel fogjuk használni, hogy

$$\sum_{i=1}^{n}k_i = \frac{1}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}\sum_{i=1}^{n}(x_i - \bar{x}) = 0,$$

$$\sum_{i=1}^{n}k_i x_i = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})\,x_i}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})\,x_i - \bar{x}\sum\limits_{i=1}^{n}(x_i - \bar{x})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = 1,$$

$$\sum_{i=1}^{n}k_i^2 = \sum_{i=1}^{n}\left(\frac{(x_i - \bar{x})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}\right)^2 = \frac{1}{\left(\sum\limits_{i=1}^{n}(x_i - \bar{x})^2\right)^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}.$$

A torzítatlanság igazolása:

$$\mathbf{E}a^* = \mathbf{E}\left(\sum_{i=1}^{n}k_i Y_i\right) = \sum_{i=1}^{n}k_i\mathbf{E}Y_i = \sum_{i=1}^{n}k_i\,(ax_i + b) = a\sum_{i=1}^{n}k_i x_i + b\sum_{i=1}^{n}k_i = a.$$

$$\mathbf{E}b^* = \sum_{i=1}^{n}l_i\mathbf{E}Y_i = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{x}k_i\right)(ax_i + b) = \frac{1}{n}a\sum_{i=1}^{n}x_i - \bar{x}a\sum_{i=1}^{n}k_i x_i + \frac{1}{n}\sum_{i=1}^{n}b - \bar{x}b\sum_{i=1}^{n}k_i = b.$$

$$\boldsymbol{\sigma}^2 a^* = \boldsymbol{\sigma}^2 \left(\sum_{i=1}^n k_i Y_i \right) = \sum_{i=1}^n k_i^2 \boldsymbol{\sigma}^2 Y_i = \sigma^2 \sum_{i=1}^n k_i^2 = \frac{\sigma^2}{\sum\limits_{i=1}^n (x_i - \bar{x})^2}.$$

$$\boldsymbol{\sigma}^2 b^* = \boldsymbol{\sigma}^2 \left(\sum_{i=1}^n l_i Y_i \right) = \sum_{i=1}^n l_i^2 \boldsymbol{\sigma}^2 Y_i = \sigma^2 \sum_{i=1}^n l_i^2 = \sigma^2 \sum_{i=1}^n \left(\frac{1}{n} - \bar{x} k_i \right)^2 =$$

$$= \sigma^2 \left(\frac{1}{n} + \bar{x}^2 \sum_{i=1}^n k_i^2 - 2\,\bar{x}\,\frac{1}{n} \sum_{i=1}^n k_i \right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum\limits_{i=1}^n (x_i - \bar{x})^2} \right).$$

Legyen most $\tilde{a} = \sum\limits_{i=1}^n c_i Y_i$ az a torzítatlan, lineáris becslése, azaz

$$a = \mathbf{E}\tilde{a} = \sum_{i=1}^n c_i \mathbf{E} Y_i = \sum_{i=1}^n c_i\ (a\,x_i + b) = a \sum_{i=1}^n c_i x_i + b \sum_{i=1}^n c_i.$$

Ez csak úgy lehet, ha $\sum\limits_{i=1}^n x_i\,c_i = 1$ és $\sum\limits_{i=1}^n\ c_i = 0$. Legyen $d_i = c_i - k_i$.

$$\boldsymbol{\sigma}^2 \tilde{a} = \sigma^2 \sum_{i=1}^n c_i^2 = \sigma^2 \sum_{i=1}^n (k_i + d_i)^2 = \sigma^2 \sum_{i=1}^n k_i^2 + 2\sigma^2 \sum_{i=1}^n k_i d + \sigma^2 \sum_{i=1}^n d_i^2 =$$

$$= \boldsymbol{\sigma}^2 a^* + \sigma^2 \underbrace{\sum_{i=1}^n d_i^2}_{\geq 0} + 2\,\sigma^2\ \underbrace{\sum_{i=1}^n k_i d_i}_{=0} \geq \boldsymbol{\sigma}^2 a^*.$$

$$\sum_{i=1}^n k_i d_i = \sum_{i=1}^n k_i (c_i - k_i) = \sum_{i=1}^n c_i \frac{(x_i - \bar{x})}{\sum\limits_{j=1}^n (x_j - \bar{x})^2} - \sum_{i=1}^n k_i^2 =$$

$$\frac{1}{\sum\limits_{j=1}^n (x_j - \bar{x})^2} \underbrace{\sum_{i=1}^n c_i\,x_i}_{=1} - \bar{x}\,\frac{1}{\sum\limits_{i=1}^n (x_i - \bar{x})^2} \underbrace{\sum_{i=1}^n c_i}_{=0} - \frac{1}{\sum\limits_{i=1}^n (x_i - \bar{x})^2} = 0.$$

Másrészt, ha $\tilde{b} = \sum\limits_{i=1}^n w_i Y_i$ a b torzítatlan, lineáris becslése, azaz

$$b = \mathbf{E}\tilde{b} = \sum_{i=1}^n w_i \mathbf{E} Y_i = \sum_{i=1}^n w_i\ (a\,x_i + b) = a \sum_{i=1}^n w_i x_i + b \sum_{i=1}^n w_i.$$

Ez csak úgy lehet, ha $\sum\limits_{i=1}^n x_i\,w_i = 0$ és $\sum\limits_{i=1}^n\ w_i = 1$. Legyen $d_i = w_i - l_i$.

$$\boldsymbol{\sigma}^2 \tilde{b} = \sigma^2 \sum_{i=1}^n w_i^2 = \sigma^2 \sum_{i=1}^n (l_i + d_i)^2 = \sigma^2 \sum_{i=1}^n l_i^2 + 2\sigma^2 \sum_{i=1}^n l_i d_i + \sigma^2 \sum_{i=1}^n d_i^2 =$$

$$= \boldsymbol{\sigma}^2 b^* + \sigma^2 \underbrace{\sum_{i=1}^n d_i^2}_{\geq 0} + 2\,\sigma^2\ \underbrace{\sum_{i=1}^n l_i d_i}_{=0} \geq \boldsymbol{\sigma}^2 b^*.$$

$$\sum_{i=1}^{n} l_i d_i = \sum_{i=1}^{n} l_i (w_i - l_i) = \sum_{i=1}^{n} l_i w_i - \sum_{i=1}^{n} l_i^2 = \sum_{i=1}^{n} w_i \left(\frac{1}{n} - \bar{x}\, k_i\right) - \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} \cdot k_i\right)^2 =$$

$$= \frac{1}{n} \sum_{i=1}^{n} w_i - \bar{x} \frac{1}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \sum_{i=1}^{n} (x_i - \bar{x})\, w_i - \frac{1}{n} - \bar{x}^2 \sum_{i=1}^{n} k_i^2 + 2\frac{1}{n}\bar{x} \sum_{i=1}^{n} k_i =$$

$$= \frac{1}{n} + \frac{\bar{x}^2}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} - \frac{1}{n} - \frac{\bar{x}^2}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} = 0.$$

■

4.2.2. tétel: Ha $Y_i = ax_i + b + \varepsilon_i$ $(i = 1, 2, \ldots, n)$, ahol az $\varepsilon_i \in N(0, \sigma^2)$ teljesen független valószínűségi változók ($\Longrightarrow$ $Y_i \in N(ax_i + b, \sigma)$ és teljesen függetlenek), akkor az előbbiek mellett még az is állítható, hogy a^* és b^* az a, b paramétereknek maximum-likelihood becslései is.

Bizonyítás: Mivel $Y_i \in N(ax_i + b, \sigma)$, teljesen függetlenek, ezért a minta együttes sűrűségfüggvénye, a likelihood függvény:

$$L(y_1, y_2, \ldots, y_n; a, b, \sigma) = (2\pi)^{-\frac{n}{2}}\, \sigma^{-n} \exp\left(-\frac{1}{2\,\sigma^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2\right),$$

a log-likelihood függvény pedig:

$$\ln\, L = l(y_1, y_2, \ldots, y_n; a, b, \sigma) = -\frac{n}{2} \ln\, (2\pi) - n \ln \sigma\, - \frac{1}{2\,\sigma^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2\,.$$

$$\frac{\partial l}{\partial a} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i\, (Y_i - ax_i - b) = 0,$$

$$\frac{\partial l}{\partial b} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - ax_i - b) = 0,$$

$$\frac{\partial l}{\partial (\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (Y_i - ax_i - b)^2 = 0,$$

amiből a maximum-likelihood becslésekre:

$$a^* = \sum_{i=1}^{n} k_i Y_i, \qquad b^* = \bar{Y} - \bar{x}\, a^*, \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i a^*\, x_i - b^*)^2$$

adódnak, azaz a^* és b^* megegyezik a legkisebb négyzetek módszerével kapott becslésekkel!

■

4.3. Sztochasztikus approximáció

A többváltozós lineáris- és nemlineáris regressziós feladat együtthatóinak meghatározását, ha a változók száma nagy, gradiens módszerrel szokás megoldani. Ez a probléma elvezet a sztochasztikus approximáció problémaköréhez, melyet ebben a szakaszban tárgyalunk. Az alapprobléma itt az, hogy az $\mathbf{r}(\mathbf{c}) = \mathbf{0}$ iterációs gyökkeresési algoritmus milyen feltételek mellett állít elő a gyökhöz konvergáló sorozatot, ha az $\mathbf{r}$ függvény értékeinek kiszámítását valamilyen véletlen zavaró körülmény lehetetlenné teszi. Először bebizonyítjuk az alábbi tételt, amely a Robins–Monroe-féle sztochasztikus approximációs algoritmust alapozza meg.

Legyen $\mathbf{f} : \mathbb{R}^{k+1} \times \mathbb{R}^M \to \mathbb{R}^{k+1}$ mérhető függvény. Adottak a $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n, \ldots$ és $\mathbf{Z}$ azonos eloszlású $\mathbb{R}^M$ értékű valószínűségi vektorváltozók. Tegyük fel, hogy $\forall \mathbf{c} \in \mathbb{R}^{k+1}$ esetén létezik $\mathbf{Ef}(\mathbf{c}, \mathbf{Z})$ az és legyen

$$\mathbf{r}(\mathbf{c}) = \mathbf{E}f(\mathbf{c}, \mathbf{Z}),$$

és

$$\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}.$$

(Nyilván $\mathbf{r}$ Borel-mérhető, melyet szokás regressziós függvénynek is nevezni.) Legyen

$$\mathbf{C}_{n+1} = \mathbf{C}_n - \gamma_n \mathbf{f}(\mathbf{C}_n, \mathbf{Z}_{n+1}), \quad n = 0, 1, \ldots$$

rekurzív összefüggéssel definiált valószínűségi vektorváltozó sorozat, ahol $\mathbf{C}_0 = \mathbf{c}_0$ tetszőleges.

4.3.1. tétel: Tegyük fel, hogy tetszőleges $\varepsilon > 0$-ra:

$$(*) \qquad \inf_{\|\mathbf{c}-\boldsymbol{\theta}\|>\varepsilon} (\mathbf{c} - \boldsymbol{\theta})^T \mathbf{r}(\mathbf{c}) > 0,$$

$$(**) \qquad \exists K > 0: \quad \mathbf{E}\|f(\mathbf{c}, \mathbf{Z})\|^2 < K\left(1 + \|\mathbf{c} - \boldsymbol{\theta}\|^2\right), \quad \forall \mathbf{c} \in \mathbb{R}^{k+1},$$

$$(***) \qquad \gamma_n > 0, \quad \sum_{n=0}^{\infty} \gamma_n = \infty, \ \sum_{n=0}^{\infty} \gamma_n^2 < \infty.$$

Akkor $\lim\limits_{n\to\infty} \|\mathbf{C}_n - \boldsymbol{\theta}\| = 0$ 1-valószínűséggel.

Bizonyítás: Szükségünk lesz két lemmára.

4.3.1. lemma: A 4.3.1. tétel feltételei mellett teljesül, hogy létezik egy C valószínűségi változó, melyre $\lim\limits_{n\to\infty} \|\mathbf{C}_n - \boldsymbol{\theta}\| = C$ 1-valószínűséggel.

Bizonyítás: A rekurzív egyenlet mindkét oldalából vonjunk le $\boldsymbol{\theta}$-t, majd emeljük négyzetre:

$$\|\mathbf{C}_{n+1} - \boldsymbol{\theta}\|^2 = \|\mathbf{C}_n - \boldsymbol{\theta}\|^2 - 2\gamma_n (\mathbf{C}_n - \boldsymbol{\theta})^T \mathbf{f}(\mathbf{C}_n, \mathbf{Z}_{n+1}) + \gamma_n^2 \|\mathbf{f}(\mathbf{C}_n, \mathbf{Z}_{n+1})\|^2.$$

Ezután vegyük mindkét oldalnak a $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n$ valószínűségi változók által generált σ-algebrára vett feltételes várható értékeit:

$$\mathbf{E}\left(\|\mathbf{C}_{n+1} - \boldsymbol{\theta}\|^2 \mid \mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n\right) = \mathbf{E}\left(\|\mathbf{C}_n - \boldsymbol{\theta}\|^2 \mid \mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n\right) -$$

$$-2\gamma_n \mathbf{E}\left((\mathbf{C}_n - \boldsymbol{\theta})^T \mathbf{f}(\mathbf{C}_n, \mathbf{Z}_{n+1}) \mid \mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n\right) + \gamma_n^2 \mathbf{E}\left(\|\mathbf{f}(\mathbf{C}_n, \mathbf{Z}_{n+1})\|^2 \mid \mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n\right).$$

A függetlenség miatt, és amiatt, hogy $\mathbf{C}_n$ csak $\mathbf{Z}_n$-től függ:

$$\mathbf{E}\left(\|\mathbf{C}_n - \boldsymbol{\theta}\|^2 \mid \mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n\right) = \|\mathbf{C}_n - \boldsymbol{\theta}\|^2,$$

és

$$\mathbf{E}\left((\mathbf{C}_n-\boldsymbol{\theta})^T \mathbf{f}(\mathbf{C}_n,\mathbf{Z}_{n+1}) \mid \mathbf{Z}_1,\mathbf{Z}_2,\dots,\mathbf{Z}_n\right) = \int\limits_{\mathbb{R}^M} (\mathbf{C}_n-\boldsymbol{\theta})^T \mathbf{f}(\mathbf{C}_n,\mathbf{y})\,\mu(\mathrm{d}\mathbf{y}) =$$

$$= (\mathbf{C}_n-\boldsymbol{\theta})^T \int\limits_{\mathbb{R}^M} \mathbf{f}(\mathbf{C}_n,\mathbf{y})\,\mu(\mathrm{d}\mathbf{y}) = (\mathbf{C}_n-\boldsymbol{\theta})^T \mathbf{r}(\mathbf{C}_n),$$

és

$$\mathbf{E}\left(\|\mathbf{f}(\mathbf{C}_n,\mathbf{Z}_{n+1})\|^2 \mid \mathbf{Z}_1,\mathbf{Z}_2,\dots,\mathbf{Z}_n\right) = \int\limits_{\mathbb{R}^M} \|\mathbf{f}(\mathbf{C}_n,\mathbf{y})\|^2 \mu(\mathrm{d}\mathbf{y}) \le K\left(1+\|\mathbf{C}_n-\boldsymbol{\theta}\|^2\right),$$

ahol μ jelöli $\mathbf{Z}$ eloszlását. Így

$$\mathbf{E}\left(\|\mathbf{C}_{n+1}-\boldsymbol{\theta}\|^2 \mid \mathbf{Z}_1,\mathbf{Z}_2,\dots,\mathbf{Z}_n\right) \le$$

$$\le \|\mathbf{C}_n-\boldsymbol{\theta}\|^2 - 2\gamma_n (\mathbf{C}_n-\boldsymbol{\theta})^T \mathbf{r}(\mathbf{C}_n) + \gamma_n^2 K\left(1+\|\mathbf{C}_n-\boldsymbol{\theta}\|^2\right).$$

Mivel

$$\inf_{\|\mathbf{c}-\boldsymbol{\theta}\|>\varepsilon} (\mathbf{c}-\boldsymbol{\theta})^T \mathbf{r}(\mathbf{c}) > 0,$$

ezért tovább növeljük a baloldalt, ha elhagyjuk a középső tagot:

$$\mathbf{E}\left(\|\mathbf{C}_{n+1}-\boldsymbol{\theta}\|^2 \mid \mathbf{Z}_1,\mathbf{Z}_2,\dots,\mathbf{Z}_n\right) \le \|\mathbf{C}_n-\boldsymbol{\theta}\|^2\left(1+\gamma_n^2 K\right) + \gamma_n^2 K.$$

Most tekintsük azt a $\{V_n\}_{n=1,2,\dots}$ valószínűségi változó sorozatot, melynek definíciója

$$V_{n+1} = \|\mathbf{C}_{n+1}-\boldsymbol{\theta}\|^2 \delta_{n+1} + K\sum_{j=n+1}^{\infty} \gamma_j^2 \delta_{j+1},$$

ahol $\delta_k = \prod_{j=k}^{\infty}\left(1+\gamma_j^2 K\right)$. Megmutatjuk, hogy $\{(V_n,\mathcal{F}_n)\}_{n=1,2,\dots}$ szupermartingál, ahol $\mathcal{F}_n = \sigma(\mathbf{Z}_1,\mathbf{Z}_2,\dots,\mathbf{Z}_n)$.

$$\mathbf{E}[V_{n+1} \mid \mathcal{F}_n] \le \|\mathbf{C}_n-\boldsymbol{\theta}\|^2\left(1+\gamma_n^2 K\right)\delta_{n+1} + \gamma_n^2 K\delta_{n+1} + K\sum_{j=n+1}^{\infty}\gamma_j^2\delta_{j+1} =$$

$$= \|\mathbf{C}_n-\boldsymbol{\theta}\|^2 \delta_n + K\sum_{j=n}^{\infty}\gamma_j^2\delta_{j+1} = V_n.$$

Még az is igaz, hogy az $\{\mathbf{E}|V_n|\}_{n=1,2,\dots}$ számsorozat korlátos. Ugyanis $V_n \ge 0$, ami a definíciójából látszik, és így $\mathbf{E}|V_n| = \mathbf{E}V_n$. Másrészt a szupermartingál tulajdonság miatt $\{\mathbf{E}|V_n|\}_{n=1,2,\dots}$ monoton fogyó. Végül $\mathbf{E}V_1 = \mathbf{E}|V_1| < \infty$ miatt $\exists M : \mathbf{E}|V_n| \le M < \infty,\quad n = 1,2,\dots$. Alkalmazható a szupermartingálok konvergenciatétele, miszerint létezik egy C valószínűségi változó, melyre $\lim_{n\to\infty} V_n = C$ 1 valószínűséggel. Mivel megmutatható, hogy $\lim_{n\to\infty}\delta_n = 1$ és $\lim_{n\to\infty}\sum_{j=n}^{\infty}\gamma_j^2\delta_{j+1} = 0$, ezért nyilván az is teljesül, hogy $\lim_{n\to\infty}\|\mathbf{C}_n-\boldsymbol{\theta}\|^2 = C$ 1 valószínűséggel.

■

4.3.2. lemma: A 4.3.1. tétel feltevései mellett

$$\sum_{n=1}^{\infty} \gamma_n \left(\mathbf{C}_n - \boldsymbol{\theta}\right)^T \mathbf{r}\left(\mathbf{C}_n\right) < \infty$$

1 valószínűséggel.

Bizonyítás: A $\{\mathbf{C}_n\}_{n=1,2,\ldots}$ rekurzív sorozatot definiáló egyenlet mindkét oldalából levonva $\boldsymbol{\theta}$-t, négyzetre emelünk, majd kiszámítjuk a várható értékeket. Az $a_n = \mathbf{E}\left\|\mathbf{C}_n - \boldsymbol{\theta}\right\|^2, b_n = 2\mathbf{E}\left[\left(\mathbf{C}_n - \boldsymbol{\theta}\right)^T \mathbf{f}\left(\mathbf{C}_n, \mathbf{Z}_{n+1}\right)\right], d_n = \mathbf{E}\left\|\mathbf{f}\left(\mathbf{C}_n, \mathbf{Z}_{n+1}\right)\right\|^2$ jelölésekkel azt kapjuk, hogy $a_{n+1} = a_n - \gamma_n b_n + \gamma_n^2 d_n$. Megmutatható, hogy

$$\sum_{n=1}^{\infty} \gamma_n b_n = 2\sum_{n=1}^{\infty} \gamma_n \mathbf{E}\left[\left(\mathbf{C}_n - \boldsymbol{\theta}\right)^T \mathbf{f}\left(\mathbf{C}_n, \mathbf{Z}_{n+1}\right)\right] < \infty,$$

ahonnan a 4.3.1. lemma eredményét felhasználva következik a monoton konvergencia tételt alkalmazva, hogy

$$\sum_{n=1}^{\infty} \gamma_n \mathbf{E}\left[\left(\mathbf{C}_n - \boldsymbol{\theta}\right)^T \mathbf{f}\left(\mathbf{C}_n, Z_{n+1}\right)\right] = \sum_{n=1}^{\infty} \mathbf{E}\left[\gamma_n \left(\mathbf{C}_n - \boldsymbol{\theta}\right)^T \mathbf{r}\left(\mathbf{C}_n\right)\right] =$$

$$= \mathbf{E}\left[\sum_{n=1}^{\infty} \gamma_n \left(\mathbf{C}_n - \boldsymbol{\theta}\right)^T \mathbf{r}\left(\mathbf{C}_n\right)\right] < \infty.$$

Ezzel a lemmát bebizonyítottuk hiszen, ha a várható érték mögötti sor 1 valószínűséggel ∞ lenne, akkor a várható érték nem lehetne véges.

■

A 4.3.1. tétel bizonyítása: A 4.3.1. lemmában bizonyítottakból következik, hogy létezik egy $\mathbf{C}$ valószínűségi vektorváltozó és egy Ω^* 1 valószínűségű esemény, hogy $\forall\omega \in \Omega^*$ elemi eseménynél fennáll a $\lim_{n\to\infty} \mathbf{C}_n(\omega) = \mathbf{C}(\omega)$ konvergencia. Másrészt jelölje Ω^{**} azt az 1 valószínűségű eseményt, amelyen fennáll $\sum_{n=1}^{\infty} \gamma_n \left(\mathbf{C}_n(\omega) - \boldsymbol{\theta}\right)^T \mathbf{r}\left(\mathbf{C}_n(\omega)\right) < \infty$. (Ilyen Ω^{**} a 4.3.2. lemma értelmében létezik.) Ekkor a (***) feltétel miatt megadható olyan $\{m_n\}_{n=1,2,\ldots}$ indexsorozat, hogy $\lim_{n\to\infty} \left(\mathbf{C}_{m_n}(\omega) - \boldsymbol{\theta}\right)^T \mathbf{r}\left(\mathbf{C}_{m_n}(\omega)\right) = 0$. Ebből a tétel $(**)$ feltétele miatt teljesül, hogy $\lim_{n\to\infty} \left\|\mathbf{C}_{m_n}(\omega) - \boldsymbol{\theta}\right\| = 0, \quad \forall\omega \in \Omega^{**}$. Tehát, $\forall\omega \in \Omega^* \cap \Omega^{**}$ elemi eseményre $\mathbf{C}(\omega) = \boldsymbol{\theta}$ áll, ami $P\left(\Omega^* \cap \Omega^{**}\right) = 1$ miatt a tétel állítását jelenti: $\lim_{n\to\infty} \left\|\mathbf{C}_n - \boldsymbol{\theta}\right\| = 0$ 1-valószínűséggel.

■

Megjegyzés: A 4.3.1. tétel segítségével igazolható a nagy számok törvényének alábbi erős alakja:

Ha $\{\mathbf{X}_n\}_{n=1,2,\ldots}$ független, azonos eloszlású a várható értékű, véges szórású valószínűségi változó sorozat, melynek tagjai az $\mathbb{R}^N$ térből veszik fel az értékeiket, akkor a $Z_{n+1} = Z_n - \lambda_n \left[Z_n - \mathbf{X}_n\right] \quad n = 0, 1, \ldots$ rekurzív valószínűségi vektorváltozó sorozatra — ha $Z_0 = z_0 \in \mathbb{R}^N$ tetszőleges és a λ_n sorozat kielégíti a 4.3.1. tétel $(* * *)$ feltételeit — teljesül, hogy $\lim_{n\to\infty} \left\|Z_n - a\right\| = 0$ 1 valószínűséggel.

$$Z_n = a_n + \sum_{i=1}^{n} c_i \mathbf{X}_i,$$

ahol

$$a_n = z_0 \prod_{i=0}^{n-1} (1-\lambda_i) \quad \text{és} \quad c_i = \lambda_{i-1}(1-\lambda_i)\cdots(1-\lambda_{n-1}).$$

Ha $z_0 = \mathbf{0}$ és $\lambda_n = \frac{1}{n+1}$, akkor $Z_n = \bar{\mathbf{X}}_n$ az átlagstatisztika.

4.3.1. Lineáris regressziós feladat

Ebben a szakaszban az előző pont eredményeinek egy különlegesen fontos speciális alkalmazását tárgyaljuk. Azzal az esettel foglalkozunk, amikor az $\mathbf{r}$ regressziós függvény $\mathbf{r}(\mathbf{c}) = \underline{\underline{A}}\mathbf{c} - \mathbf{m}$ lineáris függvény, ahol $\underline{\underline{A}}$ $(k+1)\times(k+1)$-es kvadratikus mátrix, $\mathbf{m}$ pedig $\mathbb{R}^{k+1}$ -beli vektor.

Legyen $\mathbf{Z} = \left(\underline{\underline{\Xi}}, \boldsymbol{\beta}\right)$ melyre $\mathbf{E}\underline{\underline{\Xi}} = \underline{\underline{A}}$ és $\boldsymbol{E\beta} = \mathbf{m}$ és $\mathbf{f}(\mathbf{c},\mathbf{Z}) = \underline{\underline{\Xi}}\mathbf{c} - \boldsymbol{\beta}$. Tegyük fel, hogy az $\underline{\underline{A}}$ mátrix szimmetrikus, pozitív definit és invertálható. (Ekkor a regressziós függvény $\mathbf{r}(\mathbf{c}) = \mathbf{Ef}(\mathbf{c},\mathbf{Z}) = \mathbf{E}\left(\underline{\underline{\Xi}}\mathbf{c} - \boldsymbol{\beta}\right) = \underline{\underline{A}}\mathbf{c} - \mathbf{m}$.)

4.3.2. tétel: Jelölje $\underline{\underline{\Xi}} = (V_{ij})_{\substack{i=0,1,\dots,k\\ j=0,1,\dots,k}}$ és $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)$ és tegyük fel, hogy $\exists M_1, M_2 > 0$, hogy $\mathbf{E}V_{ij}^2 \le M_1 < \infty$ és $\mathbf{E}\sum_{i=1}^{N}\beta_i^2 = M_2 < \infty$. Jelölje $\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_n, \dots$ a $\mathbf{Z}$-vel azonos eloszlású, teljesen független elemekből álló sorozatot! Tegyük fel továbbá, hogy $\{\gamma_n\}_{n=1,2,\dots}$ olyan pozitív tagú számsorozat, melyre $\sum_{n=0}^{\infty}\gamma_n = \infty$, és $\sum_{n=0}^{\infty}\gamma_n^2 < \infty$ teljesül.

Akkor az $\mathbf{C}_{n+1} = \mathbf{C}_n - \gamma_n\left(\underline{\underline{\Xi}}_{n+1}\mathbf{C}_n - \boldsymbol{\beta}_{n+1}\right)$, $\quad n = 0,1,\dots$ rekurzív megadású sorozatra $\lim_{n\to\infty}\left\|\mathbf{C}_n - \underline{\underline{A}}^{-1}\mathbf{m}\right\| = 0$ teljesül 1 valószínűséggel.

Bizonyítás: Meg fogjuk mutatni, hogy teljesülnek a 4.3.1. tétel feltételei, annak a 4.3.2. tétel speciális esete. Először megmutatjuk, hogy

$(*)\quad \exists K_1 > 0: \quad \left(\mathbf{c} - \underline{\underline{A}}^{-1}\mathbf{m}\right)^T\left(\underline{\underline{A}}\mathbf{c} - \mathbf{m}\right) \ge K_1\left\|\mathbf{c} - \underline{\underline{A}}^{-1}\mathbf{m}\right\| \quad \forall\mathbf{c}\in\mathbb{R}^{k+1}$-re.

Mivel $\underline{\underline{A}}$ szimmetrikus és pozitív definit az $\underline{\underline{A}}\boldsymbol{\varphi} = \lambda\boldsymbol{\varphi}$ sajátérték egyenletnek létezik $\{\lambda_n, \boldsymbol{\varphi}_n\}_{n=0,1,\dots k}$ megoldásrendszere, ahol $\{\boldsymbol{\varphi}_n\}_{n=0,1,\dots k}$ ortonormált rendszer $\mathbb{R}^{k+1}$-ben, és $\lambda_i \ge 0, i = 0,1,\dots,k$. Sőt $\lambda_i > 0, i = 0,1,\dots,k$ is, mivel ellenkező esetben az $\underline{\underline{A}}\boldsymbol{\varphi} = \mathbf{0}$ egyenletnek lenne nem triviális megoldása, ami ellentmond annak, hogy $\underline{\underline{A}}^{-1}$ létezik. Ekkor tehát vehetjük a következő sorfejtéseket: $\mathbf{c} = \sum_{i=0}^{k} c_i\boldsymbol{\varphi}_i, \mathbf{m} = \sum_{i=0}^{k} m_i\boldsymbol{\varphi}_i$. Ezekkel $\underline{\underline{A}}^{-1}\mathbf{m}.= \sum_{i=0}^{k}\frac{1}{\lambda_i}m_i\boldsymbol{\varphi}_i$. Ezt alkalmazva nyerjük, hogy

$$\left(\mathbf{c} - \underline{\underline{A}}^{-1}\mathbf{m}\right)^T\left(\underline{\underline{A}}\mathbf{c} - \mathbf{m}\right) = \sum_{i=0}^{k}\left(c_i - \frac{1}{\lambda_i}m_i\right)(\lambda_i c_i - m_i) = \sum_{i=1}^{N}\lambda_i\left(c_i - \frac{1}{\lambda_i}m_i\right)^2 \ge$$

$$\ge \left(\min_{0\le i\le k}\lambda_i\right)\sum_{i=0}^{k}\left(c_i - \frac{1}{\lambda_i}m_i\right)^2 = \left(\min_{0\le i\le k}\lambda_i\right)\left\|\mathbf{c} - \underline{\underline{A}}^{-1}\mathbf{m}\right\|^2.$$

A

$$K_1 = \left(\min_{0\le i\le k}\lambda_i\right)$$

választással igazoltuk a $(*)$ egyenlőtlenséget, ahonnan már következik a 4.3.1. tétel $(*)$ feltételének teljesülése. A 4.3.2. tétel bizonyítását tehát befejezhetjük, ha még megmutatjuk, hogy $\forall\mathbf{c}\in\mathbb{R}^{k+1}$-re

$$\mathbf{E}\left\|\mathbf{f}(\mathbf{c},\mathbf{Z})\right\|^2 = \mathbf{E}\left\|\left(\underline{\underline{\Xi}}\mathbf{c} - \boldsymbol{\beta}\right)\right\|^2 \le K\left(1 + \left\|\mathbf{c} - \underline{\underline{A}}^{-1}\mathbf{m}\right\|^2\right)$$

valamely $K > 0$-ra. Ugyanis a fenti reláció éppen a 4.3.1. tétel $(**)$ feltételének teljesülésével ekvivalens. Először is

$$\left\|\underline{\underline{\Xi}}\mathbf{c}-\boldsymbol{\beta}\right\|^2 = \left\|\underline{\underline{\Xi}}\left(\mathbf{c}-\underline{\underline{A}}^{-1}\mathbf{m}+\underline{\underline{A}}^{-1}\mathbf{m}\right)-\boldsymbol{\beta}\right\|^2 \le$$

$$\le \left\|\underline{\underline{\Xi}}\right\|^2\left(\left\|\mathbf{c}-\underline{\underline{A}}^{-1}\mathbf{m}\right\|^2+\left\|\underline{\underline{A}}^{-1}\mathbf{m}\right\|^2\right)+\|\boldsymbol{\beta}\|^2 \le$$

$$\le \left(\sum_{i=1}^{N}\sum_{j=1}^{N}V_{ij}^2\right)\left(\left\|\mathbf{c}-\underline{\underline{A}}^{-1}\mathbf{m}\right\|^2+\left\|\underline{\underline{A}}^{-1}\mathbf{m}\right\|^2\right)+\|\boldsymbol{\beta}\|^2 .$$

Innen

$$\left\|\underline{\underline{\Xi}}\mathbf{c}-\boldsymbol{\beta}\right\|^2 \le (k+1)^2M_1\left(\left\|\mathbf{c}-\underline{\underline{A}}^{-1}\mathbf{m}\right\|^2+\left\|\underline{\underline{A}}^{-1}\mathbf{m}\right\|^2\right)+M_2 \le K\left(1+\left\|\mathbf{c}-\underline{\underline{A}}^{-1}\mathbf{m}\right\|^2\right)$$

adódik, ahol

$$K = \max\left\{(k+1)^2M_1,(k+1)^2M_1\left\|\underline{\underline{A}}^{-1}\mathbf{m}\right\|^2+M_2\right\}.$$

■

4.3.2. Négyzetes hiba minimalizálása

A négyzetes hiba minimalizálásának problémája a lineáris regressziós problémára vezethető vissza. A probléma most az, hogy az $\mathbf{X} = (X_0, X_1, X_2, \ldots, X_N)^T, X_0 \equiv 1$ valószínűségi vektorváltozó komponenseinek milyen lineáris kombinációjával közelíthető legjobban az Y célváltozó, azaz milyen $\mathbf{c} = (c_0, c_1, \ldots, c_k)^T \in \mathbb{R}^{k+1}$ súlyok esetén lesz minimális az $m(\mathbf{c}) = \mathbf{E}\left[\sum\limits_{i=0}^{k} c_iX_i - Y\right]^2$ átlagos négyzetes hiba. Tekintsük most az $l(\mathbf{c},\mathbf{y}) = \left[\sum\limits_{i=0}^{k} c_iy_i - y_{k+1}\right]^2$ függvényt! A minimalizáláshoz származtatnunk kell az $l(\mathbf{c},\mathbf{y})$ függvény $\mathbf{c}$ vektor szerinti gradiensvektorát:

$$f(\mathbf{c},\mathbf{y}) = \operatorname*{grad}_{\mathbf{c}} l(\mathbf{c},\mathbf{y}) =$$

$$= 2\left(\left(\sum_{i=0}^{k}c_iy_i\right)y_0,\left(\sum_{i=0}^{k}c_iy_i\right)y_1,\ldots,\left(\sum_{i=0}^{k}c_iy_i\right)y_k\right)^T - 2\left(y_{k+1}y_0, y_{k+1}y_1,\ldots,y_{k+1}y_k\right)^T =$$

$$= \underline{\underline{B}}\mathbf{c}-\mathbf{b}, \quad \text{ahol } \underline{\underline{B}} = 2\begin{pmatrix} y_0y_0 & \cdots & y_0y_k \\ \vdots & \ddots & \vdots \\ y_ky_0 & \cdots & y_ky_k \end{pmatrix}$$

és

$$\mathbf{b} = 2\left(y_{k+1}y_0, y_{k+1}y_1,\ldots,y_{k+1}y_k\right)^T .$$

Másrészt a m négyzetes hiba gradiensére:

$$\operatorname*{grad}_{\mathbf{c}} m(\mathbf{c}) = \operatorname*{grad}_{\mathbf{c}} \mathbf{E}\left[\sum_{i=0}^{k}c_iX_i - Y\right]^2 = \operatorname*{grad}_{\mathbf{c}} \mathbf{E}\left[\sum_{i=0}^{k}\sum_{j=0}^{k}c_ic_jX_iX_j - 2\sum_{i=0}^{k}c_iX_iY - Y^2\right] =$$

$$= \operatorname*{grad}_{\mathbf{c}}\left[\sum_{i=0}^{k}\sum_{j=0}^{k}c_ic_j\mathbf{E}X_iX_j - 2\sum_{i=0}^{k}c_i\mathbf{E}X_iY - \mathbf{E}Y^2\right].$$

Innen, tekintettel arra, hogy

$$\frac{\partial}{\partial c_j}\sum_{i=0}^{k}\sum_{j=0}^{k} c_i c_j \mathbf{E} X_i X_j = 2\sum_{i=0}^{k} c_i \mathbf{E} X_i X_j$$

és

$$\frac{\partial}{\partial c_j}\sum_{i=0}^{k} c_i \mathbf{E} X_i Y = \mathbf{E} X_j Y$$

éppen az adódik, hogy

$$\operatorname*{grad}_{\mathbf{c}} m(\mathbf{c}) = 2\left(\sum_{i=0}^{k} c_i \mathbf{E} X_i X_0, \sum_{i=0}^{k} c_i \mathbf{E} X_i X_1, \ldots, \sum_{i=0}^{k} c_i \mathbf{E} X_i X_k\right)^T -$$

$$-2\left(\mathbf{E} X_0 Y, \mathbf{E} X_1 Y, \ldots, \mathbf{E} X_k Y\right) = \underline{\underline{A}}\mathbf{c} - \mathbf{m},$$

ahol

$$\underline{\underline{A}} = \begin{pmatrix} \mathbf{E} X_0 X_0 & \cdots & \mathbf{E} X_0 X_k \\ \vdots & \ddots & \vdots \\ \mathbf{E} X_k X_0 & \cdots & \mathbf{E} X_k X_k \end{pmatrix} \quad \text{és} \quad \mathbf{m} = 2\left(\mathbf{E} X_0 Y, \mathbf{E} X_1 Y, \ldots, \mathbf{E} X_k Y\right)^T.$$

Végül, ha

$$\underline{\underline{\Xi}} = \mathbf{X}\mathbf{X}^T = 2\begin{pmatrix} X_0 X_0 & \cdots & X_0 X_k \\ \vdots & \ddots & \vdots \\ X_k X_0 & \cdots & X_k X_k \end{pmatrix} \quad \text{és} \quad \boldsymbol{\beta} = \left(X_0 Y, X_1 Y, \ldots, X_k Y\right)^T,$$

akkor a $\mathbf{Z} = \left(\underline{\underline{\Xi}}, \boldsymbol{\beta}\right)$ és $f(\mathbf{c}, \mathbf{Z}) = \underline{\underline{\Xi}}\mathbf{c} - \boldsymbol{\beta}$ jelölésekkel, $\mathbf{r}(\mathbf{c}) = \mathbf{E} f(\mathbf{c}, \mathbf{Z}) = \mathbf{E}\left[\underline{\underline{\Xi}}\right]\mathbf{c} - \mathbf{E}\boldsymbol{\beta} = \underline{\underline{A}}\mathbf{c} - \mathbf{m}$ a regressziós függvény. Keresendő az $\mathbf{r}(\mathbf{c}) = \mathbf{0}$ egyenlet gyöke, ahol a $m(\mathbf{c})$ négyzetes hiba minimális lesz.

4.3.3. tétel: Tegyük fel, hogy $\underline{\underline{A}}^{-1}$ létezik és $\mathbf{E}\left[X_i^4\right] < \infty, \quad i = 0, 1, \ldots, k, \quad \mathbf{E} Y^4 < \infty$. Ha $\{\mathbf{Z}_n\}_{n=1,2,\ldots}$ $\mathbf{Z}$-vel azonos eloszlású valószínűségi vektorváltozó sorozat ($\mathbf{Z}_i = \left(\underline{\underline{\Xi}}_i, \boldsymbol{\beta}_i\right)$), akkor a

$$\mathbf{C}_{n+1} = \mathbf{C}_n - \gamma_n\left(\underline{\underline{\Xi}}_{n+1}\mathbf{C}_n - \boldsymbol{\beta}_{n+1}\right), \quad n = 0, 1, 2, \ldots,$$

$$\mathbf{C}_0 = \mathbf{c}_0 \in \mathbb{R}^{k+1}$$

rekurzív képlettel definiált valószínűségi vektorváltozó sorozatra $\lim_{n\to\infty}\left\|\mathbf{C}_n - \underline{\underline{A}}^{-1}\mathbf{m}\right\| = 0$ 1 valószínűséggel, feltéve, hogy $\sum_{n=0}^{\infty}\gamma_n = \infty$, és $\sum_{n=0}^{\infty}\gamma_n^2 < \infty$ teljesül.

Bizonyítás: $\underline{\underline{A}}$ definíciójából láthatóan szimmetrikus és pozitív szemidefinit, hiszen tetszőleges $\mathbf{t} \in \mathbb{R}^{k+1}$-re $\mathbf{t}^T\underline{\underline{A}}\mathbf{t} = \mathbf{E}\left(\mathbf{t}^T\underline{\underline{\Xi}}\mathbf{t}\right) = \mathbf{E}\left(\mathbf{t}^T\mathbf{X}\mathbf{X}^T\mathbf{t}\right) = \mathbf{E}\left(\mathbf{X}^T\mathbf{t}\right)^2 \geq 0$. Megmutatjuk, hogy $\exists M_1, M_2 > 0$, amivel $\mathbf{E}\left[X_i^2 X_j^2\right] \leq M_1$ és $\mathbf{E}\|\boldsymbol{\beta}\|^2 = \sum_{i=0}^{k}\mathbf{E} X_i^2 Y^2 \leq M_2$. A Cauchy–Schwarz-egyenlőtlenséget felhasználva ez azonnal adódik, hiszen

$$\mathbf{E}\left[X_i^2 X_j^2\right] \leq \sqrt{\mathbf{E}\left[X_i^4\right]\mathbf{E}\left[X_j^4\right]} < \infty$$

és

$$\sum_{i=0}^{k} \mathbf{E} X_i^2 Y^2 \le \sum_{i=0}^{k} \sqrt{\mathbf{E}\left[X_i^4\right] \mathbf{E}\left[Y^4\right]} < \infty.$$

Tehát teljesülnek a 4.3.2. tétel feltételei, akkor az állítás is igaz lesz.

■

Megjegyzés: A sztochasztikus approximáció módszerével tárgyalható a nemlineáris regresszió feladatának gradiens vektoros megoldási módja is, amikor az $l(\mathbf{c}, \mathbf{Z}) = \|\mathbf{f}(\mathbf{c}, \mathbf{X}) - Y\|^2$ alakú és $\mathbf{f}(\mathbf{c}, \mathbf{x})$ nemlineáris $\mathbf{x}$-ben.

5. fejezet

Eloszlásbecslés

Nemparaméteres statisztika esetén nem áll rendelkezésre semmilyen előzetes információ a valószínűségi változó eloszlásáról, így nem használhatjuk azt a tudást — mint paraméteres esetben —, hogy az eloszlás egy paraméteres osztály eleme lenne. Így a szabályok alapvető tulajdonságainak is eloszlásfüggetlennek kell lenniük.

5.1. Eloszlásfüggvény becslése

Legyen X valós értékű valószínűségi változó. A feladat az X valószínűségi változó $F(x)$ eloszlásfüggvényének becslése független, azonos eloszlású $X_1, X_2, \ldots, X_n$ mintákból. Mint korábban láttuk, az $F_n(x) = \frac{1}{n}\sum\limits_{i=1}^{n} I_{\{X_i<x\}}$ empirikus eloszlásfüggvény konstruálása egyrészt eloszlásfüggetlen, másrészt egyenletes a konvergenciája minden $F(x)$-re (Glivenko–Cantelli-tétel):

$$\lim_{n\to\infty}\sup_{x\in\mathbb{R}} |F(x) - F_n(x)| = 0$$

1 valószínűséggel. A konvergencia sebességéről a Glivenko–Cantelli-tétel nem ad felvilágosítást. A következő tételek azt mondják, hogy n minta körülbelül $\frac{1}{\sqrt{n}}$ nagyságrendű közelítéshez elegendő:

5.1.1. tétel: *(Szmirnov)*
Ha az $F(x)$ eloszlásfüggvény folytonos, akkor

$$\lim \mathbf{P}\left(\sqrt{n}\sup_{x\in\mathbb{R}}(F_n(x) - F(x)) < y\right) = \begin{cases} 1 - \mathrm{e}^{-2y^2}, & \text{ha } y > 0 \\ 0 & \text{különben.} \end{cases}$$

5.1.2. tétel: *(Kolmogorov)*
Ha $F(x)$ folytonos, akkor

$$\lim_{n\to\infty} \mathbf{P}\left(\sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x) - F(x)| < y\right) = \begin{cases} K(y), & \text{ha } y > 0 \\ 0 & \text{különben,} \end{cases}$$

ahol

$$K(y) = \sum_{k=-\infty}^{\infty} (-1)^k \mathrm{e}^{-2k^2y^2}$$

Vegyük észre, hogy az előbbi tételekben a határeloszlás nem függ az elméleti eloszlásfüggvénytől.

Most adunk egy alternatív bizonyítást az empirikus eloszlásfüggvény egyenletes konvergenciájára, amely sok hasznos öteletet tartalmaz és segít a következő fejezet fontos tételének, a Vapnik–Chervonenkis-egyenlőtlenségnek a bizonyításában.

5.1.3. tétel: *(Glivenko–Cantelli)*
Legyen $X_1, \dots, X_n$ független, azonos eloszlású valós értékű valószínűségi változó $F(x) = \mathbf{P}(X_1 \le x)$ eloszlásfüggvénnyel. Ekkor

$$\mathbf{P}\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \varepsilon\right) \le 8(n+1)\mathrm{e}^{-n\varepsilon^2/32}$$

és így a Borel–Cantelli-lemma miatt

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0$$

1 valószínűséggel.

A tétel bizonyításához szükségünk lesz a Hoeffding-egyenlőtlenségre.

5.1.4. tétel: *(Hoeffding)*
Legyenek $X_1, \dots, X_n$ független korlátos valószínűségi változók úgy, hogy $X_i \in [a_i, b_i]$ egy valószínűséggel. Jelölje az összegüket S_n, vagyis $S_n = \sum_{i=1}^{n} X_i$. Ekkor minden $\varepsilon > 0$-ra

$$\mathbf{P}\{S_n - \mathbf{E}S_n \ge \varepsilon\} \le \mathrm{e}^{-2\varepsilon^2/\sum_{i=1}^{n}(b_i-a_i)^2}$$

és

$$\mathbf{P}\{S_n - \mathbf{E}S_n \le -\varepsilon\} \le \mathrm{e}^{-2\varepsilon^2/\sum_{i=1}^{n}(b_i-a_i)^2}.$$

Az egyenlőtlenség bizonyításához használunk egy segédegyenlőtlenséget:

5.1.1. lemma: Legyen X olyan valószínűségi változó, amelyre $\mathbf{E}X = 0$, $a \le X \le b$. Ekkor minden $s > 0$-ra,

$$\mathbf{E}\left\{\mathrm{e}^{sX}\right\} \le \mathrm{e}^{s^2(b-a)^2/8}.$$

Bizonyítás: Az exponenciális függvény konvexitásából következik, hogy

$$\mathrm{e}^{sx} \le \frac{x-a}{b-a}\mathrm{e}^{sb} + \frac{b-x}{b-a}\mathrm{e}^{sa}, \quad \text{ha } a \le x \le b.$$

Legyen $p = -a/(b-a)$, ekkor kihasználva, hogy $\mathbf{E}X = 0$

$$\begin{aligned}
\mathbf{E}\mathrm{e}^{sX} &\le \frac{b}{b-a}\mathrm{e}^{sa} - \frac{a}{b-a}\mathrm{e}^{sb} \\
&= \left(1 - p + p\mathrm{e}^{s(b-a)}\right)\mathrm{e}^{-ps(b-a)} \\
&\stackrel{\text{def}}{=} \mathrm{e}^{\phi(u)},
\end{aligned}$$

ahol $u = s(b-a)$, és $\phi(u) = -pu + \log(1-p+pe^u)$. Mivel ϕ deriváltja

$$\phi'(u) = -p + \frac{p}{p+(1-p)e^{-u}},$$

ezért $\phi(0) = \phi'(0) = 0$. A második derivált, pedig

$$\phi''(u) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} \leq \frac{1}{4}.$$

Így a Taylor-sorfejtés szerint valamely $\theta \in [0,u]$-ra,

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}.$$

■

Az 5.1.4 tétel bizonyítása:
A bizonyítás az úgynevezett *Chernoff-technikán* alapul. A Markov-egyenlőtlenségből tudjuk, hogy minden X nemnegatív valószínűségi változóra és $\varepsilon > 0$-ra,

$$\mathbf{P}\{X \geq \varepsilon\} \leq \frac{\mathbf{E}X}{\varepsilon}.$$

Ezért, ha s tetszőleges pozitív szám, akkor minden X valószínűségi változóra

$$\mathbf{P}\{X \geq \varepsilon\} = \mathbf{P}\{e^{sX} \geq e^{s\varepsilon}\} \leq \frac{\mathbf{E}e^{sX}}{e^{s\varepsilon}}.$$

A Chernoff-technika lényege, hogy keresünk egy olyan $s > 0$-t, amely minimalizálja, vagy kellően kicsivé teszi a felső korlátot.

$$\begin{aligned}
&\mathbf{P}\{S_n - \mathbf{E}S_n \geq \varepsilon\} \\
&\leq\ e^{-s\varepsilon}\mathbf{E}\left\{\exp\left(s\sum_{i=1}^n (X_i - \mathbf{E}X_i)\right)\right\} \\
&=\ e^{-s\varepsilon}\prod_{i=1}^n \mathbf{E}\left\{e^{s(X_i-\mathbf{E}X_i)}\right\} \qquad (X_i\text{-k függetlensége miatt}) \\
&\leq\ e^{-s\varepsilon}\prod_{i=1}^n e^{s^2(b_i-a_i)^2/8} \quad (5.1.1 \text{ lemma miatt}) \\
&=\ e^{-s\varepsilon}e^{s^2\sum_{i=1}^n (b_i-a_i)^2/8} \\
&=\ e^{-2\varepsilon^2 / \sum_{i=1}^n (b_i-a_i)^2} \qquad \left(s = 4\varepsilon / \sum_{i=1}^n (b_i-a_i)^2\text{-t választva}\right).
\end{aligned}$$

A második egyenlőtlenség hasonlóan bizonyítható.

■

Az 5.1.4 tétel két egyenlőtlenségét összekombinálva kaphatjuk, hogy

$$\mathbf{P}\{|S_n - \mathbf{E}S_n| \geq \varepsilon\} \leq 2e^{-2\varepsilon^2 / \sum_{i=1}^n (b_i-a_i)^2}.$$

Az 5.1.3 tétel bizonyítása:

Vezessük be a következő jelöléseket: $\mu(A) = \mathbf{P}(X \in A)$ és $\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \in A\}}$ minden $A \subset \mathbb{R}$ mérhető halmazra. Legyen $\mathcal{A}$ a $(-\infty, x]$; $x \in \mathbb{R}$ alakú halmazok családja. Ezekkel a jelölésekkel

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = \sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|$$

Feltehetjük, hogy $n\varepsilon^2 \geq 2$, hiszen különben a felső korlát triviális (≥ 1).

1. LÉPÉS: Szimmetrizálás szellemmintával

   Legyenek $X_1', \ldots, X_n' \in \mathbb{R}$ valószínűségi változók úgy, hogy $X_1, \ldots, X_n, X_1', \ldots, X_n'$ mind független és azonos eloszlású. Jelölje μ_n' az új minták szerinti empirikus mértéket:

$$\mu_n'(A) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i' \in A\}}$$

   Ekkor megmutatjuk, hogy $n\varepsilon^2 \geq 2$-re

$$\mathbf{P}\left(\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| > \varepsilon\right) \leq 2\mathbf{P}\left(\sup_{A\in\mathcal{A}} \left|\mu_n(A) - \mu_n'(A)\right| > \frac{\varepsilon}{2}\right)$$

   Ehhez legyen $A^* \in \mathcal{A}$ egy olyan halmaz, amelyre $|\mu_n(A^*) - \mu(A^*)| > \varepsilon$, ha ilyen halmaz létezik, különben legyen A^* egy rögzített $\mathcal{A}$-beli halmaz. Ekkor

$$\mathbf{P}\left(\sup_{A\in\mathcal{A}} \left|\mu_n(A) - \mu_n'(A)\right| > \frac{\varepsilon}{2}\right) \geq \mathbf{P}\left(\left|\mu_n(A^*) - \mu_n'(A^*)\right| > \frac{\varepsilon}{2}\right) \geq$$
$$\geq \mathbf{P}\left(|\mu_n(A^*) - \mu(A^*)| > \varepsilon, \left|\mu_n'(A^*) - \mu(A^*)\right| < \frac{\varepsilon}{2}\right) =$$
$$= \mathbf{E}\left(I_{\{|\mu_n(A^*)-\mu(A^*)|>\varepsilon\}}\mathbf{P}\left(\left|\mu_n'(A^*) - \mu(A^*)\right| < \frac{\varepsilon}{2}\,\middle|\, X_1, \ldots, X_n\right)\right)$$

   A feltételes valószínűséget becsülhetjük a Csebisev-egyenlőtlenség segítségével a következőképpen, ha $n\varepsilon^2 \geq 2$:

$$\mathbf{P}\left(\left|\mu_n'(A^*) - \mu(A^*)\right| < \frac{\varepsilon}{2}\,\middle|\, X_1, \ldots, X_n\right) \geq$$
$$\geq 1 - \frac{\mu(A^*)(1-\mu(A^*))}{n\varepsilon^2/4} \geq 1 - \frac{1}{n\varepsilon^2} \geq \frac{1}{2}$$

   Összefoglalva tehát

$$\mathbf{P}\left(\sup_{A\in\mathcal{A}} \left|\mu_n(A) - \mu_n'(A)\right| > \frac{\varepsilon}{2}\right) \geq \frac{1}{2}\mathbf{P}\left(|\mu_n(A^*) - \mu(A^*)| > \varepsilon\right) \geq$$
$$\geq \frac{1}{2}\mathbf{P}\left(\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| > \varepsilon\right)$$

2. LÉPÉS: Szimmetrizálás véletlen előjelekkel

   Legyenek $\sigma_1, \ldots, \sigma_n$ független, azonos eloszlású $X_1, \ldots, X_n, X_1', \ldots, X_n'$-től független $\{-1, 1\}$ értékű valószínűségi változók, $\mathbf{P}(\sigma_i = -1) = \mathbf{P}(\sigma_i = 1) = \frac{1}{2}$ valószínűségekkel. Mivel $X_1, X_1', \ldots, X_n, X_n'$ mind független és azonos eloszlású,

$$\sup_{A\in\mathcal{A}} \left|\sum_{i=1}^{n}\left(I_{\{X_i\in A\}} - I_{\{X_i'\in A\}}\right)\right|$$

és

$$\sup_{A\in\mathcal{A}} \left| \sum_{i=1}^{n} \sigma_i \left(I_{\{X_i\in A\}} - I_{\{X_i'\in A\}} \right) \right|$$

azonos eloszlású. Így az 1. lépés miatt

$$\mathbf{P}\left(\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| > \varepsilon \right) \leq$$

$$\leq 2\mathbf{P}\left(\sup_{A\in\mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} (I_{\{X_i\in A\}} - I_{\{X_i'\in A\}}) \right| > \frac{\varepsilon}{2} \right) =$$

$$= 2\mathbf{P}\left(\sup_{A\in\mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i (I_{\{X_i\in A\}} - I_{\{X_i'\in A\}}) \right| > \frac{\varepsilon}{2} \right)$$

Az uniokorlátot használva megszabadulhatunk az $X_1', \ldots, X_n'$ segéd valószínűségi változóktól

$$\mathbf{P}\left(\sup_{A\in\mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i (I_{\{X_i\in A\}} - I_{\{X_i'\in A\}}) \right| > \frac{\varepsilon}{2} \right) \leq$$

$$\leq \mathbf{P}\left(\sup_{A\in\mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i I_{\{X_i\in A\}} \right| > \frac{\varepsilon}{4} \right) + \mathbf{P}\left(\sup_{A\in\mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i I_{\{X_i'\in A\}} \right| > \frac{\varepsilon}{4} \right) =$$

$$= 2\mathbf{P}\left(\sup_{A\in\mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i I_{\{X_i\in A\}} \right| > \frac{\varepsilon}{4} \right)$$

3. LÉPÉS:

A $\mathbf{P}\left(\sup_{A\in\mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i I_{\{X_i\in A\}} \right| > \frac{\varepsilon}{4} \right) = \mathbf{P}\left(\sup_{x\in\mathbb{R}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i I_{\{X_i\leq x\}} \right| > \frac{\varepsilon}{4} \right)$ valószínűség becsléséhez nézzük először a feltételes valószínűséget feltéve $X_1, \ldots, X_n$-et. Vegyük észre, hogy rögzített $x_1, \ldots, x_n \in \mathbb{R}$-re, ahogy x végigfut $\mathbb{R}$-en a különböző $(I_{\{x_1<x\}}, I_{\{x_2<x\}}, \ldots, I_{\{x_n<x\}})$ vektorok száma legfeljebb $n+1$. Ezért rögzített $X_1, \ldots, X_n$-re a szuprémum a fenti valószínűségben legfeljebb $n+1$ valószínűségi változó maximuma. Így, alkalmazva az uniokorlátot

$$\mathbf{P}\left(\sup_{A\in\mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i I_{\{X_i\in A\}} \right| > \frac{\varepsilon}{4} \,\middle|\, X_1, \ldots, X_n \right) \leq$$

$$\leq (n+1) \sup_{A\in\mathcal{A}} \mathbf{P}\left(\frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i I_{\{X_i\in A\}} \right| > \frac{\varepsilon}{4} \,\middle|\, X_1, \ldots, X_n \right)$$

Így mivel a szuprémum kívülre került, elég a

$$\mathbf{P}\left(\frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i I_{\{X_i\in A\}} \right| > \frac{\varepsilon}{4} \,\middle|\, X_1, \ldots, X_n \right)$$

feltételes valószínűségre találni egy exponenciális felső korlátot.

4. LÉPÉS:

Rögzített $x_1, \ldots, x_n$-re $\sum_{i=1}^{n} \sigma_i I_{\{x_i \in A\}}$ n darab független, 0 várható értékű, -1 és 1 közötti valószínűségi változó összege, ezért alkalmazhatjuk a Hoeffding-egyenlőtlenséget:

$$\mathbf{P}\left(\frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i I_{\{X_i \in A\}} \right| > \frac{\varepsilon}{4} \,\middle|\, X_1, \ldots, X_n \right) \leq 2\mathrm{e}^{-n\varepsilon^2/32}.$$

Így

$$\mathbf{P}\left(\sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i I_{\{X_i \in A\}} \right| > \frac{\varepsilon}{4} \,\middle|\, X_1, \ldots, X_n \right) \leq 2(n+1)\mathrm{e}^{-n\varepsilon^2/32}.$$

Mindkét oldal várható értékét véve

$$\mathbf{P}\left(\sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i I_{\{X_i \in A\}} \right| > \frac{\varepsilon}{4} \right) \leq 2(n+1)\mathrm{e}^{-n\varepsilon^2/32}.$$

Összefoglalva tehát azt kapjuk, hogy

$$\mathbf{P}\left(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| > \varepsilon \right) \leq 8(n+1)\mathrm{e}^{-n\varepsilon^2/32}.$$

■

5.2. Vapnik–Chervonenkis-elmélet

Ebben a fejezetben a Glivenko–Cantelli-tétel egy általánosítását bizonyítjuk. Legyen most X d-dimenziós valószínűségi változó, és legyenek $X_1, \ldots, X_n$ az X eloszlásából vett független minták. Használjuk a következő jelöléseket: $\mu(A) = \mathbf{P}(X \in A)$ és $\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \in A\}}$ minden mérhető $A \subset \mathbb{R}^d$ halmazra.

5.2.1. definíció: Legyen $x_1, \ldots, x_n$ n darab $\mathbb{R}^d$-beli rögzített pont, $\mathcal{A}$ pedig az $\mathbb{R}^d$-beli halmazok egy családja. Ekkor legyen $N_{\mathcal{A}}(x_1, \ldots, x_n)$ az

$$\{x_1, \ldots, x_n\} \cap A$$

alakú halmazok száma, ha $A \in \mathcal{A}$. Vagyis $N_{\mathcal{A}}(x_1, \ldots, x_n)$ azt mutatja, hogy az $\mathcal{A}$-beli halmazokkal az $x_1, \ldots, x_n$ pontoknak hányféle különböző részhalmazát lehet kimetszeni.

Az $\mathcal{A}$ halmazcsalád n-edik shatter együtthatója

$$s(\mathcal{A}, n) \stackrel{\text{def}}{=} \max_{x_1, \ldots, x_n \in \mathbb{R}^d} N_{\mathcal{A}}(x_1, \ldots, x_n).$$

Nyilvánvalóan $s(\mathcal{A}, n) \leq 2^n$, hiszen egy n pontú halmaznak összesen 2^n részhalmaza van. Ha $s(\mathcal{A}, n) = 2^n$, vagyis valamely n pontra $N_{\mathcal{A}}(x_1, \ldots, x_n) = 2^n$, akkor azt mondjuk, hogy $\mathcal{A}$ *darabokra töri* (vagy shattereli) $\{x_1, \ldots, x_n\}$-t. Ha ez nem teljesül, akkor bármely n pontnak van olyan részhalmaza, amelyet nem tudunk kiválasztani $\mathcal{A}$-beli halmazzal. Az is nyilvánvaló, hogy ha valamely n_0-ra $s(\mathcal{A}, n_0) < 2^{n_0}$, akkor már minden $n > n_0$-ra $s(\mathcal{A}, n) < 2^n$.

5.2.2. definíció: A legnagyobb n_0 számot, amelyre még van olyan n_0 pont, amelyet $\mathcal{A}$ darabokra tör, vagyis

$$s(\mathcal{A}, n_0) = 2^{n_0}$$

az $\mathcal{A}$ család Vapnik–Chervonenkis-dimenziójának (vagy VC-dimenziójának) nevezzük, és $V_{\mathcal{A}}$-val jelöljük. Ha minden n-re $s(\mathcal{A}, n) = 2^n$, akkor definíció szerint $V_{\mathcal{A}} = \infty$.

Azokat az $\mathcal{A}$ halmazcsaládokat, amelyekre $V_{\mathcal{A}} < \infty$, Vapnik–Chervonenkis- (vagy VC-) családoknak hívjuk.

5.2.1. tétel: *(Vapnik–Chervonenkis)*
Minden μ valószínűségi mértékre és $\mathcal{A}$ halmazosztályra, minden n-re és $\varepsilon > 0$-ra

$$\mathbf{P}\left(\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| > \varepsilon\right) \leq 8s(\mathcal{A}, n)\mathrm{e}^{-n\varepsilon^2/32}$$

Bizonyítás: Követjük a Glivenko–Cantelli-tétel bizonyításának menetét. Most is feltehetjük, hogy $n\varepsilon^2 \geq 2$, hiszen különben a korlát triviális (≥ 1).

Az első két lépésben teljesen ugyanúgy bebizonyítjuk, hogy

$$\mathbf{P}\left(\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| > \varepsilon\right) \leq 4\mathbf{P}\left(\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_{\{X_i\in A\}}\right| > \frac{\varepsilon}{4}\right)$$

Az egyetlen különbség a 3. lépésben van.

3. LÉPÉS:

Vegyük észre, hogy rögzített $x_1, \ldots, x_n \in \mathbb{R}^d$-re ahogy A végigfut $\mathcal{A}$-n a különböző $(I_{\{x_i\in A\}}, \ldots, I_{\{x_n\in A\}})$ vektorok száma nem más, mint az $\{X_1, \ldots, X_n\}$ különböző olyan részhalmazainak a száma, amelyeket úgy kaphatunk, hogy $\mathcal{A}$-beli halmazokkal elmetszszük, ami definíció szerint legfeljebb $s(\mathcal{A}, n)$. Ezért rögzített $X_1, \ldots, X_n$-re a szuprémum a

$$\mathbf{P}\left(\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_{\{X_i\in A\}}\right| > \frac{\varepsilon}{4}\right)$$

valószínűségben legfeljebb $N_{\mathcal{A}}(X_1, \ldots, X_n) \leq s(\mathcal{A}, n)$ valószínűségi változó maximuma. Az uniokorláttal kapjuk, hogy

$$\mathbf{P}\left(\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_{\{X_i\in A\}}\right| > \frac{\varepsilon}{4} \;\middle|\; X_1, \ldots, X_n\right) \leq$$

$$\leq s(\mathcal{A}, n) \sup_{A\in\mathcal{A}} \mathbf{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_{\{X_i\in A\}}\right| > \frac{\varepsilon}{4} \;\middle|\; X_1, \ldots, X_n\right)$$

Így, mivel a szuprémum kívülre került, elég a

$$\mathbf{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i I_{\{X_i\in A\}}\right| > \frac{\varepsilon}{4} \;\middle|\; X_1, \ldots, X_n\right)$$

feltételes valószínűségre találni egy exponenciális felső korlátot. Ezt a Glivenko–Cantelli-tétel bizonyításának 4. lépésével teljesen azonos módon tehetjük meg, és így végül kapjuk, hogy

$$\mathbf{P}\left(\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| > \varepsilon\right) \leq 8s(\mathcal{A}, n)\mathrm{e}^{-n\varepsilon^2/32}.$$

■

Ha a valószínűségi változóink valósak és az $\mathcal{A}$ halmazcsalád a $(-\infty, x]$ alakú halmazokból áll, ahol $x \in \mathbb{R}$, akkor

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$$

és $s(\mathcal{A}, n) = n + 1$, hiszen $(-\infty, x]$ félegyenesekkel n pontnak $n + 1$ különböző részhalmazát tudjuk kiválasztani: $\emptyset$, $\{x_1\}$, $\{x_1, x_2\}$, $\ldots$, $\{x_1, x_2, \ldots, x_n\}$, ha $x_1 < x_2 < \cdots < x_n$.

Ekkor tehát a fenti tétel azt mondja, hogy

$$\mathbf{P}\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \varepsilon\right) \le 8(n+1)\mathrm{e}^{-n\varepsilon^2/32}.$$

Tehát a fenti tétel valóban általánosítja az empirikus eloszlásfüggvény konvergenciájára vonatkozó korábbi eredményt.

A következő tétel megmutatja a kapcsolatot egy halmazcsalád VC-dimenziója és shatter együtthatója között.

5.2.2. tétel: Ha az $\mathcal{A}$ halmazcsalád VC-dimenziója $V_{\mathcal{A}}$, akkor minden n-re

$$s(\mathcal{A}, n) \le \sum_{i=1}^{V_{\mathcal{A}}} \binom{n}{i}.$$

Alkalmazva a binomiális tételt, ebből mindjárt az is következik, hogy $s(\mathcal{A}, n) \le (n+1)^{V_{\mathcal{A}}}$. Sőt az is bebizonyítható, hogy $V_{\mathcal{A}} > 2$-re $s(\mathcal{A}, n) \le n^{V_{\mathcal{A}}}$ és minden $V_{\mathcal{A}}$-ra $s(\mathcal{A}, n) \le n^{V_{\mathcal{A}}} + 1$.

Ez azt jelenti, hogy a shatter együtthatóra vagy az igaz, hogy $s(\mathcal{A}, n) = 2^n$ minden n-re, vagy pedig az, hogy $s(\mathcal{A}, n) \le n^{V_{\mathcal{A}}} + 1$, ami akkor teljesül, ha $\mathcal{A}$ Vapnik–Chervonenkis-család, vagyis a VC-dimenziója véges. Érdekes, hogy $s(\mathcal{A}, n)$ nem eshet a két nagyságrend közé, azaz nem lehet például $n^{\ln n}$ vagy $2^{\sqrt{n}}$ nagyságrendű. Ha $V_{\mathcal{A}} < \infty$, akkor a Vapnik–Chervonenkis-egyenlőtlenség felső korlátja exponenciális sebességgel csökken, ahogy n nő.

A fenti tétel felső korlátja éles.

5.2.3. tétel: 1. Ha $\mathcal{A}$ a félegyenesek családja, azaz $\mathcal{A} = \{(-\infty, x];\ x \in \mathbb{R}\}$, akkor $V_{\mathcal{A}} = 1$ és $s(\mathcal{A}, n) = n + 1 = \binom{n}{0} + \binom{n}{1}$.

2. Ha $\mathcal{A}$ az intervallumok családja: $\mathcal{A} = \{[x_1, x_2];\ x_1, x_2 \in \mathbb{R}\}$, akkor $V_{\mathcal{A}} = 2$ és $s(\mathcal{A}, n) = \frac{n(n+1)}{2} + 1 = \binom{n}{0} + \binom{n}{1} + \binom{n}{2}$.

Bizonyítás:

1. -et már láttuk.

2. -ben $V_{\mathcal{A}} = 2$ abból látszik, hogy ha lerögzítünk 3 pontot az egyenesen, akkor nincs olyan intervallum, amelyik tartalmazza a két szélsőt, de a középsőt nem. A shatter együtthatót megkapjuk, ha észrevesszük, hogy legfeljebb $n - k + 1$ halmaz van $\{A \cap \{X_1, \ldots, X_n\};\ A \in \mathcal{A}\}$-ban amelyre $|A \cap \{X_1, \ldots, X_n\}| = k$; $k = 1, \ldots, n$ és egy amelyre $|A \cap \{X_1, \ldots, X_n\}| = 0$. Ebből

$$s(\mathcal{A}, n) = 1 + \sum_{k=1}^{n} (n - k + 1) = \frac{n(n+1)}{2} + 1.$$

■

Általánosítsuk a fenti eredményt d dimenzióra.

5.2.4. tétel: 1. Ha $\mathcal{A} = \{(-\infty, x_1] \times \cdots \times (-\infty, x_d]\}$, akkor $V_{\mathcal{A}} = d$.

2. Ha $\mathcal{A}$ az összes $\mathbb{R}^d$-beli téglalap családja, akkor $V_{\mathcal{A}} = 2d$.

Ezek után megkaphatjuk az 5.1.3 tétel általánosítását d-dimenziós valószínűségi változókra.

5.2.5. tétel: Legyen $X_1, \ldots, X_n \in \mathbb{R}^d$ független, azonos eloszlású valós értékű valószínűségi változó $F(x) = \mathbf{P}(X_1 \leq x)$ eloszlásfüggvénnyel. Ekkor

$$\mathbf{P}\left(\sup_{x \in \mathbb{R}^d} |F_n(x) - F(x)| > \varepsilon\right) \leq 8n^d \mathrm{e}^{-n\varepsilon^2/32}$$

és így a Borel–Cantelli-lemma miatt

$$\lim_{n \to \infty} \sup_{x \in \mathbb{R}^d} |F_n(x) - F(x)| = 0$$

1 valószínűséggel.

Talán az egyik legfontosabb halmazcsalád az $\mathbb{R}^d$-beli félterek családja.

5.2.6. tétel: Legyen $\mathcal{A}$ az $\mathbb{R}^d$-beli félterek, azaz az $\{x : a^T x \geq b,\ a \in \mathbb{R}^d, b \in \mathbb{R}\}$ alakú részhalmazok családja. Ekkor $V_{\mathcal{A}} = d + 1$.

Nézzünk meg egy negatív példát:

5.2.7. tétel: Ha $\mathcal{A}$ az összes $\mathbb{R}^2$-beli konvex sokszög családja, akkor $V_{\mathcal{A}} = \infty$.

Bizonyítás: Legyenek $x_1, \ldots, x_n \in \mathbb{R}^2$ az egységkör különböző pontjai, ekkor könnyű látni, hogy bármely részhalmazukhoz létezik olyan konvex sokszög, amelyik pontosan azokat a pontokat tartalmazza.

■

6. fejezet

Sűrűségfüggvény becslése

6.1. Az L_1 hiba

Az egyenletes konvergencia ellenére az empirikus eloszlásfüggvény sokszor nem bizonyul elég jó eloszlásbecslésnek. A mértékelméletből tudjuk, hogy az $F(x)$ eloszlásfüggvény egyértelműen meghatározza a $\mu(A)$ eloszlást. A Glivenko–Cantelli-tétel azt mondja, hogy az $F_n(x)$ empirikus eloszlásfüggvény egyenletesen konvergál az $F(x)$ eloszlásfüggvényhez. Úgy tűnik, hogy ezzel megoldottuk a $\mu(A)$ eloszlásbecslésének problémáját. Sajnos a statisztikában sok probléma esetén a $\mu_n(A)$ empirikus eloszlásbecslés használhatatlan. Erősebb hibakritériumot kell keresnünk.

6.1.1. definíció: Két valószínűségi mérték, μ és ν variációs távolsága

$$V(\mu,\nu) = \sup_A |\mu(A) - \nu(A)|,$$

ahol a szuprémumot az összes Borel-halmaz felett vesszük.

6.1.1. tétel: *(Scheffé)*
Ha a μ és ν valószínűségi mérték abszolút folytonos f, illetve g sűrűségfüggvénnyel, akkor

$$V(\mu,\nu) = \frac{1}{2}\int |f(x) - g(x)| \, \mathrm{d}x.$$

Bizonyítás: Jelölje $A^* = \{x : \ f(x) \geq g(x)\}$. Ekkor egyrészt

$$V(\mu,\nu) = \sup_A |\mu(A) - \nu(A)| = \sup_A \left| \int_A f(x)\,\mathrm{d}x - \int_A g(x)\,\mathrm{d}x \right| \geq$$

$$\geq \left| \int_{A^*} f(x)\,\mathrm{d}x - \int_{A^*} g(x)\,\mathrm{d}x \right| = \int_{A^*} (f(x) - g(x)) \, \mathrm{d}x =$$

$$= \frac{1}{2} \left(\int_{A^*} (f(x) - g(x)) \, \mathrm{d}x + \int_{(A^*)^c} (g(x) - f(x)) \, \mathrm{d}x \right) =$$

$$= \frac{1}{2}\int |f(x) - g(x)| \, \mathrm{d}x.$$

Másrészt

$$\left|\int\limits_A f(x)\,\mathrm{d}x - \int\limits_A g(x)\,\mathrm{d}x\right| = \left|\int\limits_{A\cap A^*} (f(x)-g(x))\;\mathrm{d}x + \int\limits_{A\cap (A^*)^c} (f(x)-g(x))\;\mathrm{d}x\right| \leq$$

$$\leq \max\left(\int\limits_{A\cap A^*} (f(x)-g(x))\;\mathrm{d}x,\; \int\limits_{A\cap (A^*)^c} (g(x)-f(x))\;\mathrm{d}x\right) \leq$$

$$\leq \max\left(\int\limits_{A^*} (f(x)-g(x))\;\mathrm{d}x,\; \int\limits_{(A^*)^c} (g(x)-f(x))\;\mathrm{d}x\right) =$$

$$= \frac{1}{2}\int |f(x)-g(x)|\;\mathrm{d}x.$$

Tehát

$$V(\mu,\nu) = \frac{1}{2}\int |f(x)-g(x)|\;\mathrm{d}x.$$

■

Ebből a tételből az következik, hogy ha találunk egy L_1-ben konzisztens sűrűségfüggvény-becslőt, akkor abból kaphatunk egy variációs távolságban konzisztens eloszlásbecslőt.

6.1.2. definíció: Az f_n sűrűségfüggvénybecslő x-nek és az f sűrűségfüggvényből vett független, azonos eloszlású $X_1,\dots,X_n$ mintáknak Borel-mérhető függvénye:

$$f_n(x) = f_n(x, X_1,\dots,X_n).$$

Ha f_n egy L_1-ben konzisztens sűrűségfüggvénybecslő, azaz

$$\lim_{n\to\infty} ||f-f_n|| = \lim_{n\to\infty}\int |f(x)-f_n(x)|\;\mathrm{d}x = 0$$

(sztochasztikusan) 1 valószínűséggel, akkor a

$$\tilde{\mu}_n(A) = \int\limits_A f_n(x)\,\mathrm{d}x$$

eloszlásbecslőre

$$\lim_{n\to\infty} V(\mu,\tilde{\mu}_n) = 0$$

(sztochasztikusan) 1 valószínűséggel.

6.2. A hisztogram

Ha f a μ valószínűségi mérték sűrűségfüggvénye, akkor $\int\limits_A f = \mu(A)$ minden Borel-mérhető halmazra, f majdnem mindenhol egyenlő a $\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$ Radon–Nikodym-deriválttal, ahol λ a Lebesgue-mértéket jelöli. A legtöbb sűrűségfüggvény-becslő ezt a deriváltat próbálja közelíteni. Két standard L_1-ben konzisztens sűrűségbecslő a hisztogram és a magfüggvényes becslő.

Legyen $\mathcal{P}_n = \{A_{n1}, A_{n2}, \dots\}$ az $\mathbb{R}^d$ egy partíciója pozitív és véges Lebesgue-mértékű cellákra. Ekkor a hisztogram becslő az

$$f_n(x) = \frac{\mu_n(A_n(x))}{\lambda(A_n(x))}$$

függvény, ahol μ_n az empirikus mérték, és $A_n(x) = A_{nj}$, ha $x \in A_{nj}$. A cellák gyakran h_n élhosszúságú d dimenziós kockák, ebben az esetben

$$f_n(x) = \frac{\mu_n(A_n(x))}{h_n^d}$$

6.2.1. tétel: Tegyük fel, hogy μ-nek létezik f sűrűségfüggvénye. Ha a hisztogram becslőnél minden origó középpontú S gömbre

$$\lim_{n\to\infty} \sup_{j:A_{nj}\cap S\neq\emptyset} \operatorname{diam}(A_{nj}) = 0$$

és

$$\lim_{n\to\infty} \frac{|\{j : A_{nj}\cap S \neq \emptyset\}|}{n} = 0,$$

akkor

$$\lim_{n\to\infty} \int |f(x) - f_n(x)|\, \lambda(\mathrm{d}x) = 0$$

1 valószínűséggel, ahol $\operatorname{diam}(A) = \sup\limits_{x,y\in A} ||x-y||$.

Bizonyítás:

$$\int |f_n(x) - f(x)|\, \lambda(\mathrm{d}x) \le \underbrace{\int |f_n(x) - \mathbf{E}f_n(x)|\, \lambda(\mathrm{d}x)}_{\text{variációs tag}} + \underbrace{\int |\mathbf{E}f_n(x) - f(x)|\, \lambda(\mathrm{d}x)}_{\text{torzítás}},$$

ahol $\mathbf{E}f_n(x)$ a minták szerinti várható értéket jelöli.

Variációs tag:

$$\int |f_n(x) - \mathbf{E}f_n(x)|\, \lambda(\mathrm{d}x) = \sum_j \int\limits_{A_{nj}} |f_n(x) - \mathbf{E}f_n(x)|\, \lambda(\mathrm{d}x) = \sum_j |\mu_n(A_{nj}) - \mu(A_{nj})|,$$

hiszen $f_n(x)$ konstans minden cellán.

Jelölje $M_n = |\{j :\ A_{nj}\cap S \neq \emptyset\}|$ és számozzuk át a cellákat úgy, hogy $A_{n1}, A_{n2}, \dots, A_{nM_n}$ legyen az az M_n cella, amelyre $A_{nj}\cap S\neq\emptyset$. Legyen $S_n = \bigcup\limits_{j=1}^{M_n} A_{nj}$.

$$\int |f_n(x) - \mathbf{E}f_n(x)| \le \sum |\mu_n(A_{nj} - \mu(A_{nj})| + \mu_n(S_n^c) + \mu(S_n^c) \le$$

$$\le \sum |\mu_n(A_{nj}) - \mu(A_{nj})| + |\mu_n(S_n^c) - \mu(S_n^c)| + 2\mu(S_n^c) \le$$

$$\le \sum |\mu_n(A_{nj}) - \mu(A_{nj})| + |\mu_n(S_n^c) - \mu(S_n^c)| + 2\mu(S^c)$$

Legyen $\mathcal{A}$ azon halmazok családja, amelyek az $A_{n1}, A_{n2}, \ldots, A_{nM_n}, S_n^c$ véges egyesítései. Ekkor a Scheffé-tétel miatt

$$\sum_{j=1}^{M_n} |\mu_n(A_{nj}) - \mu(A_{nj})| + |\mu_n(S_n^c) - \mu(S_n^c)| = 2 \sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|,$$

tehát $2\mu(S^c) < \varepsilon$ esetén a Vapnik–Chervonenkis-egyenlőtlenség miatt

$$\mathbf{P}\left(\int |f_n(x) - \mathbf{E}f_n(x)|\,\lambda(\mathrm{d}x) > \varepsilon\right) \le \mathbf{P}\left(2\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| + 2\mu(S^c) > \varepsilon\right) =$$

$$= \mathbf{P}\left(\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| > \frac{\varepsilon}{2} - \mu(S^c)\right) \le$$

$$\le 8s(\mathcal{A}, n)\mathrm{e}^{-n\left(\frac{\varepsilon}{2} - \mu(S^c)\right)^2/32} \le$$

$$\le 8 \cdot 2^{M_n+1}\mathrm{e}^{-n\left(\frac{\varepsilon}{2} - \mu(S^c)\right)^2/32}$$

A tétel második feltétele miatt

$$\frac{M_n}{n} \to 0,$$

tehát elegendően nagy n-re a jobb oldal kisebb, mint

$$\mathrm{e}^{-n\left(\frac{\varepsilon}{2} - \mu(S^c)\right)^2/64},$$

amely összegezhető, tehát a Borel–Cantelli-lemma miatt a variációs tag tart 0-hoz 1 valószínűséggel.

Torzítás:

$$\mathbf{E}f_n(x) = \frac{\mu(A_n(x))}{\lambda(A_n(x))} = \frac{1}{\lambda(A_n(x))} \int\limits_{A_n(x)} f(z)\,\lambda(\mathrm{d}z) = \int f(z)K_n(x,z)\,\lambda(\mathrm{d}z),$$

ahol $K_n(x,z) = \frac{I_{\{z\in A_n(x)\}}}{\lambda(A_n(x))}$.

Ha f folytonos, és egy kompakt halmazon kívül 0, akkor egyenletesen folytonos, ezért a tétel első feltétele miatt a torzítás 0-hoz tart. Legyen most f tetszőleges, ekkor $\varepsilon > 0$-hoz létezik olyan $\tilde{f}$, amely folytonos, egy kompakt halmazon kívül 0, és

$$\int |f(x) - \tilde{f}(x)|\,\lambda(\mathrm{d}x) < \varepsilon.$$

Ekkor

$$\int |f(x) - \mathbf{E}f_n(x)|\,\lambda(\mathrm{d}x) = \int \left| f(x) - \int f(z)K_n(x,z)\,\lambda(\mathrm{d}z)\right| \lambda(\mathrm{d}x) \le$$

$$\le \int |f(x) - \tilde{f}(x)|\,\lambda(\mathrm{d}x) + \int \left|\tilde{f}(x) - \int \tilde{f}(z)K_n(x,z)\,\lambda(\mathrm{d}z)\right| \lambda(\mathrm{d}x) +$$

$$+ \int \left| \int \tilde{f}(z)K_n(x,z)\,\lambda(\mathrm{d}z) - \int f(z)K_n(x,z)\,\lambda(\mathrm{d}z)\right| \lambda(\mathrm{d}x) \le$$

$$\le \varepsilon + \int \left|\tilde{f}(x) - \int \tilde{f}(z)K_n(x,z)\,\lambda(\mathrm{d}z)\right| \lambda(\mathrm{d}x) +$$

$$+\int\left(\int|\tilde f(z)-f(z)|K_n(x,z)\,\lambda(\mathrm{d}x)\right)\lambda(\mathrm{d}z)=$$

$$=\varepsilon+\int\left|\tilde f(x)-\int\tilde f(z)K_n(x,z)\,\lambda(\mathrm{d}z)\right|\lambda(\mathrm{d}x)+\int|\tilde f(z)-f(z)|\,\lambda(\mathrm{d}z)\quad\to\quad 2\varepsilon$$

Itt igazából elmondtuk a Banach–Steinhaus-tétel bizonyítását, amely szerint ha egy operátorsorozat pontonként konvergál egy sűrű halmazon, és az operátornormák sorozata korlátos, akkor minden pontban konvergál.

■

6.2.2. tétel: Ha f egy origó közepű S kockán kívül 0, Lipschitz-folytonos, azaz

$$|f(x)-f(z)|\le C||x-z||,$$

akkor h_n oldalhosszúságú d-dimenziós kockákból álló partíció esetén a hisztogramra

$$\mathbf{E}\int|f-f_n|\le\frac{c_1}{\sqrt{nh_n^d}}+c_2h_n,$$

tehát

$$h_n=c_3n^{-\frac{1}{d+2}}$$

választásra

$$\mathbf{E}\int|f-f_n|\le c_nn^{-\frac{1}{d+2}}.$$

Bizonyítás:

$$\mathbf{E}\int|f(x)-f_n(x)|\,\lambda(\mathrm{d}x)\le\underbrace{\int|f(x)-\mathbf{E}f_n(x)|\,\lambda(\mathrm{d}x)}_{\text{torzítás}}+\underbrace{\mathbf{E}\int|f_n(x)-\mathbf{E}f_n(x)|\,\lambda(\mathrm{d}x)}_{\text{variáció}}$$

Variáció:
Legyen S olyan, hogy $\mu(S^c)=0$. Jelölje M_n azon cellák számát a partícióban, amelyek metszik S-et, $M_n=|\{j:\ A_{nj}\cap S\ne\emptyset\}|\le\frac{\mathrm{Vol}(S)}{h_n^d}$. Akkor

$$\mathbf{E}\int|f_n(x)-\mathbf{E}f_n(x)|\ \lambda(\mathrm{d}x)\le$$

$$\le\sum_{j:\ A_{nj}\cap S\ne\emptyset}\mathbf{E}\int_{A_{nj}}|f_n(x)-\mathbf{E}f_n(x)|\ \lambda(\mathrm{d}x)+\int_{S^c}2\mathbf{E}f_n(x)\,\lambda(\mathrm{d}x)=$$

$$=\sum_{j:\ A_{nj}\cap S\ne\emptyset}\mathbf{E}\,|\mu_n(A_{nj})-\mu(A_{nj})|\le$$

$$\le\sum_{j:\ A_{nj}\cap S\ne\emptyset}\sqrt{\mathbf{E}\,|\mu_n(A_{nj})-\mu(A_{nj})|^2}=$$

$$=\sum_{j:\ A_{nj}\cap S\ne\emptyset}\sqrt{\frac{\mu(A_{nj})(1-\mu(A_{nj}))}{n}}\le$$

$$\leq \sqrt{\frac{\sum\limits_{j:\, A_{nj}\cap S\neq\emptyset} \mu(A_{nj})(1-\mu(A_{nj}))}{n}} \cdot M_n \leq$$

(a Cauchy–Schwarz-egyenlőtlenség miatt)

$$\leq \sqrt{\frac{\mathrm{Vol}(S)}{nh_n^d}}.$$

Torzítás:

$$\int |f - \mathbf{E}f_n| = \int \left| f(x) - \int f(z)K_n(x,z)\,\lambda(\mathrm{d}z) \right| \lambda(\mathrm{d}x \leq$$

$$\leq \int\int |f(x) - f(z)| K_n(x,z)\,\lambda(\mathrm{d}z)\,\lambda(\mathrm{d}x) \leq$$

$$\leq \int\int C||x-z||K_n(x,z)\,\lambda(\mathrm{d}z)\,\lambda(\mathrm{d}x) \leq$$

$$\leq \int\int Ch_n K_n(x,z)\,\lambda(\mathrm{d}z)\,\lambda(\mathrm{d}x) = Ch_n$$

■

A magfüggvényes becslőt a nemnegatív, mérhető $K(x)$ magfüggvény és a pozitív h_n sorozat határozza meg:

$$f_n(x) = \frac{1}{nh_n^d}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$$

6.2.3. tétel: Tegyük fel, hogy μ-nek létezik f sűrűségfüggvénye. Ha a magfüggvényes becslőnél

$$\int K(x)\,\lambda(\mathrm{d}x) = 1, \qquad \lim_{n\to\infty} h_n = 0 \qquad \text{és} \qquad \lim_{n\to\infty} nh_n^d = \infty,$$

akkor

$$\lim_{n\to\infty} \int |f(x) - f_n(x)|\,\lambda(\mathrm{d}x) = 0$$

1 valószínűséggel, vagyis a magfüggvényes becslő is erősen konzisztens.

Példák magfüggvényre:

- Naiv magfüggvény
$$K(x) = I_{\{x\in S_{0,r}\}}$$
ahol $S_{0,r}$ origó közepű r sugarú gömb.
- Gauss magfüggvény
$$K(x) = \mathrm{e}^{-||x||^2}$$
- Cauchy magfüggvény
$$K(x) = \frac{1}{1+||x||^{d+1}}$$
- Epanechnikov magfüggvény
$$K(x) = (1-||x||^2)I_{\{||x||\leq 1\}}$$

6.2.4. tétel: Ha f egy origó közepű S gömbön kívül 0, f differenciálható és a gradiens Lipschitz-folytonos, azaz

$$||f'(x) - f'(z)|| \leq C||x - z||,$$

akkor a magfüggvényes becslésre

$$\mathbf{E}\int |f - f_n| \leq \frac{c_1}{\sqrt{nh_n^d}} + c_2 h_n^2,$$

tehát

$$h_n = c_3 n^{-\frac{1}{d+4}}$$

választásra

$$\mathbf{E}\int |f - f_n| \leq c_4 n^{-\frac{2}{d+4}}.$$

A paraméteres statisztikában a konzisztencia tisztázása után az a legfontosabb kérdés, hogy adott pontossághoz mekkora mintanagyság kell, azaz mekkora az illető becslés konvergenciasebessége. Sűrűségfüggvény-becslés esetében, ha nem teszünk fel semmit az f sűrűségfüggvényről, akkor nem tudunk semmit mondani a konvergenciasebességről, sűrűségbecslők minden $\{f_n\}$ sorozatára igaz az, hogy a várható L_1-hiba konvergenciasebessége tetszőlegesen kicsi lehet.

6.2.5. tétel: Sűrűségbecslők minden $\{f_n\}$ sorozatához és pozitív számok minden monoton, 0-hoz tartó $a_n < \frac{1}{32}$ sorozatához létezik f sűrűségfüggvény úgy, hogy

$$\mathbf{E}||f - f_n|| > a_n$$

minden n-re.

7. fejezet

Regresszióbecslés

7.1. A regressziós probléma

Legyen Y valós értékű valószínűségi változó és legyen X d-dimenziós véletlen vektor (megfigyelés). X koordinátái különböző eloszlásúak lehetnek, lehet némelyik diszkrét (például bináris), mások lehetnek abszolút folytonosak. Így nem teszünk fel semmit X eloszlásáról. A regresszióanalízis célja Y becslése, ha X adott, azaz olyan f függvényt keresünk, amely X értékkészletén van definiálva, és amelyre $f(X)$ „közel" van Y-hoz. Tegyük fel, hogy az analízis fő célja a négyzetesközép-hiba minimalizálása:

$$\min_f \mathbf{E}((f(X)-Y)^2).$$

Jól ismert, hogy a minimumot az

$$m(x)=\mathbf{E}(Y\mid X=x)$$

regressziófüggvény éri el, ugyanis minden f mérhető függvényre

$$\mathbf{E}((f(X)-Y)^2)=\mathbf{E}((m(X)-Y)^2)+\mathbf{E}((m(X)-f(X))^2)=$$

$$=\mathbf{E}((m(X)-Y)^2)+\int |m(x)-f(x)|^2\,\mu(\mathrm{d}x),$$

ahol μ az X eloszlását jelöli. A jobb oldal második tagját a f függvény integrált négyzetes hibájának nevezik, és $J(f)$-fel jelölik

$$J(f)=\int |m(x)-f(x)|^2\,\mu(\mathrm{d}x).$$

A négyzetes közép hiba nyilván pontosan akkor lesz közel a minimumhoz, ha a $J(f)$ közel van a 0-hoz. A sűrűségbecsléssel szemben, ahol az L_1-hiba volt a legalkalmasabb hibakritérium, itt az L_2-hiba a legfontosabb. Ráadásul a sűrűségbecslésnél az L_1-teret a Lebesgue-mértékkel definiáltuk, míg a regresszióbecslésnél az L_2-teret μ-vel definiáljuk.

A regresszióbecslés feladatánál legyenek $(X_1,Y_1),\dots,(X_n,Y_n)$ független, azonos eloszlású példányai (X,Y)-nak. Az m_n regresszióbecslő x-nek és az (X_i,Y_i) mintáknak mérhető függvénye:

$$m_n=m_n(x,(X_1,Y_1),\dots,(X_n,Y_n)).$$

Az m_n regresszióbecslés m-hez való $L_2(\mu)$ konvergenciáját vizsgáljuk.

7.1.1. definíció: Az m_n becslő gyengén univerzálisan konzisztens, ha

$$J(m_n) \to 0 \quad \text{sztochasztikusan}$$

(X, Y) minden olyan eloszlására, amelyre $\mathbf{E}|Y|^2 < \infty$.

7.1.2. definíció: Az m_n becslő erősen univerzálisan konzisztens, ha

$$J(m_n) \to 0 \quad 1 \text{ valószínűséggel}$$

(X, Y) minden olyan eloszlására, amelyre $\mathbf{E}|Y|^2 < \infty$.

7.2. Lokális átlagoláson alapuló becslők

A lokális átlagoló regresszióbecslők az

$$m_n(x) = \sum_{i=1}^{n} W_{ni}(x, X_1, \ldots, X_n) Y_i$$

alakú becslők, ahol a W_{ni} súlyok jellegzetesen nemnegatívak, összegük 1, továbbá W_{ni} nagy, ha x közel van X_i-hez, különben kicsi. Ilyen típusú regresszióbecslő a hisztogram, a magfüggvényes és a legközelebbi szomszéd becslő.

Legyen $\mathcal{P}_n = \{A_{n1}, A_{n2}, \ldots\}$ az $\mathbb{R}^d$ egy partíciója, és minden $x \in \mathbb{R}^d$-re jelölje $A_n(x)$ az x-et tartalmazó cellát. Ekkor a hisztogrambecslő

$$m_n(x) = \begin{cases} \dfrac{\sum_{i=1}^{n} Y_i I_{\{X_i \in A_n(x)\}}}{\sum_{i=1}^{n} I_{\{X_i \in A_n(x)\}}}, & \text{ha } \sum_{i=1}^{n} I_{\{X_i \in A_n(x)\}} > 0 \\ 0 & \text{különben.} \end{cases}$$

A cellák gyakran h_n élhosszúságú d-dimenziós kockák.

7.2.1. tétel: Ha minden origó közepű S gömbre

$$\lim_{n\to\infty} \sup_{j:\, A_{nj} \cap S \neq \emptyset} \operatorname{diam}(A_{nj}) = 0$$

és

$$\lim_{n\to\infty} \frac{|\{j :\ A_{nj} \cap S \neq \emptyset\}|}{n} = 0,$$

akkor a hisztogram regresszióbecslő erősen konzisztens, ha $|Y| \leq L$ valamely $L < \infty$-re 1 valószínűséggel.

Megjegyzés: Ha a cellák h_n élhosszúságú kockák, akkor a tétel feltételei: $h_n \to 0$ és $nh_n^d \to \infty$.

Mielőtt rátérnénk a tétel bizonyítására, kimondjuk és bebizonyítjuk a Hoeffding-egyenlőtlenség egy, McDiarmidtól származó általánosítását, amelyre a 7.2.1 tétel bizonyításánál szükségünk lesz. Ehhez először vezessük be a martingál fogalmát.

7.2.1. definíció: Valószínűségi változók egy $Z_1, Z_2, \ldots$ sorozatát *martingál*nak nevezzük, ha

$$\mathbf{E}\{Z_{i+1}|Z_1, \ldots, Z_i\} = Z_i \quad 1 \text{ valószínűséggel}$$

minden $i > 0$-ra.

Legyen $X_1, X_2, \ldots$ valószínűségi változók egy tetszőleges sorozata. $Z_1, Z_2, \ldots$-t az $X_1, X_2, \ldots$ sorozat szerinti *martingál*nak nevezzük, ha minden $i > 0$-ra Z_i az $X_1, \ldots, X_i$ egy függvénye és

$$\mathbf{E}\{Z_{i+1}|X_1, \ldots, X_i\} = Z_i \quad 1 \text{ valószínűséggel.}$$

Nyilvánvaló, hogy ha $Z_1, Z_2, \ldots$ az $X_1, X_2, \ldots$ sorozat szerinti martingál, akkor $Z_1, Z_2, \ldots$ martingál, hiszen

$$\begin{aligned}
\mathbf{E}\{Z_{i+1}|Z_1, \ldots, Z_i\} &= \mathbf{E}\{\mathbf{E}\{Z_{i+1}|X_1, \ldots, X_i\}|Z_1, \ldots, Z_i\} \\
&= \mathbf{E}\{Z_i|Z_1, \ldots, Z_i\} \\
&= Z_i.
\end{aligned}$$

A legfontosabb példa martingálra a független, nulla várható értékű valószínűségi változók összege. Legyen $U_1, U_2, \ldots$ független valószínűségi változó nulla várható értékkel. Ekkor az

$$S_i = \sum_{j=1}^{i} U_j, \quad i > 0,$$

martingál.

7.2.2. definíció: Valószínűségi változók egy $V_1, V_2, \ldots$ sorozatát *martingál differencia sorozat*nak nevezzük, ha

$$\mathbf{E}\{V_{i+1}|V_1, \ldots, V_i\} = 0 \quad 1 \text{ valószínűséggel}$$

minden $i > 0$-ra.

$V_1, V_2, \ldots$-t az $X_1, X_2, \ldots$ sorozat szerinti *martingál differencia sorozat*nak nevezzük, ha minden $i > 0$-ra V_i az $X_1, \ldots, X_i$ egy függvénye és

$$\mathbf{E}\{V_{i+1}|X_1, \ldots, X_i\} = 0 \quad 1 \text{ valószínűséggel.}$$

Minden $Z_1, Z_2, \ldots$ martingál természetes módon vezet egy martingál differenciához:

$$V_i = Z_i - Z_{i-1}.$$

7.2.2. tétel: *(Azuma–Hoeffding)*
Legyen $X_1, X_2, \ldots$ valószínűségi változók egy sorozata és $V_1, V_2, \ldots$ az $X_1, X_2, \ldots$ sorozat szerinti martingál differencia sorozat. Tegyük fel, hogy létezik valószínűségi változók egy $Z_1, Z_2, \ldots$ sorozata és nemnegatív $c_1, c_2, \ldots$ konstansok úgy, hogy minden $i > 0$-ra Z_i az $X_1, \ldots, X_{i-1}$ egy függvénye és

$$Z_i \le V_i \le Z_i + c_i \quad 1 \text{ valószínűséggel.}$$

Ekkor minden $\varepsilon > 0$-ra és n-re

$$\mathbf{P}\left\{\sum_{i=1}^{n} V_i \ge \varepsilon\right\} \le \mathrm{e}^{-2\varepsilon^2 \sum\limits_{i=1}^{n} c_i^2}$$

és

$$\mathbf{P}\left\{\sum_{i=1}^{n} V_i \leq -\varepsilon\right\} \leq \mathrm{e}^{-2\varepsilon^2 \sum_{i=1}^{n} c_i^2}.$$

A bizonyítás a Hoeffding-egyenlőtlenség bizonyításának kiterjesztése. Szükségünk lesz az 5.1.1 lemma analógjára:

7.2.1. lemma: Tegyük fel, hogy a V és Z valószínűségi változókra 1 valószínűséggel igaz, hogy $\mathbf{E}\{V|Z\} = 0$ és, hogy valamely f függvényre és $c \geq 0$ konstansra

$$f(Z) \leq V \leq f(Z) + c.$$

Ekkor minden $s > 0$-ra

$$\mathbf{E}\left\{\mathrm{e}^{sV}|Z\right\} \leq \mathrm{e}^{s^2c^2/8}.$$

A 7.2.2 tétel bizonyítása:
A Hoeffding-egyenlőtlenség bizonyításához hasonlóan most is a Chernoff-technikát használjuk. Legyen $S_k = \sum_{i=1}^{k} V_i$. Ekkor minden $s > 0$-ra

$$\begin{aligned}
\mathbf{P}\{S_n \geq \varepsilon\} &\leq \mathrm{e}^{-s\varepsilon}\mathbf{E}\left\{\mathrm{e}^{sS_n}\right\} \\
&= \mathrm{e}^{-s\varepsilon}\mathbf{E}\left\{\mathrm{e}^{sS_{n-1}}\mathbf{E}\left\{\mathrm{e}^{sV_n}|X_1,\ldots,X_{n-1}\right\}\right\} \\
&\leq \mathrm{e}^{-s\varepsilon}\mathbf{E}\left\{\mathrm{e}^{sS_{n-1}}\right\}\mathrm{e}^{s^2c_n^2/8} \quad \text{(7.2.1 lemma miatt)} \\
&\leq \mathrm{e}^{-s\varepsilon}\mathrm{e}^{s^2\sum_{i=1}^{n} c_i^2/8} \quad \text{(ismételve az előző lépéseket)} \\
&= \mathrm{e}^{-2\varepsilon^2 \sum_{i=1}^{n} c_i^2} \quad \left(\text{ha } s = 4\varepsilon/\sum_{i=1}^{n} c_i^2\right).
\end{aligned}$$

A második egyenlőtlenség hasonlóan bizonyítható.

■

7.2.3. tétel: *(McDiarmid)*
Legyenek $X_1,\ldots,X_n$ független valószínűségi változók, amelyek értéküket egy A halmazból veszik, és tegyük fel, hogy $f : A^n \to \mathbb{R}$ függvényre teljesül, hogy

$$\sup_{\substack{x_1,\ldots,x_n,\\ x_i'\in A}} |f(x_1,\ldots,x_n) - f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \leq c_i, \quad 1 \leq i \leq n.$$

Ekkor minden $\varepsilon > 0$-ra

$$\mathbf{P}\left\{f(X_1,\ldots,X_n) - \mathbf{E}f(X_1,\ldots,X_n) \geq \varepsilon\right\} \leq \mathrm{e}^{-2\varepsilon^2/\sum_{i=1}^{n} c_i^2},$$

és

$$\mathbf{P}\left\{\mathbf{E}f(X_1,\ldots,X_n) - f(X_1,\ldots,X_n) \geq \varepsilon\right\} \leq \mathrm{e}^{-2\varepsilon^2/\sum_{i=1}^{n} c_i^2}.$$

Bizonyítás: Legyen $V = f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n)$, $V_1 = \mathbf{E}\{V|X_1\} - \mathbf{E}V$, és $k > 1$-re,

$$V_k = \mathbf{E}\{V|X_1 \ldots, X_k\} - \mathbf{E}\{V|X_1, \ldots, X_{k-1}\}.$$

Így $V = \sum_{k=1}^{n} V_k$. Világos, hogy $V_1, \ldots, V_n$ az $X_1, \ldots, X_n$ szerinti martingál differencia sorozatot alkot. Definiáljuk a következő valószínűségi változókat

$$H_k(X_1, \ldots, X_k) = \mathbf{E}\left\{f(X_1, \ldots, X_n)|X_1, \ldots, X_k\right\},$$

ekkor

$$V_k = H_k(X_1, \ldots, X_k) - \int H_k(X_1, \ldots, X_{k-1}, x)F_k(\mathrm{d}x),$$

ahol F_k az X_k eloszlásfüggvénye. Vezessük be a

$$W_k = \sup_u \left(H_k(X_1, \ldots, X_{k-1}, u) - \int H_k(X_1, \ldots, X_{k-1}, x)F_k(\mathrm{d}x) \right),$$

és a

$$Z_k = \inf_v \left(H_k(X_1, \ldots, X_{k-1}, v) - \int H_k(X_1, \ldots, X_{k-1}, x)F_k(\mathrm{d}x) \right).$$

valószínűségi változókat. Világos, hogy $Z_k \leq V_k \leq W_k$ 1 valószínűséggel. Mivel minden k-ra Z_k az $X_1, \ldots, X_{k-1}$ egy függvénye, alkalmazhatjuk a 7.2.1 lemmát $V = \sum_{k=1}^{n} V_k$-ra, ha meg tudjuk mutatni, hogy $W_k - Z_k \leq c_k$. De ez következik a tétel feltételei miatt

$$W_k - Z_k = \sup_u \sup_v \left(H_k(X_1, \ldots, X_{k-1}, u) - H_k(X_1, \ldots, X_{k-1}, v)\right) \leq c_k,$$

■

Megjegyzés: Ha az X_i-k korlátosak, akkor az $f(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i$ választással a Hoeffding-egyenlőtlenséghez jutunk.

A 7.2.1 tétel bizonyítása:
A tételt arra az esetre bizonyítjuk, amikor $m(x)$ folytonos.

A célunk azt bebizonyítani, hogy

$$\lim_{n\to\infty} \int (m_n(x) - m(x))^2 \, \mu(\mathrm{d}x) = 0$$

1 valószínűséggel.

Legyen

$$m_n^*(x) = \frac{\sum_{i=1}^{n} Y_i I_{\{X_i \in A_n(x)\}}}{n\mu(A_n(x))},$$

ekkor

$$\int (m_n(x) - m(x))^2 \, \mu(\mathrm{d}x) \leq$$

$$\leq 2 \left(\int (m_n(x) - m_n^*(x))^2 \, \mu(\mathrm{d}x) + \int (m_n^*(x) - m(x))^2 \, \mu(\mathrm{d}x) \right) =$$

$$= 2J_1 + 2J_2$$

J_2 1 valószínűségű konvergenciájához megmutatjuk, hogy

$$\mathbf{P}\left(\int |m_n^*(x) - m(x)|\, \mu(\mathrm{d}x) > \varepsilon\right) \le \mathrm{e}^{-n\varepsilon^2/(32L^2)},$$

amiből

$$J_2 = \int |m_n^*(x) - m(x)|^2\, \mu(\mathrm{d}x) \le 2L \int |m_n^*(x) - m(x)|\, \mu(\mathrm{d}x)$$

és a Borel–Cantelli-lemma miatt következik J_2 1 valószínűségű konvergenciája.

$$\int |m_n^*(x) - m(x)|\, \mu(\mathrm{d}x) = \mathbf{E}\int |m_n^*(x) - m(x)|\, \mu(\mathrm{d}x) +$$

$$+\left(\int |m_n^*(x) - m(x)|\, \mu(\mathrm{d}x) - \mathbf{E}\int |m_n^*(x) - m(x)|\, \mu(\mathrm{d}x)\right) \qquad (*)$$

A háromszög egyenlőtlenségből

$$\mathbf{E}\int |m_n^*(x) - m(x)|\, \mu(\mathrm{d}x) \le$$

$$\le \int |\mathbf{E}m_n^*(x) - m(x)|\, \mu(\mathrm{d}x) + \mathbf{E}\int |m_n^*(x) - \mathbf{E}m_n^*(x)|\, \mu(\mathrm{d}x)$$

ahol az első tag a torzítás, a második a variáció. A középértéktétel miatt

$$\mathbf{E}m_n^*(x) = \int \frac{I_{\{z\in A_n(x)\}}}{\mu(A_n(x))} m(z)\mu(\mathrm{d}z) = \frac{\int_{A_n(x)} m(z)\,\mu(\mathrm{d}z)}{\mu(A_n(x))} = m(a_n(x))$$

valamely $a_n(x) \in A_n(x)$-re.

Torzítás:

Mivel $m(x)$ folytonos, egyenletesen is folytonos az origó közepű S gömbön: $\forall\delta > 0$-hoz $\exists\delta' > 0$: $||x - z|| < \delta'$ esetén $|m(x) - m(z)| < \delta$.

$$\int |\mathbf{E}m_n^*(x) - m(x)|\, \mu(\mathrm{d}x) =$$

$$= \int_S |\mathbf{E}m_n^*(x) - m(x)|\, \mu(\mathrm{d}x) + \int_{S^c} |\mathbf{E}m_n^*(x) - m(x)|\, \mu(\mathrm{d}x) =$$

$$= \int_S |m(a_n(x)) - m(x)|\, \mu(\mathrm{d}x) + \int_{S^c} |\mathbf{E}m_n^*(x) - m(x)|\, \mu(\mathrm{d}x) \le$$

$$\le \delta + 2L\mu(S^c) \le \frac{\varepsilon}{4},$$

hiszen ha $|Y| \le L$ 1 valószínűséggel, akkor $|m(\cdot)| \le L$. Tehát a torzítás 0-hoz tart.

Variáció:

Jelölje M_n azon cellák számát a partícióban, amelyek metszik S-et, $M_n = |\{j :\ A_{nj} \cap S \neq \emptyset\}|$.

$$\mathbf{E}\int |m_n^*(x) - \mathbf{E}m_n^*(x)|\, \mu(\mathrm{d}x) =$$

$$= \mathbf{E}\int_S |m_n^*(x) - \mathbf{E}m_n^*(x)|\, \mu(\mathrm{d}x) + \mathbf{E}\int_{S^c} |m_n^*(x) - \mathbf{E}m_n^*(x)|\, \mu(\mathrm{d}x) \le$$

$$\leq \sum_{j:\, A_{nj}\cap S\neq\emptyset} \mathbf{E}\int_{A_{nj}} |m_n^*(x) - \mathbf{E}m_n^*(x)|\ \mu(\mathrm{d}x) + 2L\mu(S^c) \leq$$

$$\leq \sum_{j:\, A_{nj}\cap S\neq\emptyset} \int_{A_{nj}} \sqrt{\mathbf{E}\,|m_n^*(x) - \mathbf{E}m_n^*(x)|^2}\,\mu(\mathrm{d}x) + 2L\mu(S^c) \leq$$

$$\leq \sum_{j:\, A_{nj}\cap S\neq\emptyset} \int_{A_{nj}} \sqrt{\frac{nL^2\mu(A_{nj})}{(n\mu(A_{nj}))^2}}\,\mu(\mathrm{d}x) + 2L\mu(S^c) \leq$$

$$\leq \sum_{j:\, A_{nj}\cap S\neq\emptyset} L\sqrt{\frac{\mu(A_{nj})}{n}} + 2L\mu(S^c) \leq$$

$$\leq LM_n\sqrt{\frac{\frac{1}{M_n}\sum_{j:\, A_{nj}\cap S\neq\emptyset}\mu(A_{nj})}{n}} + 2L\mu(S^c) \leq$$

(Jensen-egyenlőtlenség)

$$\leq L\sqrt{\frac{M_n}{n}} + 2L\mu(S^c) \leq \frac{\varepsilon}{4}$$

hiszen a tétel második feltétele szerint $\frac{M_n}{n} \to 0$, és $\mu(S^c)$ tetszőlegesen kicsivé tehető.

Így tehát elég nagy n-re

$$\mathbf{E}\int |m_n^*(x) - m(x)|\ \mu(\mathrm{d}x) \leq \frac{\varepsilon}{2},$$

tehát (*) miatt

$$\mathbf{P}\left(\int |m_n^*(x) - m(x)|\ \mu(\mathrm{d}x) > \varepsilon\right) \leq$$

$$\leq \mathbf{P}\left(\int |m_n^*(x) - m(x)|\ \mu(\mathrm{d}x) - \mathbf{E}\int |m_n^*(x) - m(x)|\ \mu(\mathrm{d}x) > \frac{\varepsilon}{2}\right)$$

A jobb oldalon álló valószínűségre a McDiarmid-egyenlőtlenséggel kaphatunk exponenciális felső korlátot.

Rögzítsük le az $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times [-L, L]$ mintáinkat és cseréljük ki (x_i, y_i)-t (x_i', y_i')-re. Jelöljük m_{ni}^*-vel az így kapott becslőt. Ekkor $m_n^*(x)$ és $m_{ni}^*(x)$ maximum két cellán, $A_n(x_i)$-n és $A_n(x_i')$-n, különbözik, így

$$\int |m_n^*(x) - m(x)|\ \mu(\mathrm{d}x) - \int |m_{ni}^*(x) - m(x)|\ \mu(\mathrm{d}x) \leq$$

$$\leq \int |m_n^*(x) - m_{ni}^*(x)|\ \mu(\mathrm{d}x) \leq$$

$$\leq \frac{2L}{n\mu(A_n(x_i))}\mu(A_n(x_i)) + \frac{2L}{n\mu(A_n(x_i'))}\mu(A_n(x_i')) \leq \frac{4L}{n}$$

Tehát a McDiarmid-egyenlőtlenség feltétele $c_i = \frac{4L}{n}$-nel teljesül, így

$$\mathbf{P}\left(\int |m_n^*(x) - m(x)|\ \mu(\mathrm{d}x) - \mathbf{E}\int |m_n^*(x) - m(x)|\ \mu(\mathrm{d}x) > \frac{\varepsilon}{2}\right) \leq \mathrm{e}^{-n\varepsilon^2/(32L^2)}$$

Tehát elég nagy n -re

$$\mathbf{P}\left(\int |m_n^*(x) - m(x)|\, \mu(\mathrm{d}x) > \varepsilon\right) \leq \mathrm{e}^{-n\varepsilon^2/(32L^2)},$$

amiből következik, hogy $J_2 \to 0$ 1 valószínűséggel.

J_1 1 valószínűségű konvergenciájának belátásához vegyük észre, hogy ha $\mu_n(A_n(x)) \neq 0$, akkor

$$|m_n^*(x) - m(x)| = \left| \frac{\sum_{i=1}^n Y_i I_{\{X_i \in A_n(x)\}}}{n\mu(A_n(x))} - \frac{\sum_{i=1}^n Y_i I_{\{X_i \in A_n(x)\}}}{\sum_{i=1}^n I_{\{X_i \in A_n(x)\}}} \right| \leq$$

$$\leq L \sum_{i=1}^n I_{\{X_i \in A_n(x)\}} \cdot \left| \frac{1}{n\mu(A_n(x))} - \frac{1}{\sum_{i=1}^n I_{\{X_i \in A_n(x)\}}} \right| =$$

$$= L \left| \frac{\sum_{i=1}^n I_{\{X_i \in A_n(x)\}}}{n\mu(A_n(x))} - 1 \right| = L\,|M_n^*(x) - 1|\,,$$

ahol $M_n^*(x)$ az $m_n^*(x)$ speciális alakja, ha $Y \equiv 1$.

Ha $\mu_n(A_n(x)) = 0$, akkor

$$|m_n^*(x) - m_n(x)| = 0 \leq L\,|M_n^*(x) - 1|\,.$$

Így tehát

$$J_1 = \int (m_n(x) - m_n^*(x))^2\, \mu(\mathrm{d}x) \leq L^2 \int (M_n^*(x) - 1)^2\, \mu(\mathrm{d}x) \to 0$$

1 valószínűséggel J_2 1 valószínűségű konvergenciája miatt.

Tehát a tételt bebizonyítottuk.

■

A magfüggvényes regresszióbecslőt, a sűrűségbecslőhöz hasonlóan, az abszolút integrálható $K(x)$ magfüggvény és a pozitív h_n simítótényező határozza meg

$$m_n(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x - X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right)}$$

7.2.3. definíció: A $K(x)$ magfüggvény reguláris, ha nemnegatív és létezik $S_{0,r}$ origó középpontú, $r > 0$ sugarú gömb és $b > 0$ konstans úgy, hogy

$$K(x) \geq b I_{\{x \in S_{0,r}\}}$$

és

$$\int \sup_{u \in x + S_{0,r}} K(u)\, \mathrm{d}x < \infty.$$

7.2.4. tétel: Ha a $K(x)$ magfüggvény reguláris, $h_n \to 0$, $nh_n^d \to \infty$ és $|Y| \le L$ valamely $L < \infty$-re 1 valószínűséggel, akkor a magfüggvényes regresszióbecslő erősen konzisztens.

A k_n-legközelebbi szomszéd becslő az x-hez legközelebbi k_n darab X_i mintához tartozó Y_i-ket átlagolja. Legyen $(X_{(1,n)}(x), Y_{(1,n)}(x)), \dots, (X_{(n,n)}(x), Y_{(n,n)}(x))$ az X_i-k x-től vett távolsága szerint növekedően rendezett minta. $X_{(i,n)}(x)$ az x i-edik legközelebbi szomszédja. Ha $||X_i - x|| = ||X_j - x||$, akkor X_i közelebbi, ha $i < j$. Ekkor

$$m_n(x) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i,n)}(x)$$

7.2.5. tétel: Ha az $||X - x||$ valószínűségi változó abszolút folytonos minden x-re, $k_n \to \infty$, $k_n/n \to 0$ és $|Y| \le L$ valamely $L < \infty$-re 1 valószínűséggel, akkor a k_n-legközelebbi szomszéd regresszióbecslő erősen konzisztens.

Tehát léteznek univerzálisan konzisztens regresszióbecslők, de a konvergenciasebesség, a sűrűségfüggvény-becsléshez hasonlóan, itt is tetszőlegesen kicsi lehet.

7.2.6. tétel: Regresszióbecslők minden $\{m_n\}$ sorozatához és pozitív számok minden monoton 0-hoz tartó $a_n < 1/64$ sorozatához létezik (X, Y)-nak olyan eloszlása, amelyre X egyenletes eloszlású $[0, 1]$-en, $Y = m(X)$ és

$$\mathbf{E}J(m_n) = \mathbf{E} \int (m_n(x) - m(x))^2 \, \mu(\mathrm{d}x) > a_n$$

minden n-re.

7.3. Empirikus hibaminimalizálás

Az eddig ismertetett regresszióbecslési módszerek a lokális átlagolás elvén alapulnak. Létezik egy másik, hasonlóan természetes alapelv, az empirikus hibaminimalizálás, amely szintén elvezethet univerzálisan konzisztens becslésekhez.

Választunk egy $\mathcal{F}_n$ függvénycsaládot, és a regresszióbecslés ebből a családból vehet függvényeket. Az $\mathcal{F}_n$ kiválasztásakor vagy az $m(x)$ regressziófüggvényről szerzett ismereteink játszhatnak szerepet, vagy $\mathcal{F}_n$ olyan függvényekből áll, amelyek számítógéppel bizonyos számítási bonyolultsággal realizálhatók.

Korábban már láttuk, hogy az $m(x)$ regressziófüggvény minimalizálja az L_2-hibát. Tehát mondhatnánk azt, hogy minimalizáljuk $\mathbf{E}(f(X) - Y)^2$-t az $\mathcal{F}_n$ családon. Ez azonban nyilvánvalóan lehetetlen, mert a minimalizálandó függvény függ az (X, Y) ismeretlen eloszlásától.

7.3.1. definíció: Az empirikus L_2-hiba a mintákon elkövetett hibák négyzeteinek átlaga:

$$\frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$

Az empirikus hibaminimalizálás során azt a függvényt választjuk ki $\mathcal{F}_n$-ből, amelynek az empirikus hibája minimális:

$$m_n = \operatorname*{argmin}_{f \in \mathcal{F}_n} \left(\frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2 \right)$$

A kérdés, hogy mekkora az így választott $m_n(x)$ becslő L_2 hibája.

$$\int |m_n(x)-m(x)|^2\ \mu(\mathrm{d}x) =$$

$$= \mathbf{E}\left(|m_n(X)-Y|^2 \,|\, D_n\right) - \mathbf{E}\left(|m(X)-Y|^2\right) =$$

$$= \left(\mathbf{E}\left(|m_n(X)-Y|^2 \,|\, D_n\right) - \inf_{f\in\mathcal{F}_n}\mathbf{E}\left(|f(X)-Y|^2\right)\right) +$$

$$+ \left(\inf_{f\in\mathcal{F}_n}\mathbf{E}\left(|f(X)-Y|^2\right) - \mathbf{E}\left(|m(X)-Y|^2\right)\right)$$

A jobb oldalon szereplő első tag a becslési hiba, a második tag pedig az approximációs hiba. A becslési hiba azt méri, hogy a becslő L_2-hibája mennyire tér el a függvénycsaládbeli legjobb függvény L_2-hibájától, az approximációs hiba pedig azt, hogy mennyire jól lehet a regressziófüggvényt $\mathcal{F}_n$-beli függvényekkel közelíteni L_2 értelemben. Ha az $\mathcal{F}_n$ család nagy, akkor az approximációs hiba ugyan lehet nagyon közel 0-hoz, de lehet, hogy nincs elég mintánk ahhoz, hogy $\mathcal{F}_n$-ből jó becslőt válasszunk, azaz a becslési hiba nagy lehet. Ha $\mathcal{F}_n$ kicsi, akkor pedig az approximációs hiba lehet nagyon nagy. Ahhoz, hogy univerzálisan konzisztens becslőt kapjunk, azt kell megmutatni, hogy mindkét tag 0-hoz tart.

Az approximációs hibára ez gyakran elég egyszerű. Először is könnyen látható, hogy

$$\inf_{f\in\mathcal{F}_n}\mathbf{E}\,|f(X)-Y|^2 - \mathbf{E}\,|m(X)-Y|^2 = \inf_{f\in\mathcal{F}_n}\int |f(x)-m(x)|\ \mu(\mathrm{d}x).$$

Ha például $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ minden n-re, akkor az, hogy minden μ mértékre és $m \in L_2$-re

$$\lim_{n\to\infty}\inf_{f\in\mathcal{F}_n}\int |f(x)-m(x)|^2\ \mu(\mathrm{d}x) = 0$$

egyszerűen azt jelenti, hogy $\bigcup\limits_{n=1}^{\infty}\mathcal{F}_n$ sűrű $L_2(\mu)$-ben minden μ mértékre. Ez igaz például, ha $\bigcup\limits_{n=1}^{\infty}\mathcal{F}_n$ sűrű $C_0^{\infty}(\mathbb{R}^d)$-ben a szuprémum norma szerint, mivel $C_0^{\infty}(\mathbb{R}^d)$ sűrű $L_2(\mu)$-ben minden μ eloszlásra és

$$\int |f(x)-m(x)|^2\ \mu(\mathrm{d}x) \le ||f-m||_{\infty}^2$$

Most nézzük tehát a becslési hibát.

7.3.1. lemma:

$$\mathrm{E}\left(|m_n(X)-Y|^2 \,|\, D_n\right) - \inf_{f\in\mathcal{F}_n}\mathbf{E}\,|f(X)-Y|^2 \le$$

$$\le 2\sup_{f\in\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2 - \mathbf{E}\,|f(X)-Y|^2\right|$$

Bizonyítás:

$$\mathrm{E}\left(|m_n(X)-Y|^2 \,|\, D_n\right) - \inf_{f\in\mathcal{F}_n}\mathbf{E}\,|f(X)-Y|^2 =$$

$$= \sup_{f\in\mathcal{F}_n}\left(\mathbf{E}\left((m_n(X)-Y)^2 \,|\, D_n\right) - \frac{1}{n}\sum_{i=1}^{n}|m_n(X_i)-Y_i|^2 + \right.$$

$$+\frac{1}{n}\sum_{i=1}^{n}|m_n(X_i)-Y_i|^2-\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2+$$

$$+\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2-\mathbf{E}\,|f(X)-Y|^2\Bigg)\leq$$

$$\leq\sup_{f\in\mathcal{F}_n}\left(\mathbf{E}\left((m_n(X)-Y)^2\,|\,D_n\right)-\frac{1}{n}\sum_{i=1}^{n}|m_n(X_i)-Y_i|^2+\right.$$

$$\left.+\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2-\mathbf{E}\,|f(X)-Y|^2\right)\leq$$

$$\leq 2\sup_{f\in\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2-\mathbf{E}\,|f(X)-Y|^2\right|$$

Az első egyenlőtlenség m_n választásából adódik, m_n minimalizálja az empirikus L_2-hibát $\mathcal{F}_n$-ben, így $\forall f\in\mathcal{F}_n$-re

$$\frac{1}{n}\sum_{i=1}^{n}|m_n(X_i)-Y_i|^2-\frac{1}{n}\sum_{i=1}^{n}|f(X_i)-Y_i|^2\leq 0$$

■

Tehát ahhoz, hogy megmutassuk, hogy a becslési hiba 0-hoz tart, a lemma jobb oldalán álló kifejezést kell vizsgálnunk.

Legyen $Z=(X,Y)$, $Z_i=(X_i,Y_i)$, $i=1,\dots,n$, $g_f(x,y)=|f(x)-y|^2$ minden $f\in\mathcal{F}_n$-re és $\mathcal{G}_n=\{g_f:\ f\in\mathcal{F}_n\}$. Ekkor a fenti kifejezés a következő alakban írható

$$\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n}g(Z_i)-\mathbf{E}g(Z)\right|.$$

Tehát egy átlag és a várható értéke közötti különbséget akarjuk felülről becsülni egyenletesen egy függvénycsalád felett.

Ha g korlátos, azaz $g:\ \mathbb{R}^d\times\mathbb{R}\to[0,M]$, akkor a Hoeffding-egyenlőtlenségből kapjuk, hogy

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}g(Z_i)-\mathbf{E}g(Z)\right|>\varepsilon\right)\leq 2\mathrm{e}^{-2n\varepsilon^2/M^2}$$

7.3.2. lemma:

$$\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n}g(Z_i)-\mathbf{E}g(Z)\right|\leq M\sup_{\substack{g\in\mathcal{G}_n\\ t>0}}\left|\frac{1}{n}\sum_{i=1}^{n}I_{\{g(Z_i)>t\}}-\mathbf{P}\left(g(Z)>t\right)\right|$$

Bizonyítás: Használjuk a nemnegatív valószínűségi változókra érvényes

$$\int_0^\infty\mathbf{P}(X>t)\,\mathrm{d}t=\mathbf{E}X$$

azonosságot.

$$\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n}g(Z_i)-\mathbf{E}g(Z)\right|=$$

$$= \sup_{g\in\mathcal{G}_n} \left| \int_0^\infty \left(\frac{1}{n}\sum_{i=1}^n I_{\{g(Z_i)>t\}} - \mathbf{P}\left(g(Z)>t\right) \right) \mathrm{d}t \right| \le$$

$$\le M \sup_{\substack{g\in\mathcal{G}_n\\ t>0}} \left| \frac{1}{n}\sum_{i=1}^n I_{\{g(Z_i)>t\}} - \mathbf{P}\left(g(Z)>t\right) \right|$$

■

Legyen

$$\hat{\mathcal{G}}_n = \{\{z:\ g(z)>t\}:\ g\in\mathcal{G}_n,\ t\in[0,M]\}\,.$$

Bebizonyítható a Vapnik–Chervonenkis-egyenlőtlenség általánosítása, amiből következik, hogy

$$\mathbf{P}\left(\sup_{g\in\mathcal{G}_n} \left| \frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbf{E}g(Z) \right| > \varepsilon \right) \le 8s(\hat{\mathcal{G}}_n, n)\mathrm{e}^{-n\varepsilon/(32M^2)}.$$

Összefoglalva az eddigieket, a következőt kapjuk:

7.3.1. tétel: Tegyük fel, hogy $|Y|\le L$ valamely $L<\infty$-re 1 valószínűséggel. Legyen $\mathcal{F}_n$ olyan f függvények családja, amelyekre $|f(x)|\le\beta_n$ minden x-re. Ekkor elég nagy n-re

$$\mathbf{P}\left(\mathbf{E}\left(|m_n(X)-Y|^2 \mid D_n \right) - \inf_{f\in\mathcal{F}_n} \mathbf{E}\left(f(X)-Y\right)^2 > \varepsilon \right) \le 8n^{V_{\hat{\mathcal{G}}_n}} \mathrm{e}^{-n\varepsilon^2/128(4\beta_n^2)^2}.$$

Ahol a felső korlát exponenciálisan tart 0-hoz, ha $\hat{\mathcal{G}}_n$ Vapnik–Chervonenkis dimenziója véges és $\frac{n}{\beta_n^4}\to\infty$. Ebben az esetben tehát a becslési hiba 1 valószínűséggel 0-hoz tart. Ahhoz, hogy az approximációs hiba 0-hoz tartson viszont kell az, hogy $\beta_n\to\infty$.

Legyen például $\mathcal{F}_n$ a

$$\sum_{j=1}^{K_n} a_j\Psi_j(x):\ a_1,\dots,a_{K_n}\in\mathbb{R}$$

alakú lineáris kombinációk családja, ahol a Ψ_j-k $\mathbb{R}^d$-ből $\mathbb{R}$-be képező korlátos függvények. Ha ezen a családon minimalizáljuk az empirikus négyzetes hibát, akkor konzisztens becslőt kapunk, ha $K_n\to\infty$, $\beta_n\to\infty$ és $\frac{\beta_n^4 K_n}{n}\to 0$. Ha még $\frac{\beta_n^4}{n^{1-\delta}}\to 0$ is teljesül valamilyen $\delta>0$-ra, akkor a becslő erősen univerzálisan konzisztens.

A korábban látott hisztogram becslő is egy emipirikus hibaminimalizáló becslő. Legyen $\mathcal{P}_n=\{A_{n1},A_{n2},\dots\}$ az $\mathbb{R}^d$ egy partíciója, és legyen $\mathcal{F}_n$ azon függvények családja, amelyek minden cellán konstansok. Ekkor a legkisebb empirikus négyzetes hibájú becslőt úgy kapjuk, ha cellánként minimalizálunk. Egy cellán pedig a minimumot az odaeső Y_i-k átlaga adja, ami nem más, mint a hisztogrambecslő értéke az adott cellán.

8. fejezet

Alakfelismerés

8.1. A Bayes-döntés és közelítése

Az alakfelismerésben Y két értéket vehet fel, 0-t vagy 1-et (például, hogy egy páciens szenved-e egy adott betegségben vagy nem). Az Y címke értékére szeretnénk következtetni adott $X \in \mathbb{R}^d$ megfigyelésvektor alapján (ami tartalmazhatja pl. a páciens hőmérsékletét, vérnyomását stb.). A döntés vagy osztályozási szabály egy

$$g: \mathbb{R}^d \to \{0,1\}$$

függvény, amelynek a minőségét az

$$L(g) = \mathbf{P}\left(g(X) \neq Y\right)$$

hibavalószínűség méri. A cél $L(g)$ minimalizálása.

8.1.1. definíció: Bayes-döntés:

$$g^*(x) = \begin{cases} 1, & \text{ha } \mathbf{P}\left(Y = 1 \,|\, X = x\right) \geq \frac{1}{2} \\ 0 & \text{különben.} \end{cases}$$

$L^* = L(g^*)$ az ún. Bayes-hiba.

A Bayes-döntés optimális.

8.1.1. tétel: Minden $g: \mathbb{R}^d \to \{0,1\}$ döntésfüggvényre

$$\mathbf{P}\left(g^*(X) \neq Y\right) \leq \mathbf{P}\left(g(X) \neq Y\right).$$

Bizonyítás: Igazak az alábbi egyszerű átalakítások:

$$\mathbf{P}\left(g(X) \neq Y \,|\, X = x\right) = 1 - \mathbf{P}\left(Y = g(X) \,|\, X = x\right) =$$

$$= 1 - \left(\mathbf{P}\left(Y = 1, g(X) = 1 \,|\, X = x\right) + \mathbf{P}\left(Y = 0, g(X) = 0 \,|\, X = x\right)\right) =$$

$$= 1 - \left(I_{\{g(x)=1\}}\mathbf{P}\left(Y = 1 \,|\, X = x\right) + I_{\{g(x)=0\}}\mathbf{P}\left(Y = 0 \,|\, X = x\right)\right)$$

Így tehát minden $x \in \mathbb{R}^d$-re

$$\mathbf{P}\left(g(X) \neq Y \,|\, X = x\right) - \mathbf{P}\left(g^*(X) \neq Y \,|\, X = x\right) =$$

$$= \mathbf{P}\left(Y = 1 \,|\, X = x\right)\left(I_{\{g^*(x)=1\}} - I_{\{g(x)=1\}}\right) +$$

$$+\mathbf{P}\left(Y = 0 \,|\, X = x\right)\left(I_{\{g^*(x)=0\}} - I_{\{g(x)=0\}}\right) =$$

$$= \left(2\mathbf{P}\left(Y = 1 \,|\, X = x\right) - 1\right)\left(I_{\{g^*(x)=1\}} - I_{\{g(x)=1\}}\right) \geq 0$$

g^* definíciója alapján. Mindkét oldalt μ szerint integrálva kapjuk a tétel állítását.

■

A $\mathbf{P}(Y=1\,|\,X=x)$ és $\mathbf{P}(Y=0\,|\,X=x)$ valószínűségek az ún. a posteriori valószínűségek. Vegyük észre, hogy

$$\mathbf{P}(Y=1\,|\,X=x)=\mathbf{E}(Y\,|\,X=x)=m(x).$$

Tehát a Bayes-döntés a következőképpen írható

$$g^*(x)=\begin{cases}1, & \text{ha } m(x)\geq\frac{1}{2}\\ 0 & \text{különben.}\end{cases}$$

A Bayes-döntéshez ismernünk kellene a regressziófüggvényt, ami tipikusan ismeretlen, ezért most is a $D_n=\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ független, azonos eloszlású mintákat használhatjuk. Olyan $g_n(x)=g_n\left((X_1,Y_1),\ldots,(X_n,Y_n),x\right)$ osztályozási szabályt szeretnénk találni, amelynek az

$$L(g_n)=\mathbf{P}\left(g_n(X)\neq Y\,|\,(X_1,Y_1),\ldots,(X_n,Y_n)\right)$$

hibavalószínűsége közel van L^*-hoz.

8.1.2. definíció: A g_n osztályozási szabály gyengén univerzálisan konzisztens, ha (X,Y) minden eloszlására

$$\mathbf{E}L(g_n)=\mathbf{P}(g_n(X)\neq Y)\to L^*$$

ha $n\to\infty$, és erősen univerzálisan konzisztens, ha

$$\lim_{n\to\infty}L(g_n)=L^*$$

1 valószínűséggel.

Természetes gondolat, hogy a minták segítségével becsüljük az m regressziófüggvényt az m_n regresszióbecslővel, és vegyük a Bayes-döntés mintájára az ún. „plug-in” döntésfüggvényt.

$$g_n(x)=\begin{cases}1, & \text{ha } m_n(x)\geq\frac{1}{2}\\ 0 & \text{különben.}\end{cases}\tag{*}$$

A következő tétel azt állítja, hogy ha az m_n közel van a valódi m regressziófüggvényhez, akkor a g_n hibavalószínűsége közel lesz az optimális g^* hibavalószínűséghez.

8.1.2. tétel: A fent definiált g_n döntésfüggvényre

$$0\leq\mathbf{P}\left(g_n(X)\neq Y\,|\,D_n\right)-\mathbf{P}\left(g^*(X)\neq Y\right)\leq 2\int\left|m_n(x)-m(x)\right|\,\mu(\mathrm{d}x)\leq$$

$$\leq 2\left(\int\left|m_n(x)-m(x)\right|^2\,\mu(\mathrm{d}x)\right)^{\frac{1}{2}}$$

Bizonyítás: Tetszőleges $x\in\mathbb{R}^d$-re

$$\mathbf{P}\left(g_n(X)\neq Y\,|\,D_n,X=x\right)-\mathbf{P}\left(g^*(X)\neq Y\,|\,X=x\right)=$$

$$=I_{\{g^*(x)=1\}}m(x)+I_{\{g^*(x)=0\}}(1-m(x))-\big(I_{\{g_n(x)=1\}}m(x)+I_{\{g_n(x)=0\}}(1-m(x))\big)=$$

$$=I_{\{g^*(x)=1\}}m(x)+I_{\{g^*(x)=0\}}(1-m(x))-\big(I_{\{g^*(x)=1\}}m_n(x)+I_{\{g^*(x)=0\}}(1-m_n(x))\big)+$$

$$+\left(I_{\{g^*(x)=1\}}m_n(x)+I_{\{g^*(x)=0\}}(1-m_n(x))\right)-\left(I_{\{g_n(x)=1\}}m_n(x)+I_{\{g_n(x)=0\}}(1-m_n(x))\right)+$$
$$+\left(I_{\{g_n(x)=1\}}m_n(x)+I_{\{g_n(x)=0\}}(1-m_n(x))\right)-\left(I_{\{g_n(x)=1\}}m(x)+I_{\{g_n(x)=0\}}(1-m(x))\right)\le$$
$$\le I_{\{g^*(x)=1\}}(m(x)-m_n(x))+I_{\{g^*(x)=0\}}(m_n(x)-m(x))+$$
$$+I_{\{g_n(x)=1\}}(m_n(x)-m(x))+I_{\{g_n(x)=0\}}(m(x)-m_n(x))\le$$
$$\le 2\left|m_n(x)-m(x)\right|,$$

ahol az utolsó előtti egyenlőtlenségnél azt használtuk, hogy g_n definíciója miatt

$$I_{\{g_n(x)=1\}}m_n(x)+I_{\{g_n(x)=0\}}(1-m_n(x))=\max\{m_n(x),1-m_n(x)\}$$

amiből

$$I_{\{g^*(x)=1\}}m_n(x)+I_{\{g^*(x)=0\}}(1-m_n(x))-I_{\{g_n(x)=1\}}m_n(x)+I_{\{g_n(x)=0\}}\le 0.$$

Így

$$0\le \mathbf{P}\left(g_n(X)\ne Y\mid D_n\right)-\mathbf{P}\left(g^*(X)\ne Y\right)=$$
$$=\int\left(\mathbf{P}\left(g_n(X)\ne Y\mid D_n,X=x\right)-\mathbf{P}\left(g^*(X)\ne Y\mid X=x\right)\right)\mu(\mathrm{d}x)\le$$
$$\le 2\int\left|m_n(x)-m(x)\right|\mu(\mathrm{d}x)\le 2\left(\int\left|m_n(x)-m(x)\right|^2\mu(\mathrm{d}x)\right)^{\frac{1}{2}}$$

a Cauchy–Schwartz-egyenlőtlenség miatt.

■

Így egy L_2-ben konzisztens regresszióbecslőből automatikusan kaphatunk konzisztens döntésfüggvényt. De ahhoz, hogy (*) a Bayes-döntést jól közelítse egyáltalán nem fontos, hogy $m_n(x)$ közel legyen $m(x)$-hez. Csak az a lényeges, hogy a döntési határ ugyanazon oldalán legyenek, azaz hogy $m_n(x)\ge\frac{1}{2}$ legyen, ha $m(x)\ge\frac{1}{2}$ és legyen $<\frac{1}{2}$, ha $m(x)<\frac{1}{2}$. Mégis gyakran használják a plug-in döntéseket az $\mathbf{E}\left(|m_n(X)-Y|^2\mid D_n\right)$ L_2-hiba minimalizálásával kapott regresszióbecslővel, mert az L_2-hiba minimalizálása hatékonyan számítható becslőkhöz vezet.

8.2. Lokális többségen alapuló döntések

A regresszióbecsléshez hasonlóan itt is definiálhatjuk a három lokális átlagoláson alapuló osztályozási szabályt a hisztogram, a magfüggvényes és a legközelebbi szomszéd döntést.

Legyen $\mathcal{P}_n=\{A_{n1},A_{n2},\ldots\}$ az $\mathbb{R}^d$ egy partíciója, és minden $x\in\mathbb{R}^d$-re jelölje $A_n(x)$ az x-et tartalmazó cellát. Ekkor a hisztogram szabály

$$g_n(x)=\begin{cases}1, & \text{ha } \sum_{i=1}^n I_{\{Y_i=1\}}I_{\{X_i\in A_n(x)\}}\ge\sum_{i=1}^n I_{\{Y_i=0\}}I_{\{X_i\in A_n(x)\}}\\ 0 & \text{különben.}\end{cases}$$

Azaz többségi döntést hoz az $A_n(x)$-be eső minták címkéi alapján. Látható, hogy ha $m_n(x)$ a hisztogram regresszióbecslő, azaz

$$m_n(x)=\begin{cases}\dfrac{\sum_{i=1}^n Y_iI_{\{X_i\in A_n(x)\}}}{\sum_{i=1}^n I_{\{X_i\in A_n(x)\}}}, & \text{ha } \sum_{i=1}^n I_{\{X_i\in A_n(x)\}}>0\\ 0 & \text{különben,}\end{cases}$$

akkor a hisztogram szabály nem más, mint egy plug-in osztályozási szabály, amelyben az $m(x)$ regressziófüggvényt az $m_n(x)$ becslővel becsüljük. Tehát a hisztogram szabály konzisztenciája a korábbiak miatt következik a hisztogram regresszióbecslő konzisztenciájából.

8.2.1. tétel: Ha minden origó közepű S gömbre

$$\lim_{n\to\infty} \sup_{j:\, A_{nj}\cap S\neq\emptyset} \operatorname{diam}(A_{nj}) = 0$$

és

$$\lim_{n\to\infty} \frac{|\{j:\ A_{nj}\cap S\neq\emptyset\}|}{n} = 0$$

akkor a hisztogram osztályozási szabály erősen univerzálisan konzisztens.

A magfüggvényes osztályozási szabályt a

$$g_n(x) = \begin{cases} 1, & \text{ha } \sum_{i=1}^{n} I_{\{Y_i=1\}} K\left(\frac{x-X_i}{h_n}\right) \geq \sum_{i=1}^{n} I_{\{Y_i=0\}} K\left(\frac{x-X_i}{h_n}\right) \\ 0 & \text{különben.} \end{cases}$$

függvény adja meg, ahol a $K(x) : \mathbb{R}^d \to \mathbb{R}$ egy nemnegatív, integrálható magfüggvény és h_n pedig egy n-től függő simító tényező.

Ez a szabály is egy plug-in szabály, ahol az $m_n(x)$ regresszióbecslő most a magfüggvényes becslő, tehát a konzisztencia itt is a regresszióbecslő konzisztenciájából következik.

8.2.2. tétel: Ha a $K(x)$ magfüggvény reguláris, $h_n \to 0$ és $nh_n^d \to \infty$, akkor a magfüggvényes osztályozási szabály erősen univerzálisan konzisztens.

A k_n-legközelebbi szomszéd szabály az x-hez legközelebbi k_n darab X_i címkéi alapján hoz többségi döntést. Legyen $\left(X_{(1,n)}(x), Y_{(1,n)}(x)\right), \ldots, \left(X_{(n,n)}(x), Y_{(n,n)}(x)\right)$ az X_i-k x-től vett távolsága alapján rendezett minta. Ekkor

$$g_n(x) = \begin{cases} 1, & \text{ha } \sum_{i=1}^{k_n} I_{\{Y_{(i,n)}(x)=1\}} \geq \sum_{i=1}^{k_n} I_{\{Y_{(i,n)}(x)=0\}} \\ 0 & \text{különben.} \end{cases}$$

Könnyen látható, hogy ez is egy plug-in szabály, ahol most $m_n(x)$ a k_n-legközelebbi szomszéd regresszióbecslő, tehát a konzisztencia itt is a regresszióbecslő konzisztenciájából következik.

8.2.3. tétel: Ha az $||X-x||$ valószínűségi változó abszolút folytonos minden x-re, $k_n \to \infty$ és $\frac{k_n}{n} \to 0$, akkor a k_n-legközelebbi szomszéd osztályozási szabály erősen univerzálisan konzisztens.

A sűrűségfüggvény-becsléshez és a regresszióbecsléshez hasonlóan itt sem lehet általában semmit mondani a konvergenciasebességről, a konvergencia tetszőlegesen lassú lehet.

8.2.4. tétel: Osztályozási szabályok minden $\{g_n\}$ sorozatához és pozitív számok minden monoton, 0-hoz tartó $a_n < \frac{1}{16}$ sorozatához létezik (X,Y)-nak olyan eloszlása, amelyre X egyenletes eloszlású $[0,1]$-en, $L^* = 0$ és

$$\mathbf{P}\left(g_n(X) \neq Y\right) > a_n$$

minden n-re.

8.3. Empirikus hibaminimalizálás

Hasonlóan a regresszióbecsléshez, választhatunk döntésfüggvényt az empirikus hibaminimalizálás módszerével.

8.3.1. definíció: A g szabály empirikus hibavalószínűségén a mintákon elkövetett átlagos hibát értjük, azaz

$$L_n(g) = \frac{1}{n}\sum_{i=1}^{n} I_{\{g(X_i)\neq Y_i\}}$$

Az empirikus hibavalószínűség nyilván torzítatlan becslése a valódi hibavalószínűségnek, azaz $\mathbf{E}L_n(g) = L(g)$.

Legyen $\mathcal{C}$ a $g : \mathbb{R}^d \to \{0,1\}$ döntésfüggvények egy családja. A feladat az, hogy válasszunk ki $\mathcal{C}$-ből egy olyan döntésfüggvényt, amelynek hibavalószínűsége közel van a $\mathcal{C}$-beli legjobb döntés hibavalószínűségéhez. A $\mathcal{C}$ család megállapításában sokféle szempont játszhat szerepet, például az osztályozandó adat eloszlásáról rendelkezésre álló előzetes információ, számítási megfontolások.

Válasszuk a $\mathcal{C}$ családból azt a döntést, amelynek az empirikus hibavalószínűsége minimális, azaz legyen

$$g_n = \operatorname*{argmin}_{g\in\mathcal{C}} L_n(g)$$

Azt várjuk, hogy g_n hibavalószínűsége közel lesz a családbeli optimumhoz, azaz $L(g_n) - \inf_{g\in\mathcal{C}} L(g)$ becslési hiba kicsi.

$$L(g_n) - L^* = \left(L(g_n) - \inf_{g\in\mathcal{C}} L(g)\right) + \left(\inf_{g\in\mathcal{C}} L(g) - L^*\right),$$

ahol $L(g_n) - \inf_{g\in\mathcal{C}} L(g)$ a becslési hiba és $\inf_{g\in\mathcal{C}} L(g) - L^*$ az approximációs hiba. Előfordulhat azonban, hogy a becslési hiba kicsi, de $L(g_n)$ mégis távol van az L^* Bayes-hibától. Tehát lehet, hogy az $\inf_{g\in\mathcal{C}} L(g) - L^*$ approximációs hiba nagy.

A $\mathcal{C}$ család tehát elég nagy kell, hogy legyen ahhoz, hogy jó közelítést adjon az optimális megoldásra, de nem lehet túl nagy sem, mert akkor az adatok mennyisége nem elegendő arra, hogy jó döntést válasszunk ki belőle.

A továbbiakban a becslési hibát vizsgáljuk. (Az approximációs hiba a család kiválasztásától függ csak, és attól nem, hogy a családból hogyan választunk döntést.)

8.3.1. lemma:

$$L(g_n) - \inf_{g\in\mathcal{C}} L(g) \le 2\sup_{g\in\mathcal{C}} |L_n(g) - L(g)|$$

Bizonyítás:

$$L(g_n) - \inf_{g\in\mathcal{C}} L(g) = \sup_{g\in\mathcal{C}} \left(L(g_n) - L_n(g_n) + L_n(g_n) - L_n(g) + L_n(g) - L(g)\right) \le$$

$$\le \sup_{g\in\mathcal{C}} \left(L(g_n) - L_n(g_n) + L_n(g) - L(g)\right) \le 2\sup_{g\in\mathcal{C}} |L_n(g) - L(g)|$$

Az első egyenlőtlenség g_n választásából adódik, $L_n(g_n) \le L_n(g)$.

■

Tehát $\sup_{g\in\mathcal{C}} |L_n(g) - L(g)|$-re kell felső korlátot találnunk.

8.3.1. tétel: Tegyük fel, hogy $\mathcal{C}$ véges sok döntésfüggvényt tartalmaz, ekkor

$$\mathbf{P}\left(\sup_{g\in\mathcal{C}} |L_n(g) - L(g)| > \varepsilon\right) \leq 2|\mathcal{C}|\mathrm{e}^{-2n\varepsilon^2}$$

Bizonyítás:

$$\mathbf{P}\left(\sup_{g\in\mathcal{C}} |L_n(g) - L(g)| > \varepsilon\right) \leq \sum_{g\in\mathcal{C}} \mathbf{P}\left(|L_n(g) - L(g)| > \varepsilon\right) \leq 2\cdot|\mathcal{C}|\mathrm{e}^{-2n\varepsilon^2}$$

a Hoeffding-egyenlőtlenség miatt, hiszen $nL_n(g)$ binomiális eloszlású valószínűségi változó n és $L(g)$ paraméterekkel.

■

Mégjobb felső korlátot kaphatunk, ha feltesszük, hogy a $\mathcal{C}$-beli döntések között van olyan, amelyik hibavalószínűsége nulla.

8.3.2. tétel: Tegyük fel, hogy $|\mathcal{C}| < \infty$ és $\min_{g\in\mathcal{C}} L(g) = 0$. Ekkor minden n-re és $\varepsilon > 0$-ra

$$\mathbf{P}\left(L(g_n) > \varepsilon\right) \leq |\mathcal{C}|\mathrm{e}^{-n\varepsilon},$$

és

$$\mathbf{E}\left(L(g_n)\right) \leq \frac{1+\log|\mathcal{C}|}{n}.$$

Bizonyítás:

$$\mathbf{P}\left(L(g_n) > \varepsilon\right) \leq \mathbf{P}\left(\max_{g\in\mathcal{C}:\, L_n(g)=0} L(g) > \varepsilon\right) =$$

$$= \mathbf{E}\left(I_{\left\{\max_{g\in\mathcal{C}:\, L_n(g)=0} L(g)>\varepsilon\right\}}\right) = \mathbf{E}\left(\max_{g\in\mathcal{C}} I_{\{L_n(g)=0\}} I_{\{L(g)>\varepsilon\}}\right) \leq$$

$$\leq \sum_{g\in\mathcal{C}:\, L(g)>\varepsilon} \mathbf{P}\left(L_n(g) = 0\right) \leq |\mathcal{C}|(1-\varepsilon)^n,$$

mivel annak a valószínűsége, hogy egy (X_i, Y_i) sem esik az $\{(x,y):\ g(x) \neq y\}$ halmazba, kevesebb, mint $(1-\varepsilon)^n$, ha a halmaz valószínűsége nagyobb, mint ε. Innen a tétel első állítása következik, ha használjuk az $1 - x \leq \mathrm{e}^{-x}$ egyenlőtlenséget.

A várható hibavalószínűség becsléshez vegyük észre, hogy minden $u > 0$-ra

$$\mathbf{E}\left(L(g_n)\right) = \int_0^\infty \mathbf{P}\left(L(g_n) > t\right)\,\mathrm{d}t \leq$$

$$\leq u + \int_u^\infty \mathbf{P}\left(L(g_n) > t\right)\,\mathrm{d}t \leq u + |\mathcal{C}|\int_u^\infty \mathrm{e}^{-nt}\,\mathrm{d}t = u + \frac{|\mathcal{C}|}{n}\mathrm{e}^{-nu}.$$

Mivel u tetszőleges, választhatjuk úgy, hogy minimalizálja a felső korlátot. Az optimális választás $u = \frac{\log|\mathcal{C}|}{n}$, amivel a felső korlát

$$u + \frac{|\mathcal{C}|}{n}\mathrm{e}^{-nu} = \frac{\log|\mathcal{C}|}{n} + \frac{|\mathcal{C}|}{n}\cdot\mathrm{e}^{-n\frac{\log|\mathcal{C}|}{n}} = \frac{\log|\mathcal{C}|+1}{n}$$

■

Most térjünk vissza az általános esethez, azaz felejtsük el a feltételezéseinket, hogy $|\mathcal{C}| < \infty$ és $\min\limits_{g\in\mathcal{C}} L(g) = 0$.

Legyen μ (X, Y) valószínűségi mértéke $\mathbb{R}^d \times \{0,1\}$-en, és legyen μ_n a mintáinkon alapuló empirikus mérték. Tehát egy $A \subset \mathbb{R}^d \times \{0,1\}$ mérhető halmazra $\mu(A) = \mathbf{P}\left((X,Y) \in A\right)$ és $\mu_n(A) = \frac{1}{n}\sum\limits_{i=1}^{n} I_{\{(X_i,Y_i)\in A\}}$. Ekkor

$$L(g) = \mu\left(\{(x,y):\, g(x) \neq y\}\right),$$

azaz $L(g)$ a μ-mértéke az

$$\{\{x:\, g(x) = 1\} \times \{0\}\} \cup \{\{x:\, g(x) = 0\} \times \{1\}\}$$

halmaznak. Hasonlóan

$$L_n(g) = \mu_n\left(\{(x,y):\, g(x) \neq y\}\right),$$

így

$$\sup_{g\in\mathcal{C}} |L_n(g) - L(g)| = \sup_{\bar{A}\in\bar{\mathcal{A}}} \left|\mu_n(\bar{A}) - \mu(\bar{A})\right|,$$

ahol $\bar{\mathcal{A}}$ az összes

$$\{\{x:\, g(x) = 1\} \times \{0\}\} \cup \{\{x:\, g(x) = 0\} \times \{1\}\}, \quad g \in \mathcal{C}$$

alakú halmaz családja.

Emlékezzünk vissza, hogy a

$$\sup_{\bar{A}\in\bar{\mathcal{A}}} \left|\mu_n(\bar{A}) - \mu(\bar{A})\right|$$

kifejezésre a Vapnik–Chervonenkis-egyenlőtlenség ad felső korlátot. Most vezessük be döntések családjainak Vapnik–Chervonenkis-dimenzióját is.

8.3.2. definíció: Legyen $\mathcal{C}$ a $g:\ \mathbb{R}^d \to \{0,1\}$ döntésfüggvények egy családja, és tartalmazza $\bar{\mathcal{A}}$ az összes

$$\{\{x:\, g(x) = 1\} \times \{0\}\} \cup \{\{x:\, g(x) = 0\} \times \{1\}\}, \quad g \in \mathcal{C}$$

alakú halmazt. A $\mathcal{C}$ család shatter együtthatója és VC-dimenziója egyezzen meg az $\bar{\mathcal{A}}$ halmazcsalád shatter együtthatójával és VC-dimenziójával.

$$S(\mathcal{C}, n) = s(\bar{\mathcal{A}}, n)$$

$$V_{\mathcal{C}} = V_{\bar{\mathcal{A}}}$$

Ekkor tehát a Vapnik–Chervonenkis-egyenlőtlenség és a 8.3.1. lemma miatt igaz a következő:

8.3.3. tétel:

$$\mathbf{P}\left(\sup_{g\in\mathcal{C}} |L_n(g) - L(g)| > \varepsilon\right) \leq 8S(\mathcal{C}, n)\mathrm{e}^{-n\varepsilon^2/32}$$

és

$$\mathbf{P}\left(L(g_n) - \inf_{g\in\mathcal{C}} L(g) > \varepsilon\right) \leq 8S(\mathcal{C}, n)\mathrm{e}^{-n\varepsilon^2/128},$$

ahol g_n az empirikus hibát minimalizáló döntés.

Ebből a 8.3.2 tétel bizonyításában látott módon kaphatunk felső korlátot a várható hibavalószínűségre

$$\mathbf{E}L(g_n) - \inf_{g\in\mathcal{C}} L(g) \le 16\sqrt{\frac{\log(8eS(\mathcal{C},n))}{2n}},$$

illetve mivel $V_\mathcal{C} > 2$-re $S(\mathcal{C},n) \le n^{V_\mathcal{C}}$

$$\mathbf{E}L(g_n) - \inf_{g\in\mathcal{C}} L(g) \le 16\sqrt{\frac{V_\mathcal{C}\log n + 4}{2n}}.$$

Ha feltesszük, hogy $\inf_{g\in\mathcal{C}} L(g) = 0$, azaz, hogy a Bayes-döntés benne van $\mathcal{C}$-ben és $L^* = 0$, akkor egy gyorsabban 0-hoz tartó felső korlátot kapunk.

8.3.4. tétel: A fenti esetben

$$\mathbf{P}\left(L(g_n) > \varepsilon\right) \le 2S(\mathcal{C},2n)2^{-n\varepsilon/2}.$$

A döntéscsaládok Vapnik–Chervonenkis-dimenziójának vizsgálatát könnyíti meg az alábbi tétel.

8.3.5. tétel: Ha $\bar{\mathcal{A}} = \{A\times\{0\}\cup A^c\times\{1\};\ A\in\mathcal{A}\}$, akkor $s(\bar{\mathcal{A}},n) = s(\mathcal{A},n)$ minden n-re és ezért $V_{\bar{\mathcal{A}}} = V_\mathcal{A}$.

A döntéscsaládok shatter együtthatójának definíciójában az A halmazok $\{x:\ g(x)=1\}$ alakúak, míg $\bar{A}$ olyan (x,y) párok halmaza, amelyekre $g(x)\ne y$. A fenti tétel azt jelenti, hogy $S(\mathcal{C},n) = s(\mathcal{A},n)$, tehát elég az $\mathcal{A}$ tulajdonságait vizsgálni, ami egyszerűbb, hiszen $\mathbb{R}^d$ részhalmazainak a családja.

Ha például $x\in\mathbb{R}$ és $\mathcal{C}$ a

$$g(x) = \begin{cases} 1, & \text{ha } x\le a \\ 0 & \text{különben}\end{cases}$$

alakú döntések családja, akkor az $\{x:\ g(x)=1\}$ halmazok a félegyenesek, tehát ekkor $V_\mathcal{C} = V_{\{\text{félegyenesek}\}} = 1$.

Lehet $\mathcal{C}$ például a lineáris döntések családja, azaz a

$$g(x) = \begin{cases} 1, & \text{ha } a^Tx > b \\ 0 & \text{különben}\end{cases}$$

alakú döntésfüggvényeket tartalmazó család. Ekkor az $\{x:\ g(x)=1\}$ halmazok pont az $\mathbb{R}^d$-beli $\{x:\ a^Tx > b\}$ félterek, amelyek családjáról korábban láttuk, hogy a VC-dimenziója $d+1$.

Ajánlott irodalom

[1] H. Cramer: Mathematical methods of Statistics
Princeton University Press, Princeton, 1946.

[2] E.L. Lehman: Testing Statistical Hipotheses
Wiley & Sons, New York, 1959.

[3] E.L. Lehman: Theory of Point Estimation
Chapman & Hall, New York, 1991.

[4] Mogyoródi József (szerk.): Matematikai statisztika (ELTE jegyzet)
Tankönyvkiadó, Budapest, 1990.

[5] Vincze István: Matematikai statisztika (ELTE jegyzet)
Tankönyvkiadó, Budapest, 1974.