æ

MATEMATIKAI STATISZTIKA

Dr. Bolla Marianna, Matematika Intézet, Sztochasztika Tanszék

Leíró statisztika

$(\Omega, \mathcal{A}, \mathcal{P})$ *statisztikai mező*, ahol a $\mathcal{P}$ mértékcsalád olyan $\mathbb{P}$ eloszlásokból áll, melyekkel $(\Omega, \mathcal{A}, \mathbb{P})$ valószínűségi mezőt alkot. A probléma éppen a megfelelő eloszlás kiválasztása. Általában *paraméteres* a mező: $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, ahol $\Theta \subset \mathbb{R}^k$ a *paramétertér*.

Vizsgálódásaink középpontjában egy X valószínűségi változó áll (pl. az egyetemista fiúk testmagassága, a tanszékre délelőtt 10 és 10:30 közt befutó telefonhívások száma), melynek pontos $\mathbb{P}$ eloszlását nem ismerjük, csak annyit tudunk, hogy $\mathbb{P} \in \mathcal{P}$. Itt az első példában $\mathcal{P}$ a normális, a másodikban pedig a Poisson eloszláscsalád, azaz problémánk paraméteres. Célunk a paraméterek becslése, esetleg hipotézisek vizsgálata (pl. igaz-e, hogy az egyetemista fiúk testmagasságának várható értéke mondjuk 175 cm, vagy szignifikánsan különbözik-e ez a 10 évvel ezelőtti egyetemistákétól).

Mindehhez megfigyeléseket végzünk, azaz mintát veszünk. *Statisztikai minta* alatt értjük független, azonos eloszlású valószínűségi változók egy $X_1, X_2, \ldots, X_n$ véges sorozatát, ahol az X_i valószínűségi változók eloszlása megegyezik az X háttérváltozóéval.

Az $X_1, \ldots, X_n$ mintát röviden jelölje $\mathbf{X}$, azaz $\mathbf{X} = (X_1, \ldots, X_n)$ n-dimenziós, független komponensű véletlen vektor (vektor értékű valószínűségi változó), egy konkrét kimenetelt pedig jelöljön $\mathbf{x} = (x_1, \ldots, x_n)$, ezt a minta *realizáció*jának nevezzük.

A mintaelemek egy $T = T(\mathbf{X}) = T(X_1, \ldots, X_n)$ mérhető függvényét *statisztikának* hívjuk. Egy statisztikában információt tömörítünk. Az lesz majd a "jó" statisztika, mely nem veszít lényeges információt a tömörítés által. Bevezetjük a következő *alapstatisztikák*at.

Legyen $X_1, \ldots, X_n$ független azonos eloszlású n-elemű minta.

Definíció. Az

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

statisztikát *mintaátlag*nak nevezzük.

Ha hangsúlyozni szeretnénk a mintaelemszámot, akkor az $\bar{X}_n$ jelölést használjuk, ha pedig a konkrét realizációkkal számolunk, akkor $\bar{x}$-t vagy $\bar{x}_n$-t írunk.

Steiner-tétel. *Az $x_1, \ldots, x_n \in \mathbb{R}$ rögzített értékekkel és tetszőleges $c \in \mathbb{R}$ valós számmal*

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

teljesül.

Definíció. Az

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

statisztikát *empirikus (tapasztalati) szórásnégyzet*nek nevezzük, az

$$S^{*2} = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

statisztikát pedig *korrigált empirikus (tapasztalati) szórásnégyzet*nek. A fenti mennyiségek gyöke az *empirikus (tapasztalati) szórás* illetve a *korrigált empirikus (tapasztalati) szórás*, melyeket S illetve S^* jelöl.

Ha hangsúlyozni szeretnénk a mintaelemszámot, akkor az S_n^2 illetve S_n^{*2} jelölést használjuk, ha pedig a konkrét realizációkkal számolunk, akkor s_n^2-t vagy s_n^{*2}-t írunk.

Következmény. A Steiner tételből $c = 0$ választással következik, hogy az empirikus szórásnégyzetet a következőképpen is számolhatjuk:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}^2 = \overline{X^2} - \bar{X}^2.$$

Definíció. A $\bar{X}\sqrt{n}/S^*$ mennyiséget a *mintaátlag standardizált hibájának* (standard error of mean = S.E.M.) nevezzük. Pozitív minta esetén az $S/\bar{X}$ mennyiséget *szórási együtthatónak* hívják. Mérések esetében ez utóbbi a relatív hibát jelenti.

Definíció. Legyen k rögzített pozitív egész. Az

$$M_k = \frac{1}{n}\sum_{i=1}^{n}X_i^k$$

statisztikát *k-adik empirikus (tapasztalati) momentum*nak nevezzük, az

$$M_k^c = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$$

statisztika pedig a *k-adik empirikus (tapasztalati) centrális momentum.*

Nyilván $S^2 = M_2^c = M_2 - M_1^2$.

Definíció. Az $M_3^c/(M_2^c)^{3/2}$ valószínűségi változó a *ferdeség* (skewness), az $M_4^c/(M_2^c)^2 - 3$ valószínűségi változó pedig a *lapultság* (curtosis).

Előbbi az eloszlás szimmetriáját fejezi ki (szimmetrikus eloszlásoknál elméleti értéke 0), utóbbi a sűrűségfüggvény laposságát méri (a standard normális eloszlás lapultsága zérus).

Definíció. Legyen $(X,Y)^T$ 2-dimenziós valószínűségi változó, $(X_1,Y_1)^T,\dots,(X_n,Y_n)^T$ pedig vele azonos eloszlású független azonos eloszlású n-elemű minta. Jelölje S_X illetve S_Y a komponensek empirikus szórását! A

$$C=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})=\frac{1}{n}\sum_{i=1}^{n}X_iY_i-\bar{X}\bar{Y}$$

statisztikát *empirikus (tapasztalati) kovariancia*nak, az

$$R=\frac{C}{S_XS_Y}=\frac{\sum_{i=1}^{n}X_iY_i-n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)\left(\sum_{i=1}^{n}Y_i^2-n\bar{Y}^2\right)}}$$

statisztikát pedig *empirikus (tapasztalati) korreláció*nak nevezzük.

Definíció. Az $X_1,\dots,X_n$ mintaelemek értékeit nem-csökkenő sorrendben felvevő $X_1^*\le X_2^*\le\dots\le X_n^*$ valószínűségi változókat *n-elemű rendezett mintának* nevezzük, így a rendezett mintaelemek sem nem függetlenek, sem nem azonos eloszlásúak.

Tehát minden kontrét $x_1,x_2,\dots,x_n$ realizáció esetén ezt az n valós számot kell nagyság szerint nem csökkenő sorrendbe rendezni, és a nagyság szerint i-ediket x_i^*-gal jelölni. Természetesen a szorzattér különböző elemeire más és más lesz a mintaelemek sorrendje, és így a rendezés is.

Definíció. Az $X_n^*-X_1^*$ statisztikát *mintaterjedelemnek* (range) nevezzük.

Definíció. Empirikus (tapasztalati) medián alatt értjük páratlan n ($n=2k+1$) esetén X_{k+1}^*-ot, páros n ($n=2k$) esetén pedig $(X_k^*+X_{k+1}^*)/2$-t.

Ez valójában a középső mintaelem, és amennyiben a realizációból számolt értékét m jelöli, ezzel teljesül a Steiner-tétel L_1- normában vett megfelelője:

Állítás.

$$\min_{c\in\mathbb{R}}\frac{1}{n}\sum_{i=1}^{n}|x_i-c|=\frac{1}{n}\sum_{i=1}^{n}|x_i-m|.$$

A fenti minimumot a minta *átlagos abszolút eltérés*ének is szokták nevezni.

A következőkben egy n-elemű minta alapján kívánjuk közelíteni a háttéreloszlást, ezért megkonstruáljuk az ún. empirikus eloszlásfüggvényt, amiről belátjuk, hogy "elég nagy" n-re jól rekonstruálja az ismeretlen eloszlásfüggvényt, akármi is legyen a véletlen minta. Ezt a tényt fogalmazza meg precízen a Glivenko–Cantelli-tétel, melyet a statisztika egyik alaptételének is szoktak tekinteni.

Definíció. Empirikus (tapasztalati) eloszlásfüggvény alatt a következő véletlen függvényt értjük: tetszőleges $x\in\mathbb{R}$ számra legyen

$$F_n^*(x):=\frac{\sum_{i=1}^{n}I(X_i<x)}{n}=\begin{cases}0, & \text{ha} \quad x\le X_1^*,\\ \frac{k}{n}, & \text{ha} \quad X_k^*<x\le X_{k+1}^* \quad (k=1,\dots,n-1)\\ 1, & \text{ha} \quad x>X_n^*.\end{cases}$$

Itt $I(\cdot)$ az argumentumban álló esemény indikátorváltozója. Könnyű látni, hogy az $I(X_i < x)$ indikátorváltozók független azonos eloszlásúak (Bernoulli eloszlásúak $F(x)$ paraméterrel, ahol F az X háttérváltozó eloszlásfüggvénye).

Megjegyezzük, hogy F_n^* az $x_1, \ldots, x_n$ realizációra olyan, mint egy $Y \sim \mathcal{U}(x_1, \ldots, x_n)$ diszkrét egyenletes eloszlású valószínűségi változó eloszlásfüggvénye. Nyilván $\mathbb{E}(Y) = \bar{X}$ és $\mathbb{D}^2(Y) = S^2$.

Állítás. *Legyen $F(x)$ az elméleti eloszlásfüggvény és $x \in \mathbb{R}$ rögzített. Akkor*

$$\mathbb{E}(F_n^*(x)) = F(x), \quad \mathbb{D}^2(F_n^*(x)) = \frac{F(x)(1-F(x))}{n},$$

és $\lim_{n\to\infty} F_n^*(x) = F(x)$, 1 valószínűséggel.

A következő tétel ennél még erősebb állítást fogalmaz meg: $n \to \infty$ estén az empirikus eloszlásfüggvények F_n^* sorozata nemcsak rögzített x-re, hanem az egész valós számegyenesen egyenletesen is tart F-hez, 1 valószínűséggel.

Glivenko–Cantelli tétel. *$n \to \infty$ esetén*

$$\sup_{x\in\mathbb{R}} |\, F_n^*(x) - F(x)| \to 0, \qquad \textit{1 valószínűséggel.}$$

A tétel a mintavételen alapuló eljárások jogosságát támasztja alá.

Amennyiben abszolút folytonos az eloszlásunk, az elméleti sűrűségfüggvényt is közelíteni szeretnénk. A tapasztalati eloszlásfüggvény bármilyen jól közelíti is a fenti tétel értelmében az elméletit, mégiscsak egy szakaszonként konstans függvény, így deriváltja nem adhat a problémára megoldást. Szokták az empirikus eloszlásfüggvényt ún. magfüggvény segítségével "simítani", amely már folytonos, sőt differenciálható lesz és deriváltja "jól" közelíti az elméleti sűrűséget (magfüggvényes becslők):

$$\hat{f}_n(x) := \frac{d}{dx} \int_{-\infty}^{\infty} F_n^*(x) \cdot m(x-y)\, dy,$$

ahol az m *magfüggvény* egy kellően sima valószínűségi sűrűségfüggvény. A fenti konvolúció tulajdonképpen azt jelenti, hogy az eredeti valószínűségi változónkra egy "zaj" rakódik rá.

Most csak egy egyszerűbb konstrukciót mutatunk be. n elemű mintánkhoz osszuk fel a számegyenest a h_n hosszúságú Δ_j diszjunkt intervallumokra, és jelölje ν_j a Δ_j-be eső mintaelemek számát!

Definíció. Az

$$f_n^*(x) = \frac{\nu_j}{n h_n}, \qquad x \in \Delta_j$$

összefüggéssel definiált függvényt a minta *sűrűséghisztogram*jának nevezzük.

Mivel a mintaelemek befoglalhatók egy véges intervallumba, nyilván ezen kívül $f_n^*(x) = 0$ lesz, és ezen belül véges sok különböző $f_n^*(x)$ érték alakul ki. A sűrűséghisztogram

is szakaszonként konstans függvény, és az alatta levő összterület 1. Belátható, hogy amennyiben x a valódi f sűrűségfüggvény folytonossági pontja és $n \to \infty$ olyan módon, hogy még $\lim_{n\to\infty} h_n = 0$ és $\lim_{n\to\infty} nh_n = \infty$ is teljesül, akkor $\lim_{n\to\infty} f_n^*(x) = f(x)$, 1 valószínűséggel. (Pl. ha mintánk az $[a, b]$ intervallumba foglalható be és $h_n = (b-a)/n$, akkor a feltétel nem teljesül, viszont $h_n = (b-a)/n^{1-\alpha}$, $0 < \alpha < 1$ esetén teljesül.)

A Glivenko–Cantelli tétel arról szól, hogy az empirikus eloszlásfüggvény 1 valószínűséggel (majdnem minden realizációra) az egész számegyenesen egyenletesen tart az elméleti eloszlásfüggvényhez. Tehát kellő számú mintát véve tetszőleges pontossággal közelíteni tudjuk a valódi eloszlásfüggvényt. De adott pontossághoz vajon hány elemű mintát kell vennünk? A konvergencia sebességére vonatkozóan újabb tételeket fogunk kimondani. Ezek azt jelzik, hogy n kísérlet kb. $1/\sqrt{n}$ nagyságrendű közelítéshez elegendő.

Legyen a háttéreloszlás F eloszlásfüggvénye folytonos, F_n^* pedig jelölje az n-elemű mintához tartozó empirikus eloszlásfüggvényt. Akkor

Tétel (Szmirnov).

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\sup_{x\in\mathbb{R}}(F_n^*(x) - F(x)) < z\right) = S(z), \qquad \forall z \in \mathbb{R},$$

ahol

$$S(z) = \begin{cases} 0, & ha \quad z \le 0, \\ 1 - e^{-2z^2}, & ha \quad z > 0, \end{cases}$$

az ún. Szmirnov-eloszlásfüggvény

Tétel (Kolmogorov).

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\sup_{x\in\mathbb{R}}|F_n^*(x) - F(x)| < z\right) = K(z), \qquad \forall z \in \mathbb{R},$$

ahol

$$K(z) = \begin{cases} 0, & ha \quad z \le 0, \\ \sum_{i=-\infty}^{\infty}(-1)^i e^{-2i^2z^2} = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2z^2}, & ha \quad z > 0, \end{cases}$$

az ún. Kolmogorov-eloszlásfüggvény.

Legyen most az X illetve Y háttérváltozó (nem feltétlenül ismert) eloszlásfüggvénye a folytonos F illetve G függvény, F_n^* illetve G_m^* pedig jelölje az n-elemű $X_1, \ldots, X_n$ illetve az m-elemű $Y_1, \ldots, Y_m$, egymástól is független mintákhoz tartozó empirikus eloszlásfüggvényeket. Tegyük fel továbbá, hogy $F(x) = G(x)$, $\forall x \in \mathbb{R}$. Akkor

Tétel (Szmirnov).

$$\lim_{n,m\to\infty} \mathbb{P}\left(\sqrt{\frac{nm}{n+m}}\sup_{x\in\mathbb{R}}(F_n^*(x) - G_m^*(x)) < z\right) = S(z), \qquad \forall z \in \mathbb{R}.$$

Tétel (Szmirnov).

$$\lim_{n,m\to\infty} \mathbb{P}\left(\sqrt{\frac{nm}{n+m}}\sup_{x\in\mathbb{R}}|F_n^*(x)-G_m^*(x)|<z\right)=K(z),\qquad \forall z\in\mathbb{R}.$$

A Kolmogorov–Szmirnov tételeket használni fogjuk a hipotézisvizsgálatban annak eldöntésére, hogy mintánk egy adott F eloszlásfüggvényű eloszlásból származike, vagy pedig két minta származhat-e ugyanabból az eloszlásból. Vegyük észre, hogy a határeloszlások függetlenek a valódi háttéreloszlástól, így ún. nem-paraméteres próbák definiálhatók segítségükkel.

Most az ún. "jó" statisztika fogalmát pontosítjuk.

Definíció. Likelihood-függvény alatt értjük a mintaelemek együttes valószínűség illetve sűrűségfüggvényét. Legyen $\mathbf{x}=(x_1,\dots,x_n)\in\mathbb{R}^n$ rögzített, és $L_\theta(\mathbf{x})$ a likelihood-függvény az $\mathbf{x}$ helyen. Ha a háttéreloszlás diszkrét p_θ valószínűségfüggvényel, akkor

$$L_\theta(\mathbf{x})=\mathbb{P}_\theta(\mathbf{X}=\mathbf{x})=\prod_{i=1}^n\mathbb{P}_\theta(X_i=x_i)=\prod_{i=1}^n p_\theta(x_i),$$

ha pedig abszolút folytonos f_θ sűrűségfüggvénynyel, akkor

$$L_\theta(\mathbf{x})=\prod_{i=1}^n f_\theta(x_i).$$

Vagyis a likelihood-függvény az $\mathbf{x}$ helyen diszkrét esetben annak a valószínűségét adja, hogy a realizáció éppen $\mathbf{x}$, abszolút folytonos esetben pedig annak a valószínűségével arányos, hogy a realizáció $\mathbf{x}$ "kis" környezetébe esik.

Neyman–Fisher Faktorizációs Tétel. *Egy* $\mathbf{X}$ *minta* $T(\mathbf{X})$ *statisztikája pontosan akkor elégséges, ha létezik olyan* $g_\theta(t)$ $(\theta\in\Theta,\, t\in\mathcal{T}\,(=T$ *értékkészlete))* *és* $h(\mathbf{x})$ $(\mathbf{x}\in\mathcal{X})$ *mérhető függvény, hogy*

$$L_\theta(\mathbf{x})=g_\theta(T(\mathbf{x}))\cdot h(\mathbf{x})$$

teljesül minden $\theta\in\Theta$, $\mathbf{x}\in\mathcal{X}$ *esetén.*

Azaz a likelihood-függvény csak a T statisztikán keresztül függ a paramétertől.

Keressünk elégséges statisztikákat a faktorizációs tétel alapján!

1. *Példa*: Legyen $X_1,\dots,X_n\sim\mathcal{P}(\lambda)$ független azonos eloszlású!

$$Ł_\lambda(\mathbf{x})=\prod_{i=1}^n\frac{\lambda^{x_i}}{x_i!}e^{-\lambda}=\left(\lambda^{\sum_{i=1}^n x_i}e^{-n\lambda}\right)\cdot\left(\prod_{i=1}^n\frac{1}{x_i!}\right)=g_\lambda(\sum_{i=1}^n x_i)\cdot h(\mathbf{x}),$$

   így $\sum_{i=1}^n X_i$ elégséges statisztika, és nyilván $\bar{X}$ is az.

2. *Példa*: Legyen $X_1,\dots,X_n\sim\mathcal{E}xp(\lambda)$ független azonos eloszlású!

$$L_\lambda(\mathbf{x})=\prod_{i=1}^n\lambda e^{-\lambda x_i}=\lambda^n e^{-\lambda\sum_{i=1}^n x_i},$$

ami megfelel $g_\lambda(T(\mathbf{x}))$-nek, és $h(\mathbf{x}) = 1$. Ezért $\sum_{i=1}^n X_i$ elégséges statisztika, és $\bar{X}$ is az.

Nyilván egy elégséges statisztika invertálható függvénye is elégséges lesz. Nézzünk most példákat többdimenziós paramétertér esetén elégséges statisztikára (ilyenkor persze a statisztika is többdimenziós).

3. *Példa*: Legyen $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ független azonos eloszlású! Itt $\theta = (\mu, \sigma^2)$.

$$L_\theta(\mathbf{x}) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right) =$$
$$= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right]\right),$$

ami megfelel $g_\theta(T(\mathbf{x}))$-nek a $T(\mathbf{X}) = (\bar{X}, S^2)$ elégséges statisztikapárral, $h(\mathbf{x}) = 1$. Nyilván az $(\bar{X}, S^{*2})$ statisztikapár, vagy a $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ statisztikapár is elégséges lesz.

4. *Példa*: Legyen $X_1, \ldots, X_n \sim \mathcal{U}[a, b]$ független azonos eloszlású! Itt $\theta = (a, b)$.

$$L_\theta(\mathbf{x}) = \prod_{i=1}^n f_\theta(x_i) = \begin{cases} \frac{1}{(b-a)^n}, & \text{ha} \quad x_1, \ldots, x_n \in [a, b] \\ 0, & \text{különben.} \end{cases}$$

Azaz $L_\theta(\mathbf{x}) = (b-a)^{-n} I(x_1^* \geq a,\, x_n^* \leq b) = g_\theta(x_1^*, x_n^*)$ és $h(\mathbf{x}) = 1$ választással a faktorizáció teljesül. Ezért az (X_1^*, X_n^*) pár elégséges statisztikát ad az (a, b) paraméterpárra.

Definíció. A T elégséges statisztikát *minimális elégséges statisztiká*nak nevezzük, ha függvénye bármely más elégséges statisztikának.

Ez a legtömörebb, és ekvivalencia erejéig már egyertelmű.

BECSLÉSELMÉLET

Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ paraméteres statisztikai mező, ahol $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. A θ paramétert vagy annak valamely $\psi(\theta)$ függvényét szeretnénk becsülni az $\mathbf{X} = (X_1, \ldots, X_n)$ független azonos eloszlású minta alapján konstruált $T(\mathbf{X})$ statisztika segítségével. Jelölje $\hat{\theta}$ ill. $\hat{\psi}$ az így kapott becslést!

Egy becslés jóságát különböző kritériumokkal mérjük. Ezekről, továbbá arról lesz szó, mikor található legjobb becslés, és n növekedésével hogyan javul a becslés.

Definíció. $T(\mathbf{X})$ *torzítatlan becslés* $\psi(\theta)$-ra, ha

$$\mathbb{E}_\theta(T(\mathbf{X})) = \psi(\theta), \qquad \forall \theta \in \Theta.$$

Állítás. *$\bar{X}$ mindig torzítatlan becslés $m(\theta) = \mathbb{E}_\theta(X)$-re, ha ez véges.*

Bizonyítás. Vegyük az $X_1, \ldots, X_n$ független azonos eloszlású mintát! Feltettük, hogy a közös várható érték létezik: $\mathbb{E}_\theta(X_i) = m(\theta)$, $i = 1, \ldots, n$. Így

$$\mathbb{E}_\theta(\bar{X}_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta(X_i) = m(\theta), \qquad \forall \theta \in \Theta.$$

□

Könnyű látni, hogy a mintaelemek bármely konvex lineáris kombinációja is torzítatlan becslés a fenti véges várható értékre, tehát a torzítatlanság önmagában még nem teszi egyértelművé a becslést.

A fenti állításból következik, hogy a $\mathcal{B}_n(p)$ binomiális eloszlás p paraméterére rögzített n esetén a relatív gyakoriság torzítatlan becslés, ugyanis $Y \sim \mathcal{B}_n(p)$ előáll $Y = \sum_{i=1}^n X_i$ alakban, ahol $X_1, \ldots, X_n \sim \mathcal{I}(p)$ független azonos eloszlású Bernoulli-változók p várható értékkel, $\bar{X} = Y/n$ pedig a relatív gyakoriság.

A torzítatlanságnál gyengébb követelmény a következő:

Definíció. A $T(\mathbf{X}_n)$ statisztikasorozat *aszimptotikusan torzítatlan becslés* $\psi(\theta)$-ra, ha

$$\lim_{n \to \infty} \mathbb{E}_\theta(T(\mathbf{X}_n)) = \psi(\theta), \qquad \forall \theta \in \Theta.$$

Állítás. *Legyen $X_1, \ldots, X_n$ független azonos eloszlású minta egy tetszőleges olyan eloszlásból, melyre minden $\theta \in \Theta$ esetén $\sigma^2(\theta) = \mathbb{D}^2_\theta(X) < \infty$. Akkor S_n^2 aszimptotikusan torzítatlan, $S_n^{*\,2}$ pedig torzítatlan becslése a szórásnégyzetnek.*

Célunk az, hogy a torzítatlan becslések között minél kisebb szórásúakat találjunk.

Definíció. Legyen a T_1 és T_2 statisztika torzítatlan becslés a θ paraméterre, vagy annak valamely $\psi(\theta)$ függvényére. Azt mondjuk, hogy T_1 *hatásosabb (efficiensebb) becslés*, mint T_2, ha

$$\mathbb{D}^2_\theta(T_1) \leq \mathbb{D}^2_\theta(T_2), \qquad \forall \theta \in \Theta,$$

és legalább egy $\theta_0 \in \Theta$ esetén (2)-ben $<$ teljesül. Egy torzítatlan becslés *hatásos (efficiens) becslés*, ha bármely más torzítatlan becslésnél hatásosabb.

Hatásos becslés nem mindig létezik, de ha van hatásos becslés, az egyértelmű. Tételek alapján majd el tudjuk dönteni egy torzítatlan becslésről, hogy hatásos-e, néhány esetben pedig garantálni tudjuk hatásos becslés létezését.

A konzisztencia azt jelenti, hogy a megfigyelések számának növelésével javul a becslés pontossága.

Definíció. A $T(\mathbf{X}_n)$ statisztikasorozat *(gyengén/erősen) konzisztens becslés* $\psi(\theta)$-ra, ha minden $\theta \in \Theta$-ra $n \to \infty$ esetén $T(\mathbf{X}_n) \to \psi(\theta)$ valószínűségben/1 val.séggel.

Állítás. *Ha $X_1, \dots, X_n$ független azonos eloszlású minta X-re és $m(\theta) = \mathbb{E}_\theta(X)$ létezik, akkor akkor $\bar{X}_n$ (gyengén és erősen is) konzisztens becslés $m(\theta)$-ra.*

Az állítás nem más, mint a nagy számok gyenge és erős törvénye.

Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ paraméteres statisztikai mező, ahol $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. Célunk az, hogy a θ paraméterre vagy annak valamely $\psi(\theta)$ függvényére konstruált torzítatlan becslések szórásnégyzetére alsó korlátot adjunk. Ha egy torzítatlan becslésre ez a korlát eléretik, akkor biztosak lehetünk abban, hogy hatásos becslésünk van, ami 1 val.séggel egyértelmű.

Szükségünk lesz a következő, R. A. Fishertől származó fogalomra.

Definíció. Legyen $\mathbf{X} = (X_1, \dots, X_n)$ független azonos eloszlású minta az X háttérváltozó eloszlásából, amely a θ paramétertől függ ($\theta \in \Theta$), itt csak a $\dim(\Theta) = 1$, Θ konvex esettel foglalkozunk. A fenti minta *Fisher-féle információ*ja az

$$I_n(\theta) = \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \ln L_\theta(\mathbf{X}) \right)^2 \geq 0$$

mennyiséggel van definiálva.

Tétel (Cramér–Rao-egyenlőtlenség). *Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ paraméteres statisztikai mező, ahol $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, $dim(\Theta) = 1$. Legyen $\mathbf{X} = (X_1, \dots, X_n)$ független azonos eloszlású minta a $\mathbb{P}_\theta$ eloszlásból, amiről most tegyük fel, hogy abszolút folytonos. Tegyük fel továbbá, hogy a $T(\mathbf{X})$ statisztika valamely deriválható ψ függvénnyel képzett $\psi(\theta)$ paraméterfüggvény torzítatlan becslése,*

$$\mathbb{D}^2_\theta(T) < +\infty, \qquad \forall \theta \in \Theta$$

továbbá teljesülnek az alábbi bederiválhatósági feltételek:

$$\frac{\partial}{\partial \theta} \int \cdots \int L_\theta(\mathbf{x})\, d\mathbf{x} = \int \cdots \int \frac{\partial}{\partial \theta} L_\theta(\mathbf{x})\, d\mathbf{x}, \qquad \forall \theta \in \Theta$$

és

$$\frac{\partial}{\partial \theta} \int \cdots \int T(\mathbf{x}) L_\theta(\mathbf{x})\, d\mathbf{x} = \int \cdots \int T(\mathbf{x}) \frac{\partial}{\partial \theta} L_\theta(\mathbf{x})\, d\mathbf{x}, \qquad \forall \theta \in \Theta,$$

ahol $\int \cdots \int$ n-dimenziós integrálást jelent a likelihood-függvénytartóján. Akkor

$$\mathbb{D}^2_\theta(T) \geq \frac{(\psi'(\theta))^2}{I_n(\theta)}, \qquad \forall \theta \in \Theta.$$

A következő tétel arról szól, hogyan lehet egy torzítatlan becslés hatásosságát javítani egy elégséges statisztika segít ségével.

Rao–Blackwell–Kolmogorov Tétel. *Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ paraméteres statisztikai mező, ahol $\mathcal{P} = \{\mathbb{P}_\theta\,;\,\theta \in \Theta\}$. Legyen $\mathbf{X} = (X_1, \ldots, X_n)$ független azonos eloszlású minta valamely $\mathbb{P}_\theta$ eloszlásból. Legyen továbbá*

(a) $T(\mathbf{X})$ elégséges statisztika,

(b) $S(\mathbf{X})$ torzítatlan becslés a $\psi(\theta)$ paraméterfüggvényre.

Akkor T-nek van olyan $U = g(T)$ függvénye, amely

(1) szintén torzítatlan becslése a $\psi(\theta)$ paraméterfüggvénynek: $\mathbb{E}_\theta(U) = \psi(\theta)$, $\forall \theta \in \Theta$,

(2) U legalább olyan hatásos becslése $\psi(\theta)$-nak, mint S: $\mathbb{D}^2_\theta(U) \le \mathbb{D}^2_\theta(S)$, $\forall \theta \in \Theta$.

(3) U konstrukciója a következő: $U := \mathbb{E}_\theta(S|T) = g(T(\mathbf{X}))$, $\forall \theta \in \Theta$ (ezt nevezzük "blackwellizálásnak").

A tétel üzenete: a hatásos becsléseket a minimális elégséges statisztika függvényei közt kell keresni.

Becslési módszerek

Maximum likelihood elv

Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ dominált statisztikai mező, ahol $\mathcal{P} = \{\mathbb{P}_\theta\,;\,\theta \in \Theta\}$ (a paramétertér lehet többdimenziós és legyen konvex). Vegyünk egy $X_1, \ldots, X_n$ független azonos eloszlású mintát a $\mathbb{P}_\theta$ eloszlásból (θ ismeretlen). Az $x_1, \ldots, x_n$ realizáció birtokában a paraméter becslésének azt a $\hat\theta$-ot fogadjuk el, amely mellett annak a valószínűsége, hogy az adott realizációt kapjuk, maximális. Mivel ezt a valószínűséget a likelihood-függvény tükrözi, a módszer ezt maximalizálja. A maximumhely csak a realizációtól függ, tehát statisztikát kapunk becslésként.

Definíció. Legyen $L_\theta(\mathbf{x}) : \mathcal{X} \times \Theta \to \mathbb{R}_+$ egy n-elemű mintához tartozó likelihood-függvény, tfh. L a szorzattéren mérhető. A $\hat\theta : \mathcal{X} \to \Theta$ statisztikát a θ paraméter *maximum likelihood (ML-)becslés*ének nevezzük, ha $\hat\theta$ globális maximumhelye a likelihood-függvénynek, azaz

$$L_{\hat\theta(x_1,\ldots,x_n)}(x_1, \ldots, x_n) \ge L_\theta(x_1, \ldots, x_n)$$

teljesül $\forall \theta \in \Theta$ és $(x_1, \ldots, x_n) \in \mathcal{X}$ esetén.

Amennyiben Θ konvex, nyílt halmaz és L differenciálható θ szerint, akkor a globális max. helyet a stacionárius pontok közt keressük. Ilyenkor az $L_\theta(\mathbf{x})$ likelihood-függvény helyett az $l_\theta(\mathbf{x}) = \ln L_\theta(\mathbf{x})$ loglikelihood-függvényt deriválják θ szerint, ugyanis a log-függvény monotonitása miatt a két függvény lokális max. helyei megegyeznek. Több paraméter esetén parciális deriváltakat veszünk. Ezután ellenőrizzük, hogy tényleg lokális maximumot kaptunk-e, és kiválasztjuk a globálisat.

1. *Példa*: Legyen $X_1, \ldots, X_n \sim \mathcal{P}(\lambda)$ független azonos eloszlású!

$$l_\lambda(\mathbf{x}) = \ln\left[\prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}\right] = \ln\lambda \sum_{i=1}^n x_i - \sum_{i=1}^n \ln x_i! - \lambda n,$$

melynek λ szerinti deriválásával a

$$\frac{\partial l_\lambda(\mathbf{x})}{\partial \lambda} = \frac{1}{\lambda}\sum_{i=1}^n x_i - n = 0$$

likelihood-egyenlet adódik, melynek megoldása $\hat{\lambda} = \bar{x}$. Ezen a helyen a loglikelihood-függvény λ szerinti második deriváltja negatív, így tényleg lokális maximumhelyet kapunk, ami egyben globális maximumhely is. Tehát a $T(\mathbf{X}) = \bar{X}$ statisztika a λ paraméter ML-becslése.

2. *Példa*: Legyen $X_1, \ldots, X_n \sim \mathcal{E}xp(\lambda)$ független azonos eloszlású!

$$l_\lambda(\mathbf{x}) = \ln\left[\prod_{i=1}^{n} \lambda e^{-\lambda x_i}\right] = n\ln\lambda - \lambda\sum_{i=1}^{n} x_i,$$

melynek λ szerinti deriválásával a likelihood-egyenlet adódik, melynek megoldása $\hat{\lambda} = 1/\bar{x}$. Ezen a helyen a loglikelihood-függvény λ szerinti második deriváltja negatív, így tényleg lokális maximumhelyet kapunk, ami egyben globális maximumhely is. Tehát a $T(\mathbf{X}) = 1/\bar{X}$ statisztika a λ paraméter ML-becslése.

3. *Példa*: Legyen $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ független azonos eloszlású, $\theta = (\mu, \sigma^2)$.

$$l_\theta(\mathbf{x}) = \ln\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \sum_{i=1}^{n}\left[-\ln(\sqrt{2\pi\sigma^2}) - \frac{(x_i-\mu)^2}{2\sigma^2}\right] =$$

$$= -\frac{n}{2}(\ln(2\pi) + \ln\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

$$\frac{\partial l_\theta(\mathbf{x})}{\partial\mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i-\mu)(-1) = 0 \Longrightarrow \hat{\mu} = \bar{x}.$$

$$\frac{\partial l_\theta(\mathbf{x})}{\partial\sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0.$$

Mivel a $\hat{\mu} = \bar{x}$ szélsőértékhely nem függ a σ^2 paramétertől, ezért $\hat{\mu} = \bar{x}$-ot a második egyenletbe helyettesítve $\hat{\sigma^2} = S_n^2$ adódik, ami torzított, de aszimptotikusan torzítatlan becslése a szórásnégyzetnek. Most vizsgáljuk meg a második deriváltakból álló Hesse-mátrixot a stacionárius $(\bar{x}, s_n^2)$ helyen:

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

ez negatív definit, tehát tényleg lokális maximumhelyet kaptunk, ami a paramétertertomány nyitott volta miatt egyben globális maximumhely is.

4. *Példa*: Legyen $X_1, \ldots, X_n \sim \mathcal{U}[a, b]$ független azonos eloszlású! Itt $\theta = (a, b)$. Az

$$L_\theta(\mathbf{x}) = \left(\frac{1}{b-a}\right)^n \cdot I(a \le x_1^*, b \ge x_n^*)$$

likelihood-függvény nyilván csak akkor különbözik 0-tól, ha az $a \le x_1^*$ és $b \ge x_n^*$ feltételek teljesülnek. Ilyen feltételek mellett viszont az $1/(b-a)^n$ tényező a lehető legrövidebb $[a, b]$ intervallum választása esetén lesz maximális, azaz az intervallum "ráhúzódik" a mintára. Tehát $(\hat{a}, \hat{b}) = (X_1^*, X_n^*)$ lesz a paraméterpár ML-becslése.

Momentumok módszere

A módszert általában több paraméter együttes becslésére használják. Legyen $X_1, \dots, X_n$ független azonos eloszlású minta egy $\mathbb{P}_\theta$ eloszlásból, $\theta = (\theta_1, \dots, \theta_k)$. Válasszunk k db. momentumot (általában az első k-t), amelyek a $\theta_1, \dots, \theta_k$ paramétereket már egyértelműen meghatározzák:

$$m_j = \mathbb{E}_\theta(X^j) = g_j(\theta_1, \dots, \theta_k), \qquad j = 1, \dots, k.$$

Tfh. a $(g_1, \dots, g_k) : \mathbb{R}^k \to \mathbb{R}^k$ leképezésnek létezik inverze, jelölje ezt $(h_1, \dots, h_k) : \mathbb{R}^k \to \mathbb{R}^k$, ahol tehát $h_i(m_1, \dots, m_k) = \theta_i$.

Definíció. A fenti jelölésekkel θ_i *momentum becslés*e alatt a

$$\hat{\theta}_i = h_i(\hat{m}_1, \dots, \hat{m}_k), \qquad i = 1, \dots, k$$

statisztikát értjük, ahol

$$\hat{m}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j$$

a minta j-edik empirikus momentuma.

Legkisebb négyzetes becslések, regresszió

Az alapprobléma a következő: Az X, Y v.v. együttes eloszlásának ismeretében közelíteni szeretnénk Y-t X mérhető t fv.-ével legkisebb négyzetes értelemben:

$$\mathbb{E}(Y - t(X))^2 \to \min. \quad t - \text{ben.}$$

Tudjuk, hogy az optimumot az ún. *regressziós görbe* szolgáltatja, melynek egyenlete:

$$t_{opt}(x) = \mathbb{E}(Y \mid X = x),$$

azaz Y feltételes várható értéke a $X = x$ feltétel mellett. Amennyiben X, Y együttes eloszlása 2-dimenziós normális, a regressziós görbe egyenes lesz. Egyéb esetekben is szokták a a legkisebb négyzetes értelemben legjobb lineáris közelítést keresni, különösen ha az elméleti együttes eloszlás nem ismert, csak egy 2-dimenziós minta áll rendelkezésünkre.

1. *Elméleti megoldás*
   Tegyük fel, hogy az X, Y v.v.-k (általában ismeretlen) együttes eloszlása abszolút folytonos, továbbá a változók első, második és vegyes második momentumai léteznek, ezeket külön jelöljük is:

   $$\mathbb{E}(X) = m_1, \quad \mathbb{E}(Y) = m_2, \quad \mathrm{D}^2(X) = \sigma_1^2, \quad \mathrm{D}^2(Y) = \sigma_2^2, \quad \mathrm{Cov}\,(X, Y) = c, \quad \mathrm{Corr}\,(X, Y) = r,$$

   feltehető, hogy $\sigma_1 > 0$. Keressük az $l(x) = ax + b$ *regressziós egyenes*t, mellyel

   $$h(a, b) = \mathbb{E}(Y - l(X))^2 = \mathbb{E}(Y - aX - b)^2 \to \min. \quad a, b - \text{ben.}$$

   Ez egy kétváltozós szélsőérték feladat, a stacionárius megoldás az alábbi egyenletrendszerből kapható:

   $$\frac{\partial h}{\partial a} = -2\mathbb{E}[(Y - aX - b)X] = 0$$
   $$\frac{\partial h}{\partial b} = -2\mathbb{E}[Y - aX - b] = 0$$

(ui. a fenti feltételek mellett a paraméter szerinti deriválás és az integrálást jelentő várható érték képzés felcserélhető), vagy ami ezzel ekvivalens:

$$a \cdot \mathbb{E}(X^2) + b \cdot \mathbb{E}(X) = \mathbb{E}(XY)$$

$$a \cdot \mathbb{E}(X) + b = \mathbb{E}(Y).$$

Az ismeretlenek a és b, az együtthatómátrix:

$$H = \begin{pmatrix} \mathbb{E}(X^2) & \mathbb{E}(X) \\ \mathbb{E}(X) & 1 \end{pmatrix},$$

melynek determinánsa: $|H| = \mathbb{E}(X^2) - \mathbb{E}^2(X) = \sigma_1^2 > 0$, így a Cramer-szabállyal:

$$a = \frac{\mathbb{E}(XY) - \mathbb{E}(X) \cdot \mathbb{E}(Y)}{\sigma_1^2} = \frac{c}{\sigma_1^2} = \frac{r\sigma_1\sigma_2}{\sigma_1^2} = r\frac{\sigma_2}{\sigma_1},$$

$$b = \mathbb{E}(Y) - a\mathbb{E}(X) = m_2 - \frac{c}{\sigma_1^2} m_1.$$

A másodrendű deriváltakat tartalmazó Hesse-mátrix szintén H, ennek mindkét főminora pozitív, így a fenti a, b valóban lokális minimumot szolgáltat, ami a tartományok nyíltsága, és a differenciálhatósági feltételek teljesülése miatt globális minimumot is ad. A regressziós egyenes egyenlete tehát:

$$y = ax + b = \frac{c}{\sigma_1^2}(x - m_1) + m_2,$$

vagy még könnyebben megjegyezhető formában:

$$\frac{y - m_2}{\sigma_2} = r\frac{x - m_1}{\sigma_1}.$$

Az is látható, hogy a kovariancia (korreláció) előjele adja meg a regressziós egyenes iránytangensének előjelét.

Néhány szó a regresszió (=visszatérés) fogalom jelentéséről. Sir Francis Galton brit orvos a XIX. század második felében szülő–gyerek testmagasság kapcsolatát vizsgálta. Feltételezte, hogy $\sigma_1 = \sigma_2 = \sigma$. Akkor a gyerek testmagassága (Y) a szülő testmagasságával (X) a következőképpen predikálható lineárisan:

$$Y = m_2 + r(X - m_1),$$

ahol r az X és Y közti korrelációt jelöli. Ha $|r| < 1$, akkor nyilván

$$|Y - m_2| < |X - m_1|.$$

Ebből látható, hogy az $r > 0$ esetben: amennyiben a szülő az átlagnál magasabb, a gyerek is az lesz, de az utód magassága kevesebbel múlja felül az átlagot, mint a szülőé. Hasonlóan, ha a szülő az átlagnál alacsonyabb, a gyerek is az lesz, de az utód magassága kevesebbel van alatta az átlagnak, mint a szülőé. (Az átlagtól való abszolút eltérésre negatív korreláció esetén is hasonló mondható.)

Ezt a jelenséget nevezte el Galton az átlaghoz való "visszatérés"nek, latinul regressziónak.

2. *A regressziós együtthatók becslése mintából*
Legyen most $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. minta az (X, Y) háttérváltozóra. A fenti modell a, b együtthatóit becsüljük a *legkisebb négyzetek módszerével*:

$$h(a,b) = \sum_{i=1}^{n} (Y_i - aX_i - b)^2 \to \min . \quad a, b - \text{ben.}$$

Miután az a, b szerinti parciális deriváltakat 0-val tesszük egyenlővé, a következő egyenletrendszert kapjuk:

$$a \cdot \sum_{i=1}^{n} X_i^2 + b \cdot \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} X_i Y_i$$

$$a \cdot \sum_{i=1}^{n} X_i + b \cdot n = \sum_{i=1}^{n} Y_i.$$

A Cramer-szabály itt is alkalmazható, hiszen feltehető, hogy az együtthatómátrix determinánsa $n^2 S_X^2 > 0$. Teljesen hasonló számolással, mint az 1. részben kijön, hogy

$$\hat{a} = \frac{C}{S_X^2} = R\frac{S_Y}{S_X}, \quad \hat{b} = \bar{Y} - \hat{a}\bar{X} = \bar{Y} - R\frac{S_Y}{S_X}\bar{X},$$

ahol S_X ill. S_Y jelöli X ill. Y (korrigálatlan) empírikus szórását, C ill. R pedig az X és Y közti empírikus kovarianciát ill. korrelációt jelöli. Mivel az egyenletrendszer megoldásakor ugyanazokat a lépéseket követjük el, mint az 1. részben, nem meglepő, hogy a és b becslésénél az elméleti első és második momentumok helyébe a mintából számolt empírikus momentumok lépnek, azaz momentum becslést kapunk.

– Megjegyezzük, hogy lineáris regresszióra vezethetők vissza a következő approximációs feladatok:

a. $Y \sim ae^{bX} \iff \ln Y \sim \ln a + bX$

b. $Y \sim aX^b \iff \ln Y \sim \ln a + b \ln X$

c. $Y \sim 1/(aX + b) \iff 1/Y \sim aX + b$

Mintából becslésnél a. esetben az $(X_i, \ln Y_i)$, b. esetben az $(\ln X_i, \ln Y_i)$, c. esetben az $(X_i, 1/Y_i)$ $(i = 1, \ldots, n)$ 2-dimenziós mintákon hajtjuk végre a 2. részben leírt lineáris regressziót, és a végén néha még a becsült paramétert is transzformálni kell.

– *Polinomiális regresszió*
r-edfokú polinomiális regressziónál keressük az $Y \sim a_r X^r + \cdots + a_1 X + a_0$ közelítést legkisebb négyzetes értelemben:

$$\mathbb{E}(Y - a_r X^r - \cdots - a_1 X - a_0)^2 \to \min . \quad a_i - \text{kben.}$$

Az $a_r, \ldots, a_1, a_0$ együtthatók meghatározásához deriváljuk célfv.-ünket mindegyik együttható szerint parciálisan. A deriváltakat 0-val egyenlővé téve $r+1$ db.

lineáris egyenletből álló egyenletrendszert kapunk, mely megoldható Cramer-szabállyal. A megoldásokba $2r$ rendig jönnek be momentumok (ezek létezését fel kell tenni). Amennyiben 2-dimenziós minta alapján szeretnénk becsülni az együtthatókat, a becslésekbe a megfelelő empírikus momentumok jönnek be ($2r$ rendig). Megjegyezzük, hogy itt az $r \geq 1$ egész szám értékét előre meg kell adni, bár egyes programcsomagokban elég a szóbajöhető maximális r-t megadni, és automatikusan megtörténik az ennél alacsonyabb fokú polinomokhoz való illesztés is az illeszkedés szignifikanciájának vizsgálatával együtt, ha a felhasználó kéri. (Az $r = 1$ eset a lineáris regresszió.)

Intervallumbecslések

Az eddigiekben ún. *pontbecslés*ekkel foglalkoztunk, vagyis a becsülendő paramétert v. paraméterfüggvényt a mintaelemekből képzett egyetlen statisztikával becsültük. Most becslésként egy egész intervallumot – melynek határait természetesen statisztikák jelölik ki – fogunk használni. A módszer egyben átvezet bennünket a hipotézisvizsgálatok elméletébe.

Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ paraméteres statisztikai mező, ahol $\mathcal{P} = \{\mathbb{P}_\theta \,;\, \theta \in \Theta\}$, $\dim(\Theta) = 1$! Legyen továbbá $\mathbf{X} = (X_1, \ldots, X_n)$ független azonos eloszlású minta a $\mathbb{P}_\theta$ sokaságból (θ ismeretlen)!

Definíció. A $(T_1(\mathbf{X}), T_2(\mathbf{X}))$ statisztikapárral definiált intervallum legalább $1-\varepsilon$ *szintű konfidenciaintervallum* a $\psi(\theta)$ paraméterfüggvényre, ha

$$\mathbb{P}_\theta(T_1(\mathbf{X}) < \psi(\theta) < T_2(\mathbf{X})) \geq 1 - \varepsilon, \qquad \forall \theta \in \Theta,$$

ahol ε előre adott "kis" pozitív szám (például $\varepsilon = 0.05$, $\varepsilon = 0.01$, a hozzájuk tartozó szignifikanciaszint pedig 95%, 99%).

Abszolút folytonos eloszlásoknál egyenlőség is elérhető, ekkor értelemszerűen *pontosan* $1-\varepsilon$ *szintű konfidenciaintervallum*ról beszélünk. Diszkrét eloszlásoknál nem mindig érhető el az egyenlőség.

1. *Példa*: **Konfidenciaintervallum szerkesztése a normális eloszlás várható értékére ismert szórás esetén**
   Legyen $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma_0^2)$ független azonos eloszlású minta, ahol σ_0^2 ismert, μ (a várható érték) ismeretlen paraméter. Tudjuk, hogy $\bar{X}$ torzítatlan, erősen konzisztens és hatásos pontbecslés μ-re. Keressünk μ-re $1-\varepsilon$ szintű konfidenciaintervallumot az $(\bar{X} - r_\varepsilon, \bar{X} + r_\varepsilon)$ szimmetrikus alakban:

$$\mathbb{P}_\mu(\bar{X} - r_\varepsilon < \mu < \bar{X} + r_\varepsilon) = \mathbb{P}_\mu(|\bar{X} - \mu| < r_\varepsilon) = \mathbb{P}_\mu(-r_\varepsilon < \bar{X} - \mu < r_\varepsilon) =$$
$$\mathbb{P}_\mu\left(\frac{-r_\varepsilon}{\sigma_0/\sqrt{n}} < \frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} < \frac{r_\varepsilon}{\sigma_0/\sqrt{n}}\right) = \Phi\left(\frac{r_\varepsilon}{\sigma_0/\sqrt{n}}\right) - \Phi\left(\frac{-r_\varepsilon}{\sigma_0/\sqrt{n}}\right) =$$
$$= 2\Phi\left(\frac{r_\varepsilon}{\sigma_0/\sqrt{n}}\right) - 1 = 1 - \varepsilon,$$

   azaz

$$\Phi\left(\frac{r_\varepsilon}{\sigma_0/\sqrt{n}}\right) = 1 - \frac{\varepsilon}{2},$$

   ahonnan a standard normális eloszlás $1-\varepsilon/2$ kvantilisére az

$$u_{\varepsilon/2} = \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right)$$

jelölést használva adódik, hogy

$$r_\varepsilon = \frac{u_{\varepsilon/2}\sigma_0}{\sqrt{n}}.$$

Tehát a keresett $1-\varepsilon$ szintű konfidenciaintervallum:

$$\left(\bar{X} - \frac{u_{\varepsilon/2}\sigma_0}{\sqrt{n}},\, \bar{X} + \frac{u_{\varepsilon/2}\sigma_0}{\sqrt{n}}\right)$$

lesz. Vegyük észre, hogy a konfidenciaintervallum hossza n növelésével és a σ_0 szórás csökkentésével csökken, ha viszont ezeket tartjuk konstans szinten, akkor a szignifikanciaszint növelésével (ε csökkenésével) nő (lévén a standard normális eloszlásfüggvény, Φ, és inverze is szigorúan monoton növő függvények). Azaz a mintaelemszám növelésével és a szórás csökkenésével "pontosabban" be tudjuk határolni a várható értéket, viszont nagyobb biztonság csak a "pontosság rovására" érhető el.

Ismeretlen szórás esetén ez nem alkalmazható, a számolásokhoz bevezetünk néhény fogalmat.

Definíció. Legyenek $X_1, \ldots, X_n \sim \mathcal{N}(0,1)$ független azonos eloszlású valószínűségi változók! Az $X = \sum_{i=1}^n X_i^2$ valószínűségi változó eloszlását n szabadsági fokú (centrális) χ^2-*eloszlás*nak nevezzük, és $\chi^2(n)$-nel jelöljük.

Az I.3. paragrafusban meghatároztuk a $\chi^2(n)$-eloszlás sűrűségfüggvényét, továbbá láttuk, hogy

Megjegyzések:

- $\mathbb{E}(X) = n$ és $\mathbb{D}^2(X) = 2n$.
- A definícióból következik, hogy független, $n_1, \ldots, n_r$ szabadsági fokú χ^2-eloszlású valószínűségi változók összege χ^2-eloszlású lesz $n_1 + \cdots + n_r$ szabadsági fokkal.
- Ha n elég "nagy", akkor a centrális határeloszlás tétel értelmében a $\chi^2(n)$-eloszlás normális eloszlással közelíthető az (5.4)-beli paraméterekkel.

Definíció. Legyenek $Y \sim \mathcal{N}(0,1)$ és $X \sim \chi^2(n)$ független valószínűségi változók. Az

$$\frac{Y}{\sqrt{X/n}} \sim t(n)$$

valószínűségi változót n szabadsági fokú t-eloszlásúnak (vagy *Student-eloszlásúnak*) nevezzük, és a fenti módon jelöljük.

A $t(n)$-eloszlás g_n-el jelölt sűrűségfüggvénye egy páros függvény, ami $n \to \infty$ esetén a standard Gauss-görbéhez tart. Eloszlásfüggvényére $G_n(-x) = 1 - G_n(x)$.

Állítás (Lukács Tétel). *Legyen $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ független azonos eloszlású! Akkor*

(1) $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$,

(2) $nS_n^2/\sigma^2 \sim \chi^2(n-1)$,

(3) $\bar{X}$ *és* S_n^2 *függetlenek.*

Nyilvánvaló, hogy (2) és (3) helyett a következő ekvivalens állítások használhatók:

(2') $(n-1)S_n^{*2}/\sigma^2 \sim \chi^2(n-1)$,

(3') $\bar{X}$ és S_n^{*2} függetlenek.

2. *Példa*: **Konfidenciaintervallum szerkesztése a normális eloszlás várható értékére ismeretlen szórás esetén**
Legyen $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ független azonos eloszlású minta, ahol a σ szórás és a μ várható érték is ismeretlen. Szerkesszünk $(\bar{X} - r_\varepsilon, \bar{X} + r_\varepsilon)$ alakú (szimmetrikus), $1-\varepsilon$ szintű konfidenciaintervallumot μ-re!

Az 5.1. Állításból következik, hogy az

$$\frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1) \quad \text{és} \quad \frac{(n-1)S_n^{*2}}{\sigma^2} \sim \chi^2(n-1)$$

statisztikák egymástól függetlenek. Alkalmazzuk a t-eloszlás (5.5) definícióját:

$$\frac{\frac{\bar{X}-\mu}{\sigma}\sqrt{n}}{\sqrt{\frac{(n-1)S_n^{*2}}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{S_n^*}\sqrt{n} \sim t(n-1).$$

Ekkor egyrészt

$$\begin{aligned}
&\mathbb{P}_{\mu,\sigma^2}(\bar{X} - r_\varepsilon < \mu < \bar{X} + r_\varepsilon) = \mathbb{P}_{\mu,\sigma^2}(|\bar{X}-\mu| < r_\varepsilon) = \\
&= \mathbb{P}_{\mu,\sigma^2}(-r_\varepsilon < \bar{X} - \mu < r_\varepsilon) = \\
&= \mathbb{P}_{\mu,\sigma^2}\left(\frac{-r_\varepsilon\sqrt{n}}{S_n^*} < \frac{\bar{X}-\mu}{S_n^*}\sqrt{n} < \frac{r_\varepsilon\sqrt{n}}{S_n^*}\right) = 1-\varepsilon,
\end{aligned}$$

másrészt pedig a t-eloszlás eloszlásfüggvényére tett megjegyzés miatt

$$\mathbb{P}_{\mu,\sigma^2}\left(t_{\varepsilon/2}(n-1) < \frac{\bar{X}-\mu}{S_n^*}\sqrt{n} < t_{\varepsilon/2}(n-1)\right) = 1-\varepsilon,$$

ahol a $t(n-1)$-eloszlás $1-\varepsilon/2$ kvantilisére a

$$t_{\varepsilon/2}(n-1) = G_{n-1}^{-1}\left(1-\frac{\varepsilon}{2}\right)$$

jelölést vezetjük be.

A fenti képletek összevetésével így a konfidenciaintervallum sugarára

$$r_\varepsilon = \frac{t_{\varepsilon/2}(n-1)\cdot S_n^*}{\sqrt{n}}$$

adódik. Tehát a keresett $1-\varepsilon$ szintű konfidenciaintervallum:

$$\left(\bar{X} - \frac{t_{\varepsilon/2}(n-1)\cdot S_n^*}{\sqrt{n}},\ \bar{X} + \frac{t_{\varepsilon/2}(n-1)\cdot S_n^*}{\sqrt{n}}\right).$$

Vegyük észre, hogy a konfidenciaintervallum hossza annál kisebb, minél nagyobb az n mintaelemszám és minél kisebb az S_n^* korrigált empirikus szórás, továbbá minél alacsonyabb szignifikanciaszintet (biztonságot) akarunk elérni. Mivel a szórás ritkán ismert, ez a képlet $t_{\varepsilon/2}(n-1)$ helyett $u_{\varepsilon/2}$-el ismeretlen szórás esetén is alkalmazható, ha n "nagy" ($n \geq 30$), hiszen ekkor a korrigált empirikus szórás nagy pontossággal becsli a valódit.

HIPOTÉZISVIZSGÁLAT

Az alapproblémát a következő példán érzékeltetem. Vásárlói panaszok érkeznek, hogy egy élelmiszerboltban az 1 kg-os feliratú cukros zacskóban valójában kevesebb van. Szeretnénk korrekt módon kivizsgálni az ügyet. Kiszállunk az üzletbe, megmérünk n véletlenszerűen kiválasztott zacskót, $X_1, \dots, X_n$ a minta. Legyen $n = 25$, és a realizációban azt találjuk, hogy átlaguk 0.98 kg. Mit tegyünk? Az eltérést okozhatja a véletlen is, hiszen az 1 kg várható értékű, normális eloszlású mintaelemek eltérhetnek a várható értéktől. A következőképpen gondolkozunk: az ártatlanság vélelme alapján tegyük fel, hogy nem csalnak, vagyis a normális eloszlású háttérváltozó várható értéke valóban 1 kg. Szerkesszünk például 95%-os konfidenciaintervallumot a várható értékre a minta alapján! Amennyiben az 1kg hipotetikus várható érték nincsen benne ebben az intervallumban, akkor két eset lehetséges:

- Mivel az esetek 95%-ában a várható érték benne van ebben az intervallumban, a véletlen folytán lehet, hogy mégiscsak bekövetkezett az az 5% valószínűségű esemény, hogy nincsen benne.
- Nem igaz eredeti elképzelésünk, hogy 1 kg a várható érték.

Nagyon kis okunk van azt hinni, hogy bekövetkezett egy 5% valószínűségű esemény, inkább az utóbbi mellett voksolunk, hogy nem 1 kg a várható érték. Azaz 95%-os biztonsággal úgy döntünk, hogy csaltak. Ellenkező esetben, ha az 1 kg benne van a konfidenciaintervallunban, viszont 95%-os biztonsággal úgy döntünk, hogy nem csaltak. Lehet, hogy hibásan döntöttünk. Úgy is dönthettünk hibásan, hogy felmentettük a boltot a vád alól, holott az igaz volt. Vizsgáljuk meg a hibás döntések valószínűségét!

Fogalmazzuk meg a feladatot a következőképpen: a H_0 ún. *null-hipotézis* és a H_1 *alternatív hipotézis (ellen-hipotézis)* közt szeretnénk dönteni. Esetünkben az $X \sim \mathcal{N}(\mu, \sigma_0^2)$ háttérváltozó ismeretlen μ várható értékére vonatkoznak a hipotézisek (a σ_0 szórást most ismertnek vesszük).

$$H_0 : \mu = \mu_0 (= 1 \text{ kg}), \qquad H_1 : \mu \neq \mu_0.$$

(Valójában itt a $H_1 : \mu < \mu_0$ alternatívát kellene inkább vizsgálni, ezt egyoldali ellen-hipotézisnek nevezzük, és később tárgyaljuk.)

A döntést az $X_1, \dots, X_n$ független azonos eloszlású minta, illetve az ebből számolt

$$u = \frac{\bar{X} - \mu_0}{\sigma_0} \sqrt{n}$$

statisztika alapján hozzuk. Ettől függetlenül választunk egy $1 - \varepsilon$ szignifikanciaszintet (esetünkben $\varepsilon = 0.05$), és ehhez meghatározzuk az

$$u_{\varepsilon/2} = \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right)$$

ún *kritikus érték*et. A konfidenciaintervallumoknál tanultuk, hogy ez a standard normális eloszlás $1 - \varepsilon/2$ kvantilise. Azt is láttuk, hogy

$$\mathbb{P}_{\mu_0}\left(\mu_0 \in \left(\bar{X} - \frac{u_{\varepsilon/2}\sigma_0}{\sqrt{n}}, \bar{X} + \frac{u_{\varepsilon/2}\sigma_0}{\sqrt{n}}\right)\right) = \mathbb{P}_{\mu_0}\left(|u| < u_{\varepsilon/2}\right) = 1 - \varepsilon\,.$$

Tehát H_0 fennállása esetén μ_0 $1-\varepsilon$ valószínűséggel benne van a fenti, $\bar{X}$ körüli, szimmetrikus konfidenciaintervallumban. Ezzel ekvivalens, hogy $\bar{X}$ standardizáltjának, az u valószínűségi változónak az abszolút értéke kisebb, mint az $u_{\varepsilon/2}$ kritikus érték. Ezért az ún. *u-próba* a következő lépésekből áll:

1. A mintából kiszámoljuk az u próbastatisztikát.
2. Az adott $1-\varepsilon$ *szignifikancia-szint*hez táblázat alapján meghatározzuk az $u_{\varepsilon/2}$ küszöbértéket.
3. Döntünk: ha $|u| < u_{\varepsilon/2}$, akkor $1-\varepsilon$ szinten elfogadjuk H_0-t, az $|u| \geq u_{\varepsilon/2}$ esetben pedig elutasítjuk azt. Utóbbi esetben azt mondjuk, hogy a cukroszacskók tömege $(1-\varepsilon)100\%$-os szinten szignifikánsan eltér az 1 kg-tól.

Példánkban: $\bar{x} = 0.98$, $\mu_0 = 1$, $n = 25$ és legyen $\sigma_0 = 0.05$. Ekkor $u = -2$. Mivel 95%-os szignifikanciaszintnél $\varepsilon = 0.05$ és $u_{\varepsilon/2} = 1.96$, ezért 95%-os biztonsággal el kell utasítanunk a null-hipotézist, azaz megállapítjuk, hogy csaltak. 99%-os biztonság mellett ezt már nem tudjuk megtenni, ugyanis akkor $\varepsilon = 0.01$ és $u_{\varepsilon/2} = 2.58$, ezért 99%-os biztonsággal el kell fogadnunk a null-hipotézist. Ez nem meglepő, hiszen az intervallumbecsléseknél megállapítottuk, hogy a szignifikanciaszint növelése növeli a konfidenciaintervallum szélességét (a mintaelemszám növelése viszont csökkenti azt). Azt mondhatjuk tehát, hogy 95%-os biztonsággal állíthatjuk, hogy csaltak, de 99%-os biztonsággal már nem állíthatjuk ugyanezt. (Azaz a boltot "elsőfokon" elítélik, de egy szigorúbb bíróság "másodfokon" felmenti a vád alól. A szigorúság a vádlott érdekeit képviseli: minél kisebbé akarják tenni annak valószínűségégét – másodfokon ez 0.01 –, hogy ártatlanul elítéljék.)

A standard normális eloszlásfüggvény táblázatából kikereshető, hogy $\varepsilon = 0.0456$ esetén lenne $u_{\varepsilon/2} = 2$, azaz ez lenne az a legkisebb ε, ami mellett már, illetve 95.44% lenne az a legnagyobb biztonság, ami mellett még el tudnánk utasítani a null-hipotézist.

Döntésünkkor kétfajta hiba is felléphet:

I. *fajú hiba*: H_0 fennáll, mégis elutasítom.

II. *fajú hiba*: H_0 nem áll fenn, mégis elfogadom.

(A fenti példában I. fajú hibát követünk el, ha elítéljük az ártatlant, és II. fajút, ha felmentjük a bűnöst.)

Jelölje p_1 illetve p_2 az I. illetve II. fajú hiba valószínűségét. Nyilván

$$p_1 = \mathbb{P}_{\mu_0}\left(|u| \geq u_{\varepsilon/2}\right) = \varepsilon,$$

így ezt a fajta hibát uralni tudom a szignifikanciaszint megválasztásával. A másodfajú hiba azonban függ a valódi $\mu \neq \mu_0$ paraméterértéktől:

$$p_2 = \mathbb{P}_{\mu}\left(|u| < u_{\varepsilon/2}\right),$$

továbbá függ ε-tól és a mintaelemszámtól is.

Be lehet látni, hogy a

$$\beta_n(\mu, \varepsilon) = 1 - p_2 = \mathbb{P}_{\mu}\left(|u| \geq u_{\varepsilon/2}\right)$$

ún. *erőfüggvény* annál nagyobb, minél inkább eltávolodik μ a hipotetikus μ_0-tól, minél nagyobb n, illetve minél nagyobb ε. Az I. és II. fajú hiba tehát ellentétes

mozgású. A gyakorlat dönti el, mennyire érdemes kicsinek választani az uralható I. fajú hibát.

Mivel csak az elsőfajú hiba "uralható", a másodfajú változása pedig vele ellentétes, előbbit nem érdemes túlságosan leszorítani. Az is egy megoldás, hogy a H_0, H_1 szereposztást választjuk meg úgy, hogy a másodfajú hiba elkövetése ne legyen fatális, az első fajú hibáé legyen a súlyosabb vétség, ennek valószínűségét viszont tetszőlegesen kicsivé tudjuk tenni kellőképpen magas szignifikanciaszint választásával.

Például gyógyszer-hatásvizsgálatnál legyen

$$H_0: \text{ a gyógyszer hatástalan vagy káros}, \quad H_1: \text{ a gyógyszer hatásos}.$$

Ilyenkor az uralhatatlan másodfajú hiba azt jelenti, hogy egy hatásos gyógyszert nem vezetnek be, mert hatástalannak vagy károsnak minősítjük, ami azért nem okoz fatális problémákat. Az elsőfajú hiba – hogy hatásosnak minősítünk és bevezetünk egy hatástalan, netán káros készítményt – valószínűsége viszont kellően kicsivé tehető, például legyen $\varepsilon = 0.001$, így ennek bekövetkezése nagyon valószínűtlen. Általában is, az orvosi gyakorlatban a null-hipotézis gyakran a pejoratív verziót tartalmazza: nincsen hatása egy kezelésnek, egy klinikai mérésnek nincs diagnosztizáló hatása, stb., tehát örülünk, ha ezt el tudjuk utasítani minél magasabb szinten. Ezt különösen nem-paraméteres próbáknál tudjuk megtenni.

Más szituációban (paraméteres próbáknál) viszont inkább nagynak választjuk az elsőfajú hibát. Például egy szigorúan rögzített méretű alkatrész gyártásakor gyakran előfordul, hogy a gyártóberendezés kopása miatt a várható érték megváltozik (a szórás kicsi). Minőségellenőrzést végzünk arra vonatkozóan, hogy az alkatrésze megfelel-e a szabványnak. Ekkor a

$$H_0: \text{ a várható érték megegyezik a szabvány mérettel}, \quad H_1: \text{ nem egyezik meg}$$

hipotézisek közötti választásnál viszonylag nagy ε-t kell választanunk, ha szigorúak akarunk lenni: vállaljuk, hogy selejtnek minősítünk egy jó alkatrészt is, semmint véletlenül rosszat építsünk be.

Elterjedt az a gyakorlat, hogy nem adjuk meg előre ε-t, hanem nézzük, hogy mi az a legkisebb ε, amelyre $1-\varepsilon$ szignifikancia-szinten már el tudjuk utasítani a null-hipotézist. A felhasználó aztán eldönti, elég-e neki ekkora szignifikancia (a programcsomagok ezt a küszöb-ε-t írják ki, és néha ezt nevezik szignifikanciának). Amúgy, ha egy próba konzisztens, "kellően nagy" mintaelemszám esetén a másodfajú hiba tetszőlegesen kicsivé tehető, így ilyenkor nyugodtan magasra választhatjuk a szignifikanciaszintet.

Statisztikai próbák általános elméletéről csak annyit, hogy általában a mintateret kell felosztanunk egy elfogadási és egy kritikus tartományra (valamely statisztia kvantilis-értékei alapján) úgy, hogy az I. fajú hiba (vagy azok maximuma, amennyiben null-hipotézisünk összetett) adott ε legyen. Elég általános konstrukciók léteznek erre a felosztásra, melyek adott ε mellett az erőfüggvényt maximalizálják az ellenhipotézis bármely fennállása esetén.

A leggyakrabban használt *paraméteres és nemparaméteres próbák*at az órán kiosztott táblázatban foglaltuk össze. Paraméteres próbáknál a hipotézis a paraméterre

vonatkozik, míg a nemparaméteres próbák olyan kérdéseket vizsgálnak, hogy két minta azonos eloszlásból származik-e, független-e, stb. A táblázatban szereplő χ^2-próba mellett a Kolmogorov-Szmirnov tételeken alapuló Kolmogorov-Szmirnov próbák is használhatók. Vegyük észre, hogy a statisztikai próbák lényege: találunk egy statisztikát, melynek eloszlása megadható a null-hipotézis fennállása esetén. Ezután megnézzük, hogy a mintából kiszámolt ezen statisztika értéke mennyire tipikus ilyen eloszlás esetén. Ha nem az, akkor elutasítjuk, különben pedig elfogadjuk a null-hipotézist.

Ajánlott irodalom

- Bolla Marianna, Krámli András: Statisztikai következtetések elmélete. Typotex, Budapest, 2005.
- Reiman József: Valószínűségelmélet és matematikai statisztika mérnököknek. Tankönyvkiadó, Budapest, 1992.