Matematikai statisztika

A matematikai statisztikában két feladat fogalmazható meg:

1. Véletlen jelenség megértése
2. Véletlen jelenségek kapcsolatának megértése

A véletlen jelenség vagy ismert eloszláshoz tartozik és tartalmaz egy vagy több paramétert és ekkor ez (vagy ezek) meghatározása a feladat - ekkor paraméteres statisztikáról beszélünk -, ellenkező esetben nemparaméteres statisztikáról beszélünk.

A véletlen jelenség(ek) megismeréséhez n kisérlet adatai állnak rendelkezésre, ezt statisztikai mintának hívjuk.

Tipikus statisztikai kérdések

1. A mintában lévő selejtarány alapján következtetünk az egész sokaságbeli selejt-valószínűségre (becsléselmélet)
2. Statisztikai próbával megviszgáljuk: a mintaátlagnak a várható értéktől való eltérése alapján feltehető-e, hogy teljesül egy adott szabványelőírás (hipotézisvizsgálat)
3. Megvizsgáljuk, hogy a mintaátlag körül nagy valószínűséggel milyen intervallumban található az elméleti várható érték, ami az ún. megbízhatósági vagy konfidencia-intervallum keresése

1. Adott mintabeli eloszlásfüggvény milyen elméleti eloszlásfüggvényhez illeszkedik jól (tehát milyen eloszlású az egész sokaság), ez az ún. illeszkedésvizsgálat
2. Állítható-e két v.v.-ról. hogy egyforma eloszlású (homogenitásvizsgálat)
3. Mérési eredmények alapján próbáljuk eldönteni, hogy milyen kapcsolat áll fenn két v.v. között (regresszióanalízis)

Tipikus statisztikai feladatok

1. Pénzérme szabályosságának meghatározásához 100-szor dobok. Ebből 37 fejet kapok. Gondolhatom-e, hogy a kocka szabályos vagy van elég bizonyítékom ennek ellenkezőjére?
2. Kockával 600-szor dobok. Az eredmények: 86 db 1-es, 114 db 2-es, 121 db 3-as, 98 db 4-es, 91 db 5-ös, 90 db 6-os. Ez alapján gondolhatom-e, hogy a kocka szabályos?
3. Földrengésekről azt gondoljuk, hogy Poisson-eloszlást követnek. A legalább 2-es erősségű rengésekről az elmúlt 150 évből állnak rendelkezésre adatok 10 éves bontásban $(X_1, X_2, \ldots, X_{15})$. Mi a a valószínűsége, hogy egy 10 éves periodusban 3 db legalább 2 erősségű földrengés lesz?

1. Törőkocka szilárdságára vonatkozó adatai normális eloszlás követnek. A normális eloszlás paramétereinek: (m, σ) meghatározásához 10 db kisérletet végzünk. Tippeljük meg az (m, σ) paramétereket! Mondjuk meg 95%-os biztonsággal, hogy milyen értéket vesz fel (m, σ)!
2. Ismerjük a Duna bécsi és budapesti vízállására vonatkozó adatokat. Milyen kapcsolat van a kettő között? Tippeljük meg a mai bécsi érték alapján a holnaputáni budapesti vízállást!
3. Hagyományos beton szilárdságára rendelkezésre állnak adatok. Egy újítás során azt kell eldöntenünk, hogy szilárdabb betont kapunk-e. Ehhez az új betonból is törőkockákat készítenek és meghatározzák ezek adatait. Összevetve a korábbi adatokkal hogy döntsünk? Sikeres a javulás, vagy a kis javulás csak a véletlennek tulajdonítható?

Statisztikai alapfogalmak

A statisztikában a megérteni akart X valószínűségi változót <u>háttérváltozónak</u> hívjuk, ennek eloszlása a háttéreloszlás ($F(x) = P(X < x)$). Ennek megismeréséhez n db független kisérletet végzünk így jutunk az ún. <u>statisztikai mintá</u>hoz: $X_1, X_2, \ldots, X_n$, ahol tehát X_i egy valószínűségi változó, melyenek eloszlása megegyezik X háttérváltozó eloszlásával. A jelölés:

$$\underline{X} = (X_1, X_2, \ldots, X_n)$$

egy konkrét megvalósulása, realizációja

$$\underline{x} = (x_1, x_2, \ldots, x_n),$$

ahol tehát $X_i = x_i$

Példák:

1. Beton szilárdságának megismeréséhez 10 db törőkocka szilárdsága.
2. Az egyetemista fiúk magasságának meghatározásához 40 egyetemista fiú megmérése.

A korábbi $(X_1, X_2, \ldots, X_n)$ mintát nagyság szerint sorba rendezzük:

$$X_1^* \leq X_2^* \leq \cdots \leq X_n^*$$

így jutunk az ún. rendezett mintához.

Példa: Ha az *RND* (azaz a $[0,1]$-ben egyenletes eloszlás) esetén veszünk egy $X_1, X_2, \ldots, X_{10}$ mintát, akkor a rendezés után

$$X_1^* \leq X_2^* \leq \cdots \leq X_{10}^*$$

mintát kapjuk. Van értelme arról beszélni, hogy pl. X_1^*-nak vagy X_{10}^*-nak mi az eloszlása. Ekkor X_{10}^* és X_1^* eloszlását a következő képpen kapjuk: legyen $0 < x < 1$

$$P(X_{10}^* < x) = P(X_1 < x, X_2 < x, \ldots, X_{10} < x) =$$

$$P(X_1 < x)P(X_2 < x)\ldots P(X_{10} < x) = x^{10}$$

$$P(X_1^* < x) = 1 - P(X_1^* > x) = 1 - P(X_1 > x, X_2 > x, \ldots, X_{10} > x) =$$

$$1 - P(X_1 > x)P(X_2 > x)\ldots P(X_{10} > x) = 1 - (1-x)^{10}$$

A rendezett $(X_1^*, X_2^*, \ldots, X_n^*)$ minta esetén az $X_n^* - X_1^*$ a minta terjedelme.

Empirikus medián $n = 2k + 1$ esetén: X_{k+1}^* és $n = 2k$ esetén $\frac{X_k^* + X_{k+1}^*}{2}$.

Empirikus vagy tapasztalati várható érték az átlag:

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

Empirikus vagy tapasztalati szórásnégyzet:

$$S^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n}.$$

Ekkor

$$E(\overline{X}) = E(\frac{X_1 + \cdots + X_n}{n}) = \frac{1}{n}E(X_1 + \cdots + X_n) =$$

$$\frac{1}{n}(E(X_1) + E(X_2) + \cdots + E(X_n)) = \frac{1}{n}nE(x) = E(X)$$

Megmutatható, hogy

$$E(S^2) = \frac{n-1}{n}D^2(X),$$

ezért a továbbiakban az ún
korrigált empirikus szórásnégyzettel:

$$S^{*2} = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}$$

számolunk, mert erre

$$E(S^{*2}) = D^2(X),$$

Az $(X_1, X_2, \ldots, X_n)$ minta alapján beszélhetünk tapasztalati eloszálásfüggvényről:

$$F_n^*(x) = \begin{cases} 0 & \text{ha } x \leq X_1^* \\ \frac{k}{n} & \text{ha } X_k^* < x \leq X_{k+1}^* \\ 1 & \text{ha } x > X_n^* \end{cases}$$

Megmutatható, hogy a fenti tapasztalti eloszlásfüggvény nagy n esetén jól közelíti a háttéreloszást ($F(x)$-et). Ez pontosan a matematikai statitisztika alaptételének tartott Glivenko-Cantelli tétel: Minden $x \in \mathbb{R}$ esetén egyszerre teljesül, hogy

$$P(F_n(x) \to F(x)) = 1$$

Folytonos háttérváltozó esetén természetes feladat a minta alapján a sűrűségfüggvény közelítése. Ezt az ún. hisztogrammal érhetjük le: befoglaljuk a mintát egy $[a, b]$ intervallumba, ezt felbontjuk h hosszú egyenlő részre és egy intervallum felé $\frac{k}{nh}$ nagyságú téglalapot rajzolunk, ha az adott intervallumba k db mintelem esik.

Megmutatható, hogy ha n nagy és h pici, továbbá nh nagy, akkor a fenti hisztogram jól közelíti a sűrűségfüggvényt.

Becsléselmélet

A feladat az, hogy adott egy X háttérváltozó és a hozzá tartozó $(X_1, X_2, \ldots, X_n)$ n elemű minta. Tudjuk, hogy az X háttérváltozó milyen eloszlásba tartozik, csak a pontos paramétereket nem ismerjük. A mintából szeretnénk a paraméter(ek)re információt nyerni.

Jelölje θ a paramétert vagy a paramétereket.

Példa

1. X Poisson eloszású, akkor $\theta = \lambda$ (egydimenziós paraméter)
2. X normális eloszlású, akkor $\theta = (m, \sigma)$ (kétdimenziós paraméter).

Pontbecslés

Feladat: Tippeljük meg a paraméter(eke)t egy értékkel, tehát adjunk meg egy

$$\widehat{\theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$$

értéket, ami közelíti a θ paramétert!

Erre több módszert is használnak (pl. Bayes-becslés, momentum-becslés), most csak a legegyszerűbbet az ún. maximum-likelihood becslést ismertetjük.

Maximum likelihood becslés

Legyen X egy diszkrét v.v., ekkor az $(x_1, \ldots, x_n)$ minta valószínűsége

$$P_\theta(X = x_1)P_\theta(X = x_2)\ldots P_\theta(X = x_n)$$

A feladat olyan θ választása, ahol a fenti szorzat maximális.

Ha X folytonos v.v, akkor a sűrűségfüggvények szorzatát szeretnénk maximalizálni θ-ban:

$$f_\theta(x_1)f_\theta(x_2)\ldots f_\theta(x_n)$$

Ehhez bevezetjük az ún. likelihood függvényt, ami diszkrét esetben:

$$L_\theta(x_1, \ldots, x_n) = L_\theta(\underline{x}) = P_\theta(X = x_1)P_\theta(X = x_2)\ldots P_\theta(X = x_n)$$

és folytonos esetben

$$L_\theta(x_1, \ldots, x_n) = L_\theta(\underline{x}) = f_\theta(x_1)f_\theta(x_2)\ldots f_\theta(x_n)$$

A feladat annak meghatározása, hogy az $L_\theta(\underline{x})$ likelihood-függvény milyen θ esetén lesz maximális.

Ennek meghatározására gyakran célszerű venni a likelihood-függvény logaritmusát az ún log-likelihood függvényt:

$$l_\theta(\underline{x}) = ln(L_\theta(\underline{x}))$$

Feladat: Legyen $X_1, X_2, \ldots, X_n$ egy (N, p) paraméterű binomiális eloszlásból származó minta. Adjon ML-becslést!

Megoldás: Ekkor

$$L_p(\underline{x}) = P_p(X = x_1)P_p(X = x_2)\ldots P_p(X = x_n) =$$

$$\binom{N}{x_1} p^{x_1}(1-p)^{N-x_1} \binom{N}{x_2} p^{x_2}(1-p)^{N-x_2} \ldots \binom{N}{x_n} p^{x_1}(1-p)^{N-x_n} =$$

$$Cp^{\sum_{k=1}^{n} x_k}(1-p)^{Nn-\sum_{k=1}^{n} x_k}$$

$$l_p(\underline{x}) = ln(L_p(\underline{x})) = c + (\sum_{k=1}^{n} x_k) \ln p + (Nn - \sum_{k=1}^{n} x_k) \ln(1-p)$$

Ennek szélsőértéket p-szerinti deriválással kapjuk meg:

$$\frac{d}{dp} l_p(\underline{x}) = \frac{\sum_{k=1}^{n} x_k}{p} - \frac{Nn - \sum_{k=1}^{n} x_k}{1-p} = 0$$

$$\widehat{p} = \frac{\sum_{k=1}^{n} x_k}{nN}$$

Feladat: Legyen $X_1, X_2, \ldots, X_n$ egy λ paraméterű Poisson-eloszlásból származó minta. Adjon ML-becslést!

Megoldás: Ekkor

$$L_\lambda(\underline{x}) = P_\lambda(X = x_1)P_\lambda(X = x_2)\ldots P_\lambda(X = x_n) =$$

$$e^{-\lambda}\frac{\lambda^{x_1}}{x_1!}e^{-\lambda}\frac{\lambda^{x_2}}{x_2!}\ldots e^{-\lambda}\frac{\lambda^{x_n}}{x_n!} =$$

$$Ce^{-n\lambda}\lambda^{\sum_{k=1}^{n} x_k}$$

$$l_\lambda(\underline{x}) = ln(L_\lambda(\underline{x})) = c - n\lambda + (\sum_{k=1}^{n} x_k)\ln\lambda$$

Ennek szélsőértéket λ-szerinti deriválással kapjuk meg:

$$\frac{d}{d\lambda}l_p(\underline{x}) = -n + \frac{\sum_{k=1}^{n} x_k}{\lambda} = 0$$

$$\widehat{\lambda} = \frac{\sum_{k=1}^{n} x_k}{n} = \overline{x}$$

Feladat: Legyen $X_1, X_2, \ldots, X_n$ egy λ paraméterű exponenciális eloszlásból származó minta. Adjon ML-becslést!

Megoldás: Ekkor

$$L_\lambda(\underline{x}) = f_\lambda(x_1) f_\lambda(x_2) \ldots f_\lambda(x_n) =$$

$$\lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2} \ldots \lambda e^{-\lambda x_n}$$

$$\lambda^n e^{-\lambda \sum_{k=1}^n x_k}$$

$$l_\lambda(\underline{x}) = ln(L_\lambda(\underline{x})) = -\lambda \sum_{k=1}^n x_k + n \ln \lambda$$

Ennek szélsőértéket λ-szerinti deriválással kapjuk meg:

$$\frac{d}{d\lambda} l_\lambda(\underline{x}) = -\sum_{k=1}^n x_k + \frac{n}{\lambda} = 0$$

$$\widehat{\lambda} = \frac{n}{\sum_{k=1}^n x_k} = \frac{1}{\overline{x}}$$

Feladat: Legyen $X_1, X_2, \ldots, X_n$ egy (m, σ) paraméterű normális eloszlásból származó minta. Adjon ML-becslést!

Megoldás: Ekkor

$$L_{(m,\sigma)}(\underline{x}) = f_{(m,\sigma)}(x_1) f_{(m,\sigma)}(x_2) \ldots f_{(m,\sigma)}(x_n) =$$

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_1-m)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_2-m)^2}{2\sigma^2}} \ldots \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_n-m)^2}{2\sigma^2}}$$

$$\frac{C}{\sigma^n} e^{-\frac{\sum_{k=1}^n (x_k-m)^2}{2\sigma^2}}$$

$$l_{(m,\sigma)}(\underline{x}) = c - n \ln \sigma - \frac{\sum_{k=1}^n (x_k - m)^2}{2\sigma^2}$$

A szélsőérték ott lesz, ahol a parciális deriváltak 0-val egyenlő:

$$\frac{\partial}{\partial m} l_{(m,\sigma)}(\underline{x}) = \frac{\sum_{k=1}^n (x_k - m)}{\sigma^2} = 0$$

$$\widehat{m} = \frac{\sum_{k=1}^n x_k}{n} = \overline{x}$$

$$\frac{\partial}{\partial \sigma} l_{(m,\sigma)}(\underline{x}) = -\frac{n}{\sigma} + \frac{\sum_{k=1}^{n}(x_k - m)^2}{\sigma^3} = 0$$

$$\sigma^2 = \frac{\sum_{k=1}^{n}(x_k - m)^2}{n}$$

$$\widehat{\sigma} = \sqrt{\frac{\sum_{k=1}^{n}(x_k - \overline{x})^2}{n}} = S$$

Feladat: Legyen $X_1, X_2, \ldots, X_n$ egy $(0, \theta)$ egyenletes eloszlásból származó minta. Adjon ML-becslést θ-ra!

Megoldás: Ekkor

$$L_\theta(\underline{x}) = f_\theta(x_1) f_\theta(x_2) \ldots f_\theta(x_n) = \frac{1}{\theta^n},$$

ha minden x_i esetén $0 \leq x_i \leq \theta$. Ennek maximuma a

$$\widehat{\theta} = X_n^*$$

esetén adódik.