Valószínűségszámítás B

11. előadás

Tóth Dávid (BME SZIT)

2023. május 15.

Statisztikai alapfogalmak

A statisztika (mint tudomány) célja:

- kísérleti megfigyelések eredményei alapján, a valószínűségszámítás eszközeit felhasználva minél többet feltárni a háttérben zajló véletlen jelenség természetéről,
- ezek alapján segíteni például a kockázatelemzést vagy jövőbeni döntések meghozalatalát.

Statisztikai alapfogalmak

Kiindulópont:

- valamilyen megfigyelés,
- a megfigyelés tárgyát képező egyedek összességét *statisztikai sokaságnak* nevezzük,
- ez lehet pl. egy egyetem hallgatóinak halmaza vagy egy terület időjárása egyes napokon, stb.
- a sokaságot alkotó egyedeket tipikusan vagy véletlenszerűen választjuk, vagy pedig előfordulásuk számos tényező által befolyásolt és így véletlen jelenségnek tekinthető,

$\Rightarrow$ az egész sokaságot a statisztikában egy valószínűségi mező, tehát az egyedek által alkotott Ω eseménytér, az azon kijelölt események, illetve egy valószínűségi mérték modellezi.

Statisztikai alapfogalmak

Statisztikai ismérv:

- a statisztikai sokaság egyedeire vonatkozó tulajdonság,
- ezek különféle kategóriákba sorolhatók attól függően, hogy milyen típusú jellemzőit írják le az egyedeknek,
- szokás például időbeli, területi, minőségi, mennyiségi, stb. ismérvekről beszélni,
- ezen tulajdonságokat a sokaságot modellező valószínűségi mezőn értelmezett valószínűségi változóként kezelhetjük,

Statisztikai alapfogalmak

Statisztikai ismérv:

- a mennyiségi ismérveket leíró változók értlemezése általában kézenfekvő (pl. egyedek magasassága egy populációban, egy terület napi középhőmérséklete, stb.),
- a modell a minőségi jellemzők leírásánál sem jelent korlátot: ekkor az egyedeket különféle kategóriákba soroljuk, melyek jelölhetők számokkal is, ún. kategorikus valószínűségi változókat adva,
- pl. egy populáció egyedeinek szemszínét a kék, zöld, barna, stb. kategóriákba sorolhatjuk, ezeket az $1, 2, 3, \ldots$ számokkal azonosítva egy $X : \Omega \to \mathbb{R}$ valószínűségi változót kapunk,
- mostantól kizárólag mennyiségi ismérveket modellező változókkal foglalkozunk.

Statisztikai alapfogalmak

Matematikai alapprobléma:

- adott egy valószínűségi mező és egy azon értelmezett valószínűségi változó, amelynek eloszlása (az ún. *háttéreloszlás*) nem ismert,
- feladat: ennek (vagy az eloszlás jellemzőinek) meghatározása,
- sokszor előfordul, hogy a háttéreloszlás egy adott eloszláscsalád tagjának tekinthető (pl. exponenciális, normális), ilyenkor a feladat az eloszlás paraméterének meghatározására redukálódik.

Statisztikai alapfogalmak

Definíció. Egy eseménytér, az azon kijelölt események, továbbá valószínűségi mértékek egy családja együttesen *statisztikai mezőt* alkot, ha a mértékcsalád bármely tagja az eseménytérrel és az eseményekkel együtt egy valószínűségi mezőt ad meg.

Ha a mértékcsalád tagjai valós paraméterekkel paramétezhetők, akkor *paraméteres statisztikai mezőről* beszélünk.
A paramétereket tipikusan egy valós vektorként adjuk meg, k különböző paraméter esetében tehát ezek $\mathbb{R}^k$ elemei.
A mértékcsalád tagjait leíró lehetséges paraméterek $\Theta \subset \mathbb{R}^k$ halmazát *paramétertérnek* nevezzük.

Statisztikai alapfogalmak

Statisztikai minta:

- egy teljes sokaságot tipikusan nem tudunk áttekinteni, ezért a fenti feladat megoldásához a gyakorlatban egy ún. *statisztikai mintát* használunk,
- ez a sokaság egyedeinek egy részhalmazából és az ezen egyedekhez tartozó jellemzőkből áll (pl. egy csoport egyes tagjainak magassága),
- egymástól független mérések vagy az egyedek egymástól független választásai adják,
- ha X jelöli a *háttérváltozót*, aminek eloszlása a háttéreloszlás, akkor egy n elemű minta elemeit az $X_1, \ldots, X_n$ együttesen független valószínűségi változókkal modellezzük, amelyek eloszlása azonos az X eloszlásával.

Statisztikai alapfogalmak

Minta realizációja:

- egy konkrét mintánál számadatokkal dolgozunk, nem pedig valószínűségi változókkal,
- egy ilyen konkrét mérés eredményéből származó adatsor a minta *realizációja*,
- az elméleti modellben a mintavételt változókkal írjuk le, és így egy realizáció ezen változók egy kiértékelését jelenti,
- ez garantálja, hogy az elméleti modellből kapott eredmények minden egyes realizációra érvényesek legyenek,
- a minta egy realizációját néha (az egyszerűség kedvéért pongyolán) szintén mintának nevezzük.

Statisztikai alapfogalmak

Különböző realizációk különböző eredményeket szolgáltatnak.

$\Rightarrow$ Nem mondhatjuk, hogy a háttéreloszlást egyik vagy másik kísérletsorozat eredménye adja.

$\Rightarrow$ Véletelen eloszlásokra vagy véletlen menynyiségekre általában nem tudunk a minták alapján pontosan következtetni.

$\Rightarrow$ A statisztikai következtetések (bár logikai következtetések, de) természetükből adódóan becslések.

- A becslések hibáját is kezelni kell, ez a hiba pedig kontrollálható.
- A teljes bizonyosságot itt is fel kell áldoznunk, de előírhatjuk, hogy a kapott eredményünk nagy valószínűséggel egy bizonyos hibahatáron belül közelítse a keresett értéket.
- Példákon keresztül látni fogjuk, hogy hogyan érhetünk el pontosabb becslést, vagy hogy milyen mértékben kell feláldoznunk a pontosságot a nagyobb bizonyosságért cserébe.

Statisztikai alapfogalmak

A minta általában kis elemszámű a sokaság elemszámához mérten.

- Nem feltétlenül van lehetőségünk vagy kapacitásunk a teljes sokaság áttekintésére.
- De arra is ügyelni kell, hogy ez az elemszám elegendően nagy legyen ahhoz, hogy a kapott eredményt kellően megalapozottnak tekinthessük.
- Előfordulhat, hogy nincs lehetőségünk elég nagy mintával dolgozni, ezért fontos, hogy az eredményeink az elemszám függvényében adjanak becslést a pontosságra.
- A mintának *reprezentatívnak* kell lennie abban az értelemben, hogy a vizsgálat tárgyát képező sokasághoz szerkezetében, főbb jellemzőiben hasonlatosnak kell lennie.

Statisztikai alapfogalmak

Rendezett minta:

- A mintából kapott adatsort sokszor érdemes úgy átalakítani, hogy az praktikus legyen a feldolgozásnál.
- Az alább definiált taptasztalati eloszlásfüggvény kiszámolásához például érdemes a minta elemeit nagyság szerint növekvő sorrendbe rendezni.
- Így kapjuk az ún. *rendezett mintát*.
- Elemeit egy kiindulásul szolágó $X_1, \ldots, X_n$ minta esetén $X_1^*, \ldots, X_n^*$ jelöli, ahol az X_i^* változók értékeinek halmaza megegyezik az X_i-k értékeinek halmazával (egy fix realizációnál), továbbá $X_1^* \leq X_2^* \leq \cdots \leq X_n^*$ teljesül.
- Pl. ha a minta a 3, 2, 5, 2 számokból áll, akkor a rendezett minta 2, 2, 3, 5.

Statisztikai alapfogalmak

Definíció. Legyen $X_1, \ldots, X_n$ egy független, azonos eloszlású minta, ekkor a mintához tartozó F_n^* *tapasztalalti (vagy empirikus) eloszlásfüggvény* értéke egy $t \in \mathbb{R}$ helyen

$$F_n^*(t) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{X_i < t\}}}{n}.$$

A tapasztalati eloszlásfüggvény t helyen vett értékének meghatározásához tehát azt kell megszámolni, hogy a minta hány elemének értéke esik t alá, ennek a számnak és a minta elemszámának az aránya adja a függvény értékét.

Statisztikai alapfogalmak

Ha pontosan i darab érték esik t alá, az azt jelenti, hogy $X_i^* < t$, de $X_{i+1}^* \geq t$. Külön kezelendő az az eset, amikor már a legkisebb érték is legalább t, illetve amikor a legnagyobb érték is kisebb t-nél.

$$F_n^*(t) = \begin{cases} 0, & \text{ha } t \leq X_1^*, \\ \dfrac{i}{n}, & \text{ha } X_i^* < t \leq X_{i+1}^* \qquad (1 \leq i \leq n-1), \\ 1, & \text{ha } t > X_n^*. \end{cases}$$

Statisztikai alapfogalmak

Megjegyzés. A tapasztalati eloszlásfüggvény egy valószínűségi változó, de a minta egy konkrét realizációját behelyettesítve már egy $\mathbb{R} \to \mathbb{R}$ függvényt kapunk.

Példa. Ötször dobunk egy dobókockával, jelölje a dobások eredményét $X_1, \ldots, X_5$. (Ezek független, azonos eloszlású valószínűségi változók, amelyek egy 5 elemű mintát alkotnak.) Tegyük fel, hogy egy konkrét kísérlet során a 3, 1, 5, 5, 1 eredmények adódnak. (Ez a minta egy realizációja.)

Erre a realizációra a rendezett minta értékei:

$$X_1^* = 1, \qquad X_2^* = 1, \qquad X_3^* = 3, \qquad X_4^* = 5, \qquad X_5^* = 5.$$

Statisztikai alapfogalmak

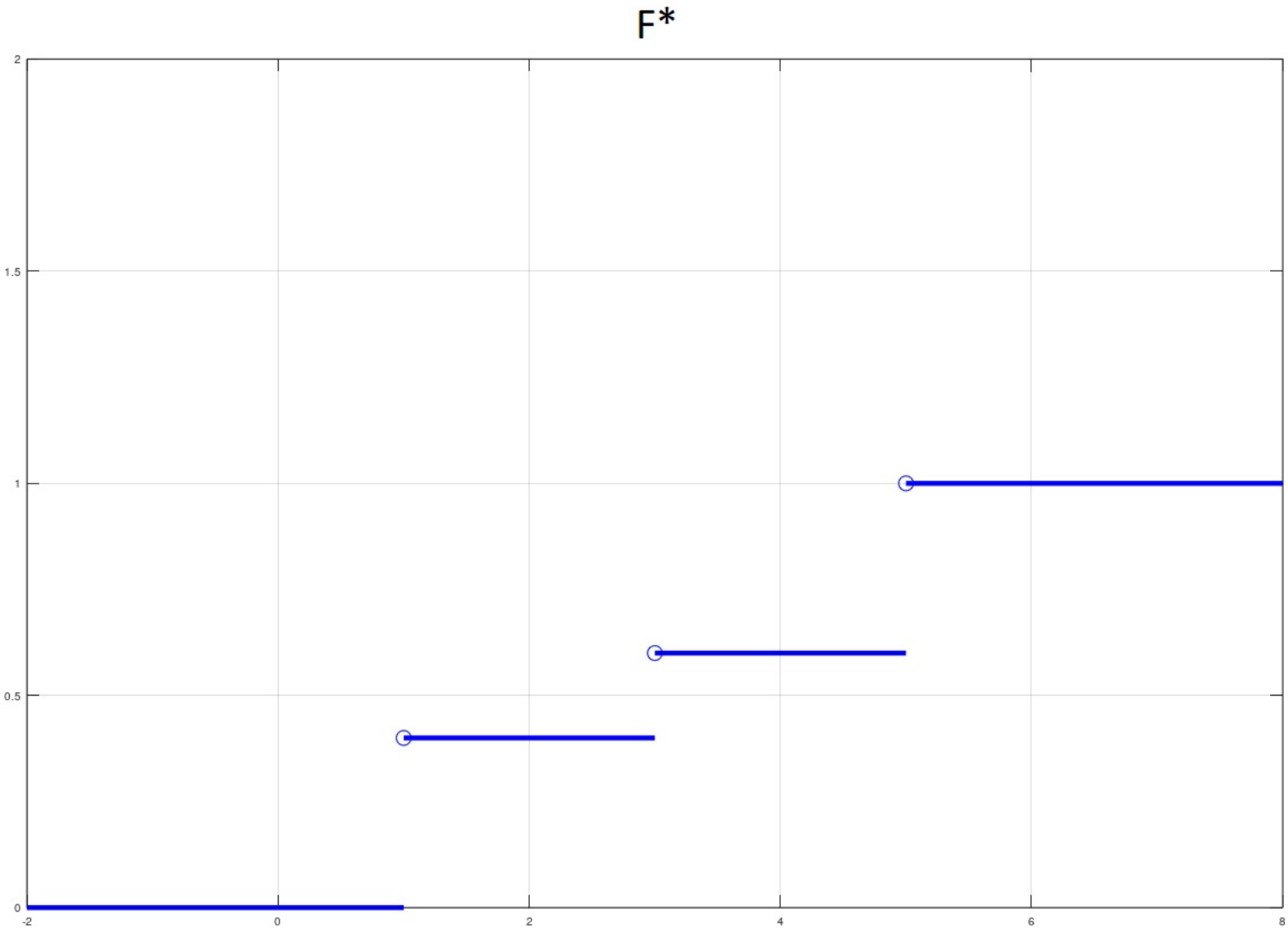
Példa. Ötször dobunk egy dobókockával, a rendezett minta értékei:

$$X_1^* = 1, \qquad X_2^* = 1, \qquad X_3^* = 3, \qquad X_4^* = 5, \qquad X_5^* = 5.$$

Az eloszlásfüggvény tehát erre a realizációra a következő:

$$F_5^*(t) = \begin{cases} 0, & \text{ha } t \leq 1, \\ \frac{2}{5}, & \text{ha } 1 < t \leq 3, \\ \frac{3}{5}, & \text{ha } 3 < t \leq 5, \\ 1, & \text{ha } t > 5. \end{cases}$$

Statisztikai alapfogalmak

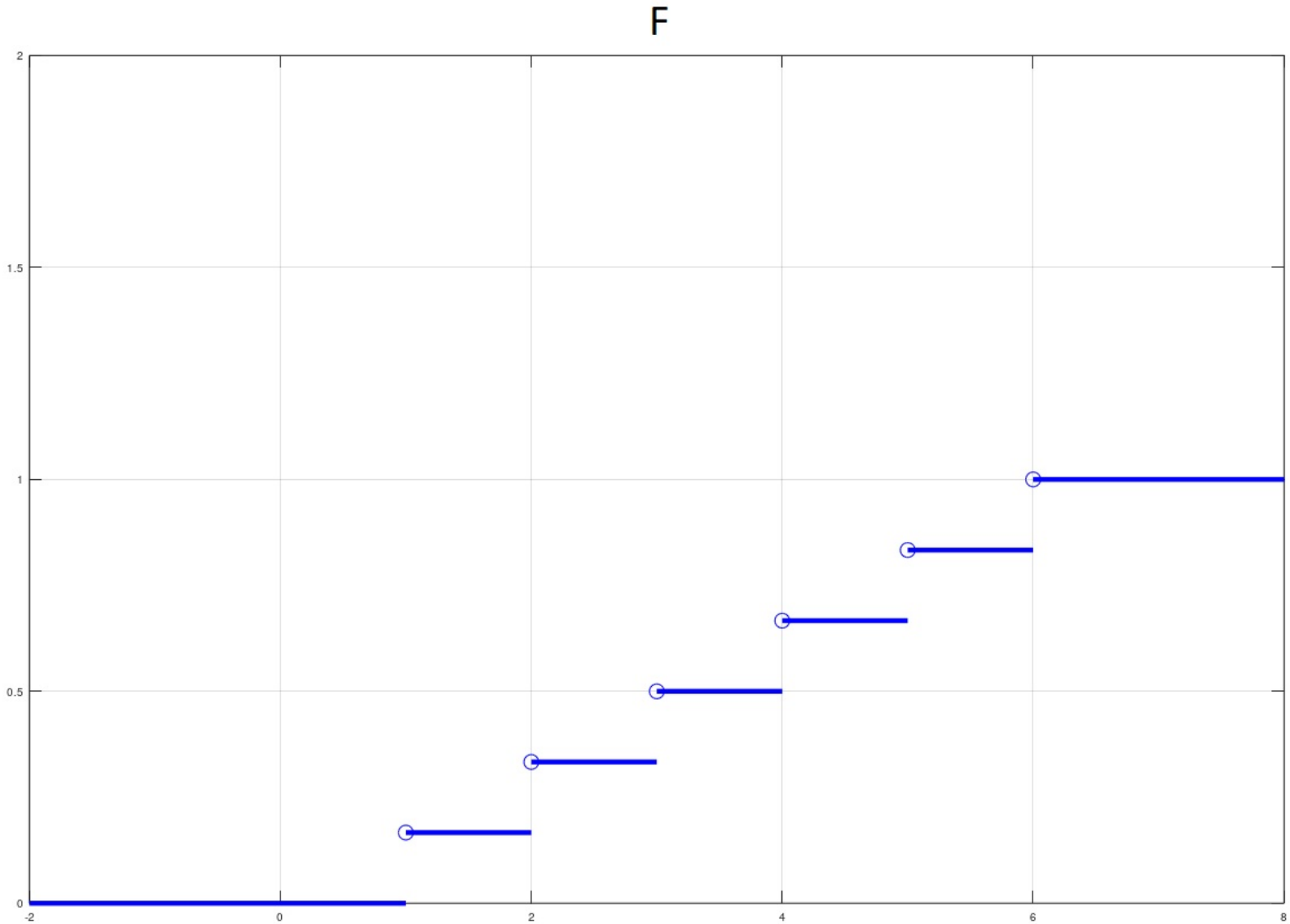


Statisztikai alapfogalmak

Ha a kocka szabályos, akkor magának a háttéreloszlásnak az eloszlásfüggvénye

$$F(t) = \begin{cases} 0, & \text{ha } t \leq 1, \\ \dfrac{i}{6}, & \text{ha } i < t \leq i+1 \qquad (1 \leq i \leq 5), \\ 1, & \text{ha } t > 6. \end{cases}$$

Statisztikai alapfogalmak



Statisztikai alapfogalmak

Kis mintaelemszám esetén a tapasztalati és a tényleges eloszlásfüggvény közt nagy különbség lehet, de ez a különbség a mintaelemszám növelésével fokozatosan csökken: a tapasztalati eloszlásfüggvény értéke az eloszlásfüggvény értékéhez tart.

Állítás. Minden $t \in \mathbb{R}$ esetén $\mathbb{P}\big(\lim_{n\to\infty} F_n^*(t) = F(t)\big) = 1$, ahol F_n^* a tapasztalati eloszlásfüggvény, F pedig a háttéreloszlás eloszlásfüggvénye.

Statisztikai alapfogalmak

Bizonyítás.

- Az $F_n^*(t)$ valószínűségi változó nem más, mint az $\mathbb{1}_{\{X_i<t\}}$ indikátorok átlaga.
- Mivel az X_i-k azonos eloszlásúak, és eloszlásuk megegyezik az X háttérváltozó eloszlásával, így

$$\mathbb{P}\big(\mathbb{1}_{\{X_i<t\}}=1\big)=\mathbb{P}(X_i<t)=\mathbb{P}(X<t),$$
$$\mathbb{P}\big(\mathbb{1}_{\{X_i<t\}}=0\big)=1-\mathbb{P}(X_i<1)=1-\mathbb{P}(X<t),$$

  azaz a fenti indikátorváltozók is azonos eloszlásúak.
- Továbbá az X_i-k együttes függetlensége miatt a fenit események és így ezek az indikátorok is együttesen függetlenek,
- valamint a szórásuk is véges.

Statisztikai alapfogalmak

Bizonyítás.

$\Rightarrow$ Alkalmazható az indikátorokra a nagy számok erős törvénye: az átlaguk 1 valószínűséggel tart a (közös) várható értékükhöz, és

$$\mathbb{E}(\mathbb{1}_{\{X_i<t\}}) = \mathbb{P}(X_i < t) = \mathbb{P}(X < t) = F(t).$$

Statisztikai alapfogalmak

A fenti állításnál több is igaz:

- látjuk, hogy minden egyes t-re a tapasztalati eloszlásfüggvény és a háttéreloszlás eloszlásfüggvénye közel lesz egymáshoz, amennyiben a minta elemszáma nagy,
- de hogy milyen nagynak kell választani az elemszámot, az elvben függhetne a t konkrét értékétől,
- ez azonban nem így van, a fenti konvergencia egyenletes, ez a *statisztika alaptételének* állítása:

Statisztikai alapfogalmak

Tétel. (Glivenko-Cantelli, a statisztika alaptétele)
Az F_n^* tapasztalati eloszlásfüggvény 1 valószínűséggel egyenletesen tart a háttéreloszlás F eloszlásfüggvényéhez.
Azaz: ha a minta elelmszáma elég nagy, akkor F_n^* értéke 1 valószínűséggel tetszőlegesen közel kerül egyenletesen (azaz egyszerre minden $t \in \mathbb{R}$-re) az F-hez.

Formalizálva:

$$\sup_{t \in \mathbb{R}} |F_n^*(t) - F(t)| \to 0, \qquad (n \to \infty)$$

teljesül 1 valószínűséggel.

Pontbecslések

- A továbbiakban paraméteres statisztikai mezőket tekintünk.
- Ha adott egy $X_1, \ldots, X_n$ minta, akkor ennek segítségével szeretnénk a háttéreloszlás paramétereit (vagy esetleg annak függvényeit) becsülni.
- Ehhez n elemű minta esetén egy $T : \mathbb{R}^n \to \mathbb{R}$ függvényt fogunk használni (amely mindig olyan lesz, hogy $T(X_1, \ldots, X_n)$ is egy valószínűségi változó).
- Ekkor a mintaelemek $T(X_1, \ldots, X_n)$ függvényét (is) *statisztikának* nevezzük.
- Az későbbiekben tipikusan az elemszámot bármilyen nagynak választhatjuk, így a T függvényt minden n-re definiálni kell.
- Megjegyezzük, hogy egy konkrét $x_1, \ldots, x_n$ realizációra $T(x_1, \ldots, x_n)$ is egy számértéket ad.

Pontbecslések

A leggyakrabban használt statisztikák közé tartozik pl. a mintaátlag, a tapasztalati szórás(négyzet) ill. a korrigált tapasztalati szórás(négyzet), melyekkel részletesen is megismerkedünk.

Definíció. Legyen $X_1, \ldots, X_n$ független, azonos eloszlású n elemű minta, ekkor az

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

statisztikát *mintaátlagnak* nevezzük.

Egy konkrét realizáció esetén a mintaátlagot $\overline{x}$ fogja jelölni. Ha hangsúlyozni szeretnénk az elemszámot, akkor $\overline{X}_n$-t ill. $\overline{x}_n$-t írunk.

Az korbbi példában a kockadobások átlaga

$$\frac{3+1+5+5+1}{5} = 3.$$

Pontbecslések

- Egy valószínűségi változó várható értéke a változó átlagos értékét hivatott jellemezni, így kézenfekvőnek tűnik a háttéreloszlás várható értékét a mintaátlaggal becsülni.
- Mennyire lesz "jó" ez a becslés?
- Egy becslés jóságát különböző kritériumokkal mérhetjük, amelyből itt most egyet tárgyalunk részletesen.

Pontbecslések

Definíció. A $T(X_1, \ldots, X_n)$ statisztika *torzítatlan becslés* a θ paraméterre, ha

$$\mathbb{E}(T(X_1, \ldots, X_n)) = \theta$$

teljesül.

Állítás. A mintaátlag torzítatlan becslés a háttéreloszlás várható értékére (amennyiben az véges).

Bizonyítás. A várható érték linearitása miatt

$$\mathbb{E}(\overline{X}_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{1}{n} \cdot n \cdot \mathbb{E}(X) = \mathbb{E}(X),$$

ahol X a háttérváltozó.

Pontbecslések

- Tehát a mintaátlag átlagosan jól viselkedik, de felmerül a kérdés, hogy mennyire lesznek közel a várható értékhez a tényleges értékek.
- Ha az X háttérváltozó szórásnégyzete véges, akkor a minta függetlensége miatt

$$\mathbb{D}^2(\overline{X}) = \mathbb{D}^2\left(\frac{X_1 + \cdots + X_n}{n}\right)$$
$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{D}^2(X_i) = \frac{1}{n}\mathbb{D}^2(X),$$

  tehát a mintaátlag szórása 0-hoz tart, ha n tart végtelenhez.
- Ez épp azt jelenti, hogy a tényleges értékek a várható érték körül koncentrálódnak, ha az elemszám nagy.

Pontbecslések

A szórásnégyzet lényegében a várható értéktől való átlagos négyzetes eltérés, így logikusnak tűnhet ezt az mintaátlagtól való átlagos négyzetes eltéréssel becsülni.

Definíció. Legyen $X_1, \ldots, X_n$ független, azonos eloszlású n elemű minta, ekkor az

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

statisztikát a minta *tapasztalati szórásnégyzetének* nevezzük. Ennek S gyöke a *tapasztalati szórás*.

Egy konkrét realizáció esetén a tapasztalati szórásnégyzetet s^2 fogja jelölni. Ha hangsúlyozni szeretnénk az elemszámot, akkor S_n^2-et ill. s_n^2-et írunk.

Pontbecslések

A gyakorlatban sokszor hasznosabb a fenti formulát átalakítani:

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

$$= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\overline{X} \cdot \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n} \cdot n \cdot \overline{X}^2$$

$$= \overline{X^2} - 2\overline{X}^2 + \overline{X}^2 = \overline{X^2} - \overline{X}^2.$$

A tapasztalati szórásnégyzetet tehát úgy kapjuk, ha a mintaelemek négyzetének átlagából kivonjuk a mintaátlag négyzetét.

Pontbecslések

A korábbi példában a mintaátlag 3, a mintaelemek négyzete pedig rendre 9, 1, 25, 25, 1, tehát

$$s^2 = \frac{9+1+25+25+1}{5} - 3^2 = \frac{61}{5} - 9 = \frac{16}{5}.$$

Pontbecslések

A tapasztalati szórásnégyzet nem torzítatlan becslése a szórásnégyzetnek:

$$\mathbb{E}(S_n^2) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \mathbb{E}\left(\overline{X}^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{D}^2(X_i) + \mathbb{E}(X_i)^2\right) - \left(\mathbb{D}^2\left(\overline{X}\right) + \mathbb{E}\left(\overline{X}\right)^2\right)$$

$$= \frac{1}{n}\cdot n\cdot\left(\mathbb{D}^2(X) + \mathbb{E}(X)^2\right) - \left(\frac{1}{n}\mathbb{D}^2(X) + \mathbb{E}(X)^2\right)$$

$$= \frac{n-1}{n}\mathbb{D}^2(X).$$

Pontbecslések

Definíció. A $T(X_1, \ldots, X_n)$ statisztika *aszimptotikusan torzítatlan* becslés θ-ra, ha

$$\lim_{n\to\infty} \mathbb{E}(T(X_1, \ldots, X_n)) = \theta.$$

A fenti definíció értelmében a tapasztalati szórásnégyzet aszimptotikusan torzítatlan becslés a szórásnégyzetre, hiszen $\lim_{n\to\infty}(n-1)/n = 1$.

Pontbecslések

Némi korrekcióval itt is torzítatlan becslést kaphatunk:

Definíció. Az $S_n^{*2} := \frac{n}{n-1} S_n^2$ statisztikát *korrigált tapasztalati szórásnégyzetnek* nevezzük. Ennek S_n^* gyöke a *korrigált tapasztalati szórás*.

Szokásos módon egy realizációra az s_n^{*2} ill. s_n^* jelöléseket használjuk, esetlegesen az elemszámot a jelölésből elhagyjuk.

A fentiek értelmében a korrigált tapasztalati szórásnégyzet torzítatlan becslése a szórásnégyzetnek, hiszen

$$\mathbb{E}\left(S_n^{*2}\right) = \frac{n}{n-1}\mathbb{E}\left(S_n^2\right) = \frac{n}{n-1} \cdot \frac{n-1}{n} \cdot \mathbb{D}^2(X) = \mathbb{D}^2(X).$$

Pontbecslések

A korrigált tapasztalati szórásnégyzet a fenti példában:

Láttuk, hogy $s^2 = \frac{16}{5}$, a mintaelemszám pedig 5, így

$$s^{*2} = \frac{5}{4} \cdot \frac{16}{5} = 4.$$

Intervallumbecslések

- A fent bemutatott pontbecslésektől természetesen nem várhatjuk, hogy a háttéreloszlás paraméterét pontosan megadják.
- Most azt fogjuk megvizsgálni, hogy mit mondhatunk a becslésk tényleges értéktől való eltérésről.
- Kényelmesebb a becsült pontot tekinteni kiindulópontnak, és azt kérdezni, hogy ettől milyen messze esik a tényleges érték.
- Vagyis: a becsült érték körül mekkora intervallumot kell venni, hogy abba a tényleges érték nagy valószínűséggel beleessen?

Intervallumbecslések

Az intervallum két végpontját egy-egy statisztika segítségével jelöljük ki:

Definíció. A $(T_1(X_1, \ldots, X_n); T_2(X_1, \ldots, X_n))$ statisztikapárral definiált intervallum *legalább* $1 - \varepsilon$ *szintű konfidenciaintervallum* a θ paraméterre, ha

$$\mathbb{P}(T_1(X_1, \ldots, X_n) < \theta < T_2(X_1, \ldots, X_n)) \geq 1 - \varepsilon$$

teljesül. Ha itt a fenti egyelnőtlenség helyett szigorú egyenlőség teljesül, akkor pontosan $1 - \varepsilon$ szintű konfidenciaintervallumról beszélünk.

Intervallumbecslések

- Az $\varepsilon > 0$ értéket az eljárás során előre rögzítjük, és úgy konstruáljuk az intervallum két végpontját, hogy $1 - \varepsilon$ valószínűséggel abba essen a tényleges paraméter.
- Ez a gyakorlatban akkor ad jól használható eljárást, ha az ε-t elég kicsinek választjuk.
- Pl. gyakori az $\varepsilon = 0{,}05$ vagy az $\varepsilon = 0{,}01$ választás, ekkor azt mondjuk, hogy 95%-os ill. 99%-os szintű konfidenciaintervallumot keresünk.
- A bizonyosság növeléséért cserébe a pontosságot kell feláldoznunk, nagyon kicsi ε esetén nagyon nagy lehet az intervallum hossza.
- De: a minta elemszámának növelése eszköz lehet a pontosság növelésére.

Intervallumbecslések

Konfidenciaintervallum szerkesztése normális eloszlás várható értékére ismert szórás esetén

- $X \sim N(\mu; \sigma^2)$ a háttérváltozó, ahol σ^2 ismert,
- a konfidenciaintervallum középpontját a mintaátlagnak fogjuk választani, az intervallumot pedig $(\overline{X} - r_\varepsilon; \overline{X} + r_\varepsilon)$ alakban keressük.

Intervallumbecslések

Az r_ε értékét a következő érvelésből kaphatjuk meg: az

$$1-\varepsilon = \mathbb{P}(\overline{X} - r_\varepsilon < \mu < \overline{X} + r_\varepsilon)$$

egyenletnek kell teljesülnie. Alakítsuk át az egyenlet jobb oldalát:

$$\begin{aligned}
\mathbb{P}(\overline{X} - r_\varepsilon < \mu < \overline{X} + r_\varepsilon) &= \mathbb{P}(-r_\varepsilon < \mu - \overline{X} < r_\varepsilon) \\
&= \mathbb{P}(-r_\varepsilon < \overline{X} - \mu < r_\varepsilon) \\
&= \mathbb{P}\left(-r_\varepsilon < \frac{\sum_{i=1}^n X_i - n\mu}{n} < r_\varepsilon\right) \\
&= \mathbb{P}\left(-\frac{r_\varepsilon\sqrt{n}}{\sigma} < \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} < \frac{r_\varepsilon\sqrt{n}}{\sigma}\right).
\end{aligned}$$

Intervallumbecslések

Felhasználjuk a következő, bizonyítás nélkül közölt tételt:

Tétel. Legyenek $X_1 \sim N(\mu_1; \sigma_1^2)$ és $X_2 \sim N(\mu_2; \sigma_2^2)$ független, normális eloszlású valószínűségi változók, ekkor

$$X_1 + X_2 \sim N(\mu_1 + \mu_2; \sigma_1^2 + \sigma_2^2).$$

Intervallumbecslések

$$\mathbb{P}(\overline{X} - r_\varepsilon < \mu < \overline{X} + r_\varepsilon) = \mathbb{P}\left(-\frac{r_\varepsilon\sqrt{n}}{\sigma} < \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} < \frac{r_\varepsilon\sqrt{n}}{\sigma}\right).$$

A fenti állítás értelmében a mintaelemek $\sum_{i=1}^n X_i$ összege normális eloszlású $n\mu$ várható értékkel és $n\sigma^2$ szórással, tehát az utolsó valószínűségnél ennek éppen a sztenderdizáltja szerepel, ami így sztenderd normális eloszlású.

Ezért ez a valószínűség felírható a Φ eloszlásfüggvény segítségével, és a következőt kapjuk:

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

Intervallumbecslések

Átrendezve:

$$1 - \frac{\varepsilon}{2} = \Phi\left(\frac{r_\varepsilon \sqrt{n}}{\sigma}\right).$$

Mivel a Φ függvény deriváltja a sztenderd normális eloszlás φ sűrűségfüggvénye, ez utóbbi pedig minden valós t-re pozitív, így a Φ függvény szigorúan monoton növő az $\mathbb{R}$-en, ebből kifolyólag pedig kölcsönösen egyértelmű leképezés $\mathbb{R}$ és a $(0;1)$ intervallum között.

Tehát létezik a $\Phi^{-1} : (0;1) \to \mathbb{R}$ inverzfüggvény, amit a fenti egyenletre alkalmazva

$$\frac{r_\varepsilon \sqrt{n}}{\sigma} = \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right), \quad \text{azaz} \quad \boxed{r_\varepsilon = \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right) \cdot \frac{\sigma}{\sqrt{n}}}$$

Intervallumbecslések

$$\boxed{r_\varepsilon = \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right) \cdot \frac{\sigma}{\sqrt{n}}}$$

- A fenti képletből látszik, hogy rögzített ε mellett a mintaelemszám növelésével az intervallum sugara csökken.
- Ha a mintaelemszám fix, akkor az ε csökkentésével $1 - \varepsilon/2$ nő. Mivel Φ szigorúan monoton növő, így könnyen meggondolható, hogy Φ^{-1} is az, vagyis ε csökkentése az intervallum sugarának növekedését eredményezi.

Megjegyzés. A centrális határeloszlás tétele szerint a

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$$

sztenderdizált (véges és pozitív szórás mellett) közelítőleg sztenderd normális eloszlású lesz tetszőleges háttéreloszlás esetén, ha a mintaelemszám elegendően nagy. Így tehát a fenti eredmény ekkor is jól használható.

Köszönöm a figyelmet!