

Matematikai statisztika szorgalmi feladatok

1. Feltételes várható érték és konvolúció

1. Legyen X és Y független és azonos eloszlású valószínűségi változó véges második momentummal. Mutassuk meg, hogy ekkor

$$E[X^2|XY] = E[Y^2|XY] .$$

Lehet esetleg gyengíteni a változókra tett feltételt? (Precíz, formális levezetést kérek, nem csak annyit, hogy „látszik”.)

2. Legyen az X változó egyenletes eloszlású a $[-1,1]$ intervallumon. Mutassuk meg, hogy tetszőleges $y \in (0,1]$ esetén

$$P(X = -y \mid |X| = y) = 1/2 = P(X = y \mid |X| = y) .$$

3. Legyen X és Y független és normális eloszlású véletlen változó. Mit mondhatunk az $X - Y$ különbség eloszlásáról? A feladatot legalább két különböző módszerrel kérem megoldani.

2. Nevezetes statisztikai eloszlások

1. Legyen $X, X_1, \dots, X_n$ független standard normális eloszlású véletlen változó. Ekkor azt mondjuk, hogy a

$$Z_n = \frac{\sqrt{n}X}{\sqrt{X_1^2 + \dots + X_n^2}}$$

változó n szabadsági fokú Student-eloszlást követ. Határozzuk meg a Z_n változó sűrűségfüggvényét.

2. Bizonyítsuk be az alábbi azonosságokat. ($a, b > 0$)

$$\Gamma(1/2) = \sqrt{\pi}, \quad \Gamma(a+1) = a\Gamma(a), \quad \Gamma(a)\Gamma(b) = B(a,b)\Gamma(a+b) .$$

3. A rendezett minta eloszlása

Tétel. (A sűrűségfüggvények lineáris transzformációs tétele.) Legyen X egy d -dimenziós vektorváltozó $f(x)$, $x \in \mathbb{R}^d$, sűrűségfüggvénnyel, továbbá legyen $A \in \mathbb{R}^{d \times d}$ egy nemelfajuló mátrix. Ekkor az AX vektorváltozó abszolút folytonos, és sűrűségfüggvénye

$$g(x) = f(A^{-1}x) \det(A^{-1}), \quad x \in \mathbb{R}^d .$$

1. Legyen $X_1^* \leq \dots \leq X_n^*$ a $[0,1]$ intervallumon egyenletes eloszlásból származó rendezett minta. Írjuk fel az X_k^* és az X_n^* változó együttes sűrűségfüggvényét, majd ennek segítségével adjuk meg az $E[X_k^* | X_n^* = x]$ feltételes várható értéket.

2. Legyen $X_1^* \leq \dots \leq X_n^*$ a λ paraméteres exponenciális eloszlásból származó rendezett minta. A transzformációs tétel segítségével írjuk fel az $X_1^*, X_2^* - X_1^*, \dots, X_n^* - X_{n-1}^*$ változók együttes sűrűségfüggvényét, majd határozzuk meg a változók marginális sűrűségfüggvényeit. Mit mondhatunk a változókról függetlenség szempontjából?
3. Az 1. feladatsor 1. feladatára több olyan – egyébként rossz – levezetést is kaptam, ahol csak arra volt szükség, hogy a változók azonos eloszlásúak legyenek, de olyan feltételre már nem, ami a változók közötti kapcsolatot kontrollálná. Ezt szeretném megcáfolni az alábbi ellenpéldával.

Legyen X olyan változó, mely az 1, 2 és 3 értéket $1/3$ - $1/3$ valószínűséggel veszi fel, továbbá legyen Y az X érték hárommal vett maradéka plusz 1. Mutassuk meg, hogy X és Y azonos eloszlású, de 1 valószínűséggel

$$E[X^2|XY] \neq E[Y^2|XY] .$$

4. Elégséges statisztikák

1. Adjunk meg $T(\underline{X}) = T(X_1, \dots, X_n)$ elégséges statisztikát az alábbi eloszláscsaládok paraméterére nézve.
- a. Azon eloszláscsalád, mely az $\alpha > 0$ paraméterrel van paraméterezve, és melynek paraméteres sűrűségfüggvénye

$$f_\alpha(x) = 2\alpha x(1-x^2)^{\alpha-1}, \quad 0 < x < 1 .$$

- b. A Bernoulli eloszláscsalád, ahol a paraméter $q \in (0,1)$, és a család súlyfüggvénye

$$p_q(x) = P_q(X=x) = \begin{cases} q, & x=1, \\ 1-q, & x=0. \end{cases}$$

Tipp: Próbáljuk egy olyan egyszerű formulát adni a $p_q(x)$ valószínűségre, mely egyszerre alkalmazható az $x=1$ és az $x=0$ esetben.

2. Mutassuk meg, hogy az egyenletes eloszlások

$$f_\theta(x) = \begin{cases} 1/(b-a), & a \leq x \leq b, \\ 0, & \text{különben}, \end{cases}$$

paraméteres sűrűségfüggvénnyel definiált családjában a $T(\underline{X}) = (X_1^*, X_n^*)$ statisztika minimális elégséges statisztika a $\theta = (a, b)$ paraméterre nézve.

3. Tekintsük az

$$f_a(x) = \begin{cases} 2a^2/x^3, & x \geq a, \\ 0, & x < a, \end{cases}$$

paraméteres sűrűségfüggvénnyel definiált eloszláscsaládot, ahol a paraméter $a > 2$, valamint a $T(\underline{X}) = (X_1^*, X_n^*)$ statisztikát, ahol $\underline{X} = (X_1, \dots, X_n)$. Mit állíthatunk, elégséges az a paraméterre nézve a $T(\underline{X})$ statisztika? Ha elégséges statisztika, akkor minimális elégséges statisztika? Ha nem minimális, akkor adjunk meg egy minimális elégséges statisztikát.

5. Becslések tulajdonságai

1. Legyen (X, Y) olyan vektorváltozó, melyre a komponensek $\text{Cov}(X, Y)$ kovarianciája jól definiált. Tekintsünk egy n elemű statisztikai mintát, tehát $(X_1, Y_1), \dots, (X_n, Y_n)$ független és az (X, Y) vektorváltozóval azonos eloszlású változókat. Ekkor a

$$\text{Cov}_n(X, Y) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

empirikus kovariancia alkalmazható, mint az (elméleti) kovariancia becslése. Mit állíthatunk az empirikus kovarianciáról torzítatlanság és konzisztencia szempontjából? Ha az empirikus kovariancia nem torzítatlan és/vagy nem erősen konzisztens, akkor adjuk meg egy olyan korrigált változatát, mely már mindkét tulajdonsággal rendelkezik.

2. Legyen $X_1, \dots, X_n$, $n \geq 2$, a λ paraméteres exponenciális eloszlásból származó statisztikai minta.

a. Tekintsük az $\bar{X}$ és az X_1 statisztikát, mint az $E_\lambda(X)$ várható érték becsléseit. Mit állíthatunk ezekről a statisztikákról torzítatlanság és konzisztencia szempontjából? Melyik a hatásosabb?

b. Adjunk torzítatlan és erősen konzisztens becslést a λ paraméterre.

3. Legyen X olyan véletlen változó, melynek folytonos az $F(x)$, $x \in \mathbb{R}$, eloszlásfüggvénye. Tegyük fel, hogy F szigorúan monoton egy $[a, b] \subseteq \mathbb{R}$ véges intervallumon, továbbá $F(x) = 0$, ha $x \leq a$, és $F(x) = 1$, ha $x \geq b$. Ekkor az X változónak létezik egyértelmű (elméleti) mediánja, tehát egy olyan $q \in \mathbb{R}$ érték, melyre

$$P(X \leq q) = 0,5.$$

Egy $n = 2k + 1$ páratlan elemszámú $X_1, \dots, X_n$ statisztikai minta alapján a medián becsülhető a $q_n = X_{k+1}^*$ középső mintaelemmel.

a. Bizonyítsuk be, hogy az F eloszlásfüggvényre tett feltételek mellett a q_n statisztika erősen konzisztens és asszimptotikusan torzítatlan.

b. Mutassunk példát olyan eloszlásra, mely teljesíti a feltételeket, de a q_n statisztika nem torzítatlan.

6. A Cramér–Rao- és a Rao–Blackwell–Kolmogorov-tétel

1. Legyen $X_1, \dots, X_n$ a p paraméteres Bernoulli-eloszlásból származó minta.

a. Ha $T_n = T_n(X_1, \dots, X_n)$ az ismeretlen paraméternek egy torzítatlan becslése, akkor mit mondhatunk a T_n szórásáról? (Ellenőrizzük is le a szükséges feltételeket.)

b. Mutassuk meg, hogy a $T_n = X_1$ statisztika torzítatlan becslés.

c. Blackwellizációval adjunk hatásosabb becslést a paraméterre. Az így kapott becslés hatásos?

2. (Ez a feladat 2 pontot ér.) Legyen $X_1, \dots, X_n$ Poisson-eloszlású minta ismeretlen λ paraméterrel, és tekintsünk a $\psi(\lambda) = P(X = 0)$ valószínűséget.

a. Tegyük fel, hogy egy $T_n = T_n(X_1, \dots, X_n)$ statisztika torzítatlan becslés a $\psi(\lambda)$ paraméterre. Adjunk alsó becslést a T_n statisztika szórására. (Ellenőrizzük is le a szükséges feltételeket.)

b. Legyen $T_n = \mathbb{1}_{\{X_1=0\}}$, és legyen T'_n az $\mathbb{1}_{\{X_k=0\}}$, $k = 1, \dots, n$, indikátorok számtani átlaga. Mutassuk meg, hogy T_n és T'_n torzítatlan becslés a $\psi(\lambda)$ paraméterre. Melyik becslés a hatásosabb? Tudunk arról nyilatkozni, hogy valamelyik becslés hatásos-e?

c. Legyen X és Y független Poisson-eloszlású változó rendre λ és μ paraméterrel. Határozzuk meg a következő valószínűségek értékét:

$$P(X + Y = m), \quad P(X = \ell \mid X + Y = m), \quad 0 \leq \ell \leq m.$$

d. Milyen becslést kapunk a $\psi(\lambda)$ paraméterre, ha blackwellizáljuk a b. pont T'_n statisztikáját? Mit mondhatunk a kapott statisztikáról torzítatlanság és konzisztencia szempontjából? Mi történik, ha a kapott statisztikát újra blackwellizáljuk? (A bátrabbak a statisztika varianciáját is kiszámolhatják, de ez nem lesz egyenlő az a. pontban kapott alsó korláttal.)

7. A maximum likelihood becslés és a momentumok módszere

1. Adott egy kísérlet, melyhez kapcsolódik egy p valószínűségű A esemény. Azt tapasztaljuk, hogy a kísérletet egymástól függetlenül n alkalommal végrehajtva az A esemény bekövetkezési gyakorisága k .

a. Tegyük fel, hogy ismerjük az n értéket. A maximum likelihood illetve a momentum módszer alkalmazásával adjunk becslést a p valószínűségre. Mit mondhatunk a becslésről konzisztencia és torzítatlanság szempontjából.

b. Tegyük fel, hogy ismerjük a p valószínűséget. Adjunk maximum likelihood és momentum becslést az n paraméter értékére.

c. A színvakság valószínűsége egy populációban genetikai tényezők miatt p a férfiaknál és p^2 a nőknél. Adjuk meg a p maximum likelihood becslését az alapján, hogy m férfiből k és n nőből ℓ volt színvak.

2. Tekintsük az $\alpha \in \mathbb{R}$ és $\lambda > 0$ paraméterekkel paraméterezett

$$F_{\alpha, \lambda}(x) = \begin{cases} 1 - \exp(\lambda(\alpha - x)), & x \geq \alpha, \\ 0, & x < \alpha, \end{cases}$$

eloszláscsaládot, és legyen $X_1, \dots, X_n$ a családból származó statisztikai minta.

a. Adjunk maximum likelihood becslést a paraméterekre. Mit állíthatunk a becslésekről konzisztencia és torzítatlanság szempontjából?

b. Adjunk becslést a paraméterekre a momentumok módszerével.

- Adott egy urna, benne pedig N golyó 1-től N -ig számozva, ahol az N ismeretlen paraméter.

a. Visszatevéssel kiválasztunk n golyót. A kapott értékek alapján adjunk maximum likelihood becslést az N paraméterre. Határozzuk meg a becslés várható értékét is.

b. Módosítsuk az a. pontot azzal, hogy visszatevés nélkül vesszük ki a golyókat. Adjunk maximum likelihood becslést az N paraméterre ezen feltételek mellett. Határozzuk meg a becslés várható értékét is, továbbá adjunk meg egy olyan korrigált változatot, ami már tozítatlan. (Figyelem: Ez most nem egy klasszikus értelemben vett statisztikai minta, hiszen a megfigyelések nem függetlenek, ezért az sem egyértelmű, hogy hogyan is kellene a likelihood függvényt definiálni.)

(Ez a statisztikai probléma a II. világháború idején merült fel, további részletek az angol nyelvű Wikipedia „German tank problem” című oldalán találhatóak.)

8. Bayes-becslések

- Egyszer feldobunk egy nem feltétlenül szabályos pénzérmét, és fejet kapunk. Adjunk Bayes-becslést a fej dobásának ismeretlen $p \in (0,1)$ valószínűségére. A p paraméter *a priori* eloszlása legyen egyenletes a $[0,1]$ intervallumon. Írjuk fel a paraméter *a posteriori* eloszlását, és adjunk pontbecslést a p értékre. Mi lenne a becslés, ha a maximum likelihood módszert alkalmaznánk?
- Tekintsük a $[-\theta, \theta]$ alakú intervallumokon értelmezett egyenletes eloszlások családját, ahol $\theta \geq 1$. Egy $X_1, \dots, X_n$ minta alapján adjunk Bayes-becslést a θ paraméterre úgy, hogy θ *a priori* eloszlása $\alpha \geq 1$ paraméteres Pareto legyen. Írjuk fel a θ *a posteriori* eloszlását, valamint adjunk pontbecslést a paraméterre. (Az α paraméteres Pareto-eloszlás sűrűségfüggvénye $q(t) = \alpha/t^{\alpha+1}$, $t \geq 1$.)
- Adott egy pénzérme, mellyel dobva ismert $p \in (0,1)$ valószínűséggel kapunk fejet. Valaki feldobja néhányszor az érmét, és elárulja nekünk, hogy k fejet kapott. Adjunk Bayes-becslést a dobások n számára úgy, hogy az n *a priori* eloszlása $\lambda > 0$ paraméteres Poisson legyen. Határozzuk meg az n *a posteriori* eloszlását, majd adjunk pontbecslést a dobások számára.

9. Konfidencia intervallumok

- Egy n elemű statisztikai minta alapján adjunk aszimptotikusan $1 - \alpha$ megbízhatóságú konfidencia intervallumot az exponenciális eloszlás mediánjára. (Tipp: Fejezzük ki a mediánt a paraméter segítségével.)
- a.** Legyen $X_1, \dots, X_n \sim N(\mu_1, \sigma_1^2)$ és $Y_1, \dots, Y_m \sim N(\mu_2, \sigma_2^2)$ két egymástól független statisztikai minta, ahol a $\mu_1, \mu_2, \sigma_1, \sigma_2$ paraméterek ismeretlenek. Adjunk $1 - \alpha$ megbízhatóságú konfidencia intervallumot a σ_1/σ_2 hányadosra.
b. Írjuk fel az a. pontban megadott konfidencia intervallumot abban az esetben, amikor $\alpha = 5\%$, $n = 3$, $m = 5$, és a mintaelemek 3, 2, 7, valamint -1, 5, 0, -3, 4.

3. a. Az emberek testsúlya és testmagassága együttesen normális eloszlású véletlen változó. Egy n elemű statisztikai minta alapján adjunk $1 - \alpha$ megbízhatóságú konfidencia intervallumot a testsúly és a testmagasság különbségének várható értékére. (Természetesen a testsúly nem független a magasságtól.)
- b. Mit állíthatunk akkor, ha nem tesszük fel a normalitást?
- c. Megkérdezzük öt ember testsúlyát és testmagasságát. A kapott értékek rendre 75, 86, 52, 90 és 63 kilogramm, valamint 184, 175, 160, 190 és 181 centiméter. Adjunk 90 százalékos megbízhatóságú konfidencia intervallumot a testsúly és a testmagasság különbségének várható értékére.

10. Paraméteres próbák

1. Egy magyarországi iskolába kínai vendégdiákok érkeznek. Ennek apropóján lemérjük néhány magyar és néhány kínai tanuló testmagasságát. A magyar diákok magassága 170, 165, 154, 172, 166 és 157, a kínai tanulóké 161, 152, 168 és 157 centiméter.
 - a. Tegyük fel, hogy mind a magyar, mind a kínai tanulók magassága normális eloszlást követ azonos szórással. Teszteljük 10 százalékos szignifikancia szinten azt a nullhipotézist, hogy a magyar és a kínai diákoknál azonos a magasság várható értéke. Adjunk 90 százalékos megbízhatóságú konfidencia intervallumot a várható értékek különbségére.
 - b. Teszteljük le 10 százalékos szignifikancia szinten a szórások egyenlőségét. (Érdeemes átnézni a tankönyv Paraméteres próbák című szekcióját.)
2. Legyen $X_1, \dots, X_n$ független és azonos eloszlású véletlen változó ismeretlen $\lambda > 0$ paraméteres exponenciális eloszlással.
 - a. Határozzuk meg a $T_n = \lambda(X_1 + \dots + X_n)$ statisztika eloszlását. Ennek segítségével konstruáljunk egy próbát a $H_0 : \lambda = \lambda_0$ nullhipotézisnek a kétoldali alternatívával szembeni tesztelésére, ahol $\lambda_0 > 0$ adott hipotetikus paraméter.
 - b. Írjuk fel a megkonstruált próba erőfüggvényét. Mit állíthatunk a próbáról konzisztencia szempontjából?
3. Adott egy nem feltétlenül szabályos pénzérme, melynél $p \in (0,1)$ a fej dobásának általunk ismeretlen valószínűsége. Tegyük fel, hogy az érmét újra és újra feldobva az X -edik dobásra kapjuk az első fejet. Szeretnénk azt a nullhipotézist tesztelni, hogy az érme szabályos, és ehhez tekintjük az X értéket, mint statisztikai mintát, és az $\{1, \dots, k_\alpha\}$ halmazt, mint elfogadási tartományt, ahol k_α egy megfelelően választott pozitív egész szám.
 Mely $\alpha \in (0,1)$ értékek esetén létezik k_α úgy, hogy a próba elsőfajú hibája pontosan α legyen? Feltéve, hogy az α ilyen érték, írjuk fel a próba erőfüggvényét. A kapott erőfüggvény alapján mit állíthatunk, torzítatlan a próba, ha a nullhipotézist a kétoldali alternatívával szemben teszteljük? És ha a tesztelést csak a $H_1 : p < 1/2$ egyoldali alternatívával szemben végezzük? Bónusz kérdés: Mit állíthatunk a próbáról konzisztencia szempontjából?

11. Randomizált próbák

1. Legyen $X_1, \dots, X_n$ statisztikai minta a $[0, b]$, $b > 0$, intervallumon definiált egyenletes eloszlásból. Tekintsük a $\mathcal{H}_0: b = b_0$ nullhipotézist és a $\mathcal{H}_1: b = b_1$ ellenhipotézist, ahol $b_1 < b_0$. Írjuk fel a Neyman–Pearson-alaplemma által definiált próbát.
2. Tekintsünk egy $\lambda > 0$ paraméteres Poisson eloszlásból származó $X_1, \dots, X_{10}$ statisztikai mintát, a $\mathcal{H}_0: \lambda = 2$ nullhipotézist és a $\mathcal{H}_1: \lambda = 1$ ellenhipotézist. Legyen továbbá $\alpha = 0,05$. Írjuk fel a Neyman–Pearson-alaplemmában szereplő tesztet, valamint határozzuk meg az ott szereplő $q = q_\alpha$ és $c = c_\alpha$ konstansok értékét. (Elég numerikusan számolni, és közelítő eredményeket megadni. Szoftvercsomag alkalmazása engedélyezett, sőt ajánlott.)
3. Bizonyítsuk be, hogy a Neyman–Pearson-alaplemmában definiált teszt torzítatlan. Tipp: Konstruáljunk egy olyan (esetleg randomizált) próbát, melynek erőfüggvénye $\beta_n(\alpha, \theta) = \alpha$ minden α, θ és n esetén.

12. Likelihood hányados próbák

1. Legyen $X_1, \dots, X_n$ Poisson eloszlású statisztikai minta ismeretlen $\lambda > 0$ paraméterrel. Tekintsük a $\mathcal{H}_0: \lambda = \lambda_0$ nullhipotézist a $\mathcal{H}_1: \lambda < \lambda_0$ alternatívával szemben. Írjuk fel a kapcsolatos likelihood hányados próbát, és az n mintaméret és az α szignifikancia szint függvényében adjuk meg a próbastatisztikát és a kritikus értéket.
2. Legyen $X_1, \dots, X_n \sim N(\mu_1, \sigma_1^2)$ és $Y_1, \dots, Y_m \sim N(\mu_2, \sigma_2^2)$ két egymástól független statisztikai minta ismert $\sigma_1, \sigma_2 > 0$ szórással, és ismeretlen $\mu_1, \mu_2 \in \mathbb{R}$ várható értékekkel. Teszteljük likelihood hányados próbával a $\mathcal{H}_0: \mu_1 = \mu_2$ nullhipotézist a kétoldali alternatívával szemben. Írjuk fel a próbastatisztikát, valamint határozzuk meg az aszimptotikus eloszlást. Van ennek a próbának valamilyen kapcsolata a kétmintás t -próbával?