Matematikai statisztika

October 6, 2024

Tartalom

1 Előadás

1.1 A statisztika fogalma és ágai

Statisztika: a valóság tömör, számszerű jellemzésére szolgáló tudományos módszertan, illetve gyakorlati tevékenység. Ágai:

1. **Leíró statisztika**: magába foglalja az információk összegyűjtését, összegzését, ábrázolását, tömör, számszerű jellemzését szolgáló módszereket

2. **Matematikai statisztika**: matematikai tudomány, adatok feldolgozásáról, érteémezéséről és felhasználásáról szóló tudományos módszertan

1.2 Leíró statisztika alapfogalmak

Statisztikai egység: a statisztikai vizsgálat tárgyát képező egyed.
Statisztikai sokaság: a megfigyelés tárgyát képező egyedek összessége, halmaza.
Statisztikai adat: valamely sokaság elemeinek száma vagy a sokaság valamilyen másféle számszerű jellemzője, mérési eredmény.
Statisztikai ismérv: a sokaság egyedeit jellemző tulajdonság.
Ismérvváltozatok: az ismérvek lehetséges kimenetelei.
Minta: a sokaság véges számosságú részhalmaza.
Statisztikai következtetés:a valóságban a teljes sokaságot nem tudjuk vagy akarjuk megfigyelni, ezért csak az egyedek egy szűkebb csoportját figyeljük meg. A viszonylag kisszámú egyedre vonatkozó információk alapján szeretnénk a teljes sokaság egészére, egyes jellemzőire, tulajdonságaira érvényes következtetéseket kimondani.

Példa:

Sokaság	most a teremben lévő homo sapiensek
Statisztikai egység	a teremben lévő oktató
Adat	a legmagasabb hallgató testtömegindexe
Ismérv	nem
Ismérvváltozatok	férfi, nő
Minta	5 véletenül választott hallgató

1.3 Csoportosítások, adatok fajtái

Sokaságok csoportosítása:

1. A sokaság egységeinek megkülönböztethetősége szerint:

   - diszkrét: a sokaság egységei elkülönülnek egymástól
   - folytonos: a sokaság egységeit nem tudjuk természetes módon elkülöníteni

2. A sokaság időpontra vagy időtartamra értelmezhető-e:

   - álló: csak egy adott időpontra értelmezhető

- mozgó: csak egy adott időtartamra értelmezhető

3. A sokaság számossága szerint:
   - véges
   - végtelen

A statisztikai adatok fajtái:

1. alapadatok: közvetlenül a sokaságból származnak
2. leszármaztatott adatok: alapadatokból műveletek eredményeként adódnak

Az ismérvek típusai:

1. minőségi: az egyedek számszerűen nem mérhető tulajdonsága
2. mennyiségi: az egyedek számszerűen mérhető tulajdonsága (diszkrét, folytonos)
3. időbeli: az egységek időbeli elhelyezésére szolgáló rendezőelvek
4. területi: az egységek térbeli elhelyezésére szolgáló rendezőelvek
5. közös: tulajdonságok, amik szerint a sokaság egyedei egyformák
6. megkülönböztető: azok a tulajdonságok, amik szerint a sokaság egyedei különböznek egymástól

Mérési skálák

1. nominális: kódszámok a sokaság egyedeinek azonosítására, pl. utasok neme
2. ordinális: valamely tulajdonság alapján való sorbarendezés, pl. az utasosztályok
3. intervallumskála: a skálaértékek különbségei is valós információt adnak a sokaság egyedeiről. A skálán a nullpont meghatározása önkényes. Ilyen skálákhoz mértékegység is tartozik. pl. hőmérséklet
4. a skálának van valódi nullpontja is. Minden matematikai művelet elvégezhető ezekkel a számokkal. pl. kor, jegy ára

Statisztikai sor: a sokaság egyes jellemzőinek felsorolása. Az ismérvek fajtája szerint beszélhetünk minőségi, mennyiségi, időbeli és területi sorokról.

1. Csoportosító sor: a sokaság egy megkülönböztető ismérv szerinti osztályozásának eredménye; az adatok összegezhetők
2. Összehasonlító sor: a sokaság egy részének a sokaságot egy megkülönböztető ismérv szerinti osztályozásának eredménye; az adatok nem összegezhetők
3. Leíró sor: különböző fajta, gyakran eltérő mértékegységű statisztikai adatokat tartalmaz

Statisztikai tábla: a statisztikai sorok összefüggő rendszere.

1. Egyszerű tábla: nem tartalmaz csoportosítást, nincs benne összegző sor
2. Csoportosító tábla: egyetlen csoportosító sort tartalmaz
3. Kombinációs tábla vagy kontingenciatábla vagy kereszttábla: legalább két csoportosító sort tartalmaz

1.4 Viszonyszámok

A statisztikai elemzések egyik legfontosabb eszközei a viszonyszámok (alias: indikátorok). A viszonyszám két statisztikai adat hányadosa. Jelölések:

$$V = \frac{A}{B}$$

ahol V: viszonyszám; A: a viszonyítás tárgya; B: a viszonyítás alapja.

A viszonyítás fajtái:

1. megoszlási: a sokaság egy részének a sokaság egészéhez való viszonyítása
2. koordinációs: a sokaság egy részének a sokaság egy másik részéhez való viszonyítása
3. dinamikus: két idopont vagy időszak adatának hányadosa
4. intenzitási: különböző fajta adatok viszonyítása egymáshoz; gyakran a mértékegységük is eltérő

2 Előadás

2.1 Tapasztalati eloszlás

Tapasztalati eloszlás: minden megfigyeléshez azonos, $\frac{1}{n}$ súlyt rendelünk. Ez egy diszkrét eloszlás.
Tapasztalati eloszlásfüggvény: a tapasztalati eloszlás eloszlásfüggvénye. Ez egy tiszta ugrófüggvény, értéke minden mintaelem helyén $\frac{1}{n}$ nagyságot ugrik felfelé.
A tapasztalati eloszlásfüggvény az x helyen:

$$\frac{I(x_1 < x) + I(x_2 < x) + \cdots + I(x_n < n)}{n} = \frac{\sum_{i=1}^{n} I(x_i < x)}{n}.$$

Azt mutatja meg, hogy a mintaelemek hányad része kisebb x-nél.

2.2 Középértékek számítása

Adott az n elemű $\underline{x} = (x_1, \ldots, x_n)$ tapasztalati minta; osztályközös gyakorisági sor esetén k jelöli az osztályok számát, x_i az ösztályközepeket, f_i pedig a gyakoriságokat.

Mintaátlag: az adatok átlagos értéke.

- számítása közvetlenül az adatokból: $\bar{x} = \dfrac{x_1 + \cdots + x_n}{n}$
- számítása osztályközös gyakorisági sorból: $\bar{x} = \dfrac{f_1 x_1 + \cdots + f_k x_k}{n}$

Módusz: a legtöbbször előforduló ismérvérték. Számítása osztályközös gyakorisági sorból:

$$\text{Mo} = x_{mo,\,a} + \frac{d_a}{d_a + d_f} \cdot h_{mo}$$

- a móduszt tartalmazó osztályköz (MTO): amelyikben egységnyi osztályköz hosszra a legnagyobb gyakoriság jut
- $x_{mo,\,a}$: a MTO alsó értéke
- h_{mo} : a MTO hossza
- d_a : a MTO korrigált gyakorisága mínusz a móduszt közvetlenül megelőző osztályköz korrigált gyakorisága
- d_f : a MTO korrigált gyakorisága mínusz a móduszt közvetlenül követő osztályköz korrigált gyakorisága

Jelölje $x_1^* \leq x_2^* \leq \cdots \leq x_n^*$ a rendezett tapasztalati mintát.

Medián: azon ismérvérték, amelynél ugyanannyi kisebb vagy egyenlő, mint nagyobb vagy egyenlő ismérvérték fordul elő a mintában. Számítása közvetlenül az adatokból:

$$\text{Me} = \begin{cases} x^*_{\frac{n+1}{2}} & \text{ha n páratlan} \\ \frac{x^*_{\frac{n}{2}} + x^*_{\frac{n}{2}+1}}{2} & \text{ha n páros} \end{cases}$$

Számítása osztályközös gyakorisági sorból - két lépésben lineáris interpolációval:

1. Melyik osztályközben van a medián: azon i, amire $f'_{i-1} \leq \frac{n}{2}$ és $f'_i \geq \frac{n}{2}$

2. $\text{Me} = x_{i,\,a} + \dfrac{\frac{n}{2} - f'_{i-1}}{f_i} \cdot h_i$, ahol

   - $x_{i,\,a}$: a mediánt tartalmazó osztályköz alsó értéke
   - h_i : a mediánt tartalmazó osztályköz hossza
   - f'_{i-1} : a mediánt közvetlenül megelőző osztályköz kumulált gyakorisága
   - f_i : a mediánt tartalmazó osztályköz gyakorisága

2.3 Tapasztalati kvantilisek számítása

Tapasztalati y-kvantilis: azon ismérvérték, amelynél a mintaelemek y-ad része kisebb vagy egyenlő, míg $(1-y)$-ad része nagyobb vagy egyenlő, $0 < y < 1$.

Számítása nem egyértelmű, mi mindig az egyik interpolációs módszert alkalmazzuk két lépésben:

1. hányadik mintaelem a keresett kvantilis $\rightarrow$ sorszám: $s := (n+1)y$

2. lineáris interpolációval a kvantilis kiszámítása
   Számítása közvetlenül az adatokból:

   - sorszám: $s = e + t$ (egész + törtrész)
   - $q_y = x^*_e + t(x^*_{e+1} - x^*_e)$

   Számítása osztályközös gyakorisági sorból:

   - melyik osztályközben van az s-edik elem: jelölje ezt i, azaz $f'_{i-1} \leq s$ és $f'_i \geq s$
   - $q_y = x_{i,\,a} + \dfrac{s - f'_{i-1}}{f_i} \cdot h_i$, ahol a szimbólumok ugyanazokat jelöli, mint az előbbiekben

2.4 Nevezetes kvantilisek

A szakirodalomban a tapasztalati és az elméleti értékek között nem tesznek különbséget, mindegyiket nagybetűvel írják. Jelölje q_y a tapasztalati y-kvantilist.

- tercilisek: $T_1 = q_{\frac{1}{3}}$, $T_2 = q_{\frac{2}{3}}$
- kvartilisek: $\mathbf{Q}_1 = q_{\frac{1}{4}}$, $\mathbf{Q}_2 = \text{Me} = q_{\frac{2}{4}}$, $\mathbf{Q}_3 = q_{\frac{3}{4}}$
- kvintilisek: $K_i = q_{\frac{i}{5}} \quad (i = 1, \ldots, 4)$
- decilisek: $D_i = q_{\frac{i}{10}} \quad (i = 1, \ldots, 9)$
- percentilisek: $P_i = q_{\frac{i}{100}} \quad (i = 1, \ldots, 99)$

2.5 Szóródási mutatók számítása

Terjedelem: $R = x_n^* - x_1^*$
Interkvantilis terjedelem: $IQR = \mathbf{Q}_3 - \mathbf{Q}_1$
Tapasztalati szórás:

- számítása közvetlenül adatokból: $s_n = \sqrt{\dfrac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}}$
- számítása osztályközös gyakorisági sorból: $s_n = \sqrt{\dfrac{f_1(x_1 - \bar{x})^2 + \cdots + f_k(x_k - \bar{x})^2}{n}}$

Korrigált tapasztalati szórás:

- számítása közvetlenül adatokból: $s_n^* = \sqrt{\dfrac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}}$
- számítása osztályközös gyakorisági sorból: $s_n^* = \sqrt{\dfrac{f_1(x_1 - \bar{x})^2 + \cdots + f_k(x_k - \bar{x})^2}{n-1}}$

Relatív szórás vagy **szórási együttható**:

$$V = \frac{s_n^*}{\bar{x}} \text{ vagy } V = \frac{s_n}{\bar{x}}.$$

2.6 Alakmutatók számítása

A szórást ezeknél is választhatjuk a tapasztalati vagy a korrigált tapasztalati szórásnak egyaránt.

Tapasztalati ferdeség:

- számítása közvetlenül az adatokból: $\dfrac{(x_1 - \bar{x})^3 + \cdots + (x_n - \bar{x})^3}{(s_n)^3}$
- számítása osztályközös gyakorisági sorból: $\dfrac{f_1(x_1 - \bar{x})^3 + \cdots + f_k(x_k - \bar{x})^3}{(s_n)^3}$

Tapasztalati csúcsosság:

- számítása közvetlenül az adatokból: $\dfrac{(x_1 - \bar{x})^4 + \cdots + (x_n - \bar{x})^4}{(s_n)^4} - 3$
- számítása osztályközös gyakorisági sorból: $\dfrac{f_1(x_1 - \bar{x})^4 + \cdots + f_k(x_k - \bar{x})^4}{(s_n)^4} - 3$

3 Előadás

3.1 Statisztikai mező

$(\Omega, \mathcal{A}, P_\theta)$, $\theta \in \Theta$ statisztikai mező, ha Θ paraméterhalmaz és $(\Omega, \mathcal{A}, P_\theta)$ minden paraméter esetén valószínűségi mező.

Definíció.

$$\underline{\xi} = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} : \Omega \to \mathcal{X} \subseteq \mathbb{R}^n$$

valószínűségi vektorváltozót mintának nevezzük. n : mintanagyság, ξ_i : i. mintaelem.

Definíció. Minta realizációja

$$\underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

a konkrét megfigyelt számsorozat.

Definíció. Legyen $\underline{\xi} : \Omega \to \mathbb{R}^n$ minta. Ekkor $\mathcal{X} := \mathcal{R}_{\underline{\xi}}$. A minta lehetséges értékeinek halmaza. Elemei a mintaértékek.

- n-elemű valós értékű minta esetén: $\mathcal{X} = \mathbb{R}^n$
- n-elemű pozitív egész értékű minta esetén: $\mathcal{X} = \mathbb{N}^n$

3.2 Minták típusai

- Független minta: a mintaelemek függetlenek.
- Független azonos eloszlású minta: a mintaelemek független és azonos eloszlásúak.
- Diszkrét minta: a mintaelemek diszkrétek.
- Abszolút folytonos eloszlású minta: a mintaelemek abszolút folytonosak.

3.3 Eloszláscsaládok

Legyen adott egy $(\Omega, \mathcal{A}, P_\theta)$ statisztikai mező és $\underline{\xi} : \Omega \to \mathbb{R}^n$ minta. Ekkor legyen a minta eloszlásfüggvénye adott $\theta \in \Theta$ mellett $F_\theta : \mathbb{R}^n \to \mathbb{R}$, ahol

$$F_\theta(\mathbf{s}) := P_\theta(\xi_1 < s_1, \ldots, \xi_n < s_n) \quad (\mathbf{s} \in \mathbb{R}^n).$$

Független minta esetén:

$$F_\theta(\mathbf{s}) = \prod_{i=1}^{n} P_\theta(\xi_i < s_i) \quad (\mathbf{s} \in \mathbb{R}^n).$$

Jelölések:

- E_θ: várható érték P_θ esetén;
- D_θ: szórás P_θ esetén;
- f_θ sűrűségfüggvény P_θ esetén
- $p_\theta(s) = P_\theta(\xi_i = s)$, $i = 1, \ldots, n$ diszkrét minta

Definíció. Egy minta függvényét statisztikának nevezzük:

$$T : \mathcal{X} \to \mathbb{R}^k.$$

Def'.: Statisztika:

$$T(\xi), \text{ ha } T : \mathcal{X} \to \mathbb{R}^k \text{ függvény.}$$

3.4 Tapasztalati momentumok

$\mathcal{X} = \mathbb{R}^n$
mintaközép:

$$T(\mathbf{x}) = \overline{x} = \frac{\sum_{i=1}^n x_i}{n}, \quad T(\xi) = \overline{\xi} = \frac{\sum_{i=1}^n \xi_i}{n},$$

tapasztalati k. momentum:

$$T(\mathbf{x}) = \frac{\sum_{i=1}^n x_i^k}{n}, \quad T(\xi) = \frac{\sum_{i=1}^n \xi_i^k}{n}.$$

3.5 Tapasztalati szórásnégyzet

$\mathcal{X} = \mathbb{R}^n$

$$T(\mathbf{x}) = \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n},$$

$$T(\xi) = s^2 = \frac{\sum_{i=1}^n (\xi_i - \overline{\xi})^2}{n},$$

4 Előadás

4.1 Becsléselmélet

A minta eloszlásának ismeretlen paraméterét közelítjük a minta függvényével.
Becslőfüggvény: $\hat{\theta} : \mathcal{X} \to \Theta$.
Becslés: $\hat{\theta}(\xi)$.

Definíció. A $\underline{\xi} = (\xi_1, \ldots, \xi_n) : \Omega \to \mathbb{R}^n$ független, azonos eloszlású minta likelihood függvénye $L : \mathcal{X} \times \Theta \to \mathbb{R}$, ahol

$$L(\mathbf{x}, \theta) \begin{cases} P_\theta(\underline{\xi} = \mathbf{x}) = \prod_{i=1}^{n} P_\theta(\xi_i = \xi_i) & \text{diszkrét minta esetén} \\ f_\theta(\mathbf{x}) = \prod_{i=1}^{n} f_\theta(x_i) & \text{abszolút folytonos minta esetén} \end{cases}$$

ahol f_θ, ξ_i sűrűségfüggvénye.

$$l(\mathbf{x}, \theta) = \ln L(\mathbf{x}, \theta)$$

a loglikelihood függvény.

Egy $\hat{\theta} \in \Theta$ maximum likelihood becslése, ha

$$L(\xi, \hat{\theta}) = \max_{\theta \in \Theta} L(\xi, \theta).$$

4.2 Likelihood egyenlet

Gyakran a loglikelihood függvény maximumhelyét keresik a

$$\partial_\theta l(\mathbf{x}, \theta) = 0$$

egyenletet (vagy egyenletrendszert) megoldva. Ez diszkrét minta esetén a

$$\sum_{i=1}^{n} \partial_\theta \ln P_\theta(\xi_i = x_i) = 0$$

egyenlet (vagy egyenletrendszert) jelenti. Abszolút folytonos minta esetén

$$\sum_{i=1}^{n} \partial_\theta \ln P_\theta(\xi_i = x_i) = 0.$$

4.3 Indikátor példa

$$L(\mathbf{x}, p) = p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i},$$

$$l(\mathbf{x}, p) = \ln L(\mathbf{x}, p) = \left(\sum_{i=1}^{n} x_i \right) \ln p + \left(n - \sum_{i=1}^{n} x_i \right) \ln(1-p).$$

Likelihood egyenlet:

$$\partial_p l(\mathbf{x},\, p) = \left(\sum_{i=1}^{n} x_i\right)\frac{1}{p} - \left(n - \sum_{i=1}^{n} x_i\right)\frac{1}{1-p} = 0.$$

Ennek megoldása:

$$p = \frac{\sum_{i=1}^{n} x_i}{n}.$$

4.4 Poisson példa

Tegyük fel, hogy $\eta_1, \ldots, \eta_n \sim \text{Poisson}(\lambda)$. Ekkor

$$L(\underline{k},\, \lambda) = P_\lambda(\eta_1 = k_1, \ldots, \eta_n = k_n) =$$

$$\prod_{i=1}^{n} \frac{\lambda^{k_i} e^{-\lambda}}{k_i!} = \left(\prod_{i=1}^{n} \frac{1}{k_i!}\right) \cdot \left(\prod_{i=1}^{n} \lambda^{k_i} e^{-\lambda}\right) = \left(\prod_{i=1}^{n} \frac{1}{k_i!}\right) \cdot \left(\lambda^{\sum_{i=1}^{n} k_i} e^{-n\lambda}\right)$$

$$l(\underline{k},\, \lambda) = \ln L(\underline{k},\, \lambda) = \left(\sum_{i=1}^{n} \ln\left(\frac{1}{k_i!}\right)\right) + \left(\sum_{i=1}^{n} k_i\right) \ln\lambda - n\lambda$$

$$\partial_\lambda l(\underline{k}, \lambda) = \frac{\sum_{i=1}^{n} k_i}{\lambda} - n = 0 \quad \Longleftrightarrow \quad \lambda = \frac{\sum_{i=1}^{n} k_i}{n}$$

4.5 Becslések tulajdonságai

Definíció. A paraméter $\hat{\theta}(\xi)$ becslése torzítatlan, ha

$$E_\theta(\hat{\theta}(\xi)) = \theta \quad (\theta \in \Theta).$$

Konzisztencia: $\hat{\theta}(\xi) \to \theta$ sztochasztikusan $(n \to \infty)$.
Elégséges feltétel: $E_\theta(\hat{\theta}_n(\xi)) \to \theta$ és $D^2_\theta(\hat{\theta}_n(\xi)) \to 0$

Definíció. Torzítatlan beclésekre: T_1 hatásosabb becslése $h(\Theta)$-nak a T_2-nél, ha

$$D^2_\theta(T_1(\underline{X})) \le D^2_\theta(T_2(\underline{X}))$$

teljesül minden θ paraméterekre.
A T torzítatlan becslés hatásos, ha minden más torzítatlan becslésnél hatásosabb.
Átlagos négyzetes eltérés:

$$E_\theta(T(\underline{X}) - \theta)^2$$

4.6 Hatásos becslés egyértelműsége

4.7 Mennyi információt hordoz a statisztika

Hatásos becslés egyértelműsége

Áll.: Amennyiben T_1 és T_2 hatásos becslései $h(\theta)$-nak, akkor 1 valószínűséggel megegyeznek minden lehetséges paraméter esetén.

$E_\theta T_1 = E_\theta T_2 = h(\theta)$, továbbá $D_\theta T_1 = D_\theta T_2$. Ebből

$$\mathrm{D}_\theta^2(T_1) \le \mathrm{D}_\theta^2\left(\frac{T_1+T_2}{2}\right) = \frac{\mathrm{D}_\theta^2(T_1)+2\,\mathrm{cov}(T_1,T_2)+\mathrm{D}_\theta^2(T_2)}{4} = \frac{\mathrm{D}_\theta^2(T_1)+\mathrm{cov}(T_1,T_2)}{2} \Rightarrow$$

$$\mathrm{D}_\theta^2(T_1) \le \mathrm{cov}(T_1,T_2) = D_\theta T_1 \square D_\theta T_2 \square R(T_1,T_2) = \mathrm{D}_\theta^2(T_1) \square R(T_1,T_2) \le \mathrm{D}_\theta^2(T_1) \Rightarrow$$

$$\mathrm{D}_\theta^2(T_1) = \mathrm{D}_\theta^2(T_2) = \mathrm{cov}(T_1,T_2) \Rightarrow \mathrm{D}_\theta^2(T_1-T_2) = \mathrm{D}_\theta^2(T_1) - 2\,\mathrm{cov}(T_1,T_2) + \mathrm{D}_\theta^2(T_2) = 0.$$

Így $E_\theta(T_1 = T_2) = 1 \;\forall \theta \in \Theta$.

Mennyi információt hordoz a statisztika?

Példa: $\xi_1,\ldots,\xi_n$ független $N(m,1)$ minta. Ekkor

$$\bar{\xi} = \frac{\sum_{i=1}^{n} \xi_i}{n} \sim N(m,\frac{1}{n})$$ eloszlású (függ m-től!),

miközben

$$s^2 = \frac{\sum_{i=1}^{n} (\xi_i - \bar{\xi})^2}{n}$$ eloszlása nem függ m-től!

4.8 Elégséges statisztika

4.9 Elégséges statisztika diszkrét minta esetén

Elégséges statisztika

- Minden információt (ugyanannyit mint az eredeti minta) tartalmaz az ismeretlen paraméterre vonatkozóan.
- "Elég" az ő értékét ismerni.
- Ismeretében már "nincs bizonytalanság" a mintában (úgy értve, hogy egyértelmű a minta eloszlása, már nem függ az ismeretlen paramétertől).

Elégséges statisztika diszkrét minta esetén

Def.: A diszkrét $\boldsymbol{\xi}$ mintából képzett $T(\boldsymbol{\xi})$ statisztika elégséges $\theta-$ra, ha a $P_\theta(\boldsymbol{\xi} = \mathbf{x} | T(\boldsymbol{\xi}) = t)$ feltételes valószínűség nem függ $\theta-$tól

4.10 Feltételes várható érték

4.11 Indikátor példa

Feltételes várható érték

Legyenek X és Y diszkrét val. változók.

$E(X|Y)$ az a val. változó, ami az $Y = y_k$ eseményen az $E(X|Y = y_k)$ értéket veszi fel.

Tulajdonságok:

- Ha $X \geq 0$, akkor $E(X|Y) \geq 0$
- $E(E(X|Y)) = EX$ (a teljes várható érték tételének általánosítasa)
- Ha X_1, X_2 várható értéke véges, akkor $E(c_1X_1 + c_2X_2|Y) = c_1E(X_1|Y) + c_2E(X_2|Y)$
- Ha X független Y −tól, akkor $E(X|Y) = E(X)$
- Ha X és $h(Y)$ várható értéke véges, akkor $E(h(Y)X|Y) = h(Y)E(X|Y)$
- Teljes szórásnégyzet tétele:

$$D^2(X) = D^2(E(X|Y)) + E(D^2(X|Y))$$

Példa (indikátor minta)

$$X_i = \begin{cases} 1, & p \text{ valószínűséggel} \\ 0, & 1-p \text{ valószínűséggel} \end{cases} \Rightarrow P_p(X_i = x) = p^x(1-p)^{1-x}, x = 0 \text{ és } 1.$$

$$P_p\left(\mathbf{X} = \mathbf{x} \middle| \sum_{i=1}^n X_i = t\right) = P_p\left(X_1 = x_1, \ldots, X_n = x_n \middle| \sum_{i=1}^n X_i = t\right) =$$

$$\frac{P_p\left(X_1 = x_1, \ldots, X_n = x_n, \sum_{i=1}^n X_i = t\right)}{P_p\left(\sum_{i=1}^n X_i = t\right)} = \begin{cases} 0 & \sum_{i=1}^n x_i \neq t \\ \dfrac{P_p(X_1 = x_1, \ldots, X_n = x_n)}{P_p\left(\sum_{i=1}^n X_i = t\right)} & \sum_{i=1}^n x_i = t \end{cases} =$$

$$\begin{cases} 0 & \sum_{i=1}^n x_i \neq t \\ \dfrac{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}}{\binom{n}{t} p^t (1-p)^{n-t}} & \sum_{i=1}^n x_i = t \end{cases} = \begin{cases} 0 & \sum_{i=1}^n x_i \neq t \\ \dfrac{1}{\binom{n}{t}} & \sum_{i=1}^n x_i = t \end{cases}$$

4.12 Neyman-féle faktorizációs

4.13 Poisson példa

Tétel (Neyman-féle faktorizációs):
A diszkrét ξ mintából képzett $T(\xi)$ statisztika pontosan akkor elégséges Θ-ra, ha $\exists\, g_\theta(t)$ és $h(\mathbf{x})$ úgy, hogy $\forall \theta \in \Theta$ és $\mathbf{x} \in \mathcal{X}$-ra
$P_\theta(\xi = \mathbf{x}) = h(\mathbf{x}) g_\theta(T(\mathbf{x}))$.
Biz.:

$\Rightarrow T(\xi)$ elégséges, ekkor $P_\theta(\xi = \mathbf{x}) = P_\theta(T(\xi) = T(\mathbf{x})) \dfrac{P_\theta(\xi = \mathbf{x}, T(\xi) = T(\mathbf{x}))}{P_\theta(T(\xi) = T(\mathbf{x}))}$

$= P_\theta(T(\xi) = T(\mathbf{x})) P_\theta(\xi = \mathbf{x} \,|\, T(\xi) = T(\mathbf{x})) = g_\theta(T(\mathbf{x})) h(\mathbf{x})$.

$\Leftarrow P_\theta(\xi = \mathbf{x} \,|\, T(\xi) = t) = 0$, ha $t \neq T(\mathbf{x})$. Amennyiben ez teljesül:

$$P_\theta(\xi = \mathbf{x} \,|\, T(\xi) = t) = \frac{P_\theta(\xi = \mathbf{x}, T(\xi) = t)}{P_\theta(T(\xi) = t)} = \frac{P_\theta(\xi = \mathbf{x}, T(\xi) = t)}{P_\theta(T(\xi) = t)} = \frac{P_\theta(\xi = \mathbf{x})}{\sum_{\mathbf{y}: T(\mathbf{y}) = t} P_\theta(\xi = \mathbf{y})}$$

$$= \frac{h(\mathbf{x}) g_\theta(T(\mathbf{x}))}{\sum_{\mathbf{y}: T(\mathbf{y}) = t} h(\mathbf{y}) g_\theta(T(\mathbf{y}))} = \frac{h(\mathbf{x}) g_\theta(t)}{\sum_{\mathbf{y}: T(\mathbf{y}) = t} h(\mathbf{y}) g_\theta(t)} = \frac{h(\mathbf{x})}{\sum_{\mathbf{y}: T(\mathbf{y}) = t} h(\mathbf{y})}.$$

Ez nem függ θ-tól!

Példa (Poisson minta)

η_i – k független λ Poissonok. Ekkor

$$P_\lambda\left(\eta_1 = k_1, \ldots, \eta_n = k_n\right) = \prod_{i=1}^{n} \frac{\lambda^{k_i} e^{-\lambda}}{k_i!} = \left(\prod_{i=1}^{n} \frac{1}{k_i!}\right) \lambda^{\sum_{i=1}^{n} k_i} e^{-n\lambda} =$$

$$= h(\mathbf{k}) g_\lambda\left(\sum_{i=1}^{n} k_i\right),$$

ahol

$$h(\mathbf{k}) = \prod_{i=1}^{n} \frac{1}{k_i!}, \quad g_\lambda(t) = \lambda^t e^{-n\lambda}.$$

4.14 Elégséges statisztika általában

4.15 Abszolút folytonos eset

Elégséges statisztika általában

Def.: A $\boldsymbol{\xi}$ mintából képzett $T(\boldsymbol{\xi})$ statisztika elégséges Θ-ra, ha minden $\mathbf{x} \in \mathbf{R}^n$-re a $P_\theta(\boldsymbol{\xi} < \mathbf{x} \,|\, T(\boldsymbol{\xi}) = t) = P_\theta(\xi_1 < x_1, \ldots, \xi_n < x_n \,|\, T(\boldsymbol{\xi}) = t)$ feltételes eloszlásfüggvény nem függ θ-tól.

Probléma: A feltételes valószínűség és várható érték fogalmát nem tanultuk általánosan!

Abszolút folytonos eset

- Definíció a faktorizációval

Def.:
Az abszolút folytonos $\boldsymbol{\xi}$ mintából képzett $T(\boldsymbol{\xi})$ statisztika elégséges $\Theta-$ra, ha $\exists\ g_\theta(t)$ és $h(\mathbf{x})$ úgy, hogy $\forall \theta \in \Theta$ és $\mathbf{x} \in \mathcal{X}-$ra a likelihood függvény felírható a következő alakban:
$L(\mathbf{x}, \theta) = h(\mathbf{x}) g_\theta(T(\mathbf{x}))$.

4.16 Példa normális eloszlásra

4.17 Pédla egyenletes eloszlásra

Példa (normális $N(m,\sigma^2)$ minta)

ξ_i – k független, $N(m,\sigma^2)$ eloszlásúak. Ekkor $\theta=(m,\sigma^2)$

$$L(\mathbf{x},(m,\sigma^2))=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-m)^2}{2\sigma^2}\right)=$$

$$\left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i^2-2x_im+m^2\right)\right)=$$

$$\left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}x_i^2-2nm\bar{x}+nm^2\right)\right)=$$

$$\left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2+n\bar{x}^2-2nm\bar{x}+nm^2\right)\right).$$

Ebből következik, hogy $\left(\sum_{i=1}^{n}x_i^2,\bar{x}\right)$ elégséges statisztika.

Hasonlóan $\left(\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2/n,\bar{x}\right)$ is.

Példa (egyenletes $E(0,a)$ minta)

ξ_i – k független, $E(0,a)$eloszlásúak.

Sűrűségfüggvényük

$$f_a(x)=\begin{cases}1/a & 0\le x\le a\\ 0 & \text{különben}\end{cases}$$

$$L(\mathbf{x},a)=\prod_{i=1}^{n}\frac{1}{a}\chi\{x_i\le a\}=\frac{1}{a^n}\chi\left\{\max_{1\le i\le n}x_i\le a\right\}$$

$\Rightarrow$

$\max_{1\le i\le n}x_i$ elégséges!

4.18 Maximum likelihood becslés

4.19 momentum módszer

Maximum likelihood becslés

- A maximum likelihood becslés mindig az elégséges statisztika függvénye
- $E(0,a)$ minta esetén a ML becslése

$$\max_{1 \le i \le n} \xi_i$$

Momentum módszer

- Ha az eloszlást k db paraméter határozza meg, akkor k db egyenletből kaphatunk rájuk becslést. Az egyenletek a tapasztalati és az elméleti momentumok egybevetéséből adódnak:

$$m_i(\underline{\theta}) = E_{\underline{\theta}}(X^i)$$

$$m_i(\underline{\theta}) = \frac{\sum_{j=1}^{n} \left(\xi_j\right)^i}{n}$$