



A matematikai statisztika alapfogalmai

Informatikai Tudományok Doktori Iskola



Adatbányászat vs Statisztika

Adatbányászat

- Valamely vizsgált populációra vonatkozólag nagymennyiségű, kontrollálatlan adathalmazból számítógépes adatkezelő technikákkal, algoritmusokkal a populációra vonatkozó hasznos információ, összefüggés kinyerése.
- Az adatok begyűjtése spontán, többnyire véletlen folyamatok eredményeképpen, nem tervezett módon történik.
- Megjelenése az informatikai világ kiteljesedése következtében történt meg. Modern tudományág. Az elméleti megalapozás napjainkban folyik.

2012. 02. 09.

Dr Ketskeméty László előadása

2



Adatbányászat vs Statisztika

Statisztika

- A vizsgált populációra vonatkozólag előre megtervezett módon, matematikai elvek figyelembe vételével beszerzett adatokkal, a minta feldolgozásával állítja elő a sokaságra vonatkozó hasznos következtetéseket.
- A statisztikai mintának reprezentatívnak kell lennie, különben a következtetések pontatlanok, megtévesztők lesznek!
- A valószínűségszámítással párhuzamosan fejlődött ki, erős matematikai elméleti háttérrel rendelkezik.

2012. 02. 09.

Dr Ketskeméty László előadása

3



A statisztika eredete

A statisztika eredetileg államszámtan volt. (Maga a „statisztika” szó is az „állam” jelentésű latin „status”-ból alakult ki.) A statisztika az ókortól kezdve arról tájékoztatta az államok vezetőit, mekkora adókat vehetnek ki alattvalóikra és hány katonára számíthatnak egy eljövendő háborúban.

A statisztika csak a polgári forradalmak után vált igazi tudománnyá. Úttörői JOHN GRAUNT (1620–1674) és WILLIAM PETTY (1623–1687). A kapitalizmusban már nemcsak az államok vezetőit, hanem a tőkés vállalkozókat is érdekelni kezdték a statisztikai felmérések, és egyre komolyabb matematikai eszközöket használtak föl adataik feldolgozására, egyre növekvő haszonnal, például a biztosításban.

A jó biztosítás alapja a pontos felmérés és a helyes matematikai következtetés. A XVII. század óta a matematikai statisztika fokozatosan a matematika önálló ágává fejlődött, amelynek fő célja: minél megbízhatóbb hasznosítható információt nyerni a felmérési, megfigyelési és mérési adatokból: a statisztikai mintából.

Székely J. Gábor „Paradoxonok a véletlen matematikájában”

2012. 02. 09.

Dr Ketskeméty László előadása

4



Statisztika bonmotok

„Csak abban a statisztikában hiszek, amit én magam hamisítok”
„A statisztika olyan, mint a bikini: sok minden megmutat, de a lényeget eltakarja.” W. Churchill

„Az éhezöket nem lehet statisztikával táplálni” Lloyd George

„Kis hazugság, nagy hazugság, statisztika!” Benjamin Disraeli

„A statisztika nem ad választ minden tudásra.”

„Az élet voltaképp nem más, mint a halál statisztikai hibája.”

„Nagy barátja vagyok a statisztikának; nem mintha azt hinném, hogy az csakugyan annyit bizonyít, mint sokan felteszik, hanem azért, mert mióta minden állításnak statisztikai adatokkal való támogatása divattá vált, a hamis tételek feldollítása valamivel több nehézséggel jár, s a tudományos paradoxonok alkotói badarságaikat legalább jobb rendszerben adják elő.”
Eötvös József

2012. 02. 09.

Dr Ketskeméty László előadása

5



Alapfogalmak

- Sokaság, populáció, véletlen kísérlet
- Statisztikai minta, minta realizáció
- Statisztikai mintavétel
- Statisztika
- Paraméter
- Statisztikai becslés

2012. 02. 09.

Dr Ketskeméty László előadása

6



Statisztikai sokaság, populáció

A vizsgálat tárgyát képező nagyszámú de véges elemszámú egyedek halmaza. A halmaz egészének kevés adattal történő tömör jellemzése, és a populáció egyedeinek leírására bevezetett változók közötti kapcsolatok leírása a célunk. Arra nincs lehetőség (erőforrás), hogy a populáció minden egyes eleméről adatokat szerezzünk be.

- Magyarország állampolgárai
- Egy egyetemi kar hallgatói
- Az érvényes forgalmival rendelkező autók halmaza
- Egy adott termék vásárlóinak halmaza
- Egy TV csatorna nézőinek halmaza

2012. 02. 09.

Dr Ketskeméty László előadása

7



Statisztikai sokaság, populáció

A statisztikai elemzés tárgya lehet egy véletlen kísérlet is, ami időben változatlan körülmények között elvileg akárhányszor lejátszódhat. A valószínűségszámítás tárgyalásában ezt **■**-val jelöltük.

- A lottóhúzás
- Egy szerver működése
- Budapest januári átlaghőmérséklete
- Egy gyümölcsös terméshozama
- Egy új gyógyszer hatása
- Egy reklámkampány hatásossága
- Egy populáció egyedének véletlen kiválasztása

2012. 02. 09.

Dr Ketskeméty László előadása

8



Statisztikai minta realizáltja

A populáció egy kis elemszámú részhalmazára vonatkozó megfigyelések adatai. A minta úgy kell, hogy tükrözze a populáció tulajdonságait, ahogy a cseppben látjuk a tengert. Azaz a minta reprezentatív kell, hogy legyen.

- Egy felmérésbe bevont magyar állampolgárok halmaza
- Egy adott előadásra belátogatott hallgatók halmaza
- Adott biztosítóval szerződött autók halmaza
- Egy adott napon megkérdezett vásárlók halmaza
- Egy nézettségi felmérésbe bevont TV nézők halmaza
- Budapest januári középhőmérsékleteinek adatai

2012. 02. 09.

Dr Ketskeméty László előadása

9



Mintavételezési eljárások

A populáció minden egyes elemének ugyanakkora esélyt kell biztosítani a mintába kerüléshez.

A minta elemszámának elég nagynek kell lennie ahhoz, hogy a következtetéseink átvihetők lehessenek a populációra is.

Rétegzett mintavételezés: A populációt adott szempontok szerint csoportokba osztjuk, és a csoportok arányait a mintában is megtartjuk

Véletlen mintavételezés: A mintába kerülő egyedeket sorsolással választjuk ki.

Cenzus: népszámlálás

2012. 02. 09.

Dr Ketskeméty László előadása

10



Alapfogalmak

Eset

A minta egy eleme, az adatmátrix egy sora.

Mintaelemszám

Az adott minta elemeinek száma. Egy adatmátrix sorainak száma.

Adatmátrix

n db eset és p db változó adatainak mátrixba rendezett alakzata

Változó

A populáció egy mérhető jellemzője. Az adatmátrix egy oszlopa.

2012. 02. 09.

Dr Ketskeméty László előadása

11



Példák változókra

- Magyarország állampolgárai: **fizetés; kor; nem; párt stb.**
- Egy egyetemi kar hallgatói: **gönygyölt tanulmányi átlag; neptun-kód; nem; szak; teljesített kreditek száma stb.**
- Az autók halmaza: **gyorsulás; fogyasztás; lóerő; típus;...**
- Egy adott termék vásárlóinak halmaza: **vélemény az árról; minőségről;...**
- Egy TV csatorna nézőinek halmaza: **kor; nem; tetszési index; iskolázottság; stb.**

2012. 02. 09.

Dr Ketskeméty László előadása

12



Statisztika

A minta realizáció adataiból adott képlettel számolt adat a statisztika számított értéke.

átlag, standard szórás, medián, kvartilis, ferdeség, lapultság, módusz, gyakoriság, próbat statisztikák, stb.

2012. 02. 09.

Dr Ketskeméty László előadása

13



A matematikai statisztika alapmodellje

$\mathcal{E}$

a véletlen kísérlet

Ω

a lehetséges kimenetek halmaza

$\mathcal{E}$

a megfigyelhető események halmaza

$\mathcal{P}$

a lehetséges valószínűségi mértékek halmaza

Az elemzésünk célja, hogy ebből a halmazból kiválasszuk a tényleges valószínűséget! Legalább is egy jó helyettesítő egyedet.

2012. 02. 09.

Dr Ketskeméty László előadása

14



A változó matematikai fogalma

$X: \Omega \rightarrow \mathbb{R}$ a vizsgált valószínűségi változó

X -nek minden $P \in \mathcal{P}$ esetén megadható az eloszlásfüggvénye!

$F_X(t) = P(X \leq t)$ minden $P \in \mathcal{P}$ -re!

$\mathcal{F} = \{F_X(t) : F_X(t) = P(X \leq t) \text{ minden } P \in \mathcal{P}\text{-re}\}$

Feladatunk tehát, ebből a halmazból kiválasztani a valószínűséget legjobban leíró eloszlásfüggvényt!

2012. 02. 09.

Dr Ketskeméty László előadása

15



A statisztikai minta fogalma

Az X valószínűségi változóval azonos eloszlású, egymással teljesen független $X_1, X_2, \dots, X_n$ valószínűségi változók együttesét *statisztikai mintának* nevezzük.

A gyakorlati alkalmazásokban pedig n db szám!

minta eloszlásfüggvénye is.

Egy mintavételkor tulajdonképpen megfigyeljük a ω véletlen kísérletet, azaz megállapítjuk melyik $\omega \in \Omega$ kimenetele realizálódott.

Az $X_1(\omega) = x_1, X_2(\omega) = x_2, \dots, X_n(\omega) = x_n$ szám n -est nevezzük a minta realizációjának.

2012. 02. 09.

Dr Ketskeméty László előadása

16



A statisztikai minta fogalma

1.1. definíció: Legyen $(\Omega, \mathcal{F})$ mérhető tér, és $\mathbb{P}$ valószínűségi mértékek egy halmaza, ahol $\forall \mathbb{P} \in \mathcal{P}$ esetén $(\Omega, \mathcal{F}, \mathbb{P})$ Kolmogorov-féle valószínűségi mező. Az $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$ statisztikai megfigyelést *statisztikai mintának* nevezzük, ha X_i -k teljesen független, azonos eloszlású valószínűségi változók $\forall \mathbb{P} \in \mathcal{P}$ esetén $(\Omega, \mathcal{F}, \mathbb{P})$ -n, azaz $\forall \mathbb{P} \in \mathcal{P}$ -re

$$\mathbb{P}(X_i < x) = F_{\mathbb{P}}(x) \quad (i = 1, 2, 3, \dots, n)$$

és

$$\mathbb{P}(X_{i_1} < x_{i_1}, X_{i_2} < x_{i_2}, \dots, X_{i_k} < x_{i_k}) = \prod_{a=1}^k F_{\mathbb{P}}(x_{i_a}) \quad (\forall 2 \leq k \leq n).$$

n a minta elemszáma, $F_{\mathbb{P}}(x)$ a minta eloszlásfüggvénye, X_i az i -edik mintaelem, $\mu_{\mathbb{P}}(A) = \mathbb{P}(X_i \in A)$, $A \in \mathcal{B}^d$ a minta eloszlása. Egy $\omega \in \Omega$ esetén az

$$X_1(\omega) = x_1, X_2(\omega) = x_2, \dots, X_n(\omega) = x_n$$

szám n -es a minta egy realizációja.

2012. 02. 09.

Dr Ketskeméty László előadása

17



Egy példa I.

Populáció

Tekintsük az USA-ban, Európában és Japánban a 70-es, 80-as években gyártott gépjárművek halmazát!

Változók

mpg	hány mérföldet tesz meg egy gallon üzemanyaggal
engine	hengerűrtartalom inch ³ -ben
horse	motorteljesítmény lóerőben
weight	az autó súlya fontban
accel	hány sec alatt éri el a 60 mph/hour sebességet
year	a gyártás éve (utolsó két számjegy: 19..)
origin	a gyártóhely: 1-USA, 2-Európa, 3-Japán
cylinder	a hengerek száma (3, 4, 5, 6, 8)

2012. 02. 09.

Dr Ketskeméty László előadása

18



Egy példa II.

A populációhoz képzünk egy $n=406$ elemű mintát!
Azaz az 1970 és 1982 között a térségekben le-
gyártott gépjárművek közül kiválasztunk 406-ot
és megmérjük a változókhoz tartozó értékeket.
Az adatokat egy mátrixba foglaljuk.
Az adatmátrixban olvasható adathalmaz lesz a **minta-
realizáció**. Tudjuk, hogy a mintavételezéskor a vélet-
lentől függött, hogy melyik autót vizsgáltuk meg, azaz
kaphattunk volna másik adatmátrixot is!

A **statisztikai minta** egy absztrakcióval nyert fogalom:
a mintarealizáció csupán egy lehetséges értékvétele.

2012. 02. 09.

Dr Ketskeméty László előadása

19



Egy példa III.

Az adatmátrix első 17 esete:

	mpg	engine	horse	weight	accel	year	origin	cylinder
1	18	307	130	3504	12	70	American	8 Cylinders
2	15	350	165	3693	12	70	American	8 Cylinders
3	18	318	150	3436	11	70	American	8 Cylinders
4	16	304	150	3433	12	70	American	8 Cylinders
5	17	302	140	3449	11	70	American	8 Cylinders
6	15	429	198	4341	10	70	American	8 Cylinders
7	14	454	220	4354	9	70	American	8 Cylinders
8	14	440	215	4312	9	70	American	8 Cylinders
9	14	455	225	4425	10	70	American	8 Cylinders
10	15	390	190	3850	9	70	American	8 Cylinders
11	.	133	115	3090	18	70	European	4 Cylinders
12	.	350	165	4142	12	70	American	8 Cylinders
13	.	351	153	4034	11	70	American	8 Cylinders
14	.	363	175	4186	11	70	American	8 Cylinders
15	.	360	175	3850	11	70	American	8 Cylinders
16	15	383	170	3563	10	70	American	8 Cylinders
17	14	340	160	3609	8	70	American	8 Cylinders

2012. 02. 09.

Dr Ketskeméty László előadása

20



Egy példa IV.

értékcimkék

Variable Values		
Value		Label
year	0*	0 (Missing)
	70	70
	71	71
	72	72
	73	73
	74	74
	75	75
	76	76
	77	77
	78	78
	79	79
	80	80
	81	81
	82	82
origin	1	American
	2	European
	3	Japanese
cylinder	3	3 Cylinders
	4	4 Cylinders
	5	5 Cylinders
	6	6 Cylinders
	8	8 Cylinders
filter_5	0	Not Selected
	1	Selected

a. Missing value

2012. 02. 09.

Dr Ketskeméty László előadása

21



Egy példa V.

Gyakoriságok

origin Country of Origin

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	American	253	62,3	62,5	62,5
	European	73	18,0	18,0	80,5
	Japanese	79	19,5	19,5	100,0
	Total	405	99,8	100,0	
Missing	System	1	,2		
Total		406	100,0		

2012. 02. 09.

Dr Ketskeméty László előadása

22



Egy példa VI.

Gyakoriságok

year Model Year (modulo 100)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	70	34	8,4	8,4	8,4
	71	29	7,1	7,2	15,6
	72	28	6,9	6,9	22,5
	73	40	9,9	9,9	32,3
	74	27	6,7	6,7	39,0
	75	30	7,4	7,4	46,4
	76	34	8,4	8,4	54,8
	77	28	6,9	6,9	61,7
	78	36	8,9	8,9	70,6
	79	29	7,1	7,2	77,8
	80	29	7,1	7,2	84,9
	81	30	7,4	7,4	92,3
	82	31	7,6	7,7	100,0
	Total	405	99,8	100,0	
Missing	0 (Missing)	1	,2		
Total		406	100,0		

2012. 02. 09.

Dr Ketskeméty László előadása

23



Egy példa VII.

Gyakoriságok

cylinder Number of Cylinders

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3 Cylinders	4	1,0	1,0	1,0
	4 Cylinders	207	51,0	51,1	52,1
	5 Cylinders	3	,7	,7	52,8
	6 Cylinders	84	20,7	20,7	73,6
	8 Cylinders	107	26,4	26,4	100,0
	Total	405	99,8	100,0	
Missing	System	1	,2		
Total		406	100,0		

2012. 02. 09.

Dr Ketskeméty László előadása

24



Egy példa VIII.

A leíró statisztikák számított értékei:

Descriptives			
		Statistic	Std. Error
Miles per Gallon	Mean	23,51	,392
	95% Confidence Interval for Mean	Lower Bound	22,74
		Upper Bound	24,28
	5% Trimmed Mean	23,22	
	Median	23,00	
	Variance	61,090	
	Std. Deviation	7,816	
	Minimum	9	
	Maximum	47	
	Range	38	
	Interquartile Range	12	
	Skewness	,457	,122
	Kurtosis	-,511	,244

2012. 02. 09.

Dr Ketskeméty László előadása

25



A statisztika matematikai fogalma

Legyen f_n egy n -változós valós függvény. Akkor a statisztikai minta $T_n = f_n(X_1, X_2, \dots, X_n)$ függvényét nevezzük *statisztikának*.

A statisztika egy valószínűségi változó, aminek eloszlásfüggvényét a minta eloszlásfüggvényéből lehet kiszámolni.

A $T_n = f_n(X_1, X_2, \dots, X_n)$ szám (amikor az argumentumba a mintarealizáció értékeit helyettesítjük, a statisztika számolt értéke.

2012. 02. 09.

Dr Ketskeméty László előadása

26



Az adatcentrum statisztikái

ÁTLAG (mean)

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

MEDIÁN (median)

$$m_n = \begin{cases} \frac{x_{\frac{n+1}{2}}^*}{2} & , \text{ha } n \text{ páratlan} \\ \frac{x_{\frac{n}{2}}^* + x_{\frac{n}{2}+1}^*}{2} & , \text{ha } n \text{ páros} \end{cases}$$

MÓDUSZ (mode)

A leggyakrabban előforduló érték a mintában

2012. 02. 09.

Dr Ketskeméty László előadása

27



A szóródást jellemző statisztikák

STANDARD SZÓRÁS (deviation)

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

VARIÁCIÓ (variance)

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

TERJEDELEM (range)

$$x_n^* - x_1^*$$

2012. 02. 09.

Dr Ketskeméty László előadása

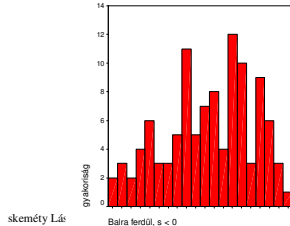
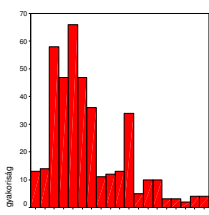
28



Az eloszlást jellemző statisztikák

FERDESÉG (skewness)

$$s = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \right)^3}$$

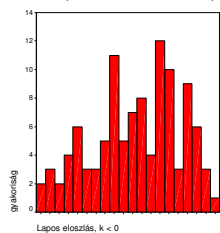
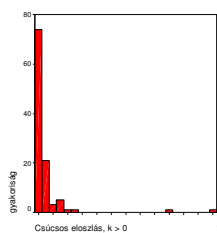




Az eloszlást jellemző statisztikák

LAPULTSÁG (curtosis)

$$k = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \right)^4} - 3$$



30



A rendezett minta statisztikák I.

Tekintsük azokat az $\text{ord}_k(x_1, x_2, \dots, x_n)$ skalár-vektor függvényeket, melyek definíciója: $x_k^* = \text{ord}_k(x_1, x_2, \dots, x_n) = x_j$, ha x_j a k -adik legnagyobb elem $x_1, x_2, \dots, x_n$ között.

rendezett mintaelem-statisztikák

$$X_k^* = \text{ord}_k(X_1, X_2, \dots, X_n) \quad (k = 1, 2, \dots, n)$$

$$X_1^* = \min \{X_1, X_2, \dots, X_n\}, \quad X_n^* = \max \{X_1, X_2, \dots, X_n\}.$$

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

2012. 02. 09.

Dr Ketskeméty László előadása

31



A rendezett minta statisztikák II.

Ha a minta eloszlásfüggvényét $F(x)$ -szel jelöljük,

$$F_k(x) = \mathbf{P}(X_k^* < x) = \sum_{i=k}^n \binom{n}{i} [F(x)]^i [1 - F(x)]^{n-i},$$

$$F_{k,l}(x, y) = \mathbf{P}(X_k^* < x, X_l^* < y) =$$

$$= \sum_{i=k}^n \sum_{j=l}^i \frac{n!}{j!(i-j)!(n-i)!} [F(x)]^j [F(y) - F(x)]^{i-j} [1 - F(y)]^{n-i},$$

$$\mathbf{P}(X_1^* < x_1, X_2^* < x_2, \dots, X_n^* < x_n) =$$

$$= \sum_{i_n=n}^n \sum_{i_{n-1}=n-1}^{i_n} \dots \sum_{i_1=1}^{i_2} \frac{n!}{i_1!(i_2-i_1)! \dots (n-i_n)!} [F(x_1)]^{i_1} [F(x_2) - F(x_1)]^{i_2-i_1} \dots [1 - F(x_n)]^{n-i_n}.$$

2012. 02. 09.

Dr Ketskeméty László előadása

32



A rendezett minta statisztikák III.

Az empirikus eloszlásfüggvény

$$F_n(x) = \begin{cases} 0, & \text{ha } x \leq X_1^* \\ \frac{k}{n}, & \text{ha } X_k^* < x \leq X_{k+1}^* \quad (k = 1, 2, \dots, n-1) \\ 1, & \text{ha } x > X_n^* \end{cases}$$

$$F_n(x) = \sum_{i=1}^n I_{\{X_i < x\}}, \text{ ahol } I_{\{X_k < x\}} = \begin{cases} 1, & \text{ha } X_k < x \\ 0, & \text{ha } X_k \geq x \end{cases}.$$

Az empirikus eloszlásfüggvény minden x helyen eloszlásfüggvény

Ugyanakkor az eloszlásfüggvény a statisztikai minta függvénye is, azaz minden x helyen valószínűségi változó lesz

2012. 02. 09.

Dr Ketskeméty László előadása

33



$$\mathbf{P} \left(\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = 0 \right) = 1.$$

Az empirikus eloszlásfüggvény 1 valószínűséggel, egyenletesen konvergál az eloszlásfüggvényhez.

Bizonyítás: Legyen $\varepsilon > 0$, $x \in \mathbb{R}$ tetszőleges! Megmutatjuk, hogy $\exists N > 0$ és $C \in \mathcal{F} : \mathbf{P}(C) = 1$, hogy $\forall \omega \in C$ esetén, ha $n > N$, úgy $|F_n(x) - F(x)| < \varepsilon$. Legyen m olyan pozitív egész szám, hogy $\frac{1}{m} < \frac{\varepsilon}{2}$, és legyenek $\mathbb{R}$ egy m intervallumból álló rendszerének osztópontjai $x_0^{(m)} = -\infty$, $x_m^{(m)} = +\infty$, $x_k^{(m)} = \sup_{x \in \mathbb{R}} \{x : F(x) \leq \frac{k}{m}\}$. Jelölje az intervallumokat: $J_k = (x_{k-1}^{(m)}, x_k^{(m)}]$, $k = 0, 1, \dots, m-1$. Tegyük fel, hogy a szóban forgó x -re éppen $x \in J_{k-1} \Rightarrow x_{k-1}^{(m)} < x \leq x_k^{(m)}$ teljesül most. Az eloszlásfüggvény tulajdonságai miatt:

$$\left. \begin{aligned} F(x_k^{(m)}) &\leq \frac{k}{m} \leq F(x_k^{(m)} + 0) \\ F(x_{k-1}^{(m)}) &\leq \frac{k-1}{m} \leq F(x_{k-1}^{(m)} + 0) \end{aligned} \right\} \Rightarrow (*) \quad F(x_k^{(m)}) \leq \frac{k}{m} \leq F(x_{k-1}^{(m)} + 0) + \frac{1}{m}.$$

A nagy számok erős törvénye értelmében a relatív gyakoriság 1 valószínűséggel közelíti az elméleti valószínűséget:

$$\exists A_k \in \mathcal{F} : \mathbf{P}(A_k) = 1 \text{ és } \forall \omega \in A_k : \lim_{n \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n I_{\{X_i < x_k^{(m)}\}}(\omega) \right) = F(x_k^{(m)}).$$

$$\exists B_k \in \mathcal{F} : \mathbf{P}(B_k) = 1 \text{ és } \forall \omega \in B_k : \lim_{n \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n I_{\{X_i \leq x_{k-1}^{(m)}\}}(\omega) \right) = F(x_{k-1}^{(m)} + 0).$$

$$\text{Legyen } C = \prod_{k=1}^m A_k B_{k-1}. \text{ Akkor } \mathbf{P}(C) = \mathbf{P} \left(\prod_{k=1}^m A_k B_{k-1} \right) = 1 \Rightarrow \mathbf{P}(C) = 1.$$



Tehát $\forall \omega \in C$ esetén $\exists N : n > N$, akkor

$$\left| \frac{1}{n} \sum_{i=1}^n I_{\{X_i < x_k^{(m)}\}} - F(x_k^{(m)}) \right| < \frac{\varepsilon}{2}, \text{ és } \left| \frac{1}{n} \sum_{i=1}^n I_{\{X_i \leq x_{k-1}^{(m)}\}} - F(x_{k-1}^{(m)} + 0) \right| < \frac{\varepsilon}{2}.$$

Így $x \in J_{k-1}$ -re

$$F(x) - F_n(x) \leq F(x_k^{(m)}) - F_n(x_k^{(m)}) \leq F(x_k^{(m)} + 0) + \frac{1}{m} - F_n(x_{k-1}^{(m)} + 0) \leq \frac{1}{m} + \frac{\varepsilon}{2}.$$

Másrészt

$$F(x) - F_n(x) \geq F(x_{k-1}^{(m)} + 0) - F_n(x_{k-1}^{(m)} + 0) \geq F(x_{k-1}^{(m)} + 0) - \frac{1}{m} - F_n(x_k^{(m)}) \geq -\frac{1}{m} - \frac{\varepsilon}{2}.$$

Azaz $|F(x) - F_n(x)| < \frac{\varepsilon}{2} + \frac{1}{m} < \varepsilon \Rightarrow$ állítás.



A paraméter

Tegyük fel, hogy a minta eloszlásfüggvénye képletét egy ν paraméter konkretizálja. Ha ismerjük az értékét, meg tudjuk pontosan adni az eloszlásfüggvényt:

$$\mathcal{F} = \{F_X(t, \nu) : \nu \in \Theta\}$$

Egy adott statisztikai minta segítségével a ν paraméter megbecslése a célunk!

2012. 02. 09.

Dr Ketskeméty László előadása

37



Példa paraméteres problémákra

1. Egy joghurt zsírtartalmát ellenőrzik. A laborban σ pontossággal meg tudják mérni a zsírtartalmat. A mérés a pontos érték körül a normális eloszlás szerint ingadozik. Ha vesznek egy mintát, akkor a minta eloszlása $N(\nu, \sigma)$!
2. Egy brókerirodában m ügyfél kötvényeit kezelik. Egy ügyfél ν valószínűséggel kér eladást/vételt az irodától. A napi tranzakciók száma $\text{Bin}(m, \nu)$ eloszlást követ.

2012. 02. 09.

Dr Ketskeméty László előadása

38



A paraméter becslése

A ν paraméter becsléséhez valamilyen alkalmas T_n statisztikát használunk: $T_n \approx \nu$.

Egy ismeretlen számot (a ν -át) egy valószínűségi változóval becsüljük!

Mikor jó egy ilyen becslés???

2012. 02. 09.

Dr Ketskeméty László előadása

39



A paraméter becslése I.

Torzítatlanság

Valószínűségszámításból tanultuk, hogy egy valószínűségi változó az összes szám közül éppen a várható értéke körül ingadozik a legkisebb mértékben.

A T_n statisztika a ϑ paraméter torzítatlan becslése, ha $E T_n = \vartheta$.

A torzítatlanság azt jelenti, hogy a becsülő statisztika éppen a becsülendő paraméterérték körül fogja felvenni az értékeit. Lövészhasonlaltal: „a találatokhoz a célkereszt jól van beállítva, nem hord félre a fegyver.”

2012. 02. 09.

Dr Ketskeméty László előadása

40



A paraméter becslése II.

A becsülendő paraméter, ϑ .

Egy nem torzítatlan becsülő statisztika realizáltjai. Ilyen statisztika *torzított*.

Egy torzítatlan becsülő statisztika realizáltjai a minta elemszám függvényében.

2012. 02. 09.

Dr Ketskeméty László előadása

41

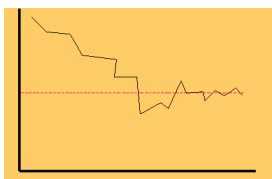


A paraméter becslése III.

Aszimptotikus torzítatlanság

Ha a torzítatlansági feltétel csak $n \rightarrow \infty$ esetben igaz:

$$\lim_{n \rightarrow \infty} E T_n = \vartheta$$



2012. 02. 09.

Dr Ketskeméty László előadása

42



A paraméter becslése IV.

Konzisztencia

Ha garancia van arra, hogy a minta elemszám növekedtével növekszik a becslés pontosságának valószínűsége, konzisztens becslésről beszélünk:

Minden $\varepsilon > 0$ szám esetén teljesül, hogy

$$\lim_{n \rightarrow \infty} \mathbf{P}(|T_n - \vartheta| > \varepsilon) = 0$$

A statisztika, mint valószínűségi változó sorozat, sztochasztikusan konvergál a ϑ konstanshoz!

2012. 02. 09.

Dr Ketskeméty László előadása

43



A paraméter becslése V.

Erős konzisztencia

Azok a torzítatlan becslések, elemszám növekedtével 0-hoz tart:

$$\mathbf{E}T_n = \vartheta \text{ és } \lim_{n \rightarrow \infty} \mathbf{D}^2T_n = 0$$

Csak a konstansnak lehet 0 a varianciája. Tehát, ha n elég nagy, a becslés gyakorlatilag a paramétert adja!

A Csebisev-egyenlőtlenségből következik, hogy az erősen konzisztens statisztikai becslések egyben konzisztensek is lesznek. A megfordítás általában nem igaz!

2012. 02. 09.

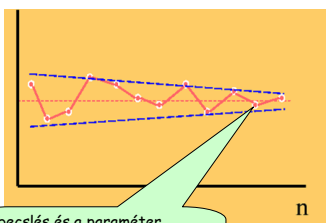
Dr Ketskeméty László előadása

44



A paraméter becslése VI.

Konzisztencia, erős konzisztencia



A becslés és a paraméter eltérése az n növekedtével csökkenni fog!

2012. 02. 09.

Dr Ketskeméty László előadása

45



Példák becslésekre II.

Ha még azt is tudjuk, hogy $D^2X < \infty$, akkor az átlag erősen konzisztens is:

$$D^2 \left(\frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n D^2 X_i = \frac{D^2 X}{n} \rightarrow 0$$

A lineáris becslések között az átlag a hatásos:

$$D^2 \left(\frac{1}{n} \sum_{i=1}^n X_i \right) \leq D^2 \left(\sum_{i=1}^n w_i \cdot X_i \right)$$

$$\sum_{i=1}^n w_i = 1 \text{ és } w_i \geq 0$$

2012. 02. 09.

Dr Ketskeméty László előadása

49



Példák becslésekre III.

Legyen a becslendő paraméter most az X varianciája:

$$D^2 X = \vartheta$$

Az empirikus szórásnégyzet aszimptotikusan torzítatlan, a korrigált empirikus szórásnégyzet pedig torzítatlan becslés!

$$E s_n^2 = E \left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right) = E \left(\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 \right) =$$

$$= \frac{1}{n} \sum_{i=1}^n E X_i^2 - E \bar{X}_n^2 = \frac{1}{n} \sum_{i=1}^n (\vartheta + m^2) - \left(\frac{\vartheta}{n} + m^2 \right) = \frac{n-1}{n} \vartheta$$

$$E s_n^{*2} = E \left(\frac{n}{n-1} s_n^2 \right) = \vartheta$$

2012. 02. 09.

Dr Ketskeméty László előadása

50



2.1.4. definíció: A $t_n(X_1, X_2, \dots, X_n) \in \mathbb{R}^k$ statisztikasorozat a $\theta \in \mathbb{R}^k$ paraméter négyzetes középben konzisztens becslése, ha $\lim_{n \rightarrow \infty} E \|t_n - \theta\|^2 = 0$.

2.1.1. tétel: Ha a t_n ($n = 1, 2, \dots$) statisztikasorozat négyzetes középben konzisztens becslése θ -nak, akkor konzisztens becslése is.

Bizonyítás: A Markov-egyenlőtlenségből:

$$P \left(\|t_n - \theta\|^2 > \varepsilon^2 \right) \leq \frac{E \|t_n - \theta\|^2}{\varepsilon^2} \rightarrow 0 \quad (n \rightarrow \infty).$$

2.1.2. tétel: Ha a t_n ($n = 1, 2, \dots$) statisztikasorozat aszimptotikusan torzítatlan becslése θ -nak és $\lim_{n \rightarrow \infty} \sigma_P^{(i)} = 0$ ($i = 1, 2, \dots, k$), akkor konzisztens becslése is.

Bizonyítás:

$$E (t_n^{(i)} - \theta_i)^2 = E (t_n^{(i)} - E_P t_n^{(i)} + E_P t_n^{(i)} - \theta_i)^2 =$$

$$= E (t_n^{(i)} - E_P t_n^{(i)})^2 + (E_P t_n^{(i)} - \theta_i)^2 + 2 E_P \left[(t_n^{(i)} - E_P t_n^{(i)}) (E_P t_n^{(i)} - \theta_i) \right] =$$

$$= \sigma_P^2(t_n^{(i)}) + (E_P t_n^{(i)} - \theta_i)^2 \rightarrow 0, \quad n \rightarrow \infty.$$

Viszont a Markov-egyenlőtlenség szerint:

$$P \left(|t_n^{(i)} - \theta_i| > \varepsilon \right) = P \left((t_n^{(i)} - \theta_i)^2 > \varepsilon^2 \right) \leq \frac{E (t_n^{(i)} - \theta_i)^2}{\varepsilon^2} \rightarrow 0,$$

amiből már következik az állítás.

2012. 02. 09.

Dr Ketskeméty László előadása

51



(ii) Ha a feltételekhez azt is hozzávesszük, hogy $\forall \mathbf{P} \in \mathcal{P}$ -re $\mathbf{E}_{\mathbf{P}} X^4$ is, úgy s_n^2 konzisztens, s_n^2 négyzetes középben konzisztens becslés is.

(ii) Belátható, hogy

$$\sigma_{\mathbf{P}, s_n^2}^2 = \frac{n^2}{(n-1)^2} \left(\frac{\mathbf{E}_{\mathbf{P}} X^4}{n} - \frac{n-3}{n(n-1)} \theta^2 \right) \rightarrow 0, \quad \sigma_{\mathbf{P}, s_n^2}^2 \rightarrow 0.$$

Összefoglalva:

- az **átlagstatisztika** a minta **várható értékének** -mint paraméternek- **torzítatlan** becslése. Ha a mintának létezik szórása, akkor ez a becslés **erősen konzisztens** is.
- A minta **tapasztalati szórásnégyzete** a minta **varianciájának** -mint paraméternek- **aszimptotikusan torzítatlan** becslése. Ha a mintának létezik negyedik momentuma, akkor a becslés **konzisztens** is.
- A minta **korrigált empirikus szórásnégyzet** statisztika a minta **varianciájának torzítatlan** becslése. Ha a minta negyedik momentuma létezik, akkor **erősen konzisztens** becslése.

2012. 02. 09.

Dr Ketskeméty László előadása

52



Az elégséges becslés

- A statisztikák elvárt, jó tulajdonságai között alapvető fontosságú az **elégségesség**. Ezen azt fogjuk érteni, hogy a statisztika a minta eloszlásának paraméterére vonatkozóan minden információt magába sűrít, egymaga képes helyettesíteni a mintát. A paraméterek becsléseihez a megfelelő statisztikákat „elégséges” az elégséges statisztika függvényei között keresni.

2012. 02. 09.

Dr Ketskeméty László előadása

53



Az elégséges becslés

Legyen adott a $\mathcal{P}$ paraméteres eloszláscsalád, és az $X_1, X_2, \dots, X_n$ statisztikai minta, amelyek eloszlásfüggvénye abszolút folytonos $\forall \mathbf{P} \in \mathcal{P}$ -re:

$$F_{\theta}(x) = \int_{-\infty}^x f_{\theta}(t) dt, \quad x \in \mathbb{R}.$$

$f_{\theta}(x)$ a minta sűrűségfüggvénye. Jelölje a $t_n(X_1, X_2, \dots, X_n)$ statisztika sűrűségfüggvényét $g_{n,\theta}(y)$, az $X_1, X_2, \dots, X_n$ és t_n együttes sűrűségfüggvényét pedig $h_{\theta}(x_1, x_2, \dots, x_n, y)$. Ha az $X_1, X_2, \dots, X_n$ mintának a t_n -re vonatkozó együttes feltételes sűrűségfüggvénye nem tartalmazza a θ paramétert, vagyis

$$f_{X_1, X_1, \dots, X_1 | t_n}(x_1, x_2, \dots, x_n | y) = \frac{h_{\theta}(x_1, x_2, \dots, x_n, y)}{g_{n,\theta}(y)},$$

nem függ θ -tól, akkor, a t_n statisztika a θ paraméter **elégséges** becslése.

2012. 02. 09.

Dr Ketskeméty László előadása

54



Rao-Blackwell-Kolmogorov tétel

Legyen adott $\mathcal{P}$, valószínűségi mértékek egy θ -paraméteres tere, és az $X_1, X_2, \dots, X_n$ statisztikai minta, amelyek eloszlásfüggvénye abszolút folytonos $\forall \mathbf{P} \in \mathcal{P}$ -re. Jelölje $T_n(X_1, X_2, \dots, X_n)$ a θ paraméter egy elégséges statisztikáját, $t_n(X_1, X_2, \dots, X_n)$ pedig a paraméter g függvényének tetszőleges torzítatlan becslését: $\mathbb{E}_\theta t_n = g(\theta)$. Akkor létezik olyan h függvény, hogy $\mathbb{E}_\theta(h(T_n)) = g(\theta)$ és $\sigma_\theta^2(h(T_n)) \leq \sigma_\theta^2(t_n)$. Továbbá $h(T_n) = \mathbb{E}_\theta(t_n | T_n)$.

Ha létezik hatásos (legjobb torzítatlan) becslés, akkor elég azt az elégséges becslés függvényei között keresni!

2012. 02. 09.

Dr Ketskeméty László előadása

55



Neymann-Fisher faktorizációs tétel

A T_n statisztika a θ paraméter elégséges becslése $\iff \exists k : \mathbb{R}^n \rightarrow \mathbb{R}$ és $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ függvények, hogy $\forall \mathbf{x} = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$ és $\forall \theta$ -ra

$$L_\theta(x_1, x_2, \dots, x_n) = k(x_1, x_2, \dots, x_n)g(T_n(x_1, x_2, \dots, x_n), \theta).$$

Ennek a tételnek a segítségével lehet ellenőrizni egy statisztikáról azt, hogy elégséges-e.

2012. 02. 09.

Dr Ketskeméty László előadása

56



Cramer-Rao-egyenlőtlenség

A Cramer-Rao-egyenlőtlenség elvi alsó korlátot ad a torzítatlan becslések szórásnégyzeteire. Ha egy statisztikára belátjuk, hogy szórásnégyzete éppen az alsó korlattal egyenlő, akkor az biztosan hatásos, sőt az egyetlen hatásos becslés!

Tegyük fel, hogy az $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$ statisztikai minta egyparaméteres $F_\mathbf{P}(x) = F_\theta(x)$ eloszlásfüggvénye abszolút folytonos: $\exists \frac{dF_\theta(x)}{dx} = f_\theta(x)$, $\theta \in (a, b)$. Jelölje

$$L_\theta(\mathbf{x}) = L_\theta(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_\theta(x_i)$$

a minta együttes sűrűségfüggvényét!

2012. 02. 09.

Dr Ketskeméty László előadása

57



Cramer-Rao-egyenlőtlenség

Feltételek:

- $I_n(\vartheta) = \int_{\mathbb{R}^n} \left(\frac{\partial L_\vartheta(\mathbf{x})}{\partial \vartheta} \right)^2 \cdot \frac{1}{L_\vartheta(\mathbf{x})} d\mathbf{x} < \infty$ (Fisher-féle információs mennyiség,)
- Legyen $g : (a, b) \rightarrow \mathbb{R}$ tetszőleges differenciálható függvény.
- Legyen a $t(\mathbf{X})$ statisztika a $g(\vartheta)$ torzítatlan becslése, azaz $\mathbb{E}_\vartheta(t) = g(\vartheta)$ ($\forall \vartheta \in (a, b)$).
- $\exists \sigma_\vartheta^2 t = \int_{\mathbb{R}^n} (t(\mathbf{x}) - g(\vartheta))^2 L_\vartheta(\mathbf{x}) d\mathbf{x}$.
- $\frac{\partial}{\partial \vartheta} \int_{\mathbb{R}^n} t^i(\mathbf{x}) L_\vartheta(\mathbf{x}) d\mathbf{x} = \int_{\mathbb{R}^n} t^i(\mathbf{x}) \frac{\partial L_\vartheta(\mathbf{x})}{\partial \vartheta} d\mathbf{x}$, ($i = 0, 1$).

Ekkor

$$\sigma_\vartheta^2 t \geq \frac{[g'(\vartheta)]^2}{I_n(\vartheta)}.$$

2012. 02. 09.

Dr Ketskeméty László előadása

58



Cramer-Rao-egyenlőtlenség

Példák:

- Az átlagstatisztika hatásos normális esetben
- Az átlagstatisztika hatásos exponenciális esetben
- Az átlagstatisztika hatásos Poisson esetben
- Az a korrigált empirikus szórásnégyzet hatásos normális esetben

2012. 02. 09.

Dr Ketskeméty László előadása

59



A maximum likelihood becslés

A módszer alapgondolatai a következők:

- A mintánk eloszlásfüggvénye a ν paramétertől függ.
- Ha egy kísérletnél több esemény is bekövetkezhet, legtöbbször a legnagyobb valószínűségű eseményt fogjuk megfigyelni.
- A sokaságra vett mintavételezés során kaptunk egy realizációt. Feltételezzük, hogy azért éppen ezt a realizációt kaptuk, és nem más, mert az összes realizációk közül ennek volt a legnagyobb a bekövetkezési valószínűsége.
- Vegyük tehát, az összes lehetséges ν paraméter közül azt, amelynél éppen kapott realizáció bekövetkezése a maximális.

2012. 02. 09.

Dr Ketskeméty László előadása

60



A maximum likelihood becslés, diszkrét eset

Legyen adott $\mathcal{P}$, valószínűségi mértékek egy tere és az $X_1, X_2, \dots, X_n$ diszkrét eloszlású statisztikai minta $E \subseteq \mathbb{R}$ értékkészlettel $\forall \mathbf{P}_\theta \in \mathcal{P}$ -re. Jelölje most

$$L(\mathbf{x}, \theta) = \mathbf{P}_\theta(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^n \mathbf{P}_\theta(X_i = x_i)$$

a minta együttes eloszlását. A θ paraméter maximum-likelihood becslésén azt a $\tau_n(X_1, X_2, \dots, X_n)$ statisztikát értjük, melyre

$$L(\mathbf{x}, \tau_n(\mathbf{x})) = \max_{\theta \in \mathbb{R}^k} L(\mathbf{x}, \theta)$$

teljesül ($\forall \mathbf{x} \in E^n$).

2012. 02. 09.

Dr Ketskeméty László előadása

61



A maximum likelihood becslés, Poisson-eloszlás

Most a minta eloszlása:

$$p_{\theta, i} = \frac{\theta^i}{i!} e^{-\theta} \quad i = 0, 1, 2, \dots$$

A likelihood függvény, a minta együttes eloszlásából számolható:

$$L(\mathbf{x}, \theta) = \prod_{i=1}^n \frac{\theta^{x_i}}{x_i!} = \frac{\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-n\theta},$$

a log-likelihood függvény pedig:

$$l(\mathbf{x}, \theta) = \ln \theta \sum_{i=1}^n x_i - n\theta - \ln \left(\prod_{i=1}^n x_i! \right).$$

2012. 02. 09.

Dr Ketskeméty László előadása

62



A maximum likelihood becslés, Poisson-eloszlás

A stacionárius helyek megkeresése:

$$\frac{\partial l(\mathbf{x}, \theta)}{\partial \theta} = \frac{1}{\theta} \sum_{i=1}^n x_i - n = 0 \implies \theta = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}_n.$$

Mivel

$$\frac{\partial^2 l(\mathbf{x}, \theta)}{\partial \theta^2} = -\frac{1}{\theta^2} \sum_{i=1}^n x_i < 0,$$

a kapott stacionáriushely maximum. Tehát a Poisson-eloszlás esetén is a paraméternek maximum-likelihood becslése az átlagstatisztika.

2012. 02. 09.

Dr Ketskeméty László előadása

63



A maximum likelihood becslés, folytonos eset

Legyen adott $\mathcal{P}$, valószínűségi mértékek egy tere és az $X_1, X_2, \dots, X_n$ statisztikai minta, amelyek eloszlásfüggvénye abszolút folytonos $\forall \mathbf{P}_\theta \in \mathcal{P}$ -re. Jelölje most

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

a minta együttes sűrűségfüggvényét. A θ paraméter maximum-likelihood becslésén azt a $\tau_n(X_1, X_2, \dots, X_n)$ statisztikát értjük, melyre

$$L(\mathbf{x}, \tau_n(\mathbf{x})) = \max_{\theta \in \mathbb{R}^k} L(\mathbf{x}, \theta)$$

teljesül ($\forall \mathbf{x} \in \mathbb{R}^n$).

2012. 02. 09.

Dr Ketskeméty László előadása

64



A maximum likelihood becslés, normális eloszlás, ismert szórás esetén

Legyen

$$f_\theta(x) = \frac{1}{\sqrt{2\pi} D_0} e^{-\frac{1}{2D_0^2}(x-\theta)^2},$$

ahol $D_0 > 0$ ismert, $\theta \in \mathbb{R}$ az ismeretlen paraméter.

Most a likelihood függvény:

$$L(\mathbf{x}, \theta) = \left(\frac{1}{\sqrt{2\pi} D_0} \right)^n e^{-\frac{1}{2D_0^2} \sum_{i=1}^n (x_i - \theta)^2},$$

a log-likelihood függvény pedig

$$l(\mathbf{x}, \theta) = n \ln \left(\frac{1}{\sqrt{2\pi} D_0} \right) - \frac{1}{2D_0^2} \sum_{i=1}^n (x_i - \theta)^2.$$

2012. 02. 09.

Dr Ketskeméty László előadása

65



A maximum likelihood becslés, normális eloszlás, ismert szórás esetén

A maximumhely keresése:

$$\frac{dl(\mathbf{x}, \theta)}{d\theta} = \frac{1}{D_0^2} \sum_{i=1}^n (x_i - \theta) = 0 \implies \theta = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}_n.$$

Mivel

$$\frac{d^2 l(\mathbf{x}, \theta)}{d\theta^2} = -\frac{1}{D_0^2} < 0,$$

a kapott stacionárius hely maximumhely. Tehát az átlagstatisztika normális esetben a várható érték maximum-likelihood becslése.

2012. 02. 09.

Dr Ketskeméty László előadása

66



A maximum likelihood becslés, normális eloszlás, két paraméteres eset

Legyen

$$f_{\theta_1, \theta_2}(x) = \frac{1}{\sqrt{2\pi} \theta_2} e^{-\frac{1}{2\theta_2^2}(x-\theta_1)^2},$$

ahol $\theta_2 > 0$ és $\theta_1 \in \mathbb{R}$ az ismeretlen paraméterek.

Most a likelihood függvény:

$$L(\mathbf{x}, \theta_1, \theta_2) = \left(\frac{1}{\sqrt{2\pi} \theta_2} \right)^n e^{-\frac{1}{2\theta_2^2} \sum_{i=1}^n (x_i - \theta_1)^2},$$

a log-likelihood függvény pedig

$$l(\mathbf{x}, \theta_1, \theta_2) = -n \ln \sqrt{2\pi} - \frac{n}{2} \ln \theta_2 - \frac{1}{2\theta_2^2} \sum_{i=1}^n (x_i - \theta_1)^2.$$

2012. 02. 09.

Dr Ketskeméty László előadása

67



A maximum likelihood becslés, normális eloszlás, két paraméteres eset

A maximumhely keresése:

$$\frac{\partial l(\mathbf{x}, \theta_1, \theta_2)}{\partial \theta_1} = \frac{1}{\theta_2} \sum_{i=1}^n (x_i - \theta_1) = 0 \implies \theta_1 = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}_n$$

$$\frac{\partial l(\mathbf{x}, \theta_1, \theta_2)}{\partial \theta_2} = -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^3} \sum_{i=1}^n (x_i - \theta_1)^2 = 0 \implies \theta_2 = \frac{1}{n} \sum_{i=1}^n (x_i - \theta_1)^2 = s_n^2$$

2012. 02. 09.

Dr Ketskeméty László előadása

68



A maximum likelihood becslés, normális eloszlás, két paraméteres eset

Mivel

$$\frac{\partial^2 l(\mathbf{x}, \theta_1, \theta_2)}{\partial \theta_1^2} = -\frac{n}{\theta_2^2},$$

$$\frac{\partial^2 l(\mathbf{x}, \theta_1, \theta_2)}{\partial \theta_2^2} = \frac{n}{2\theta_2^3} - \frac{1}{\theta_2^3} \sum_{i=1}^n (x_i - \theta_1)^2,$$

$$\frac{\partial^2 l(\mathbf{x}, \theta_1, \theta_2)}{\partial \theta_2 \partial \theta_1} = -\frac{1}{\theta_2^3} \sum_{i=1}^n (x_i - \theta_1),$$

a kapott stacionárius hely Hesse-mátrixa:

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

amiből látszik, hogy a hely maximumhely, tehát az átlagstatisztika és az empirikus szórásnégyzet statisztikák normális esetben az elméleti várható érték és szórásnégyzet maximum-likelihood becslései.

2012. 02. 09.

Dr Ketskeméty László előadása

69



A maximum likelihood becslés

Általános feltételek mellett megmutatható, hogy a maximum-likelihood becslés konzisztens, aszimptotikusan normális eloszlású, és ha van elégséges statisztika, akkor a maximum likelihood statisztika éppen azt adja meg!

2012. 02. 09.

Dr Ketskeméty László előadása

70



A maximum likelihood becslés

2.4.1. tétel: Legyen adott $\mathcal{P}$, valószínűségi mértékek egy tere és az $X_1, X_2, \dots, X_n$ statisztikai minta. Jelölje most $L(\mathbf{x}, \theta)$ a likelihood függvényt és τ_n a maximum-likelihood statisztikát!

- (i) Ha létezik hatásos becslés a θ paraméterre, akkor τ_n maga a hatásos becslés.
- (ii) Ha létezik T_n elégséges becslés a θ paraméterre, akkor megadható olyan $h(x)$ függvény, mellyel $h(\tau_n) = T_n$, azaz az elégséges becslés a maximum-likelihood statisztika függvényeként áll elő.

2012. 02. 09.

Dr Ketskeméty László előadása

71



A maximum likelihood becslés

Bizonyítás:

- (i) A Cramer-Rao-tétel után tett 2. megjegyzés szerint t_n^* hatásos becslés, ha

$$\frac{\partial l(\mathbf{x}, \theta)}{\partial \theta} = k(\theta)(t_n^*(\mathbf{x}) - \theta)$$

teljesül majdnem minden $\mathbf{x} \in \mathbb{R}^n$ vektorra. De a maximum-likelihood statisztikát éppen az $\frac{\partial l(\mathbf{x}, \theta)}{\partial \theta} = 0$ egyenlet megoldásából kapjuk, azaz

$$k(\theta)(t_n^*(\mathbf{x}) - \theta) = 0 \implies t_n^*(\mathbf{x}) = \tau_n(\mathbf{x}) = \theta \implies \text{állítás.}$$

- (ii) A Neymann-Fisher faktorizációs tételből: $\exists g, k$ függvények:

$$L(\mathbf{x}, \theta) = g(T_n(\mathbf{x}), \theta) \cdot k(\mathbf{x}).$$

Innen $\frac{\partial \ln L(\mathbf{x}, \theta)}{\partial \theta} = \frac{\partial g(T_n(\mathbf{x}), \theta)}{\partial \theta} = 0 \implies \exists h$ függvény: $h(T_n(\mathbf{x})) = \tau_n(\mathbf{x})$.

2012. 02. 09.

Dr Ketskeméty László előadása

72



A maximum likelihood becslés

2.4.2. tétel: (Cramer-Daguer)

Legyen adott $\mathcal{P}$, valószínűségi mértékek egy tere és az $X_1, X_2, \dots, X_n, \dots$ statisztikai minta, amelyek eloszlásfüggvénye abszolút folytonos $\forall \mathbf{P} \in \mathcal{P}$ -re. Tegyük fel, hogy a minta sűrűség-függvénye $f_\theta(x)$, $\theta \in (a, b)$ kielégíti az alábbi a), b), c) feltételeket:

a) $\exists \frac{\partial^i \ln f_\theta(x)}{\partial \theta^i} \quad i = 1, 2, 3 \quad \forall \theta \in (a, b)$.

b) $\exists H_1(x), H_2(x), H_3(x)$ függvények, melyekre:

$$\left| \frac{\partial f_\theta(x)}{\partial \theta} \right| < H_1(x), \quad \left| \frac{\partial^2 f_\theta(x)}{\partial \theta^2} \right| < H_2(x), \quad \left| \frac{\partial^3 f_\theta(x)}{\partial \theta^3} \right| < H_3(x).$$

$$\int_{-\infty}^{+\infty} H_1(x) dx < \infty, \quad \int_{-\infty}^{+\infty} H_2(x) dx < \infty,$$

$$\exists K : \int_{-\infty}^{+\infty} H_3(x) \cdot f_\theta(x) dx < K \quad \forall \theta \in (a, b).$$

c) $0 < I_1(\theta) = \int_{-\infty}^{+\infty} \left(\frac{\partial \ln f_\theta(x)}{\partial \theta} \right)^2 f_\theta(x) dx < \infty$.

Legyen továbbá τ_n a θ paraméter maximum-likelihood statisztikája.

2012. 02. 09.

Dr Ketskeméty László előadása

73



A maximum likelihood becslés

Ekkor

(i) τ_n az θ paraméter konzisztens becslése,

(ii) τ_n aszimptotikusan normális eloszlású, azaz $\sqrt{n}I_1^{-1}(\theta) \cdot (\tau_n - \theta) \xrightarrow{d} N(0, 1)$.

A tétel bizonyítása megtalálható az

<http://www.szit.bme.hu/~kela/stat.pdf> internetes jegyzetben a 32.-35. oldalon.

2012. 02. 09.

Dr Ketskeméty László előadása

74



A momentumok módszere

Legyen adott $\mathcal{P}$, valószínűségi mértékek egy tere és az $X_1, X_2, \dots, X_n$ statisztikai minta. Tegyük fel, hogy léteznek az

$$m_j = \mathbf{E}_\theta X_i^j = g_j(\theta) \quad (j = 1, 2, \dots, k)$$

momentumok, és

$$\exists g_j^{-1}(m_1, m_2, \dots, m_k) = \theta_j \quad (j = 1, 2, \dots, k).$$

Tekintsük az

$$\hat{m}_j = \frac{1}{n} \sum_{i=1}^n X_i^j \quad (j = 1, 2, \dots, k)$$

empirikus momentum statisztikákat. Akkor az

$$m_j = g_j^{-1}(\hat{m}_1, \hat{m}_2, \dots, \hat{m}_k) \quad (j = 1, 2, \dots, k)$$

statisztikák a θ_j paraméterek *momentumos becslései*.

2012. 02. 09.

Dr Ketskeméty László előadása

75



A momentumok módszere

A normális eloszlás paramétereinek becslése a momentumok módszerével:

$$m = g_1(m_1, m_2) = m_1, \quad D = g_2(m_1, m_2) = m_2 - m_1^2.$$

$$\hat{m}_1 = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n, \quad \hat{m}_2 = \frac{1}{n} \sum_{i=1}^n X_i^2$$

$$D \approx g_2(\hat{m}_1, \hat{m}_2) = \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2 = s_n^2.$$

2012. 02. 09.

Dr Ketskeméty László előadása

76



A momentumok módszere

A Poisson eloszlás paraméterek becslése a momentumok módszerével:

$$P_{\vartheta}(X_i = k) = \frac{\vartheta^k}{k!} e^{-\vartheta} \quad (k = 0, 1, 2, \dots).$$

A $\vartheta > 0$ paraméter éppen a várható érték

$$\vartheta \approx \hat{m}_1 = \bar{X}_n$$

2012. 02. 09.

Dr Ketskeméty László előadása

77



A normális eloszlásból származtatott folytonos eloszlások

A χ^2 -eloszlás

A Student-eloszlás

Az F-eloszlás

A Lukács-tétel

2012. 02. 09.

Dr Ketskeméty László előadása

78



A χ^2 -eloszlás

Legyenek $X_1, X_2, \dots, X_n$ standard normális eloszlású, teljesen független valószínűségi változók.

Ekkor $\sum_{i=1}^n X_i^2$ χ^2 -eloszlást követ n szabadságfokkal,

sűrűségfüggvénye

$$f(x) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} \cdot e^{-\frac{x}{2}} \cdot x^{\frac{n}{2}-1}, x > 0$$

ahol $\Gamma(s) = \int_0^{\infty} e^{-t} \cdot t^{s-1} dt$ a gamma-függvény.

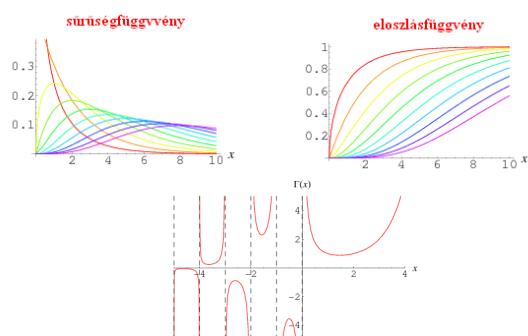
2012. 02. 09.

Dr Ketskeméty László előadása

79



A χ^2 -eloszlás



2012. 02. 09.

Dr Ketskeméty László előadása

80



A χ^2 -eloszlás és a polinomiális eloszlás kapcsolata

$\mathbf{V} \in \text{Pol}(n, p_1, p_2, \dots, p_r)$

karakterisztikus függvénye:

$$\begin{aligned} \varphi_{\mathbf{V}}(t_1, t_2, \dots, t_r) &= \mathbf{E} e^{i \mathbf{V}^T \mathbf{t}} = \\ &= \sum_{\substack{\forall k_1, k_2, \dots, k_r \\ k_1 + k_2 + \dots + k_r = n}} \frac{n!}{k_1! k_2! \dots k_r!} p_1^{k_1} p_2^{k_2} \dots p_r^{k_r} e^{i \sum_{j=1}^r k_j t_j} = \\ &= (p_1 e^{it_1} + p_2 e^{it_2} + \dots + p_r e^{it_r})^n. \end{aligned}$$

2012. 02. 09.

Dr Ketskeméty László előadása

81



A χ^2 -eloszlás és a polinomiális eloszlás kapcsolata

Ha $\mathbf{V} = (V_1, V_2, \dots, V_r)^T$ egy $n, p_1, p_2, \dots, p_r$ paraméterű polinomiális eloszlású valószínűségi vektorváltozó,

akkor $\sum_{i=1}^r \frac{(V_i - np_i)^2}{np_i} \xrightarrow{e} \chi_{r-1}^2 \quad (n \rightarrow \infty).$

Ezen a tulajdonságon alapulnak a Chi-négyzet próbák!

2012. 02. 09.

Dr Ketskeméty László előadása

82



A Student-eloszlás

Legyenek $Y, X_1, X_2, \dots, X_n$ standard normális eloszlású, teljesen független valószínűségi változók.

Ekkor $\frac{Y}{\sqrt{\frac{\sum_{i=1}^n X_i^2}{n}}}$ n szabadságfokú t-eloszlást fog követni,

melynek sűrűségfüggvénye

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})} \cdot \frac{1}{\sqrt{n}} \cdot \left(\frac{1}{1 + \frac{x^2}{n}} \right)^{\frac{n+1}{2}}, x \in \mathbb{R}$$

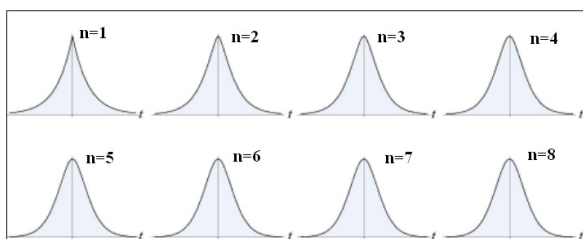
2012. 02. 09.

Dr Ketskeméty László előadása

83



A Student-eloszlás sűrűségfüggvények szabadságfokkal



2012. 02. 09.

Dr Ketskeméty László előadása

84



A Student-eloszlás

$X_1, X_2, \dots, X_n \sim N(m, D)$

teljesen független valószínűségi változók.

Ekkor

$$\frac{\bar{X}_n - m}{D} \sqrt{n} \in N(0, 1), \quad \text{és az} \quad \frac{(n-1) s_n^{*2}}{D^2} \in \chi_{n-1}^2$$

függetlenek, így

$$\frac{\frac{\bar{X}_n - m}{D} \sqrt{n}}{\sqrt{\frac{(n-1) s_n^{*2}}{D^2} \frac{1}{n-1}}} = \frac{\bar{X}_n - m}{s_n^*} \sqrt{n} \sim t_{n-1}$$

2012. 02. 09.

Dr Ketskeméty László előadása

85



Az F-eloszlás

Ha X n szabadságfokú χ^2 -eloszlású és Y tőle

független k szabadságfokú χ^2 -eloszlású, akkor $Z = \frac{X}{Y}$

n, k paraméterű F- (Fisher-) eloszlás.

Sűrűségfüggvénye:

$$f(x) = \frac{\Gamma(\frac{n+k}{2})}{\Gamma(n)\Gamma(k)} \cdot x^{\frac{k}{2}-1} \cdot (k + nx)^{-\frac{k+n}{2}}, x > 0$$

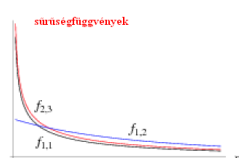
2012. 02. 09.

Dr Ketskeméty László előadása

86



Az F-eloszlás



2012. 02. 09.

Dr Ketskeméty László előadása

87



Az F -eloszlás

Ha X n szabadságfokú, Y pedig k szabadságfokú

χ^2 eloszlású változók függetlenek, akkor az

$Z = \frac{\frac{X}{n}}{\frac{Y}{k}}$ eloszlása n, k szabadságfokú F -eloszlás lesz!

2012. 02. 09.

Dr Ketskeméty László előadása

88



A Lukács-tétel

Legyenek $X_1, X_2, \dots, X_n \sim N(m, D)$ teljesen függetlenek!

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \text{ és } s_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2, s_n^{*2} = \frac{n}{n-1} s_n^2.$$

Ekkor

- (i) $\bar{X}_n \in N(m, \frac{D}{n})$,
- (ii) $\frac{ns_n^2}{D^2} \in \chi_{n-1}^2$,
- (iii) $\bar{X}_n$ és s_n^2 függetlenek

2012. 02. 09.

Dr Ketskeméty László előadása

89



A Lukács-tétel bizonyítása

(i) $\bar{X}_n$ karakterisztikus függvénye:

$$\begin{aligned} \varphi_{\bar{X}_n}(t) &= \mathbf{E} \exp \left(i \sum_{j=1}^n X_j \frac{t}{n} \right) = \prod_{j=1}^n \mathbf{E} \exp \left(i X_j \frac{t}{n} \right) = \prod_{j=1}^n \varphi_{X_j} \left(\frac{t}{n} \right) = \\ &= \left(\exp \left(i m \frac{t}{n} - \frac{D^2 t^2}{2n^2} \right) \right)^n, \end{aligned}$$

amiből leolvasható, hogy $\bar{X}_n \in N(m, \frac{D}{n})$.

Az (i) állítás abból következik, hogy normális eloszlású változók konvolúciója normális és normális változó konstans-szorosa is normális.

2012. 02. 09.

Dr Ketskeméty László előadása

90



A Lukács-tétel bizonyítása

(ii) *Segédétel:* Tekintsük a

$$\underline{H}_n = \underline{E}_n - \frac{1}{n} \mathbf{1} \mathbf{1}^T = \begin{pmatrix} \frac{n-1}{n} & -\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & \frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \vdots & \frac{n-1}{n} \end{pmatrix}$$

centráló mátrixot. A képletben az $\mathbf{1}$ olyan vektort jelöl, melynek mindegyik komponense 1-es, $\underline{E}_n$ pedig az egységmátrix.

Ekkor

- $\underline{H}_n \underline{H}_n = \underline{H}_n$ (idempotens),
- $\underline{H}_n$ szimmetrikus, pozitív szemidefinit,
- $\det(\underline{H}_n) = 0$, $\text{rank}(\underline{H}_n) = n - 1$,
- $\underline{H}_n$ sajátértékei az 1 $(n - 1)$ -szeres multiplicitással, és a 0.

2012. 02. 09.

Dr Ketskeméty László előadása

91



A Lukács-tétel bizonyítása

A segédétel bizonyítása:

- $\underline{H}_n^2 = \underline{E}_n^2 - \underline{E}_n (\frac{1}{n} \mathbf{1} \mathbf{1}^T) - (\frac{1}{n} \mathbf{1} \mathbf{1}^T) \underline{E}_n + \frac{1}{n^2} \mathbf{1} (\mathbf{1}^T \mathbf{1}) \mathbf{1}^T = \underline{E}_n - \frac{1}{n} \mathbf{1} \mathbf{1}^T = \underline{H}_n$.
- $\underline{H}_n$ szimmetrikus triviálisan. Legyen $\mathbf{x} \in \mathbb{R}^n$ tetszőleges:
 $\mathbf{x}^T \underline{H}_n \mathbf{x} = \mathbf{x}^T \underline{H}_n \cdot \underline{H}_n \mathbf{x} = \left\| \underline{H}_n \mathbf{x} \right\|^2 \geq 0$ pozitív szemidefinit, sajátértékei nemnegatívak.
- $\det(\underline{H}_n - \lambda \underline{E}_n) = \det(\underline{H}_n^2 - \lambda \underline{E}_n^2) = \det(\underline{H}_n - \sqrt{\lambda} \underline{E}_n) \cdot \det(\underline{H}_n + \sqrt{\lambda} \underline{E}_n) = 0$.
Tehát, ha λ sajátérték, akkor $+\sqrt{\lambda}$ is az. Így csak 1 és 0 lehet sajátérték! Másrészt,
 $\text{trace } \underline{H}_n = n \frac{n-1}{n} = n-1 = \lambda_1 + \lambda_2 + \cdots + \lambda_n$
csak úgy lehet, ha $\lambda_1 = \lambda_2 = \cdots = \lambda_{n-1} = 1$ és $\lambda_n = 0$.
- $\det(\underline{H}_n) = \prod_{j=1}^n \lambda_j = 0$, $\text{rank}(\underline{H}_n) = \sum_{j=1}^n \lambda_j = n-1$, mert $n-1$ darab 1 sajátértéke van.

2012. 02. 09.

Dr Ketskeméty László előadása

92



A Lukács-tétel bizonyítása

A segédételt használva bizonyíthatjuk a tétel 2. állítását.

$$\mathbf{X}^T \underline{H}_n \mathbf{X} = \mathbf{X}^T \mathbf{X} - \frac{1}{n} \mathbf{X}^T \mathbf{1} \mathbf{1}^T \mathbf{X} = \sum_{i=1}^n X_i^2 - n (\bar{X}_n)^2 = n s_n^2.$$

Legyenek $Z_i = X_i - m \in N(0, D)$, teljesen függetlenek.

$$\bar{Z}_n = \frac{1}{n} \sum_{i=1}^n Z_i = \frac{1}{n} \sum_{i=1}^n (X_i - m) = \bar{X}_n - m,$$

$$\frac{1}{n} \mathbf{Z}^T \underline{H}_n \mathbf{Z} = s_n^2 = \frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z}_n)^2 = \frac{1}{n} \sum_{i=1}^n ((X_i - m) - (\bar{X}_n - m))^2 = s_n^2.$$

Felhasználjuk $\underline{H}_n$ spektrálfelbontását:

$\underline{H}_n = \underline{G} \underline{L} \underline{G}^T$, ahol $\underline{G} \underline{G}^T = \underline{G}^T \underline{G} = \underline{E}_n$ és $\underline{L} = \text{diag}(1, 1, \dots, 1, 0)$. Így

$$n s_n^2 = \mathbf{X}^T \underline{H}_n \mathbf{X} = \mathbf{Z}^T \underline{H}_n \mathbf{Z} = \mathbf{Z}^T \underline{G} \underline{L} \underline{G}^T \mathbf{Z} = \mathbf{Y}^T \underline{L} \mathbf{Y} = \sum_{i=1}^{n-1} Y_i^2.$$

$\mathbf{Y} = \underline{G}^T \mathbf{Z} \in N_n(\underline{G} \mathbf{0}, D \underline{E}_n)$, azaz

$\frac{Y_i}{D} \in N(0, 1)$ teljesen függetlenek $\frac{n s_n^2}{D^2} = \sum_{i=1}^{n-1} \frac{Y_i^2}{D^2} \in \chi_{n-1}^2$.

2012. 02. 09.

Dr Ketskeméty László előadása

93



A Lukács-tétel bizonyítása

(iii) A $\lambda_n = 0$ sajátértékhez tartozó sajátvektor: $\mathbf{g}_n = \frac{1}{\sqrt{n}}\mathbf{1} = (\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})^T$, mert $\underline{H}_n \mathbf{g}_n = \underline{E}_n \frac{1}{\sqrt{n}}\mathbf{1} - \frac{1}{n}\mathbf{1}\mathbf{1}^T \frac{1}{\sqrt{n}}\mathbf{1} = \mathbf{g}_n - \mathbf{g}_n = \mathbf{0}$. Így $Y_n = \mathbf{g}_n^T \mathbf{Z} = \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i = \sqrt{n}Z_n$. Mivel $ns_n^2 = nd_n^2 = \sum_{i=1}^{n-1} Y_i^2$, $X_n = Z_n + m = \frac{1}{\sqrt{n}}Y_n + m$, és Y_i -k teljesen függetlenek voltak, így X_n és s_n^2 is függetlenek.

Következmény:

Ha $X_1, X_2, \dots, X_n$ $N(m, D)$ eloszlásból származó statisztikai minta, akkor az

$$\frac{X_n - m}{D} \sqrt{n} \in N(0, 1), \text{ és az } \frac{(n-1)s_n^2}{D^2} \in \chi_{n-1}^2$$

statisztikák függetlenek, így

$$\frac{\frac{X_n - m}{D} \sqrt{n}}{\sqrt{\frac{(n-1)s_n^2}{D^2} \frac{1}{n-1}}} = \frac{X_n - m}{s_n} \sqrt{n} \in t_{n-1}$$

($n-1$ szabadságfokú Student-eloszlású).

2012. 02. 09.

Dr Ketskeméty László előadása

94



Intervallumbecslések

A korábbi szakaszokban az ismeretlen paramétervektort a minta egy függvényével, azaz egyetlen statisztikával próbáltuk meg közelíteni. Konkrét realizációnál tehát, a paramétertér egy pontját egy másik ponttal becsüljük. Ezért beszélünk pontbecslésről. De tudjuk azt is, hogy folytonos eloszlásoknál, annak valószínűsége, hogy a valószínűségi változó az értékkészletének éppen egy tetszőlegesen kiválasztott pontját fogja felvenni, nulla. Tehát folytonos esetben nulla annak valószínűsége, hogy éppen a paramétert találtuk el a becsléssel. Az intervallumbecsléseknél a mintából készített tartományokat definiálunk, amely tartományok nagy valószínűséggel lefedik a kérdéses paraméterpontot.

2012. 02. 09.

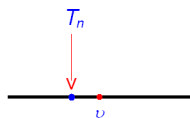
Dr Ketskeméty László előadása

95

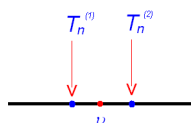


Intervallumbecslések

Pontbecslés



Intervallum-becslés



2012. 02. 09.

Dr Ketskeméty László előadása

96



Intervallumbecslések

Legyen adott $\mathcal{P}$ valószínűségi mértékek egy tere és az $X_1, X_2, \dots, X_n$ statisztikai minta. Legyen $0 < \varepsilon < 1$ rögzített. Azt mondjuk, hogy a θ paraméterhez megadtunk egy legalább $1 - \varepsilon$ szignifikanciaszintű konfidenciaintervallumot, ha $t_1(X_1, X_2, \dots, X_n)$ és $t_2(X_1, X_2, \dots, X_n)$ olyan statisztikák, hogy

$$\mathbf{P}_\theta(t_1(X_1, X_2, \dots, X_n) \leq \theta \leq t_2(X_1, X_2, \dots, X_n)) \geq 1 - \varepsilon$$

fennáll minden $P_\theta \in \mathcal{P}$ -re.

2012. 02. 09.

Dr Ketskeméty László előadása

97



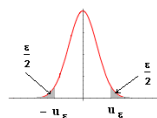
Intervallumbecslések

Konfidencia intervallum szerkesztése az ismeretlen várható értékre ismert szórású normális eloszlás esetében:

Legyen $X_1, X_2, \dots, X_n \sim N(m, D_0)$ eloszlásból származó statisztikai minta, ahol $D_0 > 0$ ismert, $m \in \mathbb{R}$ ismeretlen. Szerkesszünk m -re adott $0 < \varepsilon < 1$ mellett $(1 - \varepsilon)$ -szintű konfidencia intervallumot!

$u = \frac{\bar{X}_n - m}{D_0} \sqrt{n} \in N(0, 1)$, azaz a statisztika sűrűségfüggvénye: $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

$$\int_{-u_\varepsilon}^{+u_\varepsilon} \varphi(t) dt = \mathbf{P}(-u_\varepsilon < u < u_\varepsilon) = \Phi(u_\varepsilon) - \Phi(-u_\varepsilon) = 2\Phi(u_\varepsilon) - 1 = 1 - \varepsilon$$



2012. 02. 09.

Dr Ketskeméty László előadása

98



Intervallumbecslések

Konfidencia intervallum szerkesztése az ismeretlen várható értékre ismert szórású normális eloszlás esetében:

$$\mathbf{P}\left(\bar{X}_n - \frac{u_\varepsilon D_0}{\sqrt{n}} < m < \bar{X}_n + \frac{u_\varepsilon D_0}{\sqrt{n}}\right) = 1 - \varepsilon,$$

azaz a

$$T_1 = \bar{X}_n - \frac{u_\varepsilon D_0}{\sqrt{n}},$$

$$T_2 = \bar{X}_n + \frac{u_\varepsilon D_0}{\sqrt{n}}$$

$(1 - \varepsilon)$ -szintű konfidenciaintervallum m -re.

2012. 02. 09.

Dr Ketskeméty László előadása

99



Intervallumbecslések

Konfidencia intervallum szerkesztése az ismeretlen várható értékre ismeretlen szórású normális eloszlás esetében:

Legyen $X_1, X_2, \dots, X_n \sim N(m, D)$ eloszlásból származó statisztikai minta, ahol $D > 0$ és, $m \in \mathbb{R}$ is ismeretlen. Szerkesszünk m -re adott $0 < \varepsilon < 1$ mellett $(1 - \varepsilon)$ -szintű konfidenciaintervallumot!

$$\frac{\bar{X}_n - m}{s_n^*} \sqrt{n} \in t_{n-1},$$

Minden $\varepsilon > 0$ számhoz létezik olyan $t_\varepsilon > 0$ szám, amellyel $\mathbf{P}(|t_{n-1}| < t_\varepsilon) = 1 - \varepsilon$.

$$1 - \varepsilon = \mathbf{P}\left(-t_\varepsilon < \frac{\bar{X}_n - m}{s_n^*} \sqrt{n} < t_\varepsilon\right) = \mathbf{P}\left(\bar{X}_n - \frac{t_\varepsilon s_n^*}{\sqrt{n}} < m < \bar{X}_n + \frac{t_\varepsilon s_n^*}{\sqrt{n}}\right)$$

$$T_1 = \bar{X}_n - \frac{t_\varepsilon s_n^*}{\sqrt{n}}, \quad T_2 = \bar{X}_n + \frac{t_\varepsilon s_n^*}{\sqrt{n}}$$
