

A matematikai statisztika elemei

Mikó Teréz, dr. Szalkai István
SZALKAI@ALMOS.UNI-PANNON.HU
Pannon Egyetem, Veszprém

2017.10.03.

Tartalomjegyzék

Tartalomjegyzék	3
Bevezetés	3
Alapfogalmak	3
Intervallumbecslések	5
A valószínűség becslése	6
A várható érték becslése ismert szórás esetén	7
A várható érték becslése ismeretlen szórás esetén	8
A szórás becslése	9
Összefoglaló képletgyűjtemény	11
Mebízhatósági intervallumok	11
Táblázatok	12

Stat-intervC.tex, 2017.10.03., 19:10'

Bevezetés

Nem elméleti vizsgálatok (fejtegetések), hanem a tényleges *mért értékek* kiértékelése következik.

Alapfogalmak

0.1. Definíció. *Statisztikai minta:* egy mennyiség (ξ v.v.) többszöri mérésekor kapott eredmények.

A papíron

$$X_1, \dots, X_n \tag{1}$$

valós számok, elméletileg pedig X_1, \dots, X_n valószínűségi változók.
A fenti minta **szabadsági foka**

$$s := n - 1 \quad . \quad (2)$$

□

0.2. Definíció. (i) **Empirikus (=tapasztalati, gör.) várható érték:**

$$\bar{\xi} := \frac{X_1 + \dots + X_n}{n} \quad (3)$$

a szokásos "számtani közép".

(ii) **Empirikus (tapasztalati) szórásnégyzet:**

$$\sigma^2 := \frac{1}{n} \sum_{i=1}^n (X_i - \bar{\xi})^2 = \frac{(X_1 - \bar{\xi})^2 + \dots + (X_n - \bar{\xi})^2}{n} \quad (4)$$

(iii) **Korrigált (javított) empirikus (tapasztalati) szórásnégyzet:**

$$(\sigma^*)^2 := \frac{n}{n-1} \sigma^2 = \frac{(X_1 - \bar{\xi})^2 + \dots + (X_n - \bar{\xi})^2}{n-1} \quad (5)$$

□

0.3. Tétel. Az empirikus szórásnégyzet egyszerűbben is kiszámolható:

$$\sigma^2 = \overline{(\xi^2)} - (\bar{\xi})^2 = \frac{X_1^2 + \dots + X_n^2}{n} - (\bar{\xi})^2 \quad (6)$$

(ld. a (3) képletet), és ne feledjük: a korrigált empirikus szórásnégyzet (5) alapján

$$(\sigma^*)^2 := \frac{n}{n-1} \cdot \sigma^2 \quad . \quad \square \quad (7)$$

0.4. Példa. 12 mérést végeztünk: $\{X_1, \dots, X_{12}\} =$

$= \{20.0, 20.2, 20.4, 20.7, 20.7, 21.0, 21.1, 21.3, 21.4, 21.4, 21.4, 21.5\}$,

tehát $n = 12$ és $s = n - 1$.

Az empirikus (tapasztalati) várható érték:

$$\begin{aligned} \bar{\xi} &= \frac{20.0 + 20.2 + 20.4 + 20.7 + 20.7 + 21.0 + 21.1 + 21.3 + 21.4 + 21.4 + 21.4 + 21.5}{12} = \\ &= 20.925 \quad , \end{aligned}$$

$$a \text{ tapasztalati } \mathbf{n\acute{e}gyzetes} \text{ várható érték: } \overline{(\xi^2)} = \\ = \frac{20.0^2 + 20.2^2 + 20.4^2 + 20.7^2 + 20.7^2 + 21.0^2 + 21.1^2 + 21.3^2 + 21.4^2 + 21.4^2 + 21.4^2 + 21.5^2}{12} =$$

$$\approx 438.100833 ,$$

a tapasztalati szórásnégyzet és szórás (6) szerint:

$$\sigma^2 = \overline{(\xi^2)} - (\bar{\xi})^2 \approx 438.101 - 20.925^2 \approx 0.2454 ,$$

$$\sigma = \sqrt{\overline{(\xi^2)} - (\bar{\xi})^2} \approx \sqrt{0.2454} \approx 0.4954 ,$$

a korrigált (tapasztalati) szórásnégyzet és szórás (5) szerint

$$(\sigma^*)^2 = \frac{n}{n-1} \cdot \left(\overline{(\xi^2)} - (\bar{\xi})^2 \right) \approx \frac{12}{11} \cdot (438.101 - 20.925^2) \approx 0.2677 ,$$

$$\sigma^* = \sqrt{\frac{n}{n-1} \cdot \left(\overline{(\xi^2)} - (\bar{\xi})^2 \right)} \approx \sqrt{0.2677} = 0.5174 .$$

Intervallumbecslések

Általános probléma:

0.5. Probléma. *A kapott statisztikai minta alapján adjunk meg a valós számok egy olyan $[a, b]$ intervallumát, amelybe a vizsgált jelenség egy bizonyos γ mérőszáma (pl. valószínűség, várható érték, szórás, stb.) egy adott valószínűséggel beleesik:*

$$P(a < \gamma < b) \geq 1 - \varepsilon \quad (8)$$

ahol $0 < \varepsilon < 1$ adott (tetszőleges) szám. \square

0.6. Definíció. *A fenti $[a, b]$ intervallumot **megbízhatósági (=konfidencia) intervallumnak** nevezzük, az ε számot **hiba- vagy tűréshatárnak**, az $1 - \varepsilon$ mennyiséget pedig **megbízhatósági szintnek**. \square*

0.7. Megjegyzés. *Általában a minta elemszámának (n) növelésével az $[a, b]$ intervallum csökken, míg a hibahatár (ε) csökkentésekor az $[a, b]$ intervallum növekszik.*

Speciális eseteket és példákat az alábbi alfejezetekben látunk.

A valószínűség becslése

0.8. Probléma. Egy A esemény (kísérlet) $p = P(A)$ valószínűségére keresünk adott $\varepsilon > 0$ mellett megbízhatósági intervallumot:

$$P(a < p < b) \geq 1 - \varepsilon \quad (9)$$

0.9. Tétel. Ha akkor n független kísérletből k esetben következett be az A esemény, és n elég nagy¹⁾, akkor a keresett intervallum:

$$[a, b] = \left[\frac{k}{n} - \eta, \frac{k}{n} + \eta \right] \quad (10)$$

ahol

$$\eta = \frac{u_\varepsilon}{\sqrt{n}} \cdot \sqrt{\frac{k}{n} \cdot \left(1 - \frac{k}{n}\right)} \quad (11)$$

és az u_ε szám kielégíti a

$$\Phi(u_\varepsilon) = 1 - \frac{\varepsilon}{2} \quad (12)$$

egyenlőséget (visszakereshető a Φ táblázatból). \square

0.10. Példa. Egy mintában 30 munkadarabból 10 db volt selejtes. Adjuk meg a selejt p valószínűségének 95%-os megbízhatósági intervallumát!

Megoldás: Tehát $\varepsilon = 0.05$, u_ε -t az $\Phi(u_\varepsilon) = 1 - \frac{\varepsilon}{2} = 0.975$ összefüggés (ld. (12)) és a Φ táblázat alapján határozhatjuk meg: $u_\varepsilon = 1.96$.

(10) és (11) alapján

$$\eta = \frac{1.96}{\sqrt{30}} \cdot \sqrt{\frac{10}{30} \cdot \left(1 - \frac{10}{30}\right)} \approx 0.168690,$$

$$a \approx \frac{10}{30} - 0.168690 \approx 0.164643,$$

$$b \approx \frac{10}{30} + 0.168690 = 0.502023,$$

tehát a 0.9. Tétel alapján

$$P(0.164 < p < 0.502) \geq 0.95, \quad (13)$$

vagyis 95% biztonsággal mondhatjuk, hogy a selejt aránya (valószínűsége, p) 0.164 és 0.502 között esik. \square

¹⁾ n legyen legalább 30, de inkább $n > 200$ a tanácsos!

0.11. Megjegyzés. (i) A fenti példában n (a mérések száma) elég kicsi, továbbá a hibahatár (ε) is elég szűk (kicsi), ez magyarázza a kapott $[a, b] = [0.164, 0.502]$ intervallum aránylag nagy méretét!

(ii) $A(z)$ (12) egyenlőség pontosan azt jelenti, hogy az $\eta \in N(0, 1)$ standard normális eloszlásra teljesül az

$$P(-u_\varepsilon < \eta < u_\varepsilon) = 1 - \varepsilon \quad (14)$$

egyenlőség.

(iii) $A(z)$ 0.9. Tétel a nagy számokra vonatkozó Moivre-Laplace tételen alapul. \square

A várható érték becslése ismert szórás esetén

0.12. Probléma. Adott ε hibahatárhoz adjunk meg $[a, b]$ konfidencia intervallumot ξ **várható értékére**, m -re, HA ξ -ről tudjuk, hogy normális eloszlású ÉS adott ξ szórása σ , vagyis

$$P(a \leq m \leq b) \geq 1 - \varepsilon \quad (15)$$

0.13. Tétel. A keresett intervallum

$$[a, b] = \left[\bar{\xi} - u_\varepsilon \cdot \frac{\sigma}{\sqrt{n}}, \bar{\xi} + u_\varepsilon \cdot \frac{\sigma}{\sqrt{n}} \right] \quad (16)$$

ahol az u_ε valós szám kielégíti $a(z)$ (12) egyenlőséget (visszakereshető a Φ táblázatból). \square

0.14. Példa. Egy vegyület 1kg mennyiségében az oxigéntartalom (ξ) normális eloszlást követ, szórását tudjuk: $\sigma = 3g$.

12 mérést végeztünk: $\{X_1, \dots, X_{12}\} =$

$= \{20.0, 20.2, 20.4, 20.7, 20.7, 21.0, 21.1, 21.3, 21.4, 21.4, 21.4, 21.5\}$.

Adjunk meg egy olyan intervallumot, amelybe az oxigéntartalom 95% eséllyel belesik.

Megoldás: Az ismert adatok: $n = 12$, $D(\xi) = \sigma = 3$, $m = M(\xi) = ?$, $\varepsilon = 5\% = 0.05$.

Mivel ξ normális eloszlású és szórását ismerjük, ezért $a(z)$ 0.13. Tétel (16) képletét alkalmazzuk.

A Φ táblázatból (12) szerint olyan $u_\varepsilon = u_{0.05}$ valós számot kell (vissza)keresnünk, amelyre $\Phi(u_{0.05}) = 1 - \frac{0.05}{2} = 0.975$ ahonnan $u_{0.05} = 1.96$.

Továbbá, (3) szerint

$$\bar{\xi} = \frac{20.0 + 20.2 + 20.4 + 20.7 + 20.7 + 21.0 + 21.1 + 21.3 + 21.4 + 21.4 + 21.4 + 21.5}{12} =$$

$$= 20.925 ,$$

$$\frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{12}} \approx 0.866025 ,$$

$$(16) \text{ alapján } a \approx 20.925 - 1.96 \cdot 0.866025 \approx 19.227591$$

$$\text{és } b \approx 20.925 + 1.96 \cdot 0.866025 = 22.622409 .$$

Tehát a 12 mérés és a 0.13. Tétel alapján $m = M(\xi)$ értékére kaptuk, hogy

$$P(19.228 < m < 22.622) > 1 - \varepsilon = 0.95 , \quad (17)$$

vagyis szavakban:

" A 12 mérés és az ismert információk (ξ normális és $D(\xi) = 0.3$) alapján 95% biztonsággal állíthatjuk, hogy az oxigéntartalom ($M(\xi)$): 19.228 és 22.622 közé esik! " \square

A várható érték becslése ismeretlen szórás esetén

0.15. Probléma. Adott ε hibahatárhoz adjunk meg $[a, b]$ konfidencia intervallumot ξ várható értékére, m -re, HA ξ -ről tudjuk, hogy normális eloszlású ÉS ξ szórása ismeretlen, $a(z)$ (15) összefüggés mintájára.

0.16. Tétel. A keresett intervallum

$$[a, b] = \left[\bar{\xi} - t_\varepsilon \cdot \frac{\sigma^*}{\sqrt{n}} , \bar{\xi} + t_\varepsilon \cdot \frac{\sigma^*}{\sqrt{n}} \right] \quad (18)$$

ahol t_ε értékét az $n - 1$ -szabadságfokú **Student eloszlás** (más néven: **t-eloszlás**) táblázatából keressük ki. \square

0.17. Példa. Egy bizonyos típusú TV készülék fogyasztása (ξ) normális eloszlást követ, szórását nem tudjuk, 12 mérést végeztünk: $X_1, \dots, X_{12} = 20.0, 20.2, 20.4, 20.7, 20.7, 21.0, 21.1, 21.3, 21.4, 21.4, 21.4, 21.5$.

Adjunk meg egy olyan intervallumot, amelybe a fogyasztás 95% eséllyel belesik.

Megoldás: Az ismert adatok: $n = 12$, a szabadsági fok $s = n - 1 = 11$, $m = M(\xi) = ?$, $\varepsilon = 5\% = 0.05$.

Mivel ξ normális eloszlású és szórását nem ismerjük, ezért $a(z)$ 0.16. Tétel (18) képletét alkalmazzuk.

A táblázat szerint az $\varepsilon = 0.05$ és az $s = 11$ értékekhez $t_{0.05} = 2.201$ tartozik.

$\bar{\xi}$, $\overline{(\xi^2)}$ és σ^* értékét az 0.4. Példában már kiszámoltuk, így

$$\frac{\sigma^*}{\sqrt{n}} \approx \frac{0.5174}{\sqrt{12}} \approx 0.1494 ,$$

és végül (18) alapján

$$a \approx 20.925 - 2.201 \cdot 0.1494 \approx 20.5962 ,$$

$$b \approx 20.925 + 2.201 \cdot 0.1494 \approx 21.2538 .$$

Tehát a 12 mérés és a 0.16. Tétel alapján $m = M(\xi)$ értékére kaptuk, hogy

$$P(20.596 < m < 21.254) > 1 - \varepsilon = 0.95 , \quad (19)$$

vagyis szavakban:

" A 12 mérés és az ismert információ (ξ normális) alapján 95% biztonsággal állíthatjuk, hogy a fogyasztás ($M(\xi)$) 20.596 és 21.254 közé esik! " \square

A szórás becslése

0.18. Probléma. Adott ε hibahatárhoz adjunk meg $[a, b]$ konfidencia intervallumot ξ szórásnégyzetére illetve szórására, HA ξ -ről csak annyit tudunk, hogy normális eloszlású, $a(z)$ (15) összefüggés mintájára.

0.19. Tétel. A szórásnégyzet keresett intervalluma

$$[a^2, b^2] = \left[\frac{n \cdot (\sigma^*)^2}{\chi_{\varepsilon/2}^2}, \frac{n \cdot (\sigma^*)^2}{\chi_{1-\varepsilon/2}^2} \right] \quad (20)$$

míg a szórás intervalluma

$$[a, b] = \left[\frac{\sqrt{n} \cdot \sigma^*}{\chi_{\varepsilon/2}}, \frac{\sqrt{n} \cdot \sigma^*}{\chi_{1-\varepsilon/2}} \right] \quad (21)$$

ahol $\chi_{\varepsilon/2}^2$ és $\chi_{1-\varepsilon/2}^2$ értékeit az $n-1$ -szabadságfokú χ^2 -eloszlás táblázatából keressük ki. \square

0.20. Példa. A csimpánzkölykök testsúlya normális eloszlású, legutóbbi mérésnél a következő mintát kaptuk: $X_1, \dots, X_{12} =$

$$= 20.0, 20.2, 20.4, 20.7, 20.7, 21.0, 21.1, 21.3, 21.4, 21.4, 21.4, 21.5 .$$

Adjunk meg egy olyan intervallumot, amelybe a testsúly szórása 95% eséllyel belesik.

Megoldás: Az ismert adatok: $n = 12$, a szabadsági fok $s = n - 1 = 11$, $\varepsilon = 5\% = 0.05$, vagyis $\beta = \varepsilon/2 = 0.025$ ill. $\beta = 1 - \varepsilon/2 = 0.975$.

Mivel ξ normális eloszlású és szórását becsüljük, ezért $a(z)$ 0.19. Tétel (20) és (21) képleteit alkalmazzuk.

A χ^2 táblázat szerint a β és $s = 11$ adatokhoz a

$$\chi_{\varepsilon/2}^2 = \chi_{0.025}^2 \approx 21.920 \quad \text{és} \quad \chi_{1-\varepsilon/2}^2 = \chi_{0.975}^2 \approx 3.816 \quad (22)$$

értékek tartoznak, továbbá

$$\chi_{0.025} \approx \sqrt{21.920} \approx 4.6819 \quad \text{és} \quad \chi_{0.975} \approx \sqrt{3.816} \approx 1.9535 . \quad (23)$$

$\bar{\xi}$, $\overline{(\xi^2)}$ és σ^* értékét az 0.4. Példában már kiszámoltuk, így

$$a^2 = \frac{n \cdot (\sigma^*)^2}{\chi_{\varepsilon/2}^2} \approx \frac{12 \cdot 0.2677}{21.920} \approx 0.1466 ,$$

$$b^2 = \frac{n \cdot (\sigma^*)^2}{\chi_{1-\varepsilon/2}^2} \approx \frac{12 \cdot 0.2677}{3.816} \approx 0.8418 ,$$

továbbá $a \approx \sqrt{0.1466} \approx 0.3829$ és $b \approx \sqrt{0.8418} \approx 0.9175$,

tehát (20) alapján

$$P(0.1466 < D^2(\xi) < 0.8418) > 1 - \varepsilon = 0.95 \quad (24)$$

és (21) alapján

$$P(0.3829 < D(\xi) < 0.9175) > 1 - \varepsilon = 0.95 , \quad (25)$$

vagyis szavakban:

" A 12 mérés és az ismert információ (ξ normális) alapján 95% biztonsággal állíthatjuk, hogy a testsúlyok szórásnégyzete ($D^2(\xi)$) 0.1466 és 0.8418 közé esik, míg szórása ($D(\xi)$) 0.3829 és 0.9175 közé esik! " \square

Összefoglaló képletgyűjtemény

A vizsgán kizárólag az alábbi oldalt és a táblázatokat lehet (kinyomtatva) használni:

Megbízhatósági intervallumok

Valószínűség: $\left[\frac{k}{n} - \eta, \frac{k}{n} + \eta \right]$

ahol $\eta = \frac{u_\varepsilon}{\sqrt{n}} \cdot \sqrt{\frac{k}{n} \cdot \left(1 - \frac{k}{n}\right)}$ és $\Phi(u_\varepsilon) = 1 - \frac{\varepsilon}{2}$

Várható érték (szórás ismert): $\left[\bar{\xi} - u_\varepsilon \cdot \frac{\sigma}{\sqrt{n}}, \bar{\xi} + u_\varepsilon \cdot \frac{\sigma}{\sqrt{n}} \right]$

Várható érték (szórás ismeretlen): $\left[\bar{\xi} - t_\varepsilon \cdot \frac{\sigma^*}{\sqrt{n}}, \bar{\xi} + t_\varepsilon \cdot \frac{\sigma^*}{\sqrt{n}} \right]$

ahol t_ε a **Student** táblázatból ($s = n - 1$ szabadságfokú),

Szórás $\left[\frac{\sqrt{n} \cdot \sigma^*}{\chi_{\varepsilon/2}^2}, \frac{\sqrt{n} \cdot \sigma^*}{\chi_{1-\varepsilon/2}^2} \right]$

ahol $\chi_{\varepsilon/2}^2$ és $\chi_{1-\varepsilon/2}^2$ a χ^2 -táblázatból ($s = n - 1$ szabadságfokú).

χ^2 eloszlás

$s \backslash \beta$	0,10	0,05	0,025	0,01	0,001	0,95	0,975
1	2,706	3,841	5,024	6,635	10,827	0,004	0,001
2	4,605	5,991	7,378	9,210	13,815	0,103	0,051
3	6,251	7,815	9,348	11,345	16,268	0,352	0,022
4	7,779	9,488	11,143	13,277	18,465	0,711	0,484
5	9,236	11,070	12,833	15,086	20,517	1,145	0,831
6	10,645	12,592	14,449	16,812	22,457	1,635	1,237
7	12,017	14,067	16,013	18,475	24,322	2,167	1,690
8	13,362	15,507	17,535	20,090	26,125	2,733	2,180
9	14,684	16,919	19,023	21,666	27,877	3,325	2,700
10	15,987	18,307	20,483	23,209	29,588	3,940	3,247
11	17,275	19,675	21,920	24,725	31,264	4,575	3,816
12	18,549	21,026	23,337	26,217	32,909	5,226	4,404
13	19,812	22,362	24,736	27,688	34,528	5,892	5,009
14	21,064	23,685	26,119	29,141	36,123	6,571	5,629
15	22,307	24,996	27,488	30,578	37,697	7,261	6,262
16	23,542	26,296	28,845	32,000	39,252	7,962	6,908
17	24,769	27,590	30,191	33,409	40,790	8,672	7,564
18	25,989	28,869	31,526	34,805	42,312	9,390	8,231
19	27,204	30,144	32,852	36,191	43,820	10,117	8,901
20	28,412	31,410	34,170	37,566	45,315	10,851	9,591
21	29,615	32,671	35,479	38,932	46,797	11,591	10,283
22	30,813	33,924	36,781	40,289	48,268	12,338	10,982
23	32,007	35,172	38,076	41,638	49,728	13,091	11,689
24	33,196	36,415	39,364	42,980	51,179	13,484	12,401
25	34,382	37,652	40,646	44,314	52,620	14,611	13,120
26	35,563	38,885	41,923	45,642	54,052	15,379	13,844
27	36,741	40,113	43,194	46,963	55,476	16,151	14,573
28	37,916	41,337	44,461	48,278	56,893	16,928	15,308
29	39,087	42,557	45,772	49,558	58,302	17,708	16,047
30	40,256	43,773	46,979	50,892	59,703	18,493	16,791

Student-eloszlás (*t*-eloszlás)

$\frac{\varepsilon}{s}$	0,20	0,10	0,05	0,02	0,01	0,001
1	3,078	6,314	12,706	31,821	63,657	636,619
2	1,886	2,92	4,303	6,965	9,925	31,598
3	1,638	2,353	3,182	4,541	5,841	12,941
4	1,533	2,132	2,776	3,747	4,604	8,61
5	1,476	2,015	2,571	3,365	4,032	6,859
6	1,44	1,943	2,447	3,143	3,707	5,959
7	1,415	1,895	2,365	2,998	3,499	5,405
8	1,397	1,86	2,306	2,896	3,355	5,041
9	1,383	1,833	2,262	2,821	3,25	4,781
10	1,372	1,812	2,228	2,764	3,169	4,587
11	1,363	1,796	2,201	2,718	3,106	4,437
12	1,356	1,782	2,179	2,681	3,055	4,318
13	1,35	1,771	2,16	2,65	3,012	4,221
14	1,345	1,761	2,145	2,624	2,977	4,14
15	1,341	1,753	2,131	2,602	2,947	4,073
16	1,337	1,746	2,12	2,583	2,921	4,015
17	1,333	1,74	2,11	2,567	2,898	3,965
18	1,33	1,734	2,101	2,552	2,878	3,922
19	1,328	1,729	2,093	2,539	2,861	3,883
20	1,325	1,725	2,086	2,528	2,845	3,85
21	1,323	1,721	2,08	2,518	2,831	3,819
22	1,321	1,717	2,074	2,508	2,819	3,792
23	1,319	1,714	2,069	2,5	2,807	3,767
24	1,318	1,711	2,064	2,492	2,797	3,745
25	1,316	1,708	2,06	2,485	2,787	3,725
26	1,315	1,706	2,056	2,479	2,779	3,707
27	1,314	1,703	2,052	2,473	2,771	3,69
28	1,313	1,701	2,048	2,467	2,763	3,674
29	1,311	1,699	2,045	2,462	2,756	3,659
30	1,31	1,697	2,042	2,457	2,75	3,646
40	1,303	1,684	2,021	2,423	2,704	3,551
60	1,296	1,671	2	2,39	2,66	3,46
120	1,289	1,658	1,98	2,358	2,617	3,373
∞	1,282	1,645	1,96	2,326	2,576	3,291

eof