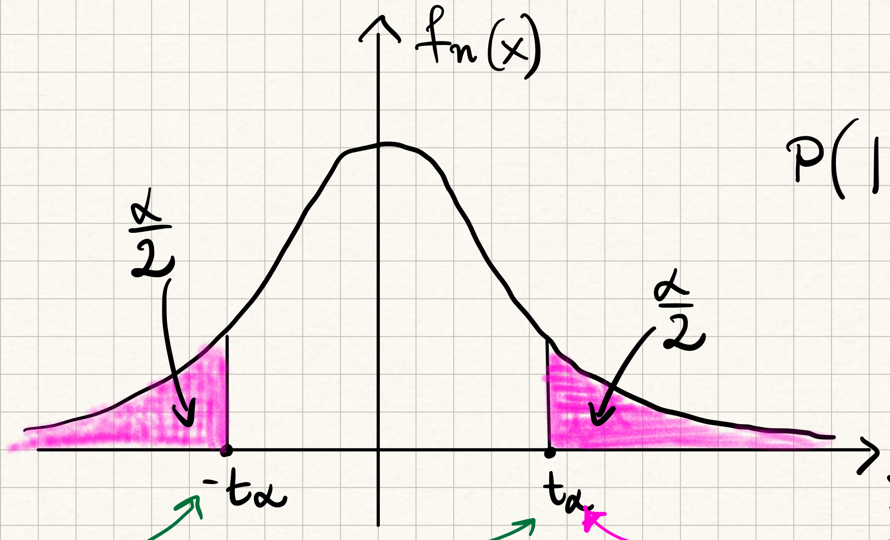
СТУДЕНТОВА РАСПОДЕЛА (НАСТАВАК)



$X \sim t(n)$

$$P(|X| > t_\alpha) = \alpha$$

Табл. за студентову t-расподелу даје вредности t_α за коју важи

$$P(|X| > t_\alpha) = \alpha,$$

где је α задато и једнако $0{,}01,\ 0.02, \ldots, 0.99$, а сл. променљива $X \sim t(n)$.

n	$\alpha = 0{,}9$	...	$\alpha = 0{,}01$
1	0,158		63,657
2	0,142		9,925
⋮			
17	0,128		2,898
⋮			
30	0,127		2,750

$X \sim t(17)$ и $\alpha = 0{,}01$.

Из таблице добијамо

$$P(|X| > 2{,}898) = 0{,}01$$

α – праг значајности

Ако је $n > 30$ тада ћемо студентову апроксимирати $\mathcal{N}(0,1)$

ЗНАЧАЈ У СТАТИСТИЦИ:

- када је узорак мали ($n<30$)
- оцену параметара
- дефиницију интервала поверења
- за оцену средњих вредности.

ФИШЕРОВА РАСПОДЕЛА

Ако су $Y \sim \chi^2(m)$ и $X \sim \chi^2(n)$ незав. сл. променљиве, тада:

$$X = \frac{\frac{Y}{m}}{\frac{X}{n}}$$

има ФИШЕРОВУ РАСПОДЕЛУ са m и n степени слободе: $X \sim \mathcal{F}(m,n)$.

Њена ф-ја густине:

$$f(x) = \frac{m}{n} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \cdot \left(\frac{m}{n} \cdot x\right)^{\frac{m}{2}-1} \cdot \left(1+\frac{m}{n} x\right)^{-\frac{m+n}{2}}.$$

$$E(X) = \frac{n}{n-2}$$

када $n \to \infty$ $E(X) \to 1$

$\downarrow$

Када $n \to \infty$ тада $\mathcal{F}_{m,n}$ можемо апрокс. расподелом сл. променљиве $\frac{Y}{n}$ где је $Y \sim \chi^2(m)$

$$D(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$$

- Ако је $X \sim F(m,n)$ тада је $Y = \frac{1}{X} \sim F(n,m)$, па се у таблицама дају само горње вредности за које је $m > n$.

ОСНОВНИ ПОЈМОВИ МАТЕМАТИЧКЕ СТАТИСТИКЕ

ПОПУЛАЦИЈА (СТАТИСТИЧКИ СКУП)

пример: скуп свих грађана Србије, скуп свих студената нашег факултета, скуп свих мерења неке величине (тежина, IQ, просечна оцена, да ли више воли coca-cola или pepsi, боја косе...)

ПОПУЛАЦИЈУ

ДЕФ: Ако .. скупа Ω придружимо један реалан број, имали смо деф. пресликавање

$$X : \Omega \to \mathbb{R} \to \text{сл. величина}$$

које се назива ОБЕЛЕЖЈЕ

ДЕФ: Било који подскуп популације Ω називамо УЗОРАК.

↘ коначног обима

Узорак мора бити:

- РЕПРЕЗЕНТАТИВАН (сваки члан популације мора да има подједнаку шансу да се ...)
- ДОВОЉНО ВЕЛИКИ
- ОБЈЕКТИВАН

ДЕФ: Свака сл. променљива која је ф-ја узорка $S = f(X_1, X_2, \ldots, X_n)$ назива се СТАТИСТИКА.

МЕРЕ ПОЛОЖАЈА

следећи примери преузети са курса

Статистика за аутоматску анализу података

др Марко Обрадовић

3 важна параметра популације који одређују положај расподеле су:

- МАТЕМАТИЧКО ОЧЕКИВАЊЕ, тј. средња вр. попул.
- МЕДИЈАНА ПОПУЛАЦИЈЕ
- МОДА попул.

СРЕДЊА ВРЕДНОСТ

Непознати параметар је мат. очекивање μ

↑

ПРОЦЕЊУЈЕМО ГА (приближно) статистиком коју називамо УЗОРАЧКОМ СРЕДЊОМ ВРЕДНОШЋУ (УЗОРАЧКА СРЕДИНА).

ДЕФ: Нека су $X_1, X_2, \ldots, X_n$ вредности сл. пр. X добијене у узорку. УЗОРАЧКОМ СРЕДИНОМ називамо аритметичку средину њих вредности:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

пример: Меримо број запамћених речи у минути

8 2 4 9 7 2 12 5 5 7

$$\bar{X} = \frac{8+2+4+\dots+5+7}{10} = \frac{61}{10} = 6,1$$

Комбиновање више узорачких средина:

пример: Број хитних сл. у једној болници је $\bar{x}_1 = 3$ за $n_1 = 5$, а у другој болници је $\bar{x}_2 = 15$ за $n_2 = 100$.

$$\bar{X} = \frac{3\cdot 5 + 15\cdot 100}{105} = 14,4$$

МЕДИЈАНА ПОПУЛАЦИЈЕ

ДЕФ: МЕДИЈАНА ПОПУЛАЦИЈЕ је непозната вел. од које је пола популације веће, а пола мање. Процењујемо је статистиком коју називамо УЗОРАЧКОМ МЕДИЈАНОМ.

ДЕФ: .. $x_1, x_2, \dots, x_n$ узорак поређан по вел. од најм. до највеће вредности. .. n - непаран, УЗОРАЧКА М. је број на средини низа. .. n - паран, узорачка мед.. аритм. средина средња два члана низа

пример: Године старости купаца у продавници Гардер.

12 15 17 20 24 | 27 30 35 42 60

$$\frac{24+27}{2} = 25,5$$

17 29 37 40 72

? Какав је однос медијане и средње вредности?

пример: Тржишна вредност 10 кућа у хиљадама долара

82 91 78,5 86 80,5 85 82.5 80 77 850

- Средња вредност $\bar{X} = \frac{82+91+\dots+77+850}{10} = 159,25$

- Медијана популације?

аутлајер

82	91	78,5	86	80,5	85	82.5	80	77	850
5		2		4		6	3	1	

поређано:

77 78,5 80 80,5 82 | 82,5 _ _ _ 850

$$\frac{82+82,5}{2} = 82.25$$

МОДА ПОПУЛАЦИЈЕ

ДЕФ: МОДА ПОПУЛАЦИЈЕ је непозната вр. која је најчешћа у популацији. Процењујемо је статистиком коју називамо УЗОРАЧКОМ МОДОМ, тј. вредношћу која се најчешће појављује.

МЕРЕ РАСЕЈАЊА

Важни парам. популације који описују расејање су:

- РАСПОН ПОПУЛАЦИЈЕ
- ДИСПЕРЗИЈА (ВАРИЈАНСА) ПОПУЛАЦИЈЕ σ^2
- СТАНДАРДНА ДЕВИЈАЦИЈА ПОП. σ
- МЕЂУКВАРТИЛНО РАСТОЈАЊЕ

РАСПОН ПОПУЛАЦИЈЕ

ДЕФ: РАСПОН ПОПУЛАЦИЈЕ ...

ДЕФ: УЗОРАЧКИ РАСПОН је разлика највећег и најмањег елемента узорка..

Пример: Резултати студената на испиту у току семестра

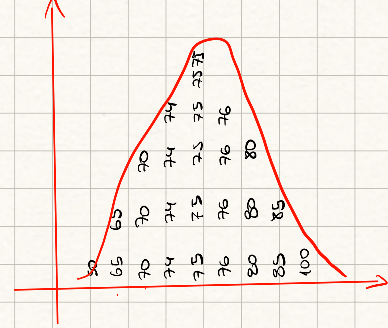
први

50 50 50 50 50 50
60 60 60
70 70
75
80 80
85 85 85
100 100 100 100 100 100

други

50
65 65
70 70 70
74 74 74 74
75 75 75 75 75 75
76 76 76 76
80 80 80
85 85
100

ОБИМ:	23	26
СРЕДЊИ БРОЈ ПОЕНА $\bar{X}$:	~ 75	75
МЕДИЈАНА:	75	75
МОДА:	50 или 100	75
РАСПОН:	50	50



ДИСПЕРЗИЈА

ДЕФ: Параметар популације — средње квадр. одступање сл. вел. X од средње вредности μ

$$D(X) = E\,(X_i - E(X))^2.$$

Прибл. је оцењујемо УЗОРАЧКОМ ДИСПЕРЗИЈОМ.

ДЕФ. Нека је $X_1, X_2, \ldots, X_n$ узорак од n-елем. УЗОРАЧКА ДИСПЕРЗИЈА деф. се као:

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

У задацима је једноставније користити ф-лу:

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

СТАНДАРДНА ДЕВИЈАЦИЈА

ДЕФ: УЗОРАЧКО СТАНД. ОДСТУПАЊЕ је једнако квадр. корену из узорачке дисперзије

$$S = \sqrt{S^2}$$

МЕЂУКВАРТИЛНО РАСТОЈАЊЕ

Мера расипања која је НЕОСЕТЉИВА на outlier-е (штрчке), за разлику од распона и дисперзије.

УЗОРАЧКИ КВАРТИЛИ

- Први квартил Q_1 је вредност од које је 1/4 узорка мање, а 3/4 веће
- Други квартил Q_2 је медијана, тј. вредност од које је 2/4 узорка мање, а 2/4 веће
- Трећи квартил Q_3 је вредност од које је 1/4 узорка веће, а 3/4 мање

МЕЂУКВАРТИЛНО РАСТОЈАЊЕ $IQR = Q_3 - Q_1$

је распон у ком се налази средњих 50% узорка

50 50 50 50 50 50 — Q_1
60 60 60
70 70
75
80 80
85 85 85
100 100 100 100 100 100 Q_3

50
65 65
70 70 70 Q_1
74 74 74 74
75 75 75 75 75 75
76 76 76 76
Q_3 80 80 80
85 85
100

Међуквартилно растојање
овде приближно = 50, а овде приближно = 10

ТАЧКАСТЕ ОЦЕНЕ

Користимо 2 оцене параметра: тачкасте и интервалне

ДЕФ: Сл. променљива $\bar{\Theta} = f(X_1, X_2, \dots, X_n)$ која се користи за оцену непознатог параметра Θ, зове се ТАЧКАСТА ОЦЕНА.

↘ конкретан број!

Оцена треба да буде:

1. ЦЕНТРИРАНА, тј. НЕПРИСТРАСНА $E(\bar{\Theta}) = \Theta$
(не постоји одступање даље процене од праве вредности)

2. ЕФИКАСНА: То је оцена која има минималну дисперзију.
→ пожељно да $\to 0$, када $n \to \infty$

Пример: Дат је узорак од 5 елемената из популације са средњом μ и дате су неке оцене. Која је од следећих оцена ЦЕНТРИРАНА, а која најефикаснија?

$$\bar{\mu}_1 = X_4 \qquad \bar{\mu}_4 = \frac{X_1 + 2X_2 + X_5}{4}$$

$$\bar{\mu}_2 = \frac{2X_1 + X_2}{2} \qquad \bar{\mu}_5 = 2X_1 - X_2$$

$$\bar{\mu}_3 = \frac{X_1 + X_2 + X_5}{5} \qquad \bar{\mu}_6 = \frac{X_2 + X_5}{2}$$

Центрирана?

$$E(\bar{\mu}_1) = E(X_4) = \mu$$ ← ЦЕНТРИРАНА!

$$E(\bar{\mu}_2) = E\left(\frac{2X_1 + X_2}{2}\right) = \frac{1}{2} E(2X_1 + X_2) = \frac{1}{2}\left[E(2X_1) + E(X_2)\right] =$$

$$= \frac{1}{2}\left[2\underbrace{E(X_1)}_{\mu} + \underbrace{E(X_2)}_{\mu}\right] = \frac{1}{2}(2\mu + \mu) = \frac{1}{2} \cdot 3\mu = \frac{3}{2}\mu$$ ↯

$$E(\bar{\mu}_3) = E\left(\frac{X_1+X_2+X_5}{5}\right) = \frac{1}{5}\left[\underbrace{E(X_1)}_{\mu} + \underbrace{E(X_2)}_{\mu} + \underbrace{E(X_5)}_{\mu}\right] = \frac{1}{5}\cdot 3\mu \ ↯$$

$$E(\bar{\mu}_4) = E\left(\frac{X_1+2X_2+X_5}{4}\right) = \frac{1}{4}\left[\underbrace{E(X_1)}_{\mu} + \underbrace{2E(X_2)}_{2\mu} + \underbrace{E(X_5)}_{\mu}\right] = \mu$$

ЦЕНТРИРАНА!

$$E(\bar{\mu}_5) = E(2X_1 - X_2) = E(2X_1) - E(X_2) = 2E(X_1) - E(X_2) =$$
$$= 2\mu - \mu = \mu \Leftarrow$$ ЦЕНТРИРАНА!

$$E(\bar{\mu}_6) = E\left(\frac{X_1+X_5}{2}\right) = \mu$$

Која је најефикаснија? $\mu_1, \mu_4, \mu_5, \mu_6$?

$$D(\mu_1) = D(X_4) = \sigma^2$$

$$D(\mu_4) = D\left(\frac{X_1+2X_2+X_5}{4}\right) = \frac{1}{16}\left[\overbrace{D(X_1)}^{\sigma^2} + \overbrace{D(2X_2)}^{4\sigma^2} + \overbrace{D(X_5)}^{\sigma^2}\right] =$$
$$= \frac{1}{16}\cdot 6\cdot\sigma^2 = \frac{3}{8}\sigma^2$$

постаје +1

$$D(\mu_5) = D(2X_1 - X_2) = D(2X_1 + (-X_2)) = D(2X_1) + D(-1\cdot X_2)$$
$$= 4\underbrace{D(X_1)}_{\sigma^2} + \underbrace{D(X_2)}_{\sigma^2} = 5\sigma^2$$

$$D(\bar{\mu}_6) = D\left(\frac{X_2+X_5}{2}\right) = \frac{1}{4}\left[D(X_2) + D(X_5)\right] = \frac{1}{4}\cdot 2\sigma^2 = \frac{\sigma^2}{2}$$

$\Rightarrow$ Најмања је $D(\mu_4) = \frac{3}{8}\sigma^2 \Rightarrow$ она је најефикаснија!

ДЕФ. Нпр. да имамо популацију и на њој деф. сл. променљиву X чија су средња вредност μ и дисперзија σ^2 непознати. Ако X има нормалну расподелу тада је СТАНДАРДНА ТАЧКАСТА ОЦЕНА математ. очекивања μ случајне променљиве X аритмет. средина: $\boxed{\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}}$

где су $X_i \sim N(\mu, \sigma^2)$. независне сл. пром.

- Да ли је центр. и ефикасна?

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{1}{n}\left[E(X_1) + E(X_2) + \dots + E(X_n)\right]$$

$$= \frac{1}{n} \cdot n \cdot \mu = \mu$$ ← ЦЕНТРИРАНА

$$D(\bar{X}) = D\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{1}{n^2}\left[D(X_1) + D(X_2) + \dots + D(X_n)\right]$$

$$= \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$$ $\to 0$ када $n \to \infty$

↓

Јесте ефикасна!

ДЕФ: Оцена дисперзије независних сл. променљивих $x_1, x_2, \dots, x_n$ где је $\bar{x}$ аритметичка средина, дата је изразом

$$\hat{S}^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\dots+(x_n-\bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

треба да добијемо дисперзију

Шта је са $E(\hat{S}^2)$, $D(\hat{S}^2)$?

Може се показати да је $E(\hat{S}^2) = \frac{n-1}{n}\sigma^2$

треба да добијем

$E(\hat{S}^2) = \sigma^2$

да би била центрирана

Због тога множимо са $\frac{n}{n-1}$

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

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

Ова тачкаста оцена се користи у пракси!

да би била центрирана!