

Теорија узорака

11. фебруар '20.

Научно истраживање

- **Научно истраживање** је систематско, планско и објективно испитивање неког проблема, према одређеним методолошким правилима, чија је сврха да се пружи поуздан и прецизан одговор на унапред постављено питање.
- Може се схватити као критички, контролисани и поновљиви процес стицања нових знања, неопходних (а понекад и довољних) за идентификовање, одређивање и решавање научних (теоријских и емпиријских) проблема.
- **Теоријско** истраживање  Емпиријско (**искуствено**) истраживање

Научно истраживање

(наставак)

- Свако научно истраживање има више међусобно логично повезаних фаза. **Фазе** су:
 - ◀ идентификовање и одређивање проблема
 - ▼ одређивање циља истраживања
 - ▼ дефинисање кључних израза
 - ▼ постављање хипотезе и извођење логичких последица из хипотезе
 - ▼ избор истраживачке стратегије и плана истраживања
 - ▼ развијање мерних и других инструмената истраживања
 - ▼ одређивање основног скупа (популације) и одабир узорка
 - ▼ спровођење истраживања и прикупљање релевантних података
 - ▼ обрађивање и анализа података добијених истраживањем
 - ▼ тумачење резултата истраживања и извођење закључ(а)ка
 - ▼ израда извештаја о обављеном истраживању
 - ▶ презентација резултата истраживања

ОСНОВНИ ПОЈМОВИ

- **Ентитет / јединица посматрања** (ен. 'observation unit') - живо биће или објекат чија су својства предмет истраживања
- **Популација** ('population') - скуп / колекција ентитета

На основу броја ентитета, тј. **обима** / величине **популације** ('population size') N , може бити:

- коначна популација – N је природан број
- бесконачна популација – $N \rightarrow +\infty$

Требало би разликовати:

- циљну популацију ('target population')
- популацију на којој се ефективно спроводи истраживање ('study population')

👉 Надаље се претпоставља: target population = study population и $N < +\infty$

- **Узорак** ('sample') - подскуп популације; садржи извесне ентитете који потичу из популације, на бази чијег проучавања ће се изводити закључци о читавој популацији

Обим узорка ('sample size') n

👉 Увек је коначна вредност

- **Јединица узорковања** ('sampling unit')

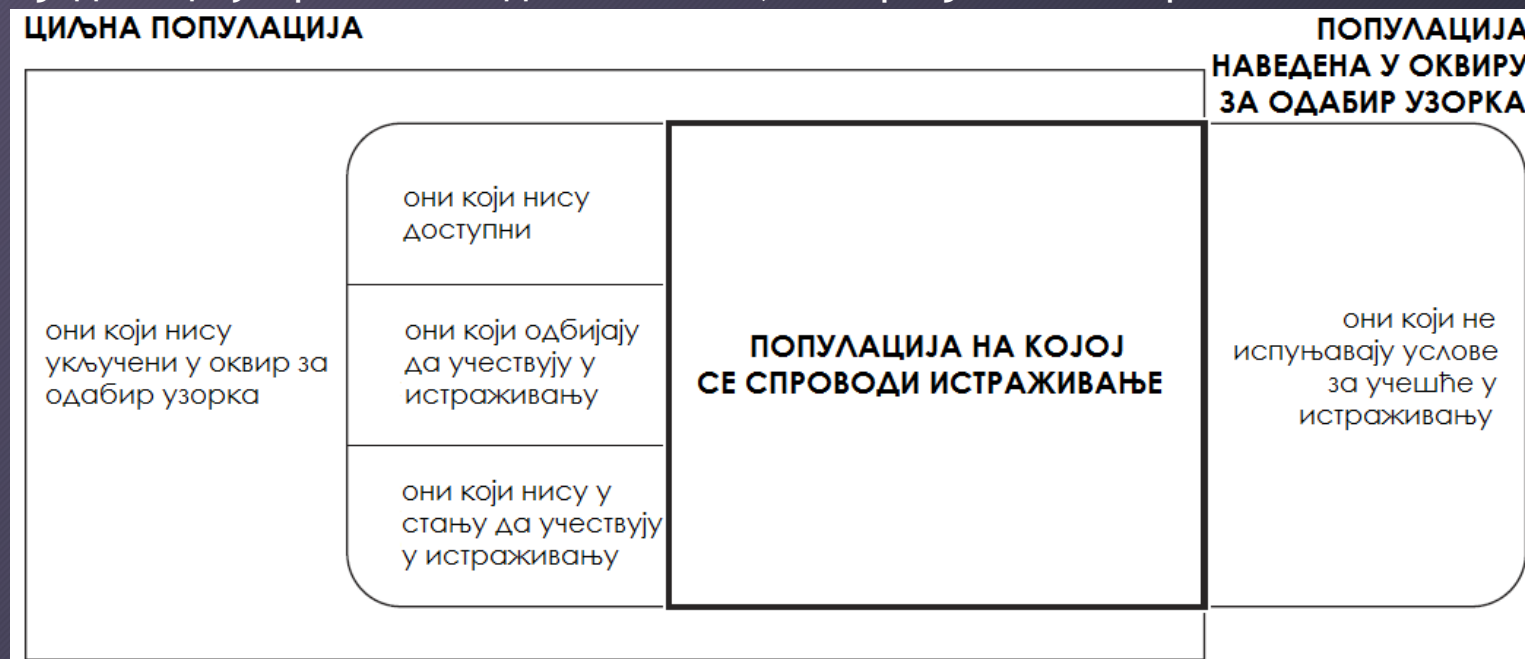
☞ У општем случају није исто што и јединица посматрања, која представља основни објект посматрања и прикупљања информација. Јединице узорковања су међусобно дисјунктни скупови ентитета

- **Оквир за одабир узорка** ('sampling frame') - попис (или нека друга спецификација) свих јединица узорковања

Нпр. свакој јединици узорковања придружи се различит природан број (почевши од 1). Ти бројеви називају се **ознаке јединица**, служе за њихово идентификовање и остају непромењени све до краја истраживања.

- Пример - телефонско истраживање бирачког тела

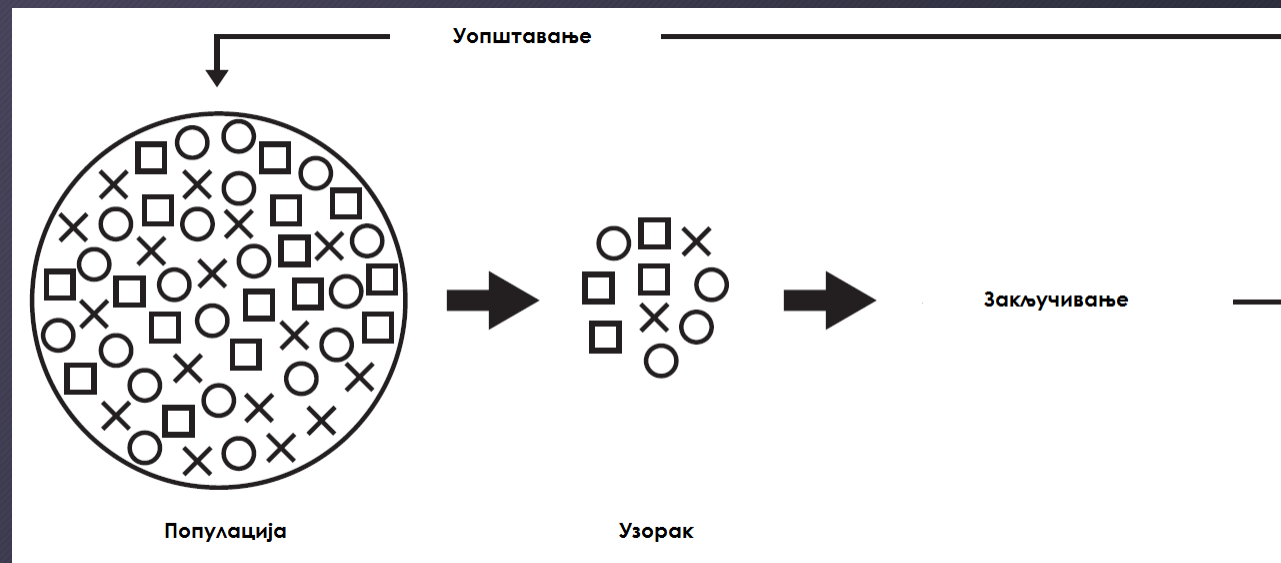
јединица узорковања - домаћинство; оквир - јавни телефонски именик



Зашто узорковање?

- **Потпуно** испитивање популације (проучавање тзв. **цензуса**) је, у многим случајевима, нерационално или чак принципијелно немогуће. Чак и онда када постоји могућност потпуног испитивања популације истраживач се обично опредељује за **делимично** испитивање (проучавање узорка) јер је (у односу на потпуно испитивање):

- јефтиније
- брже
- контрола тачности прикупљених података једноставнија и лакша



ОСНОВНИ ПОЈМОВИ

(наставак)

Термин популација односи се на скуп ентитета истоврсних у односу на једно или више заједничких својстава, која се могу посматрати. Ипак, ентитети, иако **истоврсни**, **нису истоветни**.

Одређивање популације представља значајну и, неретко, тешку фазу истраживања. Популација мора бити дефинисана: појмовно (у смислу свог садржаја - шта су ентитети, а шта јединице узорковања?), просторно и временски.

- **Обележје** ('study variable') - посматрана заједничка карактеристика свих ентитета у популацији, тј. прецизније, извесно варијабилно својство од интереса, које је одређено за сваки ентитет у популацији.

👉 Обележје најчешће није неко од дефиниционих својстава популације

Типови обележја



Обележје које се посматра може бити једнодимензионо, односно вишедимензионо.

- Пример - типови обележја

- Квалитативна:

- Номинална:

- боја очију; крвна група
 - етничка / верска припадност
 - радна места на факултету
 - расположење грађана Србије према приступању ЕУ
 - поседовање  профила

- Ординална:

- ниво академских студија
 - чин официра у војсци
 - оцена ресторана на 
 - стање пацијента
 - интензитет бола

- Пример - типови обележја (наставак)

- Квантитативна:

- Дискретна:

- број становника са правом гласа у одређеној општини
 - број близанаца рођених у току године у одређеној регији
 - број кућних љубимаца у одређеном домаћинству

- Непрекидна:

- висина; тежина; старост; IQ
 - дужина листа одређене биљне врсте
 - концентрација соли у морској води
 - трајање кувања тестенине „ал денте“

- Пример - популација и обележје
 - **Популација:** скуп студената који су уписали Увод у теорију узорака школске 2019/20. године
Обележја: пол; број положених испита, број положених ЕСПБ бодова, просечна оцена свих положених испита - закључно са роком Јануар 2 ове школске године; оцена на курсу Статистика
 - **Популација:** скуп свих пољопривредних газдинстава у Србији (референтни период - октобар / новембар 2018)
Обележја: површина коришћеног пољопривредног земљишта; број грла стоке; примењени производни методи
 - **Популација:** скуп свих домаћинстава у региону Шумадије и Источне Србије (референтни период -2017. г)
Обележје: лична потрошња домаћинства (месечни просек)
 - **Популација:** једна серија LED сијалица извесног проивођача
Обележје: дужина радног века сијалице у сатима
 - **Популација:** скуп свих месеци у периоду од 2000. до 2016. г.
Обележје: месечни број ветровитих дана у Вршцу

ОСНОВНИ ПОЈМОВИ

(наставак)

Обележје се може схватити као функција која ентитетима у популацији придружује реалне бројеве или неке друге вредности.

Нека је дата популација са N јединица, које су у оквиру за одабир узорка означене бројевима из скупа $\Omega = \{1, 2, \dots, N\}$ (и тиме једнозначно одређене) и нека је Y обележје од интереса. Нека је са y_k означена вредност обележја Y ентитета означеног са k .

- Задатак при истраживању обично се своди на доношење закључака о (непознатој) вредности реалне функције

$$\theta := f(y_1, y_2, \dots, y_N),$$

која се назива **популацијска вредност** ('population value') или **параметар популације**.

Најчешће функције које се појављују као параметри популације:

квантитативна
обележја

- популацијска средња вредност ('population mean')

$$m_Y = m = \frac{1}{N} \sum_{k=1}^N y_k$$

- популацијски тотал ('population total')

$$\tau_Y = \tau = \sum_{k=1}^N y_k = N m_Y$$

- популацијска дисперзија ('population variance') / стандардно одступање

$$\sigma_Y^2 = \sigma^2 = \frac{1}{N-1} \sum_{k=1}^N (y_k - m_Y)^2 \text{ и } \sigma_Y = \sigma = \sqrt{\sigma_Y^2}$$

квалитативна
обележја

- популацијска пропорција ('population proportion')

- популацијска медијана, квантили, мода...

Идеја је да се закључци о популацијским вредностима доносе на основу информација добијених испитивањем узорка.

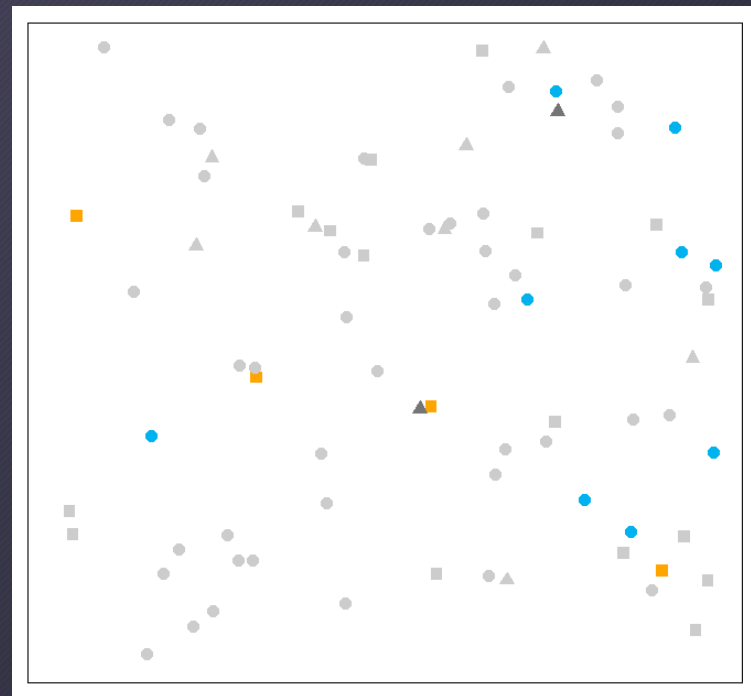
Репрезентативан узорак

„Добар“ узорак има особину **репрезентативности**.

То је узорак који представља „умањену“, а никако „искривљену“, нити „увеличану“ слику једног дела популације. Узорак са овом особином верно одсликава структуру популације коју представља, „изгледајући“ као и популација у свим аспектима релеватним за истраживање.

На репрезентативност узорка утичу:

- тип узорка (према методи одабира)
- величина узорка
- варијабилност посматраног обележја



ОСНОВНИ ПОЈМОВИ

(наставак)

- **План узорковања** ('sampling design') поседује две основне компоненте:
 - ▷ метод(и) одабира узорка
 - ▷ метод(и) закључивања
- **Метод одабира узорка** је поступак којим се бирају елементи популације у узорак, уз одређивање адекватног обима узорка.

Ови методи могу се поделити у две групе:

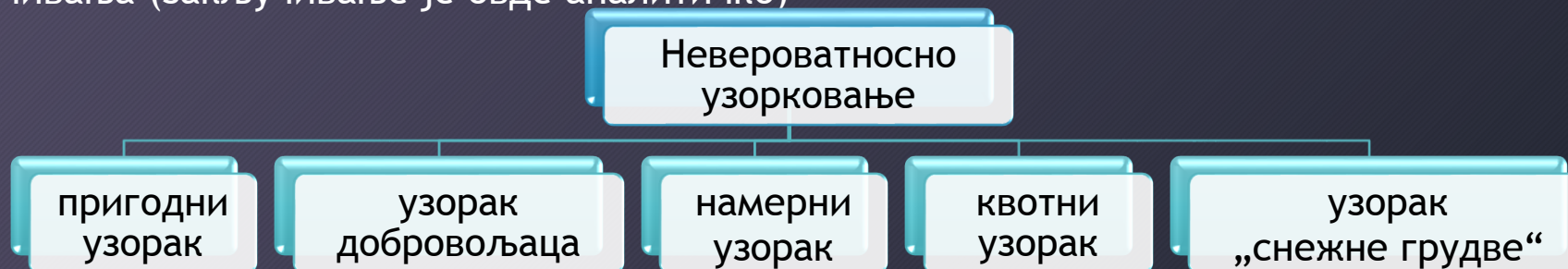
- **вероватносно узорковање** ('probability sampling')
- **невероватносно узорковање** ('nonprobability sampling')

Невероватносно узорковање

Овакви методи узорковања не заснивају се на теорији вероватноћа, него на одређеним критеријумима истраживача.

Дакле, њихова основна особина јесте да се узорковање врши на основу **субјективне процене истраживача**, а не случајним избором. Њима се прибегава онда када је, због ограничених временских рокова, износа трошкова и осетљивости предмета истраживања (етичких обзира), тешко спровести случајно узорковање.

- ▲ Предности: ефикасније се примењују код експлоративних истраживања (пилот истраживања, студије у циљу доказивања концепта, квалитативна истраживања, студије за генерисање хипотеза), чији циљ није прецизно закључивање о параметрима популације на основу репрезентативног узорка
- ▼ Мане: није могуће одређивање квалитета узорка, а самим тим ни квантификовање тачности закључивања (закључивање је овде аналитичко)



Вероватносно узорковање

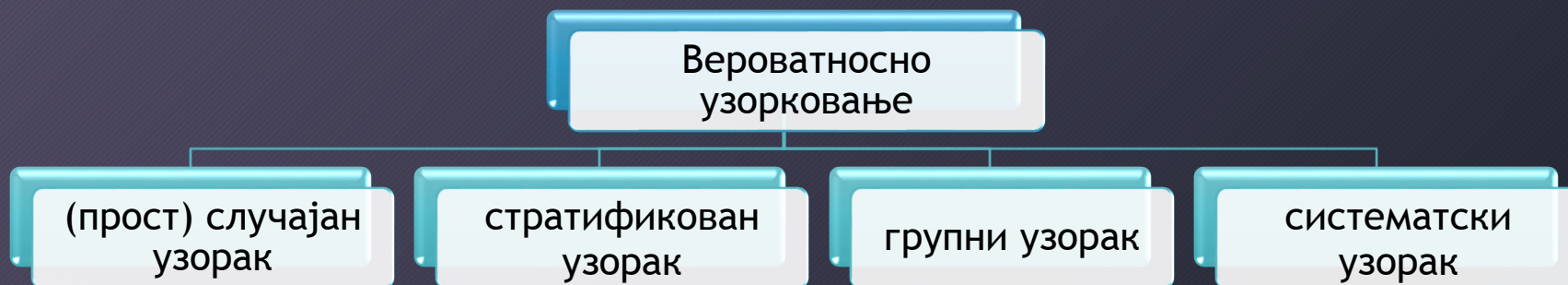
Овакви методи узорковања заснивају се на теорији вероватноћа, тј. на „планираној“ случајности.

Могући узорци су фактички математички конструисани, и за сваког од њих позната је вероватноћа да буде одабран. Дакле, узорковање се врши у складу са расподелом вероватноћа, дефинисаном на колекцији свих могућих узорака.

👉 Нека је са Ω означен скуп ознака јединица у популацији и нека $s \subset \Omega$ представља узорак. Вероватносно узорковање се заснива на познавању расподеле вероватноћа $p(\cdot)$:

$$p(s) \geq 0, \text{ за свако } s \subset \Omega, \text{ и } \sum_{s \subset \Omega} p(s) = 1$$

Случајан узорак $\mathcal{S}$ је онда случајан скуп ознака јединица са расподелом вероватноћа:
 $P\{\mathcal{S} = s\} = p(s), \text{ за свако } s \subset \Omega$



Вероватносно узорковање

▲ Предности:

- доследном применом искључује се постојање било какве пристрасности, што доприноси постизању објективности истраживања
- виши ниво поузданости резултата истраживања
- могућност процене / квантификовања узорачке грешке
- повећане су шансе за доношење ваљаних закључака о читавој популацији, уопштавањем резултата добијених испитивањем узорка

▼ Мане:

- углавном се тичу потреба за временом, ресурсима, финансијама и људством (нпр. потребно је поседовати комплетан оквир за одабир узорка)

ОСНОВНИ ПОЈМОВИ

(наставак)

Ако се на случајан начин (са унапред одређеном вероватноћом) одабере једна јединица из популације, вредност обележја коју она има није унапред позната / одређена. То значи да се вредност обележја случајно одабране јединице може схватити као реализација случајне величине. Расподела вероватноћа те случајне величина назива се **расподела обележја**.

👉 Задатак математичке статистике је одређивање расподеле обележја, или, одређивање бар неких општих нумеричких карактеристика те расподеле

- **Статистика** ('statistic') је функција вредности обележја регистрованих на јединицама из одабраног узорка, у којој евентуално могу фигулисати и неке познате константе.

👉 Статистика је случајна величина са својом расподелом вероватноћа, која се назива **узорачка расподела**

Статистике су значајне јер се често користе за формирање **оцена** ('estimator') параметара популације.

Реализоване вредности статистика су реални бројеви који тада дају **оцене** ('estimate') непознатих параметара.

Нпр. ако је θ непозната популацијска вредност онда је $\hat{\theta} = \hat{\theta}(\mathcal{S})$ статистика, која представља **тачкасту оцену** ('point estimator') параметра θ .

Често коришћене статистике ($n(\mathcal{S})$ представља обим узорка $\mathcal{S}$):

- узорачка средња вредност ('sample mean')

$$\bar{Y} = \frac{1}{n(\mathcal{S})} \sum_{k \in \mathcal{S}} y_k$$

- узорачки тотал ('sample total')

$$T = n(\mathcal{S})\bar{Y}$$

- узорачка дисперзија ('sample variance') / стандардно одступање

$$\bar{S}^2 = \frac{1}{n(\mathcal{S}) - 1} \sum_{k \in \mathcal{S}} (y_k - \bar{Y})^2 \text{ и } \bar{S} = \sqrt{\bar{S}^2}$$

- узорачка пропорција ('sample proportion')

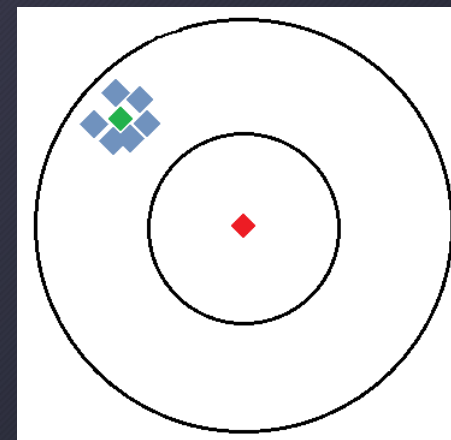
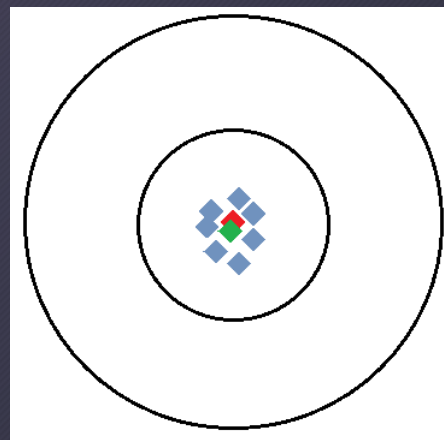
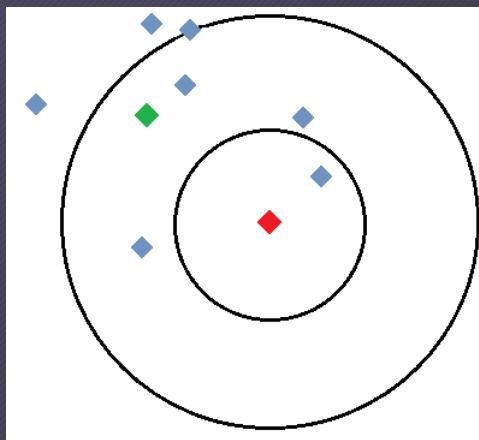
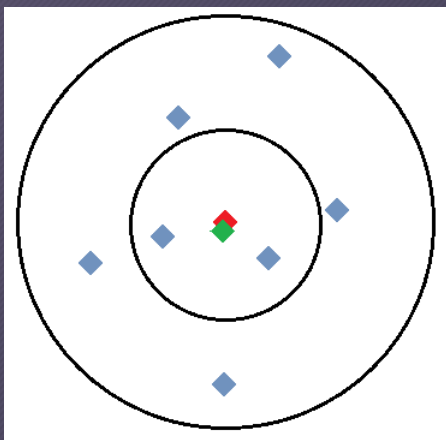
- узорачка медијана, квантили, мода...

Нека је $\hat{\theta}$ тачкаста оцена популацијске вредности θ . Она је:

- **непристрасна** ('unbiased')
ако једнакост $E\hat{\theta} = \theta$ важи за сваку вредност параметра θ ;
ако оцена $\hat{\theta}$ није непристрасна онда се она назива **пристрасна оцена**,
а вредношћу разлике $B(\hat{\theta}) := E\hat{\theta} - \theta$ мери се њена **пристрасност**
- **прецизна** ('precise')
ако је дисперзија $D\hat{\theta} = E(\hat{\theta} - E\hat{\theta})^2$ оцене $\hat{\theta}$ мала (тежи 0)
- **тачна** ('accurate')
ако је средње квадратна грешка $MSE(\hat{\theta}) := E(\hat{\theta} - \theta)^2$ оцене $\hat{\theta}$ мала

👉 Важи и једнакост: $MSE(\hat{\theta}) = D\hat{\theta} + (B(\hat{\theta}))^2$,
па је оцена тачна ако је и прецизна и непристрасна

$SE(\hat{\theta}) := \sqrt{D\hat{\theta}}$ назива се
стандардна грешка оцене $\hat{\theta}$



◆ – реализоване вредности оцене $\hat{\theta}$ на различитим узорцима; ◆ – средња вредност $E\hat{\theta}$ оцене $\hat{\theta}$; ◆ – популацијска вредност θ

Литература



Sharon L. Lohr, “*Sampling: Design and Analysis*”,

Brooks/Cole, Cengage Learning, 2010



Steven K. Thompson, “*Sampling*”,

John Wiley & Sons, Inc., 2012



Зоран Ивковић и остали, „*Статистика* – за IV разред
усмереног образовања“

Научна књига, 1980.



Љиљана Петровић, „*Теорија узорака и планирање
експеримената*“,

ЕКОФ ЦИД, 2018.