

Теорија узорака

26. ФЕБРУАР '19.

(Прост) случајан узорак

- ▶ Код (простог) случајног узорковања ('simple random sampling'):

јединица посматрања = јединица узорковања.

Нека је дата популација са N јединица, које су у оквиру за одабир узорка означене бројевима из скупа $\Omega = \{1, 2, \dots, N\}$ и нека је Y обележје од интереса. Бира се узорак обима n .

- ▶ Може бити:

- без понављања ('without replacement')
- са понављањем ('with replacement')

👉 скраћеница: SRSWOR

👉 скраћеница: SRSWR

SRSWOR

- ▶ Представља један од најједноставнијих и најстаријих метода одабира узорка
- ▶ Расподела вероватноћа $p(\cdot)$ на колекцији свих узорака $s \subset \Omega$ дата је са:

$$p(s) = \begin{cases} \binom{N}{n}^{-1}, & \text{ако је обим узорка } s \text{ једнак } n \\ 0, & \text{иначе} \end{cases}$$

Дакле, овде се сваки од $\binom{N}{n}$ могућих подскупова скупа Ω кардиналности n са подједнаком (позитивном) вероватноћом може одабрати као узорак

- ▶ Поменути план обично се у пракси имплементира једним од следећа два еквивалентна поступка:
 - одабир узорка врши се кроз n извлачења („корака“) на случајан начин, при чему је у сваком кораку вероватноћа извлачења било које од јединица, које у ранијим корацима нису одабране у узорак, иста
 - одабир узорка врши се кроз низ **независних** извлачења на случајан начин **из целе популације**, при чему је у сваком кораку вероватноћа извлачења било које од јединица иста $\left(= \frac{1}{N}\right)$, уз одбацивање јединица раније одабраних у узорак и понављање корака све док се не добије узорак обима n

☞ Узорак одабран на описани начин може се приказати и као **уређен** низ $(j_1, j_2, \dots, j_n)$ ознака јединица које су се нашле у узорку (j_k је ознака k -те јединице задржане у узорку).

Под узорком се, такође, подразумева и припадни низ $(y_{j_1}, y_{j_2}, \dots, y_{j_n})$ вредности посматраног обележја Y регистрованих на одабраним јединицама.

Парови (j_k, y_{j_k}) , $k = \overline{1, n}$, представљају **податке добијене у истраживању**.

SRSWR

- ▶ Одабир узорка врши се кроз n **независних** извлачења на случајан начин, и то увек **из комплетне популације**, при чему је у сваком кораку вероватноћа извлачења било које од јединица иста и једнака $\frac{1}{N}$
- ▶ Расподела вероватноћа $p(\cdot)$ на колекцији свих узорака $s \in \Omega^n$ као уређених низова дужине n са дозвољеним понављањем елемената дата је са:

$$p(s) = N^{-n}$$

ИЗВЛАЧЕЊЕ ЈЕДИНИЦЕ НА СЛУЧАЈАН НАЧИН

- ▶ Случајан одабир јединице (из популације у узорак) врши се коришћењем **случајних** и **псеудослучајних бројева**.
- ▶ Случајни бројеви обично се добијају помоћу тзв. **физичких генератора** (TRNG – 'true random number generator')
 - у макро свету: бацање фер новчића / коцкица, случајан избор карте из шпила / куглице из кутије, рулет итд.
 - у микро свету: природни феномени за које важе законитости квантне механике, шум итд.

Они су садржани у тзв. **таблицама случајних бројева**.

- ▶ Псеудослучајни бројеви се добијају помоћу тзв. **програмских генератора** (PRNG – 'pseudorandom number generator')
То су рачунарски програми који користе изван алгоритама за добијање низа бројева чија својства, у одређеној мери, oponaшају својства низа случајних бројева.

НОВИ ПОЈМОВИ

- ▶ **Индикатор укључења** ('inclusion indicator'):

$$I_k = \begin{cases} 1, & \text{ако је јединица означена са } k \text{ одабрана у узорак} \\ 0, & \text{иначе} \end{cases}$$

- ▶ **Вероватноћа укључења** ('inclusion probability') првог, односно другог реда:

π_k – вероватноћа да јединица означена са k буде одабрана у узорак

π_{kl} – вероватноћа да и јединица означена са k и јединица означена са l буду одабране у узорак

- ▶ **„Тежина“ узорковања** ('sampling weight') – реципрочна вредност очекиваног броја појављивања јединице означене са k у узорку (што се, код узорка без понављања, своди на реципрочну вредност вероватноће укључења првог реда π_k)
може се интерпретирати као број јединица у популацији које репрезентује јединица означена са k

SRSWOR



SRSWR

SRSWOR

Вероватноћа укључења првог реда:

$$\pi_k = \frac{n}{N} \text{ за свако } k$$

Вероватноћа да ће јединица означена са k бити одабрана у узорак у j -том кораку:

$$\frac{1}{N}$$

☝ то би био први, а уједно и једини, корак у коме је она извучена

Очекивани број појављивања јединице означене са k у узорку:

$$\pi_k$$

Вероватноћа укључења другог реда:

$$\pi_{kl} = \frac{n(n-1)}{N(N-1)} \text{ за } k \neq l$$

SRSWR

Вероватноћа укључења првог реда:

$$\pi_k = 1 - \left(\frac{N-1}{N}\right)^n \text{ за свако } k$$

Вероватноћа да ће јединица означена са k бити одабрана у узорак у j -том кораку:

$$\frac{1}{N}$$

Вероватноћа да ће јединица означена са k бити одабрана у узорак више од једанпут:

$$1 - \left(\frac{N-1}{N}\right)^{n-1} \left(\frac{N-1-n}{N}\right)$$

Очекивани број појављивања јединице означене са k у узорку:

$$\frac{n}{N}$$

Вероватноћа укључења другог реда:

$$\pi_{kl} = 1 - 2\left(\frac{N-1}{N}\right)^n + \left(\frac{N-2}{N}\right)^n \text{ за } k \neq l$$

Приступи приликом закључивања

приступ заснован на методу одабира узорка ('design-based approach')	приступ заснован на моделу ('model-based approach')
популација је фиксирана: вредности $y_1, y_2, \dots, y_N$ обележја Y на јединицима у популацији сматрају се фиксираним (детерминистичким), али непознатим вредностима теорија вероватноћа појављује се само у вези са случајношћу приликом одабира узорка, а која је планирана кроз наметнути метод одабира узорка	популација је стохастичка: вредности $Y_1, Y_2, \dots, Y_N$ обележја Y на јединицима у популацији сматрају се случајним величинама модел популације дат је заједничком расподелом вероватноћа сл. вектора $(Y_1, Y_2, \dots, Y_N)$, у којој учествује бар једна непозната константа (параметар) конкретан низ вредности $(y_1, y_2, \dots, y_N)$ обележја Y на јединицима у популацији представља само једну реализацију сл. вектора $(Y_1, Y_2, \dots, Y_N)$

Приступи приликом закључивања (наставак)

приступ заснован на методу одабира узорка ('design-based approach')	приступ заснован на моделу ('model-based approach')
узорачка расподела статистике је дискретна расподела вероватноћа: ако је $\hat{\theta} = \hat{\theta}(s)$ статистика, онда важи $P\{\hat{\theta} = m\} = \sum_{s: \hat{\theta}(s)=m} p(s)$ а њено математичко очекивање и дисперзија израчунавају се по формулама: $E\hat{\theta} = \sum_m mP\{\hat{\theta} = m\} = \sum_s \hat{\theta}(s) \cdot p(s)$ $D\hat{\theta} = \sum_s (\hat{\theta}(s) - E\hat{\theta})^2 \cdot p(s)$	узорачка расподела статистике је нека једнодимензиона расподела вероватноћа одређена заједничком расподелом вероватноћа претпостављеног модела популације
непристрасност тачкасте оцене $\hat{\theta}$ у односу на метод одабира узорка	непристрасност тачкасте оцене $\hat{\theta}$ у односу на модел

SRSWOR SRSWR

ТАЧКАСТЕ ОЦЕНЕ

Нека је са $\mathcal{S}$ означен (прост) случајан узорак обима n . Кључни резултати у вези непознатом популацијском средњом вредношћу m_Y , када је **приступ заснован на методу** SRSWOR, односно SRSWR, одабира узорка, дати су у следећој табели:

	SRSWOR	SRSWR	SRSWR (узимају се у обзир само различите јединице)
тачкаста оцена $\hat{m}_Y$	$\frac{1}{n} \sum_{k \in \mathcal{S}} y_k$	$\frac{1}{n} \sum_{k=1}^n y_{jk}$	$\frac{1}{n_D} \sum_k y_{(k)}$
$E \hat{m}_Y$	m_Y	m_Y	m_Y
$D \hat{m}_Y$	$\frac{\sigma^2}{n} \left(1 - \frac{n}{N}\right)$	$\frac{N-1}{N} \cdot \frac{\sigma^2}{n}$	$\sum_{k=1}^{N-1} \frac{k^{n-1}}{N^n} \cdot \sigma^2$
тачкаста оцена $D \hat{m}_Y$	$\frac{\bar{S}^2}{n} \left(1 - \frac{n}{N}\right)$	$\frac{\bar{S}^2}{n}$	

n_D је **ефективан обим узорка**, тј. обим редукованог узорка $(y_{(1)}, y_{(2)}, \dots, y_{(n_D)})$ у коме су изостављена евентуална понављања јединица из оригиналног узорка

♪ може се показати да је $\bar{S}^2$ непристрасна оцена σ^2

где је σ^2 (непозната) популацијска дисперзија, а $\bar{S}^2$ (позната) узорачка дисперзија.

Нови појмови

- ▶ Стопа одабира узорка, или тзв. **разломак узорковања** ('sampling fraction'), је однос обима узорка и обима популације, тј. количник $\frac{n}{N}$
- ▶ Вредност $1 - \frac{n}{N}$ назива се **фактор корекције** због коначности популације ('finite-population correction factor')
 - 👉 У пракси се често занемарује када стопа одабира узорка не прелази 5%, а у многим случајевима и када је до 10%
- ▶ Када су познати математичко очекивање и дисперзија тачкасте оцене $\hat{\theta}$ може се одредити **коефицијент варијације** оцене $\hat{\theta}$, дефинисан са:

$$CV(\hat{\theta}) := \frac{SE(\hat{\theta})}{E\hat{\theta}}$$

и који представља релативну меру варијабилности оцене

SRSW(O)R – тачкасте оцене

Нека је са $\mathcal{S}$ означен случајан узорак без понављања обима n . Када је **приступ заснован на моделу**, врло једноставан модел популације био би модел у коме су случајне величине $Y_1, Y_2, \dots, Y_N$ **независне и имају исту расподелу вероватноћа као посматрано обележје Y** . Кључни резултати у вези непознатом средњом вредношћу $m_Y := EY$ обележја Y , дати су у следећој табели:

тачкаста оцена $\hat{m}_Y$	$\bar{Y} = \frac{1}{n} \sum_{k \in \mathcal{S}} Y_k$
$E\hat{m}_Y$	m_Y
$D\hat{m}_Y$	$\frac{\sigma_Y^2}{n}$
тачкаста оцена $D\hat{m}_Y$	$\frac{\bar{S}^2}{n}$

иста оцена може се користити за оцењивање, односно **предвиђање** вредности сл. величине

$$\frac{1}{N} \sum_{k=1}^N Y_k$$

Средње квадратна грешка предвиђања једнака је:

$$\frac{\sigma_Y^2}{n} \left(1 - \frac{n}{N}\right)$$

а њена оцена:

$$\frac{\bar{S}^2}{n} \left(1 - \frac{n}{N}\right)$$

где је $\sigma_Y^2 := DY$ дисперзија обележја Y , а $\bar{S}^2$ (позната) узорачка дисперзија.

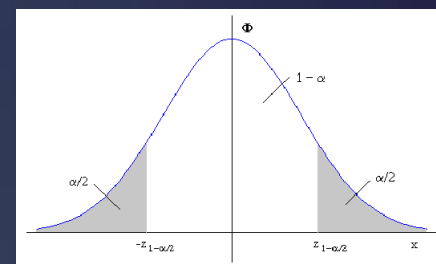
SRSW(O)R – интервалне оцене

Претпоставља се модел популације са претходног слајда, при чему обележје Y има коначну средњу вредност и дисперзију.

- Ако је обим узорка n „довољно велики“ (у пракси је довољно већ $n \geq 30$), на основу важења Централне граничне теореме, **апроксимативни** $100 \cdot (1 - \alpha)\%$ (двострани) **интервал поверења** за непознату средњу вредност m_Y обележја Y , дат је са:

$$\left[\bar{Y} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_Y^2}{n}}, \bar{Y} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_Y^2}{n}} \right]$$

↗ σ_Y^2 се оцењује са $\bar{S}^2$



где је $z_{1-\frac{\alpha}{2}}$ вредност $\left(1 - \frac{\alpha}{2}\right)$ -квантила стандардне нормалне расподеле

- Ако је обим узорка n мањи од 30, горња апроксимација не важи, па се примењује егзактан метод, који на основу претпоставки модела даје тачне интервале поверења са нивоом поверења **не мањим** од $1 - \alpha$

SRSW(O)R – интервалне оцене

Специјално, ако обележје Y има нормалну $\mathcal{N}(m_Y, \sigma_Y^2)$ расподелу **тачан** $100 \cdot (1 - \alpha)\%$ (двострани) **интервал поверења** за непознату средњу вредност m_Y :

- када је σ_Y^2 познато дат је са:

$$\left[\bar{Y} - z_{1-\frac{\alpha}{2}} \sqrt{\sigma_Y^2/n}, \bar{Y} + z_{1-\frac{\alpha}{2}} \sqrt{\sigma_Y^2/n} \right]$$

где је $z_{1-\frac{\alpha}{2}}$ вредност $\left(1 - \frac{\alpha}{2}\right)$ -квантила стандардне нормалне расподеле

- када је σ_Y^2 непознато дат је са:

$$\left[\bar{Y} - t_{n-1; 1-\frac{\alpha}{2}} \sqrt{\bar{S}^2/n}, \bar{Y} + t_{n-1; 1-\frac{\alpha}{2}} \sqrt{\bar{S}^2/n} \right]$$

где је $t_{1-\frac{\alpha}{2}}$ вредност $\left(1 - \frac{\alpha}{2}\right)$ -квантила Студентове расподеле са $(n - 1)$

степени слободе

☞ За велики обим узорка из обележја са нормалном расподелом практично нема разлике када је дисперзија обележја Y позната и када није, јер се тада Студентова расподела добро апроксимира $\mathcal{N}(0, 1)$ расподелом

SRSWR SRSWOR

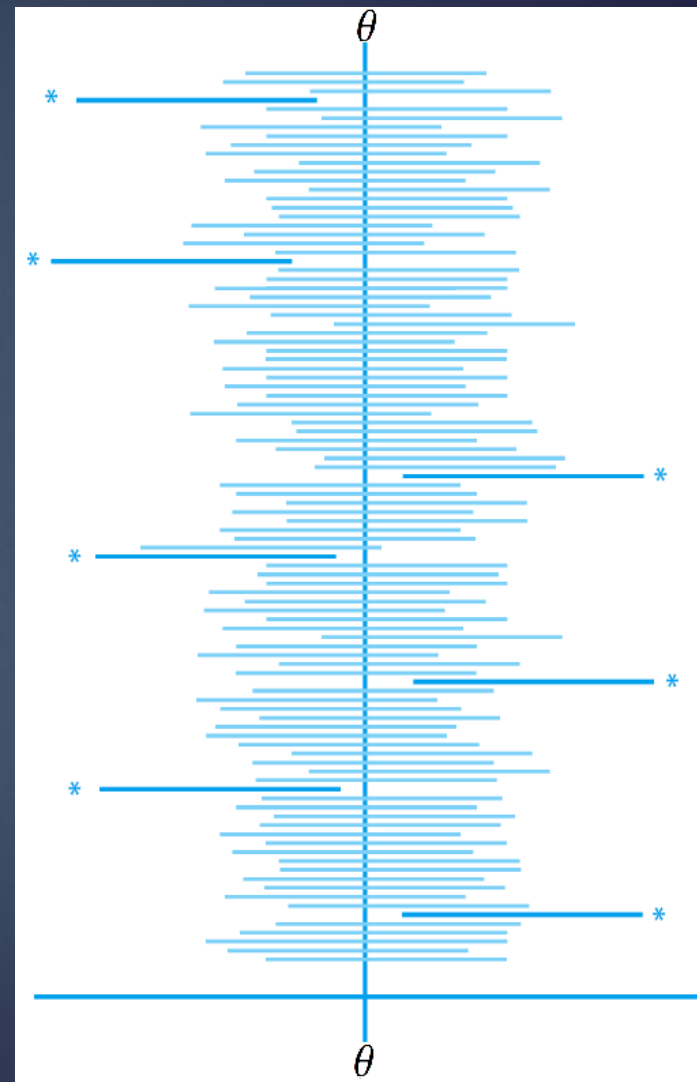
ИНТЕРВАЛНЕ ОЦЕНЕ

Нека је са $\mathcal{S}$ означен (прост) случајан узорак довољно великог обима n . Кључни **асимптотски резултати** у вези са интервалном оценом непознате популацијске средње вредности m_Y , када је **приступ заснован на методу** SRSWOR, односно SRSWR **одабира узорка**, дати су у следећој табели:

апроксимативни $100 \cdot (1 - \alpha)\%$ двострани интервал поверења	
SRSWR	$\left[\hat{m}_Y - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{S}^2}{n}}, \hat{m}_Y + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{S}^2}{n}} \right]$ код случајног узорка са понављањем чланови узорка су реализације независних и једнако расподељених случајних величина, па у основи лежи важење стандардне Централне граничне теореме (тј. узорачка средина која се појављује као тачкаста оцена за m_Y има приближно нормалну расподелу за довољно велико n)
SRSWOR	$\left[\hat{m}_Y - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{S}^2}{n} \left(1 - \frac{n}{N}\right)}, \hat{m}_Y + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{S}^2}{n} \left(1 - \frac{n}{N}\right)} \right]$ код случајног узорка без понављања чланови узорка су реализације случајних величина које нису независне, па се формулише специјална верзија Централне граничне теореме која се може применити у случају оваквог узорковања из коначне популације када су n, N и $N - n$ „довољно велики“; увођење појма „суперпопулације“

Интерпретација нивоа поверења

Интерпретација интервала поверења, односно одговарајућег нивоа поверења $(1 - \alpha)$, разликује се у зависности од приступа приликом закључивања



Одређивање обима узорка

Једно је од првих питања при планирању истраживања, а одговор на њега није увек једноставан. Суштински, ради се о одлучивању о томе колика је (узорачка) грешка прихватљива приликом закључивања, при чему се обично мора уравнотежити тачност закључивања са трошковима истраживања

- ▶ Нека је $\hat{\theta}$ тачкаста оцена непознате популацијске вредности θ . Након прецизирања **апсолутне (дозвољене) грешке** ('margin of error') Δ за задати ниво поверења $1 - \alpha$, питање се своди на одређивање вредности n тако да важи

$$P\{|\hat{\theta} - \theta| > \Delta\} < \alpha$$

Нпр. ако је $\hat{\theta}$ непристрасна, нормално расподељена оцена параметра θ онда

$$P\left\{\frac{|\hat{\theta} - \theta|}{\sqrt{D\hat{\theta}}} > z_{1-\frac{\alpha}{2}}\right\} = P\{|\hat{\theta} - \theta| > z_{1-\frac{\alpha}{2}}\sqrt{D\hat{\theta}}\} = \alpha$$

па како дисперзија оцене $\hat{\theta}$ опада са обимом узорка n , онда ће горња неједнакост бити задовољена ако се одабере довољно велико n тако да важи $z_{1-\frac{\alpha}{2}}\sqrt{D\hat{\theta}} \leq \Delta$

SRSWR SRSWOR

обим узорка

Најједноставнија једначина за одређивање обима узорка за оцењивање непознате популацијске средње вредности m_Y , тако да се постигне апсолутна грешка не већа од Δ са поверењем $1 - \alpha$, може се добити на основу апроксимативних интервала поверења:

формула за одређивање обима узорка

SRSWR

$$n_0 = \left(\frac{\sigma^2 Z_{1-\frac{\alpha}{2}}}{\Delta} \right)^2$$

SRSWOR

$$n = \frac{1}{\frac{1}{n_0} + \frac{1}{N}}$$

σ^2 је, у општем случају, непозната популацијска дисперзија; она се мора оценити на неки начин:

- спровођењем пилот истраживања на узорку „малог“ обима
- коришћењем ранијих истраживања или постојећих података у литератури
- ако не долази у обзир ништа од претходно наведеног – „погађањем“)

Одређивање обима узорка (наставак)

Поред описаног критеријума одређивања обима узорка задавањем апсолутне грешке оцено, постоје и други критеријуми и то:

- ▶ задавањем ширине интервала поверења
- ▶ задавањем горње границе дисперзије / стандардне грешке оцено
- ▶ задавањем релативне грешке оцено

Нека је $\hat{\theta}$ тачкаста оцена непознате популацијске вредности θ . Након прецизирања **релативне грешке** ρ за задати ниво поверења $1 - \alpha$, питање се се своди на одређивање вредности n тако да важи

$$P \left\{ \frac{|\hat{\theta} - \theta|}{\theta} > \rho \right\} < \alpha$$

- ▶ задавањем коефицијента варијације оцено
- ▶ задавањем трошкова узорковања

👉 Резултати који се тичу непознатог популацијског тотала τ_Y у потпуности су аналогни приказаним резултатима у вези са непознатом популацијском средњом вредношћу m_Y
(лако се добијају једни из других коришћењем везе: $\tau_Y = Nm_Y$)