Теорија узорака

Јован Самарџић, 13/2019

Професор: Милан Јовановић

- дефиниције
- ознаке
- теореме
- докази
- примери

Година курса: 2021/22

Молим да ми све грешке пријавите преко мејла или друштвених мрежа.

0. Популација и узорак

деф. Популација је скуп који се посматра. Састоји се од јединки ω.

Обим популације је број јединки популације, у ознаци N. Он је познат.

деф. Обележје је карактеристика коју придружујемо свакој јединки популације.
(најчешће нумеричка, а ако није – „кодирамо")

Са $x_1, \dots, x_N$ означавамо реализоване вредности обележја популације.

Циљ нам је да сазнамо како се обележје понаша, тј. занима нас какву расподелу има.
Пошто није практично гледати све x_i засебно, уводимо следећи појам:

деф. Параметар обележја је реална функција вредности обележја популације.

Пример: 1) укупна (тотална) сума: $t := \sum_{i=1}^{N} x_i$;

2) средња вредност: $\bar{x} := \frac{1}{N} \sum_{i=1}^{N} x_i$;

3) дисперзија/ варијанса: $s^2 := \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$;

4) стандардна девијација: $s := \sqrt{s^2}$;

5) пропорција неког својства: $p := \dfrac{\text{бр. јед. са тим својством}}{\text{укупан број јединки}} = \dfrac{A}{N}$.

Како прикупљамо податке?

деф. **Попис** је одређивање реализ. вр. за све јединке популације. (добијамо $x_1, \dots, x_N$)

У пракси, то је тешко изводљиво. Зато не испитујемо целу популацију.

деф. **Узорак** је било који подскуп популације.

Обим узорка је број јединки узорка, у ознаци n. Он је познат.

Напомена: узорак мора бити репрезентативан

Како се бира узорак?

1. корак: **узорачки оквир** – листа/списак свих јединки популације.

Нпр. држављане поређамо у листу по ЈМБГ.
Ипак, узорачки оквир и циљна популација се често не поклопе (нпр. неко нема ЈМБГ)

2. корак: **план узорковања** – правило по ком се од свих могућих узорака фиксираног обима n бира један узорак

При томе, јављају се грешке и то два типа:

а) **неузорачке грешке** – не зависе од избора плана узорковања.

Примери:
- узорачки оквир и популација се разликују;
- двосмислена питања на упитницима;
- мултипликација (телефонска анкета + више фиксних телефона у истој кући);
- проблем мерења.

б) **узорачке грешке** – настају јер закључујемо на основу узорка, а не целе популације.

Напомена: некад можемо да их контролишемо.

Постоје два типа планова узорковања:

деф. **Вероватносни план узорковања** је онај код кога је позната вероватноћа избора било ког узорка фиксиране величине.

Самим тим, позната је вероватноћа појављивања било које јединке у извученом узорку. Због тога је могуће контролисати узорачку грешку.

Постоје 4 типа:
1) **прост случајан узорак**;
2) **статификован узорак**;
3) **кластер узорак**;
4) **систематски узорак**.

У супротном, имамо **невероватносни план узорковања**

На пример:
- узорковање само из дела доступне популације;
- узорковање одокативно
- узорковање само оних који то желе, а не свих.

Пример: Имамо оквир $\mathcal{U} = \{1, 2, 3, 4\}$.

Узорци обима 2 су: $S_1 = \{1,2\}$ $S_4 = \{2,3\}$
$S_2 = \{1,3\}$ $S_5 = \{2,4\}$
$S_3 = \{1,4\}$ $S_6 = \{3,4\}$

Скуп свих узорака S се састоји од $S_1, \ldots, S_6$. Нама треба један од њих:

$$\left.\begin{aligned} &P(S_1) = 1/3 \\ &P(S_2) = 1/6 \\ &P(S_3) = P(S_4) = P(S_5) = 0 \\ &P(S_6) = 1/2 \end{aligned}\right\} \quad S: \begin{pmatrix} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \\ 1/3 & 1/6 & 0 & 0 & 0 & 1/2 \end{pmatrix}$$

– ово је пример једног вероватносног плана узорковања

$\pi_i := \sum_{S: i \in S} P(S)$ – вероватноћа да се јединка i нађе у узорку који изаберемо

$$\pi_1 = P(S_1) + P(S_2) + P(S_3) = 1/2;$$
$$\pi_2 = P(S_1) + P(S_4) + P(S_5) = 1/3;$$
$$\pi_3 = P(S_2) + P(S_4) + P(S_6) = 2/3;$$
$$\pi_4 = P(S_3) + P(S_5) + P(S_6) = 1/2.$$

Оцена и особине оцена.

деф. **Статистика** је било која Борелова функција ($\mathbb{R}^n \to \mathbb{R}$) која не зависи од непознатих параметара.

деф. **Оцена** је статистика која процењује параметар.

Пример: Настављамо претк. пример.

За оцену статистике t (укупна сума), предлаже се оцена $\hat{t} = N \cdot \bar{x}$.

Дате су реализ. вр:

i	1	2	3	4
x_i	2	5	3	6

(ми овде знамо да је $t = 16$)

$\bar{x}_{s_1} = \frac{2+5}{2} = 3.5 \Rightarrow \hat{t}_{s_1} = 4 \cdot 3.5 = 14$ и $P(S_1) = 1/3$;
$\bar{x}_{s_2} = 2.5 \Rightarrow \hat{t}_{s_2} = 10$ и $P(S_2) = 1/6$;
$\bar{x}_{s_3} = 4 \Rightarrow \hat{t}_{s_3} = 16$ и $P(S_3) = 0$;
$\bar{x}_{s_4} = 4 \Rightarrow \hat{t}_{s_4} = 16$ и $P(S_4) = 0$;
$\bar{x}_{s_5} = 5.5 \Rightarrow \hat{t}_{s_5} = 22$ и $P(S_5) = 0$;
$\bar{x}_{s_6} = 4.5 \Rightarrow \hat{t}_{s_6} = 18$ и $P(S_6) = 1/2$.

$\hat{t}: \begin{pmatrix} 10 & 14 & 16 & 18 & 22 \\ 1/6 & 1/3 & 0 & 1/2 & 0 \end{pmatrix}$ – ово је **узорачка расподела оцене** $\hat{t}$.

Приметимо $E\hat{t} = 15\frac{1}{3} \neq 16$, тј. $E\hat{t} \neq t$.

деф. Нека је $S: \begin{pmatrix} s_1 & \dots & s_\nu \\ P(s_1) & \dots & P(s_\nu) \end{pmatrix}$ један план узорковања.
Оцена $\hat{\theta}$ је **непристрасна оцена** параметра θ ако $E(\hat{\theta}) = \theta$. $\left(E(\hat{\theta}) = \sum \hat{\theta}(s_i) \cdot P(s_i) \right)$

Дакле, оцена $\hat{t}$ из примера није непристрасна, тј. **пристрасна** је.

деф. **Пристрасност оцене** је $B(\hat{\theta}) := E(\hat{\theta}) - \theta$.

деф. **Средње квадратна грешка оцене** је $MSE(\hat{\theta}) = E\left((\hat{\theta}-\theta)^2\right)$.

Приметимо: $MSE(\hat{\theta}) = E\left((\hat{\theta}-\theta)^2\right) = E\left((\underline{\hat{\theta} - E\hat{\theta} + E\hat{\theta}} - \theta)^2\right)$

$$= E\left((\hat{\theta}-E\hat{\theta})^2 + 2(\hat{\theta}-E\hat{\theta})\underbrace{(E\hat{\theta}-\theta)}_{const} + (E\hat{\theta}-\theta)^2 \right)$$

$$= E(\hat{\theta}-E\hat{\theta})^2 + 2(E\hat{\theta}-\theta)\, E(\hat{\theta}-E\hat{\theta}) + E(E\hat{\theta}-\theta)^2$$

$$= D(\hat{\theta}) + 2(E\hat{\theta}-\theta)\left(E\hat{\theta} - \underbrace{E(E\hat{\theta})}_{E\hat{\theta}}\right) + (E\hat{\theta}-\theta)^2$$

$$= D(\hat{\theta}) + (B(\hat{\theta}))^2$$

Лема 1: $MSE(\hat{\theta}) = D(\hat{\theta}) + (B(\hat{\theta}))^2$.

Доказ: управо извели

Последица: Ако је $\hat{\theta}$ непристрасна $\Rightarrow$ $MSE(\hat{\theta}) = D(\hat{\theta})$.

Напомена: Од две оцене, боља она са мањим MSE.

I део

деф. **Прост случајан узорак** има најједноставнији план узорковања.
У овом случају, вероватноће избора било ког узорка обима n су једнаке.

Одавде следи и да је једнака вероватноћа да било која јединка буде у узорку.
Другим речима, сви π_i су једнаки.

Обрнуто не важи: постоје планови узорковања са истим π_i,
али различитим вероватноћама избора било ког узорка обима n. (пример: стр. 22)

1. Прост случајан узорак без понављања

деф. **Прост случајан узорак без понављања** је онај у коме су све јединке у узорку различите.

Својства:
- укупан број узорака: $\binom{N}{n}$ ($\{1,2\}$ и $\{2,1\}$ су исти узорак)
- вероватноћа избора узорка: $\dfrac{1}{\binom{N}{n}}$
- $\pi_i \overset{(*)}{=} \dfrac{\binom{N-1}{n-1}}{\binom{N}{n}} = \dfrac{n}{N}$ не зависи од i $\Rightarrow$ сви π_i јесу једнаки.
- $\pi_{ij} = \dfrac{\binom{N-2}{n-2}}{\binom{N}{n}} = \dfrac{n(n-1)}{N(N-1)}$ ($\pi_{ij} = P\{$појављивање i-те и j-те јединке у узорку$\}$)

(*) $X_1, \dots, \boxed{X_i}, \dots, X_n$
одмах га бирамо

Како извлачимо овај узорак?

I начин: одједном од N извучемо n: $\{X_1, \dots, X_n\}$;

II начин: један по један, са избацивањем.

коректност

$$P(\{X_{i_1}, \dots, X_{i_n}\}) = \frac{n}{N} \cdot \frac{n-1}{N-1} \cdot \ldots \cdot \frac{1}{N-n+1} = \frac{n!}{\frac{N!}{(N-n)!}} = \frac{1}{\binom{N}{n}}$$

У пракси, бирање ПСУ се врши помоћу **таблица случајних бројева**.
Оне су већ генерисане (рачунарски) и прошле су све тестове случајности.

Теорема 1: 1) За ПСУ без понављања, оцена $\bar{x}_n = \frac{1}{n}\sum_{i\in S} x_i$ је непристрасна оцена за $\bar{x}$; (S-узорак)

2) За дисперзију те оцене важи: $D(\bar{x}_n) = \frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)$;

3) Непристрасна оцена те дисперзије је $\widehat{D(\bar{x}_n)} = \frac{\sigma_n^2}{n}\left(1-\frac{n}{N}\right)$, где $\sigma_n^2 = \frac{1}{n-1}\sum_{i\in S}(x_i-\bar{x}_n)^2$.

Доказ: 1) $\bar{x}_n = \frac{1}{n}\sum_{i\in S} x_i = \frac{1}{n}\sum_{i=1}^{N} x_i \cdot I_i$, при чему су индикатори зависни (ако их је n 1 остали су 0)

$I_i : \begin{pmatrix} 0 & 1 \\ 1-\pi_i & \pi_i \end{pmatrix}$, где $\pi_i = \frac{n}{N}$

$E(\bar{x}_n) = E\left(\frac{1}{n}\sum_{i=1}^{N} x_i \cdot I_i\right) = \frac{1}{n}\sum^{N} x_i \cdot E(I_i) = \frac{1}{n}\sum^{N} x_i \cdot \frac{n}{N} = \frac{1}{N}\sum^{N} x_i = \bar{x}$ $\Rightarrow$ јесте непристрасна;

2) $D(\bar{x}_n) = D\left(\frac{1}{n}\sum_{i=1}^{N} x_i\cdot I_i\right) = E\left(\left(\frac{1}{n}\sum_{i=1}^{N} x_i\cdot I_i\right)^2\right) - \left(E\left(\frac{1}{n}\sum_{i=1}^{N} x_i\cdot I_i\right)\right)^2 \overset{1)}{=} \frac{1}{n^2}E\left(\sum^{N} x_i I_i\right)^2 - \bar{x}^2$

$= \frac{1}{n^2}E\left((x_1I_1+\dots+x_NI_N)(x_1I_1+\dots+x_NI_N)\right) - \bar{x}^2$

(Δ) $(a_1+\dots+a_N)^2 = \sum^{N} a_i^2 + \sum^{N}\sum^{N}_{j\neq i} a_i a_j$

$\overset{(\Delta)}{=} \frac{1}{n^2}E\left(\sum^{N} x_i^2 I_i^2 + \sum_i^{N}\sum_{j\neq i}^{N} x_i x_j I_i I_j\right) - \bar{x}^2$

(**) 1) $I_i^2 = I_i$
2) $I_i I_j : \begin{pmatrix} 0 & 1 \\ 1-\pi & \pi_{ij} \end{pmatrix}$

$= \frac{1}{n^2}\left(\sum^{N} x_i^2 \cdot E(I_i^2) + \sum_i^{N}\sum_{j\neq i}^{N} x_i x_j E(I_i I_j)\right) - \bar{x}^2$

$\overset{(**)}{=} \frac{1}{n^2}\left(\sum^{N} x_i^2 \cdot \frac{n}{N} + \sum_i^{N}\sum_{j\neq i}^{N} x_i x_j \cdot \frac{n(n-1)}{N(N-1)}\right) - \left(\frac{1}{N}\sum^{N} x_i\right)^2$

$= \frac{1}{nN}\left(\sum^{N} x_i^2 + \frac{n-1}{N-1}\sum_i^{N}\sum_{j\neq i}^{N} x_i x_j\right) - \frac{1}{N^2}\left(\sum^{N} x_i^2 + \sum_i^{N}\sum_{j\neq i}^{N} x_i x_j\right)$

извлачимо заједнички $\rightarrow$ $= \frac{1}{nN^2}\left[(N-n)\sum^{N} x_i^2 + \left(\frac{N(n-1)}{N-1} - n\right)\sum_i^{N}\sum_{j\neq i}^{N} x_i x_j\right]$

$= \frac{1}{nN^2}\left[(N-n)\sum^{N} x_i^2 - (N-n)\frac{1}{N-1}\sum_i^{N}\sum_{j\neq i}^{N} x_i x_j\right] = \frac{N-n}{nN^2}\left[\sum^{N} x_i^2 - \frac{1}{N-1}\sum_i^{N}\sum_{j\neq i}^{N} x_i x_j\right]$

$= \frac{N-n}{nN^2}\left[\sum^{N} x_i^2 - \frac{1}{N-1}\sum_i^{N}\sum_{j\neq i}^{N} x_i x_j - \frac{1}{N-1}\sum^{N} x_i^2 + \frac{1}{N-1}\sum^{N} x_i^2\right]$

$= \frac{1}{nN}\left(1-\frac{n}{N}\right)\left[\frac{N}{N-1}\sum^{N} x_i^2 - \frac{1}{N-1}\left(\sum^{N} x_i^2 + \sum_i^{N}\sum_{j\neq i}^{N} x_i x_j\right)\right]$

$= \frac{1}{nN}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left[N\sum^{N} x_i^2 - \left(\sum^{N} x_i\right)^2\right] = \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left[\sum^{N} x_i^2 - \overset{=\frac{1}{N}}{\frac{N}{N^2}}\left(\sum^{N} x_i\right)^2\right]$

$= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left[\sum^{N} x_i^2 - N\left(\frac{1}{N}\sum^{N} x_i\right)^2\right] = \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left[\sum^{N} x_i^2 - N\bar{x}^2\right]$

(***) $N\bar{x} = \sum^{N} x_i$

Битан трик (користи се и касније)

$(N-1)s^2 = \left(\sum^{N} x_i^2 - N\bar{x}^2\right)$

$= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left[\sum^{N} x_i^2 - 2N\bar{x}^2 + \underline{N\bar{x}^2}\right] \overset{(***)}{=} \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left[\sum^{N} x_i^2 - 2\bar{x}\sum^{N} x_i + \underbrace{\sum^{N}\bar{x}^2}\right]$

N истих сабирака

$N\bar{x}^2 = \underbrace{\bar{x}^2+\dots+\bar{x}^2}_{N}$

$= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\sum^{N}\left[x_i^2 - 2x_i\bar{x} + \bar{x}^2\right]$

$= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\sum^{N}(x_i - \bar{x})^2$

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

(σ^2 је параметар популације, па је у пракси непознат, па зато га оцењујемо)

3) $E(\hat{D(\bar{x}_n)}) = E\left(\frac{s_n^2}{n}\left(1-\frac{n}{N}\right)\right) = \frac{1}{n}\left(1-\frac{n}{N}\right)E(s_n^2) \overset{\text{Л1}}{=} \frac{\sigma^2}{n}\left(1-\frac{n}{N}\right) = D(\bar{x}_n) \Rightarrow$ јесте непристрасна;

Овде смо користили непристрасност оцене s_n^2, па то морамо да докажемо:

Лема 1: За псу без понављања, s_n^2 је непристрасна оцена за σ^2.

Доказ: Доказујемо $E(s_n^2) = \sigma^2$.

$$E((n-1)s_n^2) = E\left(\sum_{i\in S}(x_i-\bar{x}_n)^2\right) \overset{\text{кв. бинома}}{=} E\left(\sum_{i\in S}(x_i^2 - 2x_i\bar{x}_n + \bar{x}_n^2)\right)$$

$$= E\left(\sum_{i\in S}x_i^2 - 2\bar{x}_n\sum_{i\in S}x_i + \underbrace{\sum_{i\in S}\bar{x}_n^2}_{n \text{ сабирака (исто као пре)}}\right)$$

$$= E\left(\sum_{i\in S}x_i^2 - 2n\frac{1}{n}\bar{x}_n\sum_{i\in S}x_i + n\cdot\bar{x}_n^2\right)$$

$$= E\left(\sum_{i\in S}x_i^2 - 2n\bar{x}_n^2 + n\bar{x}_n^2\right) = E\left(\sum_{i\in S}x_i^2 - n\bar{x}_n^2\right)$$

$$= E\left(\sum^N x_i^2 I_i - n\bar{x}_n^2\right) = E\left(\sum^N x_i^2 I_i\right) - nE(\bar{x}_n^2)$$

$$= \sum^N x_i^2 E(I_i) - nE(\bar{x}_n^2) \overset{EX^2 = DX-(EX)^2}{=} \sum^N x_i^2\frac{n}{N} - n\left(D(\bar{x}_n) + (E(\bar{x}_n))^2\right)$$

$$= \frac{n}{N}\sum^N x_i^2 - n\left(\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right) + \bar{x}^2\right) = \frac{n}{N}\sum^N x_i^2 - n\bar{x}^2 - \sigma^2\left(1-\frac{n}{N}\right)$$

$N\bar{x} = \sum^N x_i$

и пишемо $-2+1$

па наместимо на σ^2

(као у ***) уназад

$$= \frac{n}{N}\left(\sum^N x_i^2 - N\bar{x}^2\right) - \sigma^2\left(1-\frac{n}{N}\right) = \frac{n}{N}(N-1)\sigma^2 - \sigma^2\left(1-\frac{n}{N}\right)$$

$$= \sigma^2(n-1)$$

Дакле: $E((n-1)s_n^2) = \sigma^2(n-1) \Rightarrow E(s_n^2) = \sigma^2$.

деф. **Стандардна грешка** оцене $\hat{\theta}$ је $SE(\hat{\theta}) := \sqrt{MSE(\hat{\theta})}$

Последица: Стандардна грешка оцене $\bar{x}_n$ је $\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$.

Напомена: За оцену стандардне грешке, узима се $\frac{s_n}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$.

Ова оцена није непристрасна! (Без обзира што је $E(s_n^2) = \sigma^2$, не важи $E(s_n) = \sigma$)

Теорема 2: 1) За ПСУ без понављања, оцена $\hat{t} = N\cdot\bar{x}_n$ је непристрасна оцена за t;

2) За дисперзију те оцене важи: $D(\hat{t}) = N^2\cdot\frac{s^2}{n}\left(1-\frac{n}{N}\right)$;

3) Непристрасна оцена те дисперзије је $\widehat{D(\hat{t})} = N^2\cdot\frac{s_n^2}{n}\left(1-\frac{n}{N}\right)$.

Доказ: тривијално следи из Т1 и чињенице $t = N\cdot\bar{x}$.

$\boxed{2.}$ Оцена пропорције на основу ПСУ без понављања

Подсетимо се из $\boxed{0}$: $p = \frac{A}{N}$.

деф. Узорачка пропорција је $p_n := \frac{a}{n}$, a - бр. јединки узорка са датим својством.

Кодирање: Ово важи универзално! (без и са понављањем).

$$x_i = \begin{cases} 1, & \text{има то својство} \\ 0, & \text{нема} \end{cases}$$ - обележје које користимо $\Rightarrow$ $A = \sum^N x_i$

* $p = \frac{\sum^N x_i}{N} = \bar{x}$

* $$s^2 = \frac{1}{N-1} \sum^N (x_i - \bar{x})^2 = \frac{1}{N-1} \Big(\sum^N x_i^2 \overset{(-2+1)\bar{x}^2}{-} N\bar{x}^2 \Big)$$

$$\overset{x_i^2 = x_i}{=} \frac{1}{N-1} \Big(\sum^N x_i - N\bar{x}^2 \Big) = \frac{1}{N-1} (Np - Np^2)$$

$$= \frac{N}{N-1} p(1-p)$$

Теорема 1: 1) За ПСУ без понављања, оцена $p_n = \frac{a}{n}$ је непристрасна оцена за p;

2) За дисперзију те оцене важи: $D(p_n) = \frac{N-n}{N-1} \cdot \frac{p(1-p)}{n}$;

3) Непристрасна оцена те дисперзије је $\widehat{D(p_n)} = \frac{N-n}{N} \cdot \frac{p_n(1-p_n)}{n-1}$.

Доказ: 1) По $\boxed{1}$ Т1.1: $E(\bar{x}_n) = \bar{X} \overset{\text{код}}{\Rightarrow} E(p_n) = p \Rightarrow$ јесте непристрасна;

2) $D(p_n) = \frac{s^2}{n}\left(1-\frac{n}{N}\right) \overset{\text{код}}{=} \frac{1}{n} \cdot \frac{N}{N-1} p(1-p)\left(1-\frac{n}{N}\right) = \frac{N-n}{N-1} \cdot \frac{p(1-p)}{n}$;

3) $\widehat{D(p_n)} = \frac{s_n^2}{n}\left(1-\frac{n}{N}\right) \overset{\text{Л1}}{=} \frac{n}{n-1} \cdot \frac{p_n(1-p_n)}{n} \cdot \frac{N-n}{n} = \frac{N-n}{N} \cdot \frac{p_n(1-p_n)}{n-1} \Rightarrow$ јесте непристрасна.

Лема 1: $s_n^2 = \frac{n}{n-1} p_n(1-p_n)$

Доказ: $s_n^2 = \frac{1}{n-1}\sum_{i\in S}(x_i - \bar{x}_n)^2 = \frac{1}{n-1}\left(\sum_{i\in S} x_i^2 - n\bar{x}_n^2\right)$

$= \frac{1}{n-1}\left(\sum_{i\in S} x_i - n\bar{x}_n^2\right) = \frac{1}{n-1}\left(np_n - np_n^2\right)$

$= \frac{n}{n-1} p_n(1-p_n)$.

Теорема 2: 1) За ПСУ без понављања, оцена $\hat{A} = N \cdot p_n$ је непристрасна оцена за A;

2) За дисперзију те оцене важи: $D(\hat{A}) = N^2 \cdot \frac{N-n}{N-1} \cdot \frac{p(1-p)}{n}$;

3) Непристрасна оцена те дисперзије је $\widehat{D(\hat{A})} = N \cdot (N-n) \cdot \frac{p_n(1-p_n)}{n-1}$.

Доказ: тривијално следи из Т1 и чињенице $A = N \cdot p$.

3. Прост случајан узорак са понављањем

деф. Прост случајан узорак са понављањем је онај у коме се јединке у узорку могу понављати.

Својства: - укупан број узорака: N^n

- вероватноћа избора узорка: $\frac{1}{N^n}$ $\left(= \frac{1}{N} \cdot \ldots \cdot \frac{1}{N} \right)$

Напомена: Редослед је битан! (нпр. (1,3,4,5) и (5,1,3,4) су различити узорци)
Зато кажемо да су ово уређени узорци.

Такође, нпр. $P(\{1,1,1,1\}) \neq P(\{1,3,4,5\})$.
↳ само (1,1,1,1) ↳ може бити (1,3,4,5), (5,1,3,4)...

Теорема 1: 1) За ПСУ са понављањем, оцена $\bar{X}_n = \frac{1}{n}\sum^{n} X_i$ је непристрасна оцена за $\bar{X}$;

2) За дисперзију те оцене важи: $D(\bar{X}_n) = \frac{\sigma^2}{n}\left(1 - \frac{1}{N}\right)$;

3) Непристрасна оцена те дисперзије је $\hat{D}(\bar{X}_n) = \frac{\sigma_n^2}{n}$.

Доказ: Имамо узорак обима n: $(X_1, \ldots, X_n)$ - X_i су независне и имају исту расподелу као обележје X.

$X: \begin{pmatrix} x_1 & \ldots & x_N \\ 1/N & \ldots & 1/N \end{pmatrix}$.

случајно ⇒ све једнако ⇒ $\frac{1}{N}$

Рачунамо: * $EX = \frac{1}{N} \cdot \sum x_i = \bar{x}$;

* $DX = EX^2 - (EX)^2 = \frac{1}{N}(x_1^2 + \ldots + x_N^2) - \bar{x}^2$

$= \frac{1}{N}\left[\sum^{N} x_i^2 - N\bar{x}^2\right] = \frac{N-1}{N} \cdot s^2$.

опет: $(N-1)s^2 = \left(\sum^{N} x_i^2 - N\bar{x}^2\right)$

1) $E(\bar{X}_n) = E\left(\frac{1}{n}\sum^{n} X_i\right) = \frac{1}{n} \cdot \sum^{n} EX_i = \frac{1}{n}\sum^{n} EX = \frac{1}{n} \cdot n \cdot EX = \bar{x}$ ⇒ јесте непристрасна;

2) $D(\bar{X}_n) = D\left(\frac{1}{n}\sum^{n} X_i\right) \overset{\text{нез.}}{=} \frac{1}{n^2}\sum D(X_i) = \frac{n}{n^2} DX = \frac{s^2}{n}\left(1 - \frac{1}{N}\right)$;

3) $E(\widehat{D(\bar{x}_n)}) = E\left(\frac{s_n^2}{n}\right) = \frac{1}{n} E(s_n^2) \overset{\text{Л1}}{=} \frac{1}{n} \cdot \frac{N-1}{N} s^2 = D(\bar{x}_n)$ $\Rightarrow$ јесте непристрасна.

Лема 1: За ПСУ са понављањем: $E(s_n^2) = \frac{N-1}{N} \cdot s^2$

Доказ: $E(s_n^2) = E\left(\frac{1}{n-1} \sum^n (x_i - \bar{x}_n)^2\right) = \frac{1}{n-1} E\left(\sum^n x_i^2 - n \cdot \bar{x}_n^2\right) = \frac{1}{n-1}\left(n E(X^2) - n E(\bar{X}_n^2)\right)$

$$= \frac{1}{n-1}\left[n \cdot (DX + (EX)^2) - n \cdot (D\bar{X}_n + (E\bar{X}_n)^2)\right]$$

$$\overset{1),2)}{=} \frac{n}{n-1}\left[\frac{N-1}{N} s^2 + \cancel{\bar{x}^2} - \frac{s^2}{n}\left(1 - \frac{1}{N}\right) - \cancel{\bar{x}^2}\right]$$

$$= \frac{N-1}{N} \cdot \frac{n}{n-1} \cdot s^2 \left(1 - \frac{1}{n}\right) = \frac{N-1}{N} \cdot s^2.$$

Теорема 2: 1) За ПСУ са понављањем, оцена $\hat{t} = N \cdot \bar{x}_n$ је непристрасна оцена за t;

2) За дисперзију те оцене важи: $D(\hat{t}) = \frac{N(N-1)}{n} \cdot s^2$;

3) Непристрасна оцена те дисперзије је $\widehat{D(\hat{t})} = \frac{N^2 s_n^2}{n}$.

Доказ: тривијално следи из Т1 и чињенице $t = N \cdot \bar{x}$.

* Упоређујемо оцене за случајеве ПСУ без и са понављањем:

Знамо да је боља оцена која има мање MSE.
Како су наше оцене непристрасне, гледамо која има мању дисперзију.

$$D_{BP}(\bar{x}_n) = \underbrace{\frac{N-n}{N-1}}_{<1} D_{SP}(\bar{x}_n) \quad \Rightarrow \quad D_{BP}(\bar{x}_n) < D_{SP}(\bar{x}_n);$$

Закључак: Оцене без понављања су боље од оцена са понављањем код $\underbrace{\text{коначних}}_{N<\infty}$ популација.

$\boxed{4.}$ Оцена пропорције на основу ПСУ са понављањем

Кодирање: исто као у $\boxed{2}$.

Теорема 1: 1) За ПСУ са понављањем, оцена $p_n = \frac{a}{n}$ је непристрасна оцена за p;

2) За дисперзију те оцене важи: $D(p_n) = \frac{p(1-p)}{n}$;

3) Непристрасна оцена те дисперзије је $\widehat{D(p_n)} = \frac{p_n(1-p_n)}{n-1}$.

Доказ: 1) $E(\bar{x}_n) = \bar{X} \overset{\text{код}}{\Longrightarrow} E(p_n) = p$;

2) $D(p_n) = \frac{s^2}{n}\left(1-\frac{1}{N}\right) \overset{\text{код}}{=} \frac{N}{N-1} \cdot \frac{p(1-p)}{n} \cdot \frac{N-1}{N} = \frac{p(1-p)}{n}$;

3) $\widehat{D(p_n)} = \frac{s_n^2}{n} \overset{\boxed{2}\text{Л1}}{=} \frac{n}{n-1} \cdot \frac{p_n(1-p_n)}{n} = \frac{p_n(1-p_n)}{n-1}$.

Теорема 2: 1) За ПСУ са понављањем, оцена $\hat{A} = N \cdot p_n$ је непристрасна оцена за A;

2) За дисперзију те оцене важи: $D(\hat{A}) = N^2 \cdot \frac{p(1-p)}{n}$;

3) Непристрасна оцена те дисперзије је $\widehat{D(\hat{A})} = N^2 \cdot \frac{p_n(1-p_n)}{n-1}$.

Доказ: тривијално следи из Т1 и чињенице $A = N \cdot p$.

5. Интервали поверења на основу ПСУ без понављања

До сада смо гледали искључиво тачкасте оцене, а сад гледамо интервалне.

деф. Са I означавамо интервал поверења, а $1-\alpha$ нам је ниво поверења.

То значи $P\{\theta \in I\} = 1-\alpha$, где је θ непознато.

Интервал поверења за $\bar{X}$

Хајек је доказао (један облик) ЦГТ за ПСУ без понављања:

За довољно велико n, N, $N-n$ код ПСУ без понављања важи: $X^* = \dfrac{\overline{X}_n - \bar{X}}{\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}} \approx \mathcal{N}(0,1)$.

$n \geq 30$

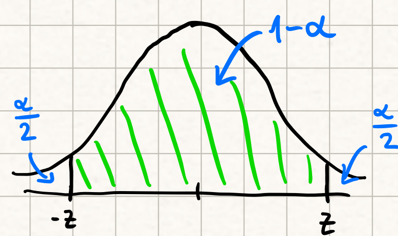
$E\overline{X}_n$, $\sqrt{D\overline{X}_n}$

Наравно, постоји ∞ интервала поверења, па морамо неки да изаберемо.
Посматрамо тзв. центрирани интервал поверења:

$$P\{|X^*| \leq z\} = P\{-z \leq X^* \leq z\} = 1-\alpha$$

$$P\left\{-z \leq \frac{\overline{X}_n - \bar{X}}{\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}} \leq z\right\} = 1-\alpha$$

$$P\left\{\overline{X}_n - z\cdot\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}} \leq \bar{X} \leq \overline{X}_n + z\cdot\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}\right\}$$



$$\Rightarrow \quad I_{\bar{X}} = \left[\overline{X}_n - z\cdot\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}},\ \overline{X}_n + z\cdot\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}\right], \quad \text{где} \quad F_{X^*}(z) = 1-\frac{\alpha}{2}.$$

Ово је интервал поверења величине $1-\alpha$ за ср. вр. обележја популације.

Проблем: σ је параметар популације и непознат је. Зато га оцењујемо са σ_n:

$$I_{\bar{X}} = \left[\overline{X}_n - z\cdot\frac{\sigma_n}{\sqrt{n}}\sqrt{1-\frac{n}{N}},\ \overline{X}_n + z\cdot\frac{\sigma_n}{\sqrt{n}}\sqrt{1-\frac{n}{N}}\right], \quad \text{где} \quad F_{X^*}(z) = 1-\frac{\alpha}{2}. \qquad (*)$$

Ово је апроксимативни интервал поверења за ср. вр. обележја популације.

За $n < 30$: користимо Студентову t_{n-1} расподелу. ($n-1$ степена слободе)

$$I_{\bar{X}} = \left[\overline{X}_n - t\cdot\frac{\sigma_n}{\sqrt{n}}\sqrt{1-\frac{n}{N}},\ \overline{X}_n + t\cdot\frac{\sigma_n}{\sqrt{n}}\sqrt{1-\frac{n}{N}}\right], \quad \text{где} \quad F_{t_{n-1}}(t) = 1-\frac{\alpha}{2}.$$

Интервал поверења за t

Користимо чињеницу да је $\hat{t} = N \cdot \bar{x}_n$.
Зато интервал (*) множимо са N: $I_{N\cdot\bar{x}} = \left[N\cdot\bar{x}_n - z\cdot \frac{N s_n}{\sqrt{n}} \sqrt{1-\frac{n}{N}} ,\ N\cdot\bar{x}_n + z\cdot \frac{N s_n}{\sqrt{n}} \sqrt{1-\frac{n}{N}} \right]$,

$$\Rightarrow I_t = \left[\hat{t} - z\cdot \frac{N s_n}{\sqrt{n}} \sqrt{1-\frac{n}{N}} ,\ \hat{t} + z\cdot \frac{N s_n}{\sqrt{n}} \sqrt{1-\frac{n}{N}} \right]$$

При томе, за $n \geq 30$: $F_{X^*}(z) = 1-\frac{\alpha}{2}$;

за $n < 30$: $F_{t_{n-1}}(t) = 1-\frac{\alpha}{2}$.

Напомена: исти резултат бисмо добили истим поступком као за $\bar{x}$, само крећемо од $\dfrac{\hat{t}-t}{\frac{s}{\sqrt{n}}\sqrt{N(N-n)}} \approx \mathcal{N}(0,1)$.

Интервал поверења за p

Аналогно почетном поступку, добија се:

$$I_p = \left[p_n - z\cdot\sqrt{\frac{p_n(1-p_n)}{n-1}\left(1-\frac{n}{N}\right)} - \frac{1}{2n} ,\ p_n + z\cdot\sqrt{\frac{p_n(1-p_n)}{n-1}\left(1-\frac{n}{N}\right)} + \frac{1}{2n} \right].$$

– корекција (x_n – непрекидно, p_n дискретно)

Множењем I_p са N добијамо интервал поверења за A:

$$I_A = \left[\hat{A} - z\cdot\sqrt{\frac{p_n(1-p_n)}{n-1}(N-n)} - \frac{N}{2n} ,\ \hat{A} + z\cdot\sqrt{\frac{p_n(1-p_n)}{n-1}(N-n)} + \frac{N}{2n} \right].$$

6. Одређивање обима ПСУ без понављања

Ако имамо већи обим, имамо и прецизније резултате.
Са друге стране, јављају се веће неузорачке грешке, као и трошкови.

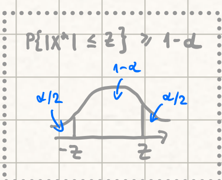
Наш циљ је да одредимо оптимално n.

Уведимо ознаке: θ: параметар који оцењујемо;
$\hat{\theta}$: оцена параметра;
d: „подношљиво" одступање $(d>0)$;
α: ризик да се премаши d (обично мало: 0.05, 0.01, 0.1).
} унапред задати

Дакле: $P\{|\hat{\theta}-\theta| > d\} \le \alpha$

Одређивање обима за оцењивање $\bar{X}$

$$P\{|\bar{X}_n - \bar{X}| > d\} \le \alpha \Rightarrow P\{|\bar{X}_n - \bar{X}| \le d\} \ge 1-\alpha \Rightarrow P\Bigg\{\underbrace{\frac{|\bar{X}_n - \bar{X}|}{\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}}}_{X^* \sim \mathcal{N}(0,1)} \le \frac{d}{\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}}\Bigg\} \ge 1-\alpha$$



$$\Rightarrow \frac{d}{\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}} \ge z, \quad \text{где } \Phi(z) = 1-\frac{\alpha}{2} \Rightarrow \frac{d}{z} \ge \frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$$

$$\Rightarrow \frac{d^2}{z^2} \ge \frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)$$

$$\Rightarrow \frac{d^2}{z^2} \ge \frac{\sigma^2}{n} - \frac{\sigma^2}{N} \Rightarrow \frac{\sigma^2}{n} \le \frac{d^2}{z^2} + \frac{\sigma^2}{N}$$

Теорема 1: $n \ge \dfrac{1}{\frac{d^2}{\sigma^2 z^2} + \frac{1}{N}}$, где $\Phi(z) = 1-\frac{\alpha}{2}$

Напомена: d, α, N су познати, z прочитамо из таблице.
Проблем: σ^2 је непознато.
Не можемо га оценити са s_n^2, јер узорак још немамо.

Решење: Узимамо узорак из сличних истраживања из прошлости или направимо прелиминарно истраживање.

Тако добијамо s_n^2.

Одређивање обима за оцењивање t

$$P\{|\hat{t}-t|>d\}\le\alpha \Rightarrow P\{|\hat{t}-t|\le d\}\ge 1-\alpha \Rightarrow P\Big\{\underbrace{\frac{|\hat{t}-t|}{\frac{sN}{\sqrt{n}}\sqrt{1-\frac{n}{N}}}}_{\sim\mathcal{N}(0,1)}\le\frac{d}{\frac{sN}{\sqrt{n}}\sqrt{1-\frac{n}{N}}}\Big\}\ge 1-\alpha$$

$$\Rightarrow \frac{d}{\frac{sN}{\sqrt{n}}\sqrt{1-\frac{n}{N}}}\ge z, \quad \text{где } \Phi(z)=1-\frac{\alpha}{2} \Rightarrow \frac{d}{z}\ge\frac{sN}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$$

$$\Rightarrow \frac{d^2}{z^2}\ge\frac{N^2s^2}{n}\cdot\frac{N-n}{N}$$

$$\Rightarrow \frac{d^2}{z^2}\ge\frac{N^2s^2}{n}-Ns^2 \Rightarrow \frac{N^2s^2}{n}\le\frac{d^2}{z^2}+Ns^2$$

Теорема 2: $n\ge\dfrac{1}{\frac{d^2}{z^2N^2s^2}+\frac{1}{N}}$, где $\Phi(z)=1-\frac{\alpha}{2}$

Напомена: Проблем непознатог s решавамо на исти начин.

Одређивање обима за оцењивање p

$$P\{|p_n-p|>d\}\le\alpha \Rightarrow P\{|p_n-p|\le d\}\ge 1-\alpha \Rightarrow P\Big\{\underbrace{\frac{|p_n-p|}{\sqrt{\frac{N-n}{N-1}\cdot\frac{p(1-p)}{n}}}}_{\sim\mathcal{N}(0,1)}\le\frac{d}{\sqrt{\frac{N-n}{N-1}\cdot\frac{p(1-p)}{n}}}\Big\}\ge 1-\alpha$$

$$\Rightarrow \frac{d}{\sqrt{\frac{N-n}{N-1}\cdot\frac{p(1-p)}{n}}}\ge z, \quad \text{где } \Phi(z)=1-\frac{\alpha}{2} \Rightarrow \frac{d}{z}\ge\sqrt{\frac{N-n}{N-1}\cdot\frac{p(1-p)}{n}}$$

$$\Rightarrow \frac{d^2}{z^2}\ge\frac{N-n}{N-1}\cdot\frac{p(1-p)}{n}$$

$$\Rightarrow \frac{d^2}{z^2}\ge\frac{N}{N-1}\cdot\frac{p(1-p)}{n}-\frac{p(1-p)}{N-1} \Rightarrow \frac{N\,p(1-p)}{(N-1)\,n}\le\frac{d^2}{z^2}+\frac{p(1-p)}{N-1}$$

Теорема 3: $n\ge\dfrac{1}{\frac{d^2}{z^2\cdot\frac{N}{N-1}p(1-p)}+\frac{1}{N}}$, где $\Phi(z)=1-\frac{\alpha}{2}$

Напомена: p нам је непознато. Може као и пре, а може и на други начин:

Искористимо да важи $p(1-p)\le 1/4$.
Мало грубља апроксимација, али олакшава рачун

Тада је: $n\ge\dfrac{1}{\frac{4d^2}{z^2\frac{N}{N-1}}+\frac{1}{N}}$, где $\Phi(z)=1-\frac{\alpha}{2}$

Овим смо завршили ПСУ.

Поставља се питање када треба користити ПСУ?

У случајевима када имамо добар узорачки оквир,
али скоро никакве додатне информације по којим бисмо јединке могли да групишемо.

Пример: Имамо списак студената целог универзитета,
али само по имену и презимену.

7. Методи избора узорка са неједнаким вероватноћама избора јединки

деф. Уводимо помоћно обележје: свакој јединки додељујемо величину M_i. $(i = \overline{1,N})$

При томе, $\sum^N M_i = M$.

деф. $p_i := \frac{M_i}{M}$ је вероватноћа избора i-те јединке у једном бирању.

Напомена: При томе, очигледно важи $\sum^N p_i = 1$.

а) са понављањем:
- узмемо, па вратимо;
- тиме почетна расподела p_i-ева остаје иста $\left(p_i = \frac{M_i}{M}\right)$

б) без понављања:
- извучена i-та $\left(p_i = \frac{M_i}{M}\right)$
- избацимо је
- извлачимо j-ту $\left(p_j = \frac{M_j}{M - M_i}\right)$
- избацимо је

⋮

Наводимо методе избора оваквог узорка:

1) Метод кумуланте:

Из таблице случ. бр.[1] бирамо број $B = \overline{1, M}$.

$B \in [1, M_1] \Rightarrow$ извукли смо јединку 1;
$B \in [M_1+1, M_1+M_2] \Rightarrow$ извукли смо јединку 2;
⋮
$B \in [M_1+\ldots+M_{k-1}+1, M_1+\ldots+M_k] \Rightarrow$ извукли смо јединку k;
⋮
$B \in [M_1+\ldots+M_{n-1}+1, \underbrace{M_1+\ldots+M_n}_{M}] \Rightarrow$ извукли смо јединку n.

Границе се зову кумуланте. $(M_1, M_1+M_2, M_1+M_2+M_3, \ldots)$

Метод је коректан: $p_i = \dfrac{M_1+\ldots+M_i - (M_1+\ldots+M_{i-1})}{M} = \dfrac{M_i}{M}$.

Напомена: Све ово је за узорак са понављањем.
За без понављања: ако извучемо неки од раније - одбацујемо га.

Мана: код великих популација, тешко рачунамо кумуланте.

Зато уводимо други метод.

2) Лахиријев метод:

Означимо са $K = \max M_i$.

Бирамо уређени пар: (i, R) – где $i \in \{1, 2, \dots, N\}$ и $R \in \{1, 2, \dots, \overset{\max M_i}{\overset{\shortparallel}{K}}\}$

Ако је $R \le M_i \Rightarrow$ i-та јединка је ушла у узорак;

$R > M_i \Rightarrow$ није ушла.

Поступак понављамо док не добијемо n јединки. (различитих, ако је без понављања)

Метод је коректан:

$p_i = \sum_{s=1}^{\infty} p_i(s)$, где је $p_i(s)$ вероватноћа да i-та јединка буде извучена у s-том покушају
($s-1$ ни једна, па s-ти покушај i-ту)

оно старо

Важи: $p_i(1) = \frac{1}{N} \cdot \frac{M_i}{K}$ (■ - морамо извући i-ту јединку од њих N; ■ - мора $R \le M_i$, а $R = \overline{1,K}$)

$$p_i(2) = \underbrace{\left(1 - \sum_{i=1}^{N} \frac{1}{N} \cdot \frac{M_i}{K}\right)}_{q} \cdot \frac{1}{N} \cdot \frac{M_i}{K}$$

(није извучена ни једна јединка) (тј. пропао покушај)

$\vdots$

Уочимо правило: $p_i(s) = q^{s-1} \cdot \frac{1}{N} \cdot \frac{M_i}{K}$.

Убацимо то:
$$p_i = \sum_{s} \left(q^{s-1} \frac{1}{N} \frac{M_i}{K}\right) \overset{\sum q^{k-1} = \frac{1}{1-q}}{=} \frac{1}{N} \frac{M_i}{K} \cdot \frac{1}{1 - \left(1 - \sum_{i=1}^{N} \frac{1}{N} \frac{M_i}{K}\right)}$$
$$= \frac{1}{N} \frac{M_i}{K} \cdot \frac{1}{\frac{1}{NK} \sum M_i}$$
$$= \frac{M_i}{M}.$$

8. Hansen - Hurwitz - ове оцене

Гледамо како изгледају оцене код јединки са неједнаким вероватноћама избора, са понављањем. Нове оцене ће у индексу имати ознаку НН.

Теорема 1: Нека је p_i вероватноћа избора i-те јединке у извлачењу код узорковања са понављањем.

1) Оцена $\hat{t}_{HH} = \frac{1}{n}\sum_{i\in S}\frac{x_i}{p_i}$ је непристрасна оцена за t;

2) За дисперзију те оцене важи: $D(\hat{t}_{HH}) = \frac{1}{n}\sum_{i=1}^{N} p_i\left(\frac{x_i}{p_i} - t\right)^2$;

3) Непристрасна оцена те дисперзије је $\widehat{D(\hat{t}_{HH})} = \frac{1}{n(n-1)}\sum_{i\in S}\left(\frac{x_i}{p_i} - \hat{t}_{HH}\right)^2$.

Доказ:

1) $E(\hat{t}_{HH}) = E\left(\frac{1}{n}\sum_{i\in S}\frac{x_i}{p_i}\right) = \frac{1}{n}E\Big(\sum_{i\in S}\underbrace{\frac{x_i}{p_i}}_{y_i}\Big) = \frac{1}{n}E\left(\sum_{i\in S} y_i\right) \overset{\substack{y_i \text{ iid} \\ \text{иста расп. као } y}}{=} \frac{1}{n}\sum_{i\in S} Ey$

$\overset{(*)}{=} \frac{1}{n}\cdot n\cdot t = t \Rightarrow$ јесте непристрасна

(*) $y: \begin{pmatrix} \frac{x_1}{p_1} & \dots & \frac{x_N}{p_N} \\ p_1 & \dots & p_N \end{pmatrix} \Rightarrow Ey = \sum^{N} x_i = t$

И имамо да $i\in S$ (узорку) (не популацији) $\Rightarrow n\cdot t$

2) $D(\hat{t}_{HH}) = D\left(\frac{1}{n}\sum_{i\in S}\frac{x_i}{p_i}\right) \overset{iid}{=} \frac{1}{n^2}\sum_{i\in S} Dy = \frac{1}{n^2}\cdot n\cdot Dy = \frac{1}{n}Dy$

$\overset{(**)\,I}{=} \frac{1}{n}\sum^{N}\left(\frac{x_i}{p_i} - t\right)^2\cdot p_i$

(**) Рачунамо на два начина:

I) $Dy = E(y - Ey)^2 = \sum\left(\frac{x_i}{p_i} - t\right)^2\cdot p_i$

II) $Dy = Ey^2 - (Ey)^2 = \sum\frac{x_i^2}{p_i} - t^2$

3) $E(\widehat{D(\hat{t}_{HH})}) = E\left[\frac{1}{n(n-1)}\sum_{i\in S}\left(\frac{x_i}{p_i} - \hat{t}_{HH}\right)^2\right] = \frac{1}{n(n-1)}E\left[\sum_{i\in S}\left(\frac{x_i}{p_i} - \frac{1}{n}\sum_{i\in S}\frac{x_i}{p_i}\right)^2\right]$

$= \frac{1}{n(n-1)}E\left[\sum_{i\in S}\left(\frac{x_i^2}{p_i^2} - 2\frac{x_i}{p_i}\frac{1}{n}\sum_{i\in S}\frac{x_i}{p_i} + \frac{1}{n^2}\left(\sum_{i\in S}\frac{x_i}{p_i}\right)^2\right)\right]$

$= \frac{1}{n(n-1)}E\left[\sum_{i\in S}\frac{x_i^2}{p_i^2} - \frac{2}{n}\left(\sum_{i\in S}\frac{x_i}{p_i}\right)^2 + \frac{n}{n^2}\left(\sum_{i\in S}\frac{x_i}{p_i}\right)^2\right] = \frac{1}{n(n-1)}E\left[\sum_{i\in S}\frac{x_i^2}{p_i^2} - \frac{1}{n}\left(\sum_{i\in S}\frac{x_i}{p_i}\right)^2\right]$

$= \frac{1}{n(n-1)}\left[\sum_{i\in S}E\left(\frac{x_i^2}{p_i^2}\right) - \frac{1}{n}E\left(\sum_{i\in S}\frac{x_i}{p_i}\right)^2\right] = \frac{1}{n(n-1)}\left[\sum_{i\in S}E(y_i^2) - \frac{1}{n}E\left(\sum_{i\in S}y_i\right)^2\right]$

$EX^2 = DX + (EX)^2$

$= \frac{1}{n(n-1)}\left[n\cdot\sum^{N}\frac{x_i^2}{p_i^2}\cdot p_i - \frac{1}{n}\left(D\left(\sum y_i\right) + E\left(\sum y_i\right)^2\right)\right]$

$\overset{iid}{=} \frac{1}{n(n-1)}\left[n\cdot\sum^{N}\frac{x_i^2}{p_i} - \frac{1}{n}\left(nDy + (nEy)^2\right)\right]$

$\overset{(*)}{=} \frac{1}{n(n-1)}\left[n\cdot\sum^{N}\frac{x_i^2}{p_i} - Dy - nt^2\right]$

$= \frac{1}{n(n-1)}\left[n\cdot\left(\sum^{N}\frac{x_i^2}{p_i} - t^2\right) - Dy\right] \overset{(**)\,II}{=} \frac{1}{n(n-1)}\left[n\cdot Dy - Dy\right]$

$= \frac{1}{n}\cdot Dy \overset{(**)\,I}{=} D(\hat{t}_{HH}) \Rightarrow$ јесте непристрасна

Теорема 2: Нека је p_i вероватноћа избора i-те јединке у извлачењу код узорковања са понављањем.

1) Оцена $\overline{X}_{HH} = \frac{1}{Nn} \sum_{i \in S} \frac{x_i}{p_i}$ је непристрасна оцена за t;

2) За дисперзију те оцене важи: $D(\overline{X}_{HH}) = \frac{1}{n} \sum_{i=1}^{N} p_i \left(\frac{x_i}{Np_i} - \overline{X} \right)^2$;

3) Непристрасна оцена те дисперзије је $\widehat{D}(\overline{X}_{HH}) = \frac{1}{n(n-1)} \sum_{i \in S} \left(\frac{x_i}{Np_i} - \overline{X}_{HH} \right)^2$.

Доказ: тривијално следи из Т1 и чињенице $\overline{X} = \frac{1}{N} \cdot t$.

Напомена: Оцене за ПСУ са понављањем су спец. случај НН.

Доказ: Користимо да је у ПСУ $p_i = \frac{1}{N}$.

$$\overline{X}_{HH} = \frac{1}{N \cdot n} \sum_{i \in S} \frac{x_i}{p_i} = \frac{1}{Nn} \sum_{i \in S} x_i \cdot N = \frac{1}{n} \sum_{i \in S} x_i = \overline{x}_n$$

Остале такође тривијално

Напомена: Предности оцене $\hat{t}_{HH}$ у односу на ПСУ са понављањем:

- није потребно знати N;
- није потребно знати све p_i, већ само за оне i које су упале у узорак.

$\boxed{9.}$ Horvitz - Thompson - ове оцене

За неједнаке вероватноће избора, без понављања, оцене нису неке + су компликоване.
Ми ћемо радити Horvitz - Thompson - ове, које могу да се користе за сваки план узорковања. (и са / и без)

Теорема 1: Нека је π_i вероватноћа избора i-те јединке.
Нека је π_{ij} $(i \neq j)$ вероватноћа избора i-те и j-те јединке.
Такође, нека $\pi_i > 0$, $\pi_{ij} > 0$. $(\forall i, j = \overline{1, N},\ i \neq j)$

1) Оцена $\hat{t}_{HT} = \sum^{\nu} \frac{x_i}{\pi_i}$ је непристрасна оцена за t;

2) За дисперзију те оцене важи: $D(\hat{t}_{HT}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} \cdot x_i^2 + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ i \neq j}}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} x_i x_j$;

3) Непристрасна оцена те дисперзије је $\widehat{D(\hat{t}_{HT})} = \sum_{i=1}^{\nu} \frac{1-\pi_i}{\pi_i^2} \cdot x_i^2 + \sum_{i=1}^{\nu} \sum_{\substack{j=1 \\ i \neq j}}^{\nu} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \cdot \frac{x_i x_j}{\pi_i \pi_j}$;

Напомена: ν је број различитих јединки у узорку.

Под $\sum^{\nu}$ мисли се да узимамо различите, што не мора да значи првих ν!

Другим речима: $\hat{t}_{HT} = \sum^{\nu} \frac{x_i}{\pi_i} = \sum_{i=1}^{N} \frac{x_i}{\pi_i} \cdot I_i$

букв. исто што и $i \in S$

ако се нека понови, због ове суме ће се рачунати само једном

$I_i : \begin{pmatrix} 0 & 1 \\ 1-\pi_i & \pi_i \end{pmatrix}$, $I_{ij} : \begin{pmatrix} 0 & 1 \\ 1-\pi_{ij} & \pi_{ij} \end{pmatrix}$

i-та јединка упала у узорак

За узорак без понављања важи $\nu = n$.

Напомена: $\widehat{D(\hat{t}_{HT})}$ има проблем: $\pi_{ij} - \pi_i \pi_j$ може бити негативно.

Тада је ова оцена бескорисна и праве се (компликованије) пристрасне модификације.

Исто ће важити у $\boxed{10}$ за $\widehat{D_{SYG}(\hat{t}_{HT})}$.

Доказ: 1) $E(\hat{t}_{HT}) = E\left(\sum_{i=1}^{N} \frac{x_i}{\pi_i}\cdot I_i\right) = \sum^{N} \frac{x_i}{\pi_i} EI_i = \sum^{N} \frac{x_i}{\pi_i}\pi_i = \sum x_i = t$ $\Rightarrow$ јесте непристрасна;

2) $$D(\hat{t}_{HT}) = D\left(\sum_{i=1}^{N} \frac{x_i}{\pi_i}\cdot I_i\right) = E\left[\left(\sum_{i=1}^{N} \frac{x_i}{\pi_i}\cdot I_i - E\left(\sum_{i=1}^{N} \frac{x_i}{\pi_i}\cdot I_i\right)\right)^2\right]$$

$$= E\left[\left(\sum^{N} \frac{x_i}{\pi_i}\cdot I_i - \sum^{N} \frac{x_i}{\pi_i} EI_i\right)^2\right] = E\left[\left(\sum_{i=1}^{N} \frac{x_i}{\pi_i}\cdot I_i - \sum_{i=1}^{N} \frac{x_i}{\pi_i}\pi_i\right)^2\right]$$

$$= E\left[\left(\sum^{N} \frac{x_i}{\pi_i}(I_i - \pi_i)\right)^2\right]$$

$$\overset{(\Delta)}{=} E\left[\sum^{N} \frac{x_i^2}{\pi_i^2}(I_i-\pi_i)^2 + \sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N} \frac{x_i}{\pi_i}\frac{x_j}{\pi_j}(I_i-\pi_i)(I_j-\pi_j)\right]$$

(Δ) $(a_1+\dots+a_N)^2 = \sum^{N} a_i^2 + \sum^{N}\sum_{j\neq i}^{N} a_i a_j$

$$= \sum^{N} \frac{x_i^2}{\pi_i^2} \underbrace{E(I_i - EI_i)^2} + \sum^{N}\sum^{N} \frac{x_i}{\pi_i}\frac{x_j}{\pi_j} E(I_iI_j - I_i\pi_j - \pi_iI_j + \pi_i\pi_j)$$

$$= \sum^{N} \frac{x_i^2}{\pi_i^2} DI_i + \sum^{N}\sum^{N} \frac{x_ix_j}{\pi_i\pi_j}\left(E(\overbrace{I_iI_j}^{I_{ij}}) - \pi_j EI_i - \pi_i EI_j + \pi_i\pi_j\right)$$

$$= \sum^{N} \frac{x_i^2}{\pi_i^2} \pi_i(1-\pi_i) + \sum^{N}\sum^{N} \frac{x_ix_j}{\pi_i\pi_j}\left(\pi_{ij} - \pi_j\pi_i - \cancel{\pi_i\pi_j} + \cancel{\pi_i\pi_j}\right)$$

$$= \sum^{N} \frac{1-\pi_i}{\pi_i}\cdot x_i^2 + \sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} x_ix_j;$$

3) $$E(\widehat{D}(\hat{t}_{HT})) = E\left[\sum_{i=1}^{\nu} \frac{1-\pi_i}{\pi_i^2}\cdot x_i^2 + \sum_{i=1}^{\nu}\sum_{\substack{j=1\\ i\neq j}}^{\nu} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\cdot\frac{x_ix_j}{\pi_i\pi_j}\right]$$

$$= E\left[\sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i^2}\cdot x_i^2\cdot I_i + \sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\cdot\frac{x_ix_j}{\pi_i\pi_j} I_iI_j\right]$$

$$= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i^2}\cdot x_i^2 \; EI_i + \sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\cdot\frac{x_ix_j}{\pi_i\pi_j} E(I_iI_j)$$

$$= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i^{\cancel{2}}}\cdot x_i^2 \,\cancel{\pi_i} + \sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\cdot\frac{x_ix_j}{\cancel{\pi_i\pi_j}}\cdot\cancel{\pi_i\pi_j}$$

$$= D(\hat{t}_{HT})$$ $\Rightarrow$ јесте непристрасна.

Теорема 2: 1) Оцена $\bar{x}_{HT} = \frac{1}{N}\sum^{\nu}\frac{x_i}{\pi_i}$ је непристрасна оцена за $\bar{x}$;

2) За дисперзију те оцене важи: $D(\bar{x}_{HT}) = \frac{1}{N^2}\sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}\cdot x_i^2 + \sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} x_ix_j$;

3) Непристрасна оцена те дисперзије је: $\widehat{D}(\bar{x}_{HT}) = \frac{1}{N^2}\left[\sum_{i=1}^{\nu}\frac{1-\pi_i}{\pi_i^2}\cdot x_i^2 + \sum_{i=1}^{\nu}\sum_{\substack{j=1\\ i\neq j}}^{\nu}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\cdot\frac{x_ix_j}{\pi_i\pi_j}\right]$;

Доказ: тривијално следи из Т1 и чињенице $\bar{x} = \frac{1}{N}\cdot t$.

10. Sen-Yates-Grundy-јеве оцене

Постоји боља оцена за $D(\hat{t}_{HT})$ од $\widehat{D(\hat{t}_{HT})}$. Доказаћемо да је и она непристрасна.
Ова оцена се користи за фиксирано $\nu = n$ (тј. код узорка без понављања, n различитих)

Лема 1: 1) $\sum_{i=1}^{N} \pi_i = n$;
2) $\sum_{\substack{j=1\\ j\neq i}} \pi_{ij} = (n-1)\pi_i$.

Доказ: 1) тривијално $\left(\sum_{i=1}^{N} \pi_i = \sum^{N} E I_i = E\left(\sum^{N} I_i\right) = E\nu = \nu = n\right)$; ($I_i$ из 9)

2) $\sum_{\substack{j=1\\ j\neq i}}^{N} \pi_{ij} = \sum_{\substack{j=1\\ j\neq i}}^{N} E(I_i I_j) = E\left(I_i \sum_{\substack{j=1\\ j\neq i}}^{N} I_j\right) = E(I_i(n-I_i))$

$= E(nI_i - I_i^2) = E(nI_i - I_i) = (n-1)EI_i = (n-1)\pi_i.$

Теорема 1: Непристрасна оцена за $D(\hat{t}_{HT})$ је: $\widehat{D_{SYG}(\hat{t}_{HT})} = \sum_{i\in S}\sum_{\substack{j\in S\\ j>i}} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^2$, $\pi_i, \pi_j, \pi_{ij} > 0$.

Доказ: Запишимо $D(\hat{t}_{HT})$ у алт. облику:

$$D(\hat{t}_{HT}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i}\cdot x_i^2 + \sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} x_i x_j = \sum_{i=1}^{N} \frac{\pi_i(1-\pi_i)}{\pi_i^2}\cdot x_i^2 - \sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_i\pi_j} x_i x_j$$

$$\overset{(*)}{=} \sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N} (\pi_i\pi_j - \pi_{ij})\left(\frac{x_i}{\pi_i}\right)^2 - \sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N} (\pi_i\pi_j - \pi_{ij})\left(\frac{x_i}{\pi_i}\cdot\frac{x_j}{\pi_j}\right)$$

$$\overset{(**)}{=} \sum_{i=1}^{N}\sum_{\substack{j=1\\ j>i}}^{N} (\pi_i\pi_j - \pi_{ij})\left(\left(\frac{x_i}{\pi_i}\right)^2 + \left(\frac{x_j}{\pi_j}\right)^2\right) - \sum_{i=1}^{N}\sum_{\substack{j=1\\ j>i}}^{N} (\pi_i\pi_j - \pi_{ij})\cdot 2\left(\frac{x_i}{\pi_i}\cdot\frac{x_j}{\pi_j}\right)$$

$$= \sum_{i=1}^{N}\sum_{\substack{j=1\\ j>i}}^{N} (\pi_i\pi_j - \pi_{ij})\left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^2$$

$(*)$ $\sum_{\substack{j=1\\ i\neq j}}^{N} (\pi_i\pi_j - \pi_{ij}) = \sum_{\substack{j=1\\ i\neq j}}^{N} \pi_{ij} - \pi_i \sum_{\substack{j=1\\ i\neq j}}^{N} \pi_j \overset{Л1}{=} (n-1)\pi_i - \pi_i(n-\pi_i) = \pi_i(1-\pi_i)$

$(**)$ нпр. $n=4$:

$\boxed{12}$	21	31	41
13	$\boxed{23}$	32	42
14	24	$\boxed{34}$	44

Остали су исти

$$E(\widehat{D_{SYG}(\hat{t}_{HT})}) = E\left[\sum_{i\in S}\sum_{\substack{j\in S\\ j>i}} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^2\right]$$

$$= \sum_{i=1}^{N}\sum_{\substack{j=1\\ j>i}}^{N} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^2 E(I_i I_j)$$

$$= \sum_{i=1}^{N}\sum_{\substack{j=1\\ j>i}}^{N} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^2 \pi_{ij} = \sum_{i=1}^{N}\sum_{\substack{j=1\\ j>i}}^{N} (\pi_i\pi_j - \pi_{ij})\left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^2 \overset{\text{алт. облик}}{=} D(\hat{t}_{HT}) \Rightarrow \text{јесте непристрасна.}$$

Теорема 2: Непристрасна оцена за $D(\bar{x}_{HT})$ је: $\widehat{D_{SYG}(\bar{x}_{HT})} = \frac{1}{N^2}\left[\sum_{i=1}^{n}\sum_{\substack{j=1\\ j>i}}^{n} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^2\right]$;

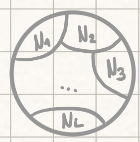
Доказ: тривијално из Т1

II део

11. Стратификован узорак.
Веза укупне оцене и оцена по стратумима

* Имамо популацију коју делимо на делове/слојеве који се зову стратуми.

У оквиру једног стратума, вредности обележја треба да буду што хомогенија, а између стратума што хетерогенија.



$N = N_1 + \dots + N_L$: популација подељена на L стратума.

Проблем: Не знамо вредност обележја. Како направити стратуме онда?
Тражи се нека особина јединки коју знамо и да је у корелацији са вредностима обележја.

(нпр. обележје: плата, особина: ниво студија)

Бр. елем. у стратуму мора бити ≥ 2.

Када направимо стратуме, вадимо узорак обима $n = n_1 + \dots + n_L$ (n_h - независни)
Тако добијамо стратификован узорак.

Стратификован случајан узорак је онај стр. уз. код ког се узорци из сваког стратума ваде као ПСУ.

Напомена: Са страт. уз. се добијају боље оцене него код ПСУ, јер имамо додатне информације.

Постављају се питања:
- на колико стратума поделити популацију, тј. које L узети?
- коју особину изабрати за поделу?
- како изабрати n_h-ове?

* деф. N - обим популације;

n - обим узорка

L - број стратума;

N_h - обим h-тог стратума;

n_h - обим узорка из h-тог стратума.

деф. Обележја јединки h-тог стратума: x_{h_i}, $i \in \{1, \dots, N_h\}$;

деф. Средња вредност обележја на h-том стратуму: $\bar{X}_h = \dfrac{\sum^{N_h} x_{h_i}}{N_h}$;

деф. Укупна сума обележја на h-том стратуму: $t_h = \sum^{N_h} x_{h_i}$;

деф. Пропорција јединки стратума са неким својством на h-том стратуму: $p_h = \dfrac{A_h}{N_h}$;

деф. Дисперзија на h-том стратуму: $s_h^2 = \dfrac{1}{N_h - 1} \sum_i^{N_h} (x_{h_i} - \bar{X}_h)^2$.

деф. Узорачка средина на h-том стратуму: $\bar{x}_{n_h} = \dfrac{\sum_{i \in S_h} x_{h_i}}{n_h}$;

деф. Оцена укупне суме на h-том стратуму: $\hat{t}_h = \sum^{n_h} x_{h_i}$;

деф. Пропорција јединки стратума са својством p_h на узорку из h-тог стратума: $p_{n_h} = \dfrac{a_{n_h}}{n_h}$.

деф. Оцена дисперзије на h-том стратуму: $s_{n_h}^2 = \dfrac{1}{n_h - 1} \sum_i^{n_h} (x_{h_i} - \bar{x}_{n_h})^2$.

Следеће две теореме су „опште": важе за сваки стратификован узорак (не само случајни).

Теорема 1: Нека је у сваком стратуму $\hat{\tau}_h$ непристрасна оцена за t_h.
и нека је у сваком стратуму $\widehat{D(\hat{\tau}_h)}$ непристрасна оцена за $D(\hat{\tau}_h)$.

1) Оцена $\hat{\tau} = \sum_h^L \hat{\tau}_h$ је непристрасна оцена за t;

2) За дисперзију те оцене важи: $D(\hat{\tau}) = \sum_h^L D(\hat{\tau}_h)$;

3) Непристрасна оцена те дисперзије је $\widehat{D(\hat{\tau})} = \sum_h^L \widehat{D(\hat{\tau}_h)}$.

Доказ: 1) $E(\hat{\tau}) = E\left(\sum^L \hat{\tau}_h\right) = \sum^L E(\hat{\tau}_h) = \sum^L t_h = t \Rightarrow$ јесте непристрасна;

2) $D(\hat{\tau}) = D\left(\sum^L \hat{\tau}_h\right) \overset{\text{нез.}}{=} \sum^L D(\hat{\tau}_h)$;

3) $E(\widehat{D(\hat{\tau})}) = E\left(\sum^L \widehat{D(\hat{\tau}_h)}\right) = \sum^L E(\widehat{D(\hat{\tau}_h)}) = \sum^L D(\hat{\tau}_h) = D(\hat{\tau}) \Rightarrow$ јесте непристрасна.

Теорема 2: Нека је у сваком стратуму $\hat{x}_h$ непристрасна оцена за $\bar{x}_h$
и нека је у сваком стратуму $\widehat{D(\hat{x}_h)}$ непристрасна оцена за $D(\hat{x}_h)$.

1) Оцена $\hat{x} = \frac{1}{N}\sum_h^L N_h \hat{x}_h$ је непристрасна оцена за $\bar{x}$;

2) За дисперзију те оцене важи: $D(\hat{x}) = \frac{1}{N^2}\sum_h^L N_h^2 D(\hat{x}_h)$;

3) Непристрасна оцена те дисперзије је $\widehat{D(\hat{x})} = \frac{1}{N^2}\sum_h^L N_h^2 \widehat{D(\hat{x}_h)}$;

Доказ: 1) $E(\hat{x}) = E\left(\frac{1}{N}\sum^L N_h \hat{x}_h\right) = \frac{1}{N}\sum^L N_h E\hat{x}_h = \frac{1}{N}\sum^L N_h \bar{x}_h = \frac{1}{N}\sum^L t_h = \frac{t}{N} = \bar{x} \Rightarrow$ јесте непристрасна;

2) $D(\hat{x}) = D\left(\frac{1}{N}\sum^L N_h \hat{x}_h\right) \overset{\text{нез.}}{=} \frac{1}{N^2}\sum^L N_h^2 D(\hat{x}_h)$;

3) $E(\widehat{D(\hat{x})}) = E\left(\frac{1}{N^2}\sum^L N_h^2 \widehat{D(\hat{x}_h)}\right) = \frac{1}{N^2}\sum^L N_h^2 E(\widehat{D(\hat{x}_h)}) = \frac{1}{N^2}\sum^L N_h^2 D(\hat{x}_h) = D(\hat{x}) \Rightarrow$ јесте непристрасна.

12. Стратификован случајан узорак са и без понављања

Сва тврђења у овом питању су последица / спец. случ. претходне две „опште" теореме.

Следеће три теореме важе за узорак без понављања.

Теорема 1: Ако се из сваког стратума вади ПСУ без понављања:

1) Оцена $\hat{t}_{STR} = \sum_h^L N_h \cdot \bar{x}_{n_h}$ је непристрасна оцена за t;

2) За дисперзију те оцене важи: $D(\hat{t}_{STR}) = \sum_h^L \frac{N_h^2 \cdot s_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)$;

3) Непристрасна оцена те дисперзије је $\widehat{D(\hat{t}_{STR})} = \sum_h^L \frac{N_h^2 \cdot s_{n_h}^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)$.

Теорема 2: Ако се из сваког стратума вади ПСУ без понављања:

1) Оцена $\bar{x}_{STR} = \frac{1}{N}\sum_h^L N_h \cdot \bar{x}_{n_h}$ је непристрасна оцена за $\bar{x}$;

2) За дисперзију те оцене важи: $D(\bar{x}_{STR}) = \frac{1}{N^2}\sum_h^L \frac{N_h^2 \cdot s_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)$;

3) Непристрасна оцена те дисперзије је $\widehat{D(\bar{x}_{STR})} = \frac{1}{N^2}\sum_h^L \frac{N_h^2 \cdot s_{n_h}^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)$.

Теорема 3: Ако се из сваког стратума вади ПСУ без понављања:

1) Оцена $\hat{p}_{STR} = \frac{1}{N}\sum_h^L N_h \cdot p_{n_h}$ је непристрасна оцена за p;

2) За дисперзију те оцене важи: $D(\hat{p}_{STR}) = \frac{1}{N^2}\sum_h^L \frac{N_h^2 (N_h - n_h)}{n_h (N_h - 1)} p_h (1 - p_h)$;

3) Непристрасна оцена те дисперзије је $\widehat{D(\hat{p}_{STR})} = \frac{1}{N^2}\sum_h^L \frac{N_h (N_h - n_h)}{n_h - 1} p_{n_h}(1 - p_{n_h})$.

Следеће три теореме важе за узорак са понављањем.

Теорема 4: Ако се из сваког стратума вади ПСУ са понављањем:

1) Оцена $\hat{t}_{STR} = \sum_{h}^{L} N_h \cdot \bar{x}_{n_h}$ је непристрасна оцена за t;

2) За дисперзију те оцене важи: $D(\hat{t}_{STR}) = \sum_{h}^{L} \frac{N_h(N_h-1)}{n_h} \cdot s_h^2$;

3) Непристрасна оцена те дисперзије је $\widehat{D}(\hat{t}_{STR}) = \sum_{h}^{L} N_h^2 \cdot \frac{s_{n_h}^2}{n_h}$.

Теорема 5: Ако се из сваког стратума вади ПСУ са понављањем:

1) Оцена $\bar{x}_{STR} = \frac{1}{N} \sum_{h}^{L} N_h \cdot \bar{x}_{n_h}$ је непристрасна оцена за $\bar{x}$;

2) За дисперзију те оцене важи: $D(\bar{x}_{STR}) = \frac{1}{N^2} \sum_{h}^{L} \frac{N_h(N_h-1)}{n_h} \cdot s_h^2$;

3) Непристрасна оцена те дисперзије је $\widehat{D(\bar{x}_{STR})} = \frac{1}{N^2} \sum_{h}^{L} N_h^2 \cdot \frac{s_{n_h}^2}{n_h}$.

Теорема 6: Ако се из сваког стратума вади ПСУ са понављањем:

1) Оцена $\hat{p}_{STR} = \frac{1}{N} \sum_{h}^{L} N_h \cdot p_{n_h}$ је непристрасна оцена за p;

2) За дисперзију те оцене важи: $D(\hat{p}_{STR}) = \frac{1}{N^2} \sum_{h}^{L} N_h^2 \cdot \frac{p_h(1-p_h)}{n_h}$;

3) Непристрасна оцена те дисперзије је $\widehat{D(\hat{p}_{STR})} = \frac{1}{N^2} \sum_{h}^{L} N_h^2 \cdot \frac{p_{n_h}(1-p_{n_h})}{n_h - 1}$.

Напомена: У наставку подразумевамо ПСУ без понављања, осим ако не нагласимо другачије.

13. Пропорционални избор обима узорка по стратумима

У наредна три питања гледамо избор обима по стратумима.

* За почетак, гледамо случај када су n_h-ови пропорционални N_h-овима: $\frac{n_1}{N_1} = \ldots = \frac{n_L}{N_L} = \frac{n}{N} \Rightarrow n_h = \frac{n}{N} \cdot N_h$

Означимо са $\hat{t}_{PROP}$ специјални случај за $\hat{t}_{STR}$ под овим условима.

Од две оцене, боља је она са мањим MSE.
Како су и $\hat{t}_{PROP}$ и $\hat{t}_{псу}$ непристрасне, довољно је да им упоредимо дисперзије.

$$D(\hat{t}_{PROP}) \overset{12\,T1.2}{=} \sum_h^L N_h^2 \cdot \frac{s_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right) \overset{n_h}{=} \sum_h^L \frac{N}{n} N_h \cdot s_h^2 \left(1-\frac{n}{N}\right) = \frac{N}{n}\left(1-\frac{n}{N}\right)\sum_h^L N_h s_h^2 ;$$

$$D(\hat{t}) \overset{11\,T2.2}{=} \frac{N^2 s^2}{n}\left(1-\frac{n}{N}\right) = \frac{N}{n}\left(1-\frac{n}{N}\right) \cdot N \cdot \frac{1}{N-1}\sum^N (x_i - \bar{x})$$

$$\overset{(*)}{=} \frac{N}{n}\left(1-\frac{n}{N}\right)\frac{N}{N-1}\sum_h^L \sum_i^{N_h}(x_{h_i}-\bar{x})^2$$

(*) Уместо да бројимо редом од x_1 до x_n, пребројавамо све у првом стратуму, па све у другом, па у трећем итд.

$$= \frac{N}{n}\left(1-\frac{n}{N}\right)\frac{N}{N-1}\sum_h^L \sum_i^{N_h}(\underbrace{x_{h_i}-\bar{x}_h}+\underbrace{\bar{x}_h-\bar{x}})^2$$

$$= \frac{N}{n}\left(1-\frac{n}{N}\right)\frac{N}{N-1}\sum_h^L \sum_i^{N_h}\left[(x_{h_i}-\bar{x}_h)^2 + 2(x_{h_i}-\bar{x}_h)(\bar{x}_h+\bar{x}) + (\bar{x}_h-\bar{x})^2\right]$$

$$= \frac{N}{n}\left(1-\frac{n}{N}\right)\frac{N}{N-1}\left[\sum_h^L \underline{(N_h-1)s_h^2} + 2\sum_h^L\left(\underline{(\bar{x}_h-\bar{x})}\sum_i^{N_h}(x_{h_i}-\bar{x}_h)\right) + \sum_h^L N_h \cdot (\bar{x}_h-\bar{x})^2\right]$$

$$= \frac{N}{n}\left(1-\frac{n}{N}\right)\frac{N}{N-1}\left[\sum_h^L (N_h-1)s_h^2 + 2\sum_h^L\left((\bar{x}_h-\bar{x})\cancel{(N_h\cdot\bar{x}_h - N_h\cdot\bar{x}_h)}^{0}\right) + \sum_h^L N_h \cdot (\bar{x}_h-\bar{x})^2\right]$$

$$= \frac{N}{n}\left(1-\frac{n}{N}\right)\frac{N}{N-1}\left[\sum_h^L (N_h-1)s_h^2 + \sum_h^L N_h \cdot (\bar{x}_h-\bar{x})^2\right]$$

$$= \frac{N}{n}\left(1-\frac{n}{N}\right)\left(1+\frac{1}{N-1}\right)\left[\sum_h^L N_h s_h^2 - \sum_h^L s_h^2 + \sum_h^L N_h(\bar{x}_h-\bar{x})^2\right]$$

$$= \frac{N}{n}\left(1-\frac{n}{N}\right)\left[\sum_h^L N_h s_h^2 - \sum_h^L s_h^2 + \frac{1}{N-1}\sum_h^L (N_h-1)s_h^2 + \frac{N}{N-1}\sum_h^L N_h(\bar{x}_h-\bar{x})^2\right]$$

$$= \frac{N}{n}\left(1-\frac{n}{N}\right)\sum_h^L N_h s_h^2 + \frac{N}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left[-\underline{(N-1)}\sum_h^L s_h^2 + \sum_h^L (N_h-1)s_h^2 + N\sum_h^L N_h(\bar{x}_h-\bar{x})^2\right]$$

$$= D(\hat{t}_{PROP}) + \frac{N}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left[N\sum_h^L N_h(\bar{x}_h-\bar{x})^2 - \sum_h^L (N-N_h)s_h^2\right]$$

Дакле: $D(\hat{t}_{PROP}) \leq D(\hat{t})$ акко $N\sum_{h}^{L} N_h(\bar{x}_h - \bar{x})^2 \geq \sum_{h}^{L}(N-N_h)s_h^2 \xrightarrow{/:N} \sum_{h}^{L} \underset{\uparrow \text{ што веће}}{N_h}\underbrace{(\bar{x}_h - \bar{x})^2}_{\text{што веће}} \geq \sum_{h}^{L}(1-\overbrace{\tfrac{N_h}{N}}^{\text{фикс.}})\underbrace{s_h^2}_{\text{што мање}}$

Закључак:

$\hat{t}_{PROP}$ ће бити боља од $\hat{t}$ када је:

→ дисперзија у сваком стратуму што мања (што хомогенији у оквиру стратума)

→ ср. вр. на стратумима што различитије од ср. вр. целе популације (што хетерогеније између стратума)

→ N_h да буде што веће.

14. Оптимални избор обима узорка по стратумима

Када су дисперзије стратума s_h^2 приближно једнаке, пропорц. избор n_h-ова ради добар посао.
Ако ипак s_h^2-ови варирају – користимо тзв. оптимални избор – тиме се сада бавимо.

* Видели смо шта се дешава ако су величине узорака из сваког стратума пропорционалне.
Сада нас занима: колики узорак узети из сваког од стратума како би оцена била што боља?

За непристрасне оцене важи да су боље ако имају мању дисперзију.
Имамо проблем условног екстремума: за унапред задате трошкове, тражимо најмању дисперзију. (самим тим и најбољу оцену)

(*) Подсетимо се: $D(\hat{t}_{STR}) \overset{[12]\ T1.2}{=} \sum_h^L \frac{N_h^2}{n_h} \cdot s_h^2 \left(1 - \frac{n_h}{N_h}\right) = \sum_h^L \frac{N_h^2}{n_h} \cdot s_h^2 - \sum_h^L N_h \cdot s_h^2$

деф. (**) $C = c_0 + \sum_{h=1}^L c_h \cdot n_h$, где: C – укупни трошкови;
c_0 – фиксни трошкови узорковања;

c_h – трошкови по јединки h-тог стратума, променљиво – познато;
n_h – обим узорка из h-тог стратум – непознато.

Дакле: тражимо минимум дисперзије, тј. функције (*) уз услов о трошковима (**).

То радимо тако што правимо Лагранжову ф-ју за тражење условног екстремума за ф-ју више променљивих.
При томе, тражимо непознате $n_1, \ldots, n_L$.

$$L(n_1, \dots, n_L, \lambda) = \sum_h^L \frac{N_h^2}{n_h} \cdot S_h^2 - \sum_h^L N_h \cdot S_h^2 + \lambda\left(\sum_h^L c_h \cdot n_h + c_0 - C\right)$$

$$= \sum_h^L \frac{N_h^2}{n_h} \cdot S_h^2 \left(1 - \frac{n_h}{N_h}\right) + \lambda\left(\sum_h^L c_h \cdot n_h + c_0 - C\right)$$

(ф-ја чији се минимум тражи $+ \ \lambda \cdot$ услов)

$$\left.\begin{aligned} \frac{\partial L}{\partial n_h} &= -\frac{N_h^2 S_h^2}{n_h^2} + \lambda c_h = 0, \quad h = 1, \dots, L \\ \frac{\partial L}{\partial \lambda} &= \sum_h^L c_h n_h + c_0 - C = 0 \end{aligned}\right\} \text{ систем } (L+1) \text{ једначина}$$

$$\lambda c_h = \frac{N_h^2 S_h^2}{n_h^2} \Rightarrow n_h = \frac{N_h S_h}{\sqrt{\lambda}\sqrt{c_h}}$$

$$\sum_h^L c_h n_h = \sum_h^L c_h \frac{N_h S_h}{\sqrt{\lambda}\sqrt{c_h}} = C - c_0 \Rightarrow \frac{1}{\sqrt{\lambda}} \cdot \sum^L \sqrt{c_h} N_h S_h = C - c_0 \Rightarrow \sqrt{\lambda} = \frac{\sum^L \sqrt{c_h} N_h S_h}{C - c_0}$$

бројач ће бити k
пошто је h заузето

$$\Rightarrow n_h = \frac{N_h S_h}{\frac{\sum_k^L \sqrt{c_k} N_k S_k}{C - c_0}\sqrt{c_h}} \Rightarrow \boxed{n_h = \frac{(C - c_0)\frac{N_h S_h}{\sqrt{c_h}}}{\sum_k^L \sqrt{c_k} N_k S_k}} \xRightarrow{n = \sum^L n_h} n = \frac{(C - c_0)\sum_h^L \frac{N_h S_h}{\sqrt{c_h}}}{\sum_h^L \sqrt{c_h} N_h S_h}$$

Ова стационарна тачка је кандидат за минимум. То морамо да проверимо.
Зато рачунамо и друге парцијалне изводе (који иду у матрицу Φ):

$$\left.\begin{aligned} \frac{\partial^2 L}{\partial n_h^2} &= \frac{2 N_h^2 S_h^2}{n_h^3} \\ \frac{\partial^2 L}{\partial n_i \partial n_j} &= 0, \quad i \neq j \end{aligned}\right\} \Rightarrow \Phi = \begin{bmatrix} \frac{2 N_1^2 S_1^2}{n_1^3} & & 0 \\ & \ddots & \\ 0 & & \frac{2 N_L^2 S_L^2}{n_L^2} \end{bmatrix}$$

(Φ поз. дефинитна $\Rightarrow$ минимум;
Φ нег. дефинитна $\Rightarrow$ максимум.)

Како је Φ позитивно дефинитна $\Rightarrow$ у тој тачки јесте минимум.

Напомена: Може се десити $n_h > N_h$.

Тада се за n_h проглашава баш $n_h = N_h$, а онда поновимо процес, али без тог стратума.

15. Neyman-ов избор обима узорка по стратумима

* Посматрамо спец. случај оптималног узорковања, када је $c_1 = \dots = c_L = c$.
  Такав избор зове се Нејманов избор узорка.

Изводимо цео поступак поново.

$C = c_0 + \sum_h^L c_h n_h$, па кад наметнемо услов: $C = c_0 + c \cdot n$.

Како су трошкови унапред познати, то значи и да ће n бити познато (тј. фиксно).

Опет тражимо минимум по n_h за $D(\hat{t}_{STR}) = \sum_h^L \frac{N_h^2}{n_h} s_h^2 \left(1 - \frac{n_h}{N_h}\right)$, уз услов $n_1 + \dots + n_L = n$.

$$L(n_1, \dots, n_L, \lambda) = \sum_h^L \frac{N_h^2}{n_h} s_h^2 \left(1 - \frac{n_h}{N_h}\right) + \lambda(n_1 + \dots + n_L - n) = \sum_h^L \frac{N_h^2}{n_h} s_h^2 - \sum_h^L N_h \cdot s_h^2 + \lambda(n_1 + \dots + n_L - n)$$

$$\frac{\partial L}{\partial n_h} = -\frac{N_h^2 s_h^2}{n_h^2} + \lambda = 0, \quad h = 1, \dots, L \quad \Rightarrow \quad n_h = \frac{N_h s_h}{\sqrt{\lambda}};$$

$$\frac{\partial L}{\partial \lambda} = n_1 + \dots + n_L - n = 0 \quad \Rightarrow \quad n = \sum_h^L \frac{N_h s_h}{\sqrt{\lambda}} \quad \Rightarrow \quad \sqrt{\lambda} = \frac{\sum^L N_h s_h}{n} \quad \Rightarrow \quad \boxed{n_h = \frac{N_h s_h}{\sum_k^L N_k s_k} \cdot n}$$

Нашли смо стационарну тачку (n_h), па проверавамо да ли је минимум:

$$\left.\begin{aligned} \frac{\partial^2 L}{\partial n_h^2} &= \frac{2 N_h^2 s_h^2}{n_h^3} \\ \frac{\partial^2 L}{\partial n_i \partial n_j} &= 0, \quad i \neq j \end{aligned}\right\} \Rightarrow \quad \phi = \begin{bmatrix} \frac{2N_1^2 s_1^2}{n_1^3} & & 0 \\ & \ddots & \\ 0 & & \frac{2N_L^2 s_L^2}{n_L^2} \end{bmatrix}$$

Како је ϕ позитивно дефинитна $\Rightarrow$ у тој тачки јесте минимум.

Када то уврстимо: $D(\hat{t}_{STR}) = \sum_h^L \frac{N_h^2}{n_h} s_h^2 - \sum_h^L N_h \cdot s_h^2 = \sum^L \frac{N_h^2 s_h^2}{\frac{N_h s_h}{\sum_k^L N_k s_k} n} - \sum_h^L N_h \cdot s_h^2$

$$D(\hat{t}_{NEY}) = \frac{1}{n}\left(\sum_h^L N_h s_h\right)^2 - \sum_h^L N_h s_h^2.$$

* Проблем: Да бисмо знали колики узорак бирамо, морамо да знамо s_h - ове.

Решење: Исто као онај пут: нпр. прелиминарно истраживање.

* Како се све ове оцене пореде међусобно?

$\hat{t}^{псу}$, $\hat{t}_{PROP}$ $\boxed{13}$, $\hat{t}_{NEY}$ $\boxed{15}$ могу да се пореде, зато што је n унапред познато.

Оцена добијена оптималним узорковањем $\boxed{14}$ не може да се пореди са њима, јер n није познато.

Лема 1: $D(\hat{t}_{NEY}) \leq D(\hat{t}_{PROP})$.

Доказ:

$$D(\hat{t}_{PROP}) - D(\hat{t}_{NEY}) = \frac{N-\cancel{n}}{n}\sum_h^L N_h s_h^2 - \frac{1}{n}\left(\sum_h^L N_h s_h\right)^2 + \cancel{\sum_h^L N_h s_h^2}$$

$$= \frac{N}{n}\sum_h^L N_h s_h^2 - \frac{N}{n}\cdot\frac{1}{N}\left(\sum_h^L N_h s_h\right)^2$$

$$\frac{N}{N^2} = \frac{\sum N_h}{N^2}$$

$$= \frac{N}{n}\left[\sum_h^L N_h s_h^2 - \frac{2}{N}\left(\sum_h^L N_h s_h\right)^2 + \frac{1}{N}\left(\sum_h^L N_h s_h\right)^2\right]$$

$$= \frac{N}{n}\left[\sum_h^L N_h s_h^2 - \frac{2}{N}\left(\sum_h^L N_h s_h\right)\left(\sum_k^L N_k s_k\right) + \frac{\sum^L N_h}{N^2}\left(\sum_k^L N_k s_k\right)^2\right]$$

$$= \frac{N}{n}\sum_h^L N_h\left[s_h^2 - 2s_h\cdot\frac{1}{N}\sum_k^L N_k s_k + \left(\frac{1}{N}\sum_k^L N_k s_k\right)^2\right]$$

$$= \frac{N}{n}\sum_h^L N_h\left[s_h - \frac{1}{N}\sum_k^L N_k s_k\right]^2 \geq 0$$

Закључак: $D(\hat{t}_{NEY}) \underset{\text{увек}}{\leq} D(\hat{t}_{PROP}) \underset{\text{углавном}}{\overset{\boxed{13}}{\leq}} D(\hat{t})$ (хомог. + хетерог.)

III neo

Количничко оцењивање

* За почетак, објаснимо шта је количничко оцењивање.

Већ код стратификовања смо имали неку помоћну променљиву / обележје.

Боље оцене можемо добити на два начина:

а) узорковање вршимо на основу те помоћне променљиве – дакле, стратификован узорак;

б) на основу те помоћне променљиве правимо боље оцене у оквиру ПСУ.
Ово се зове количничко оцењивање и линеарно-регресионо оцењивање.

Код количничког оцењивања, имамо главно обележје x и помоћно обележје y.

Оцена на основу количничког оцењивања ће бити боља ако:

→ x, y су у корелацији;

→ ако из $y=0$ следи и $x=0$.

Пример: x – бр. животиња у шуми
y – површина шуме

Контра пример: x – принос житарице
y – загађеност ваздуха

Јесу у корелацији, али ако је $y=0$, принос расте.

21. Количничко оцењивање на основу ПСУ без понављања

Имамо узорак: $(x_1, y_1), \dots, (x_n, y_n)$ обима n.

Напомена: За количничко оцењивање, увек морамо да знамо колико је $t_y = \sum^N y_i$

деф. **Количник популације** је $R := \dfrac{\sum^N x_i}{\sum y_i} = \dfrac{t}{t_y} \overset{/:N}{=} \dfrac{\bar{X}}{\bar{y}}$

деф. Количничка оцена за R је **узорачки количник** $\hat{R} := \dfrac{\sum_{i\in S} x_i}{\sum_{i\in S} y_i} = \dfrac{\bar{x}_n}{\bar{y}_n}$

деф. Количничка оцена за t је $\hat{t}_R = \hat{R}\cdot t_y$. ($t_y$ је познато)

Количничка оцена за $\bar{x}$ је $\bar{x}_R = \hat{R}\cdot\bar{y}$. ($\bar{y} = \frac{t_y}{N}$ је познато)

Када се користи количничко оцењивање?

- за оцењивање количника R;
- за оцењивање t када N није познато, али t_y јесте познато;
- некад добијемо боље оцене за t и $\bar{x}$. 22

* Испитујемо пристрасност ових оцена. Испоставиће се да су асимптотски непристрасне.
Због тога, за велико n, за меру њиховог квалитета се опет користи дисперзија, уместо MSE.

Теорема 1: $\hat{R}$ је асимптотски непристрасна.

за велико n: $\bar{y}_n \approx \bar{y}$

Доказ: $\hat{R} - R = \dfrac{\bar{x}_n}{\bar{y}_n} - R = \dfrac{\bar{x}_n - \bar{y}_n\cdot R}{\bar{y}_n} \approx \dfrac{\bar{x}_n - R\cdot\bar{y}_n}{\bar{y}}$

$$E(\hat{R}-R) \approx E\left(\frac{\bar{x}_n - R\cdot\bar{y}_n}{\bar{y}}\right) = \frac{1}{\bar{y}}E(\bar{x}_n - R\bar{y}_n) = \frac{1}{\bar{y}}\left[E(\bar{x}_n) - R\cdot E(\bar{y}_n)\right] \overset{\text{псу}}{=} \frac{\bar{x} - R\bar{y}}{\bar{y}} = 0$$

$\Rightarrow E(\hat{R}) \approx R.$

Последица: $\hat{t}_R$ и $\bar{x}_R$ су асимптотски непристрасне оцене. (јер $t_y = const$)

Теорема 2: За велико n, важе следеће апроксимације:

$\cdot\, t_y^2 = (N\cdot\bar{y})^2$

1) $$\mathrm{MSE}(\hat{R}) \approx D(\hat{R}) = \frac{\left(1-\frac{n}{N}\right)}{n\bar{y}^2}\cdot\frac{1}{N-1}\cdot\sum^{N}(x_i - Ry_i)^2;$$

2) $$\mathrm{MSE}(\hat{t}_R) \approx D(\hat{t}_R) = \frac{N^2\left(1-\frac{n}{N}\right)}{n}\cdot\frac{1}{N-1}\cdot\sum^{N}(x_i - Ry_i)^2;$$ (алтернативни облик у $\boxed{22}$)

$\cdot\,\frac{1}{N^2}$

3) $$\mathrm{MSE}(\bar{x}_R) \approx D(\bar{x}_R) = \frac{\left(1-\frac{n}{N}\right)}{n}\cdot\frac{1}{N-1}\cdot\sum^{N}(x_i - Ry_i)^2.$$

Доказ: 1) Због Т1, јасно је да важи $\mathrm{MSE} \approx D$.

ознака

$$\bar{x}_n - R\cdot\bar{y}_n = \frac{1}{n}\sum_{i\in S} x_i - R\cdot\frac{1}{n}\sum_{i\in S} y_i = \frac{1}{n}\sum_{i\in S}\underbrace{(x_i - R\cdot y_i)}_{d_i} = \bar{d}_n$$

$$\mathrm{MSE}(\hat{R}) = E(\hat{R}-R)^2 \approx E\left(\frac{\bar{x}_n - R\cdot\bar{y}_n}{\bar{y}}\right)^2 = \frac{1}{\bar{y}^2}\left[E(\bar{x}_n - R\cdot\bar{y}_n)^2\right]$$

$$= \frac{1}{\bar{y}^2}\Big[E(\bar{x}_n - R\cdot\bar{y}_n)^2 - \underbrace{[E(\bar{x}_n - R\cdot\bar{y}_n)]^2}_{=0,\ \text{па може}}\Big] = \frac{1}{\bar{y}^2}D(\bar{x}_n - R\cdot\bar{y}_n) = \frac{1}{\bar{y}^2}D(\bar{d}_n)$$

$$\overset{\boxed{1}\,T1}{=} \frac{1}{\bar{y}^2}\left(1-\frac{n}{N}\right)\frac{\boxed{s_d^2}}{n} = \frac{1}{\bar{y}^2}\left(1-\frac{n}{N}\right)\cdot\frac{1}{n}\cdot\frac{1}{N-1}\sum^{N}(d_i - \bar{d})^2 = \frac{\left(1-\frac{n}{N}\right)}{n\bar{y}^2}\cdot\frac{1}{N-1}\cdot\sum^{N}(x_i - Ry_i)^2;$$

$$\bar{d} = \bar{x} - R\bar{y} = 0$$

2) тривијално (стрелица);

3) тривијално (стрелица).

Напомена: За ове дисперзије предлажемо следеће оцене:

$$\widehat{D(\hat{t}_R)} := \frac{N^2\left(1-\frac{n}{N}\right)}{n}\cdot\frac{1}{n-1}\cdot\sum_{i\in S}(x_i - \hat{R}y_i)^2;$$

$$\widehat{D(\bar{x}_R)} := \frac{\left(1-\frac{n}{N}\right)}{n}\cdot\frac{1}{n-1}\cdot\sum_{i\in S}(x_i - \hat{R}y_i)^2;$$

Ове оцене су асимптотски непристрасне. (зато што код ПСУ без пон. важи $E(s_n^2) = s^2$, па ће важити $E(s_{n_d}^2) = s_d^2$.)

$\boxed{22.}$ Упоређивање количничке оцене и оцене на основу ПСУ без понављања

* Подсетимо се: Коефицијент корелације мери лин. зав. два обележја ($|\rho| \approx 1 \Rightarrow y \approx kx + n$)

$$\rho_{X,Y} := \frac{E[(X-EX)(Y-EY)]}{\sqrt{DX} \cdot \sqrt{DY}} = \frac{\frac{1}{N}\sum^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{N-1}{N}\cdot s_x^2}\cdot\sqrt{\frac{N-1}{N}\cdot s_y^2}} = \frac{\sum^{N}(x_i-\bar{x})(y_i-\bar{y})}{(N-1)\, s_x\, s_y}.$$

$X : \begin{pmatrix} x_1 & \cdots & x_N \\ \frac{1}{N} & \cdots & \frac{1}{N} \end{pmatrix} \Rightarrow EX = \frac{1}{N}\sum^{N} x_i, \qquad EX^2 = \frac{1}{N}\sum^{N} x_i^2$

↳ псу б.п.

$$DX = EX^2 - (EX)^2 = \frac{1}{N}\sum^{N} x_i^2 - \bar{x}^2 = \frac{1}{N}\Big(\sum^{N} x_i^2 - N\bar{x}^2\Big) \overset{(***)\text{ из } \boxed{1}}{=} \frac{N-1}{N} s^2$$

* Поредимо оцене $\hat{t}_R^{\,\boxed{21}}$ и $\hat{t}^{\,\boxed{1}}$ (псу без понављања)

Како су обе непристрасне ($\hat{t}_R$ асимптотски), поредимо им дисперзије:

$$\rightarrow D(\hat{t}) \overset{\boxed{1}}{=} N^2\Big(1-\frac{n}{N}\Big)\frac{1}{n} s_x^2$$

$$\rightarrow D(\hat{t}_R) \overset{\boxed{21}}{=} \frac{N^2(1-\frac{n}{N})}{n}\cdot\frac{1}{N-1}\cdot\sum^{N}(x_i - Ry_i)^2 \overset{\substack{\text{правимо} \\ \text{алт. облик}}}{=} \frac{N^2(1-\frac{n}{N})}{n}\cdot\frac{1}{N-1}\cdot\sum^{N}\big(x_i - Ry_i - \underbrace{(\bar{x}-R\bar{y})}_{0}\big)^2$$

$$= \frac{N^2(1-\frac{n}{N})}{n}\cdot\frac{1}{N-1}\cdot\sum^{N}\big(x_i-\bar{x} - R\cdot(y_i-\bar{y})\big)^2$$

$$= \frac{N^2(1-\frac{n}{N})}{n}\cdot\frac{1}{N-1}\Big(\sum^{N}(x_i-\bar{x})^2 + \sum^{N} R^2(y_i-\bar{y})^2 - 2R\sum^{N}(x_i-\bar{x})(y_i-\bar{y})\Big)$$

$$= \frac{N^2(1-\frac{n}{N})}{n}\big(s_x^2 + R^2 s_y^2 - 2R\rho\, s_x s_y\big)$$

погледати деф. ρ

$\sum^{N}(x_i-\bar{x})(y_i-\bar{y}) = \rho\cdot(N-1)\cdot s_x\cdot s_y$

$\hat{t}_R$ боља од $\hat{t}$ $\Leftrightarrow$ $D(\hat{t}_R) - D(\hat{t}) = N^2\Big(1-\frac{n}{N}\Big)\frac{1}{n}\big(R^2 s_y^2 - 2R\rho s_x s_y\big) < 0$

$\Leftrightarrow \quad R^2 s_y^2 - 2R\rho s_x s_y < 0$

$\overset{R>0}{\Leftrightarrow} \quad \rho > \frac{1}{2} R \frac{s_y}{s_x}$

$\Leftrightarrow \quad \rho > \frac{1}{2}\cdot\frac{CV_y}{CV_x}$, $\qquad CV_x := \frac{s_x}{\bar{x}}$ је коефицијент варијације.

Закључак: $\hat{t}_R$ је боља од $\hat{t}$ када је ρ што веће, тј. што је већа лин. зав. између x и y.

23. Количничко оцењивање на основу узорка са неједнаким вероватноћама избора

Користимо Horvitz - Thompson - ове оцене.[9]

деф. За оцену R, користимо $\hat{R}_{HT} := \dfrac{\hat{t}_{HT}(x)}{\hat{t}_{HT}(y)} = \dfrac{\sum \frac{x_i}{\pi_i}}{\sum \frac{y_i}{\pi_i}}$. (писаћемо опет само $\hat{R}$)

деф. $\hat{t}_R := \hat{R}_{HT} \cdot t_y$.

$\bar{x}_R := \hat{R}_{HT} \cdot \bar{y}$.

Радимо аналогно као у [21]:

Теорема 1: $\hat{R}_{HT}$ је асимптотски непристрасна.

за велико n: $\hat{t}_{HT}(y) \approx t_y$

Доказ: $\hat{R} - R = \dfrac{\hat{t}_{HT}(x)}{\hat{t}_{HT}(y)} - R = \dfrac{\hat{t}_{HT}(x) - R\cdot\hat{t}_{HT}(y)}{\hat{t}_{HT}(y)} \approx \dfrac{\hat{t}_{HT}(x) - R\cdot\hat{t}_{HT}(y)}{t_y}$

$$E(\hat{R}-R) \approx E\left(\frac{\hat{t}_{HT}(x) - R\cdot\hat{t}_{HT}(y)}{t_y}\right) = \frac{1}{t_y}\left[E(\hat{t}_{HT}(x)) - R\cdot E(\hat{t}_{HT}(y))\right] \overset{[9]\,T4}{=} \frac{1}{t_y}(t_x - Rt_y) = 0.$$

Последица: $\hat{t}_R$ и $\bar{x}_R$ су асимптотски непристрасне оцене.

Теорема 2: За велико n, важе следеће апроксимације:

$\cdot t_y^2$

1) $$MSE(\hat{R}_{HT}) \approx D(\hat{R}_{HT}) = \frac{1}{t_y^2}\Big[\sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}\cdot(x_i - R\cdot y_i)^2 + \sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}(x_i - R\cdot y_i)(x_j - R\cdot y_j)\Big];$$

2) $$MSE(\hat{t}_R) \approx D(\hat{t}_R) = \sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}\cdot(x_i - R\cdot y_i)^2 + \sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}(x_i - R\cdot y_i)(x_j - R\cdot y_j);$$

$\cdot \frac{1}{N^2}$

3) $$MSE(\bar{X}_R) \approx D(\bar{X}_R) = \frac{1}{N^2}\Big[\sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}\cdot(x_i - R\cdot y_i)^2 + \sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}(x_i - R\cdot y_i)(x_j - R\cdot y_j)\Big].$$

Доказ: 1) Због Т1, јасно је да важи $MSE \approx D$.

$$MSE(\hat{R}) = E(\hat{R}-R)^2 \approx E\Big(\frac{\hat{t}_{HT}(x) - R\cdot\hat{t}_{HT}(y)}{t_y}\Big)^2 = \frac{1}{t_y^2}\Big[E(\hat{t}_{HT}(x) - R\hat{t}_{HT}(y))^2\Big]$$

$$= \frac{1}{t_y^2}\Big[E(\hat{t}_{HT}(x) - R\cdot\hat{t}_{HT}(y))^2 - \underbrace{\big(E(\hat{t}_{HT}(x) - R\cdot\hat{t}_{HT}(y))\big)^2}_{=0,\ \text{па може}}\Big] = \frac{1}{t_y^2}D(\hat{t}_{HT}(x) - R\hat{t}_{HT}(y))$$

$$= \frac{1}{t_y^2}D\Big(\sum^{\nu}\frac{x_i}{\pi_i} - R\cdot\sum^{\nu}\frac{y_i}{\pi_i}\Big) = \frac{1}{t_y^2}\cdot D\Big(\sum^{\nu}\frac{x_i - Ry_i}{\pi_i}\Big) = \frac{1}{t_y^2}D(\hat{t}_{HT}(x - R\cdot y))$$

$$\overset{T1}{=} \frac{1}{t_y^2}\Big[\sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}\cdot(x_i - R\cdot y_i)^2 + \sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}(x_i - R\cdot y_i)(x_j - R\cdot y_j)\Big]$$

2) тривијално (стрелица);

3) тривијално (стрелица).

Напомена: За ове дисперзије предлажемо следеће оцене:

$$\widehat{D(\hat{t}_R)} := \sum_{i=1}^{\nu}\frac{1-\pi_i}{\pi_i^2}\cdot(x_i - \hat{R}\cdot y_i)^2 + \sum_{i=1}^{\nu}\sum_{\substack{j=1\\ i\neq j}}^{\nu}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\cdot\frac{(x_i - \hat{R}\cdot y_i)(x_j - \hat{R}\cdot y_j)}{\pi_{ij}};$$

$\cdot \frac{1}{N^2}$

$$\widehat{D(\bar{X}_R)} := \frac{1}{N^2}\Big[\sum_{i=1}^{\nu}\frac{1-\pi_i}{\pi_i^2}\cdot(x_i - \hat{R}\cdot y_i)^2 + \sum_{i=1}^{\nu}\sum_{\substack{j=1\\ i\neq j}}^{\nu}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\cdot\frac{(x_i - \hat{R}\cdot y_i)(x_j - \hat{R}\cdot y_j)}{\pi_{ij}}\Big].$$

24. Количничко оцењивање на основу стратификованог узорка без понављања комбинована оцена

деф. За оцену R, користимо $\hat{R}_C := \dfrac{\hat{t}_{STR}(x)}{\hat{t}_{STR}(y)} = \dfrac{\sum^{L} N_h \cdot \bar{x}_{n_h}}{\sum^{L} N_h \cdot \bar{y}_{n_h}} = \dfrac{\bar{x}_{STR}}{\bar{y}_{STR}}$. (писаћемо опет само $\hat{R}$)

деф. $\hat{t}_{RC} := \hat{R}_c \cdot t_y$.

$\bar{x}_{RC} := \hat{R}_c \cdot \bar{y}$.

Опет исто.

Теорема 1: $\hat{R}_c$ је асимптотски непристрасна.

за велико n: $\bar{y}_{STR} \approx \bar{y}$

Доказ: $\hat{R} - R = \dfrac{\bar{x}_{STR}}{\bar{y}_{STR}} - R = \dfrac{\bar{x}_{STR} - R \cdot \bar{y}_{STR}}{\bar{y}_{STR}} \approx \dfrac{\bar{x}_{STR} - R \cdot \bar{y}_{STR}}{\bar{y}}$

$$E(\hat{R} - R) \approx E\left(\frac{\bar{x}_{STR} - R \cdot \bar{y}_{STR}}{\bar{y}}\right) = \frac{1}{\bar{y}}\left[E(\bar{x}_{STR}) - R \cdot E(\bar{y}_{STR})\right] \overset{[12]\,T2}{=} \frac{1}{\bar{y}}(\bar{x} - R \cdot \bar{y}) = 0.$$

Последица: $\hat{t}_{RC}$ и $\bar{x}_{RC}$ су асимптотски непристрасне оцене.

Теорема 2: За велико n, важе следеће апроксимације:

1) $MSE(\hat{R}_C) \approx D(\hat{R}_C) = \frac{1}{t_y^2}\left[\sum_h^L \frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)\left(s_h^2(x) + R^2 s_h^2(y) - 2R\rho_h \cdot s_h(x)s_h(y)\right)\right];$

$\cdot t_y^2$

2) $MSE(\hat{t}_{RC}) \approx D(\hat{t}_{RC}) = \sum_h^L \frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)\left(s_h^2(x) + R^2 s_h^2(y) - 2R\rho_h \cdot s_h(x)s_h(y)\right);$

$\cdot \frac{1}{N^2}$

3) $MSE(\bar{X}_{RC}) \approx D(\bar{X}_{RC}) = \frac{1}{N^2}\left[\sum_h^L \frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)\left(s_h^2(x) + R^2 s_h^2(y) - 2R\rho_h \cdot s_h(x)s_h(y)\right)\right].$

Доказ: 1) Због Т1, јасно је да важи $MSE \approx D$.

$$MSE(\hat{R}) = E(\hat{R}-R)^2 \approx E\left(\frac{\bar{x}_{STR} - R\cdot\bar{y}_{STR}}{\bar{y}}\right)^2 = \frac{1}{\bar{y}^2}\left[E(\bar{x}_{STR} - R\cdot\bar{y}_{STR})^2\right]$$

$$= \frac{1}{\bar{y}^2}\left[E(\bar{x}_{STR} - R\cdot\bar{y}_{STR})^2 - \underbrace{\left(E(\bar{x}_{STR} - R\cdot\bar{y}_{STR})\right)^2}_{=0,\ \text{па може}}\right] = \frac{1}{\bar{y}^2} D(\bar{x}_{STR} - R\cdot\bar{y}_{STR})$$

ново обележје

$$= \frac{1}{\bar{y}^2} D\left(\frac{1}{N}\sum_h^L N_h \bar{x}_{n_h} - R\cdot\frac{1}{N}\sum_h^L N_h\bar{y}_{n_h}\right) = \frac{1}{\bar{y}^2} D\left(\frac{1}{N}\sum_h^L N_h \frac{\sum_{i\in S}(x_{hi} - Ry_{hi})}{n_h}\right) = \frac{1}{\bar{y}^2} D\left((\overline{x-Ry})_{STR}\right)$$

$$\overset{\boxed{12}}{=} \frac{1}{\bar{y}^2}\cdot\frac{1}{N^2}\sum_h^L \frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)\cdot\frac{1}{N_h-1}\sum_i^{N_h}\left(x_{hi} - R\cdot y_{hi} - (\bar{x}_h - R\bar{y}_h)\right)^2$$

$$= \frac{1}{t_y^2}\sum_h^L \frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)\frac{1}{N_h-1}\sum_i^{N_h}\left((x_{hi}-\bar{x}) - R(y_{hi}-\bar{y}_h)\right)^2$$

$$= \frac{1}{t_y^2}\sum_h^L \frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)\frac{1}{N_h-1}\sum_i^{N_h}\left((x_{hi}-\bar{x}_h)^2 + R^2(y_{hi}-\bar{y}_h)^2 - 2R(x_{hi}-\bar{x})(y_{hi}-\bar{y}_h)\right)$$

$$= \frac{1}{t_y^2}\left[\sum_h^L \frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)\left(s_h^2(x) + R^2 s_h^2(y) - 2R\rho_h\cdot s_h(x)s_h(y)\right)\right]$$

2) тривијално (стрелица);

3) тривијално (стрелица).

Напомена: Предлажемо следећу оцену:

$$\widehat{D(\hat{t}_{RC})} := \sum_h^L \frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)\left(s_{nh}^2(x) + \hat{R}^2 s_{nh}^2(y) - 2\hat{R}\rho_h\cdot s_{nh}(x)s_{nh}(y)\right).$$

$\cdot \frac{1}{N^2}$

$$\widehat{D(\bar{X}_{RC})} := \frac{1}{N^2}\cdot\left[\sum_h^L \frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)\left(s_{nh}^2(x) + \hat{R}^2 s_{nh}^2(y) - 2\hat{R}\rho_h\cdot s_{nh}(x)s_{nh}(y)\right)\right].$$

25. Количничко оцењивање на основу стратификованог узорка без понављања посебна по стратумима оцена

* На сваком стратуму оцењујемо укупну суму и онда их саберемо. (зато иде $\sum_h^L$)

деф. $\hat{R}_h := \dfrac{\hat{t}_h(x)}{\hat{t}_h(y)} = \dfrac{\bar{x}_{n_h}}{\bar{y}_{n_h}}$; – по [21], ово је асимптотски непристрасно (стратум = псу)

деф. $\hat{t}_{RS} := \sum_h^L \hat{R}_h \cdot t_h(y) = \sum_h^L \dfrac{\sum_{i\in S} x_{hi}}{\sum_{i\in S} y_{hi}} \cdot t_h(y)$

$\bar{x}_{RS} := \dfrac{1}{N} \cdot \hat{t}_{RS}$

Теорема 1: $\hat{t}_{RS}$ је асимптотски непристрасна.

Доказ: $E(\hat{t}_{RS}) = \sum_h^L t_h(y) \cdot E(\hat{R}_h) \approx \sum_h^L t_h(y) \cdot R_h = \sum_h^L t_h(x) = t$

Теорема 2: За велико n, важе следеће апроксимације:

1) $MSE(\hat{t}_{RS}) \approx D(\hat{t}_{RS}) = \sum_h^L \dfrac{N_h^2}{n_h}\left(1-\dfrac{n_h}{N_h}\right)\left(s_h^2(x) + R_h^2 s_h^2(y) - 2R_h \cdot \rho_h \cdot s_h(x)\, s_h(y)\right)$;

$\cdot \frac{1}{N^2}$

2) $MSE(\bar{x}_{RS}) \approx D(\bar{x}_{RS}) = \dfrac{1}{N^2}\left[\sum_h^L \dfrac{N_h^2}{n_h}\left(1-\dfrac{n_h}{N_h}\right)\left(s_h^2(x) + R_h^2 s_h^2(y) - 2R_h \cdot \rho_h \cdot s_h(x)\, s_h(y)\right)\right]$.

стратум = псу $\Rightarrow$ $D(\hat{t}_R)$

Доказ: 1) $MSE(\hat{t}_{RS}) \approx D(\hat{t}_{RS}) = D\left(\sum_h^L t_h(y) \cdot \hat{R}_h\right) = \sum_h^L D(t_h(y) \cdot \hat{R}_h) = \sum_h^L D(\hat{t}_{R_h}(x))$

користимо алт. облик за $D(\hat{t}_R)$ из [22] $\approx \sum_h^L \dfrac{N_h^2}{n_h}\left(1-\dfrac{n_h}{N_h}\right)\left(s_h^2(x) + R_h^2 s_h^2(y) - 2R_h \cdot \rho_h \cdot s_h(x)\, s_h(y)\right)$;

2) тривијално (стрелица).

Напомена: За ове дисперзије предлажемо следеће оцене: (s_{n_h} уместо s_h)

$\widehat{D(\hat{t}_{RS})} := \sum_h^L \dfrac{N_h^2}{n_h}\left(1-\dfrac{n_h}{N_h}\right)\left(s_{n_h}^2(x) + R_h^2 s_{n_h}^2(y) - 2R_h \cdot \rho_h \cdot s_{n_h}(x)\, s_{n_h}(y)\right)$;

$\cdot \frac{1}{N^2}$

$\widehat{D(\bar{x}_{RS})} := \dfrac{1}{N^2}\left[\sum_h^L \dfrac{N_h^2}{n_h}\left(1-\dfrac{n_h}{N_h}\right)\left(s_{n_h}^2(x) + R_h^2 s_{n_h}^2(y) - 2R_h \cdot \rho_h \cdot s_{n_h}(x)\, s_{n_h}(y)\right)\right]$.

* Када користимо комбиновану оцену, а када посебну по стратумима?

1° n велико, а нису сви n_h велики: комбинована;

2° n_h велики, а R_h-ови слични: комбинована;

3° n_h велики, а R_h-ови различити: посебна.

* Ако имамо помоћну променљиву y о којој знамо све,
да ли помоћу ње да стратификујемо или да правимо количничку оцену?

Некад једно, некад друго – нема правила.

Најчешће, ако постоји веза која није линеарна $\Rightarrow$ стратификујемо.

Ако веза јесте линеарна $\Rightarrow$ правимо количничку оцену.

због закључка у $\boxed{22}$

Линеарно-регресионо оцењивање

* Као и код количничког, и овде имамо помоћну променљиву

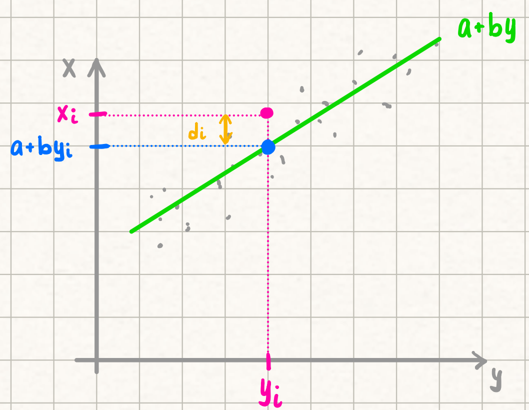
Потребно је: → x, y су у корелацији;

→ из $y=0$ не следи и $x=0$.

Пример: x - принос житарице
y - загађеност ваздуха

Јесу у корелацији, али ако је $y=0$, принос расте.

* Вадимо узорке: $(x_1, y_1), \dots, (x_n, y_n)$.



Узорачки коеф. корелације: $\hat{\rho} := \dfrac{\frac{1}{n-1}\sum_{i\in S}(x_i-\bar{x}_n)(y_i-\bar{y}_n)}{s_n(x)\cdot s_n(y)}$.

Важи: $-1 \leq \hat{\rho} \leq 1$.

Ако је близу -1 или 1,
онда има смисла правити праву лин. регресије.

Напомена: $d_i := x_i - (a+by_i)$

27. Регресионо оцењивање на основу ПСУ без понављања када се b оцењује

Конструишемо праву линеарне регресије $x = a + b\cdot y$, користећи методу најмањих квадрата.

Желимо да збир квадрата одступања (тј. $\sum_{i\in S} d_i^2$) буде што мањи.

Другим речима, тражимо минимум ф-је $f(a,b) = \sum_{i\in S}(x_i - (a+by_i))^2$. (на стд. начин)

$$\frac{\partial f}{\partial a} = 2\sum(x_i - a - by_i)\cdot(-1) = 0$$

$$\frac{\partial f}{\partial b} = 2\sum(x_i - a - by_i)\cdot(-y_i) = 0$$

$$\sum_{i\in S} x_i - \sum_{i\in S} a - \sum_{i\in S} by_i = 0 \quad\Rightarrow\quad n\cdot a = \sum_{i\in S} x_i - b\sum_{i\in S} y_i \quad\overset{/:n}{\Rightarrow}\quad \boxed{a = \bar{x}_n - b\cdot\bar{y}_n}$$

$$\sum_{i\in S} x_i y_i - \sum_{i\in S} a y_i - \sum_{i\in S} b y_i^2 = 0$$

$$\left.\begin{aligned} a\cdot n + b\cdot\sum_{i\in S} y_i &= \sum_{i\in S} x_i \\ a\cdot\sum_{i\in S} y_i + b\cdot\sum_{i\in S} y_i^2 &= \sum_{i\in S} x_i y_i \end{aligned}\right\}$$

решавамо систем по a, b (преко Крамера: $b = \frac{\Delta_b}{\Delta}$) [а | д] лева

$$b = \frac{\begin{vmatrix} n & \sum_{i\in S} x_i \\ \sum_{i\in S} y_i & \sum_{i\in S} x_i y_i \end{vmatrix}}{\begin{vmatrix} n & \sum_{i\in S} y_i \\ \sum_{i\in S} y_i & \sum_{i\in S} y_i^2 \end{vmatrix}} = \frac{n\cdot\sum_{i\in S} x_i y_i - \sum_{i\in S} x_i\cdot\sum_{i\in S} y_i}{n\cdot\sum_{i\in S} y_i^2 - \left(\sum_{i\in S} y_i\right)^2} \overset{:n^2}{\underset{:n^2}{=}} \frac{\frac{1}{n}\sum_{i\in S} x_i y_i - \bar{x}_n\bar{y}_n}{\frac{1}{n}\sum_{i\in S} y_i^2 - (\bar{y}_n)^2} = \frac{\frac{1}{n}\left[\sum_{i\in S} x_i y_i - n\cdot\bar{x}_n\bar{y}_n\right]}{\frac{1}{n}\sum_{i\in S} y_i^2 - (\bar{y}_n)^2}$$

$$\overset{(*)}{\underset{(**)}{=}} \frac{\cancel{\frac{1}{n}}\left[\sum_{i\in S}(x_i - \bar{x}_n)(y_i - \bar{y}_n)\right]}{\cancel{\frac{1}{n}}\left[\sum_{i\in S}(y_i - \bar{y}_n)^2\right]} = \frac{(n-1)\,\hat{\rho}\,s_n(x)\,s_n(y)}{(n-1)\,s_n^2(y)} \quad\Rightarrow\quad \boxed{b = \hat{\rho}\cdot\frac{s_n(x)}{s_n(y)}}$$

(*) $\sum_{i\in S}(x_i - \bar{x}_n)(y_i - \bar{y}_n) = \sum_{i\in S} x_i y_i - \bar{x}_n\sum_{i\in S} y_i - \bar{y}_n\sum_{i\in S} x_i + \sum_{i\in S}\bar{x}_n\bar{y}_n$

$= \sum_{i\in S} x_i y_i - n\bar{x}_n\bar{y}_n - \cancel{n\bar{y}_n\bar{x}_n} + \cancel{n\bar{x}_n\bar{y}_n} = \sum_{i\in S} x_i y_i - n\bar{x}_n\bar{y}_n$

(**) $\frac{1}{n}\sum_{i\in S}(y_i - \bar{y}_n)^2 = \frac{1}{n}\sum_{i\in S} y_i^2 - 2\bar{y}_n\underbrace{\frac{1}{n}\sum_{i\in S} y_i}_{\bar{y}_n} + \frac{1}{n}\underbrace{\sum_{i\in S}\bar{y}_n^2}_{n\cdot\bar{y}_n^2} = \frac{1}{n}\sum_{i\in S} y_i^2 - (\bar{y}_n)^2$

Ово је стационарна тачка, па треба проверити да ли је и минимум.

$$\left.\begin{aligned} \frac{\partial^2 f}{\partial a^2} &= 2n =: A \\ \frac{\partial^2 f}{\partial a \partial b} &= 2\sum_{i\in S} y_i =: B \\ \frac{\partial^2 f}{\partial b^2} &= 2\sum_{i\in S} y_i^2 =: C \end{aligned}\right\} \quad \begin{bmatrix} A & B \\ B & C \end{bmatrix} \implies \begin{array}{ll} AC-B^2>0 & \text{(екстремум)} \\ A>0 & \text{(минимум)} \end{array}$$

$A>0 \iff 2n>0$, тј. $n>0$ ✓

$$AC-B^2>0 \iff 4n\sum_{i\in S} y_i^2 - 4\Big(\sum_{i\in S} y_i\Big)^2 > 0 \iff n(y_1^2+\ldots+y_n^2) > (y_1+\ldots+y_n)^2$$

$$\iff (n-1)(y_1^2+\ldots+y_n^2) > \underbrace{2y_1y_2}_{<\,y_1^2+y_2^2} + \ldots + \underbrace{2y_{n-1}y_n}_{<\,y_{n-1}^2+y_n^2} \implies ✓$$

(сваки се појави n-1 пут)

Дакле, по методу најмањих квадрата, оцене за a и b су: $\hat{b} = \hat{\rho}\cdot\frac{s_n(x)}{s_n(y)}$, $\hat{a} = \bar{x}_n - \hat{b}\cdot\bar{y}_n$.

Тиме добијамо праву лин. регресије: $\underbrace{x = \hat{a} + \hat{b}\cdot y}_{\text{права}}$.

Ово све значи да $x_i \approx \hat{a} + \hat{b}y_i$.

Када сумирамо за целу популацију: $\sum^N x_i \approx N\cdot\hat{a} + \hat{b}\sum^N y_i \overset{:N}{\implies} \bar{X} \approx \hat{a} + \hat{b}\cdot\bar{Y} \overset{\hat{a}}{\implies} \bar{X} \approx \bar{x}_n - \hat{b}\bar{y}_n + \hat{b}\bar{Y}$.

Закључак: Ако се бира ПСУ без понављања, при чему b није унапред познат, имамо оцену:

$$\bar{x}_{LR} = \hat{a} + \hat{b}\cdot\bar{Y}_n = \bar{x}_n + \hat{b}(\bar{Y} - \bar{y}_n)$$

мотив

Теорема 1: Нека се бира ПСУ без понављања. Оцена $\bar{x}_{LR}$ је непристрасна оцена за $\bar{X}$

Доказ: $E(\bar{x}_{LR}) = E(\bar{x}_n) + \hat{b}\big(E(\bar{Y}) - E(\bar{y}_n)\big) \overset{\text{ПСУ}}{=} \bar{X} + \hat{b}(\bar{Y} - \bar{Y}) = \bar{X}$;

Напомена: Може се показати да за велико n: $D(\bar{x}_{LR}) \approx \left(1-\frac{n}{N}\right)\cdot\frac{1}{n}\cdot s_x^2\cdot(1-\rho^2)$;

$\boxed{26.}$ Регресионо оцењивање на основу ПСУ без понављања када је b познат

деф. Због мотива из $\boxed{27}$, за познато b_0 дефинишемо оцене:

$$\bar{x}_{LR} := \bar{x}_n + b_0(\bar{y} - \bar{y}_n);$$

$$\hat{t}_{LR} := N \cdot \bar{x}_{LR}.$$

Теорема 1: Нека се бира ПСУ без понављања и b_0 познато. (претх. искуство)

1) Оцена $\bar{x}_{LR}$ је непристрасна оцена за $\bar{x}$;

2) За дисперзију те оцене важи: $D(\bar{x}_{LR}) = \frac{1}{n}\left(1 - \frac{n}{N}\right)\left(s_x^2 - 2b_0 s_x s_y + b_0^2 s_y^2\right)$;

3) Непристрасна оцена те дисперзије је $\widehat{D(\bar{x}_{LR})} = \frac{1}{n}\left(1 - \frac{n}{N}\right)\left(s_n^2(x) - 2b_0 s_n(x) s_n(y) + b_0 s_n^2(y)\right)$.

Доказ: 1) $E(\bar{x}_{LR}) = E(\bar{x}_n) + b_0\left(E(\bar{y}) - E(\bar{y}_n)\right) \overset{\text{ПСУ}}{=} \bar{x} + b_0(\bar{y} - \bar{y}) = \bar{x}$;

2) $D(\bar{x}_{LR}) = D\left(\bar{x}_n + b_0(\bar{y} - \bar{y}_n)\right) \overset{(*)}{=} D(\bar{e}_n) \overset{\text{ПСУ}}{=} \boxed{\left(1 - \frac{n}{N}\right)\frac{1}{n} s_e^2}$

$$= \left(1 - \frac{n}{N}\right)\frac{1}{n} \cdot \frac{1}{N-1}\sum^{N}(e_i - \bar{e})^2 \overset{e_i}{=} \left(1 - \frac{n}{N}\right)\frac{1}{n} \cdot \frac{1}{N-1}\sum^{N}\left[\left(x_i + b_0(\bar{y} - y_i)\right) - \bar{x}\right]^2$$

$$= \left(1 - \frac{n}{N}\right)\frac{1}{n} \cdot \frac{1}{N-1}\sum^{N}\left[(x_i - \bar{x}) - b_0(y_i - \bar{y})\right]^2$$

$$= \left(1 - \frac{n}{N}\right)\frac{1}{n} \cdot \frac{1}{N-1}\sum^{N}\left[(x_i - \bar{x})^2 - 2b_0(x_i - \bar{x})(y_i - \bar{y}) + b_0^2(y_i - \bar{y})^2\right]$$

$$= \left(1 - \frac{n}{N}\right)\frac{1}{n}\left[s_x^2 - 2b_0\rho s_x s_y + b_0^2 s_y^2\right]$$

(*) $e_i = x_i + b_0(\bar{y} - y_i)$

$\bar{e} = \bar{x} + b_0(\bar{y} - \bar{y}) = \bar{x}$

$\bar{e}_n = \bar{x}_n - b_0(\bar{y} - \bar{y}_n)$

3) $\left(1 - \frac{n}{N}\right)\frac{1}{n} s_n^2(e) \overset{(**)}{=} \left(1 - \frac{n}{N}\right)\frac{1}{n} \cdot \frac{1}{n-1}\sum_{i \in S}(e_i - \bar{e}_n)^2$

$$= \left(1 - \frac{n}{N}\right)\frac{1}{n} \cdot \frac{1}{n-1}\sum_{i \in S}\left[\left(x_i + b_0(\bar{y} - y_i)\right) - \left(\bar{x}_n - b_0(\bar{y} - \bar{y}_n)\right)\right]^2$$

$$= \left(1 - \frac{n}{N}\right)\frac{1}{n} \cdot \frac{1}{n-1}\sum_{i \in S}\left[(x_i - \bar{x}_n) - b_0(\cancel{\bar{y}} - \bar{y}_n - \cancel{\bar{y}} + y_i)\right]^2$$

$$= \left(1 - \frac{n}{N}\right)\frac{1}{n} \cdot \frac{1}{n-1}\sum_{i \in S}\left[(x_i - \bar{x}_n)^2 - 2b_0(x_i - \bar{x}_n)(y_i - \bar{y}_n) + b_0^2(y_i - \bar{y}_n)^2\right]$$

$$= \left(1 - \frac{n}{N}\right)\frac{1}{n}\left[s_n^2(x) - 2b_0\hat{\rho}\, s_n(x)\, s_n(y) + b_0^2 s_n^2(y)\right]$$

$$= \widehat{D(x_{LR})}$$

(**) $\left(1 - \frac{n}{N}\right)\frac{1}{n} s_n^2(e)$ јесте неприс. оцена за $D(\bar{x}_{LR})$
јер је s_n^2 неприс. за s^2 код ПСУ БП
Ми доказујемо да је та оцена иста као наша

28. Упоређивање регресионе оцене

* Прво поредимо лин-рег и ПСУ оцене:

$$D(\bar{x}_{LR}) \overset{\boxed{27}}{\approx} \left(1-\frac{n}{N}\right)\cdot\frac{1}{n}\, s_x^2 \underbrace{(1-\rho^2)}_{-1\le\rho\le 1}, \quad \text{за велико } n$$

$$D(\bar{x}_n) \overset{\boxed{1}}{=} \left(1-\frac{n}{N}\right)\cdot\frac{1}{n}\, s_x^2$$

Закључак: За велико n, $\bar{x}_{LR}$ је боља од $\bar{x}_n$, осим ако је $\rho = 0$.

* Сада поредимо лин-рег и количничку:

$$D(\bar{x}_R) \overset{\boxed{22}}{\approx} \left(1-\frac{n}{N}\right)\cdot\frac{1}{n}\left(s_x^2 + R^2 s_y^2 - 2R\rho\, s_x s_y\right), \quad \text{за велико } n$$

$$\Rightarrow D(\bar{x}_R) - D(\bar{x}_{LR}) = \left(1-\frac{n}{N}\right)\cdot\frac{1}{n}\cdot\left((s_x\rho)^2 + (R s_y)^2 - 2R s_y s_x \rho\right)$$

$$= \underbrace{\left(1-\frac{n}{N}\right)\cdot\frac{1}{n}}_{>0}\cdot\underbrace{\left(s_x\rho - R s_y\right)^2}_{\geq 0} \geq 0$$

Закључак: За велико n, $\bar{x}_{LR}$ је боља од $\bar{x}_R$, осим ако је $\rho s_x = R s_y$.

Може се доказати да то важи ако је веза линеарна и пролази кроз коорд. почетак, а већ смо рекли да тада користимо количничку.

Закључак: Линеарно-регресиона оцена је боља и од ПСУ оцене и од количничке оцене.

29. Регресионо оцењивање на основу стратификованог узорка без понављања комбинована оцена

деф. **Комбинована линеарно-регресиона оцена:**

$$\bar{x}_{LRC} := \bar{x}_{STR} + b(\bar{y} - \bar{y}_{STR})$$

Теорема 1: Нека се бира стратификован узорак и b_0 познато. (претх. искуство)

1) Оцена $\bar{x}_{LRC}$ је непристрасна оцена за $\bar{X}$;

2) За дисперзију те оцене важи:

$$D(\bar{x}_{LRC}) = \frac{1}{N^2}\sum_{h}^{L} \frac{1}{n_h} \cdot N_h^2 \left(1 - \frac{n_h}{N_h}\right)\left(s_h^2(x) - 2b_0 \rho_h s_h(x) s_h(y) + b_0^2 s_h^2(y)\right);$$

3) Непристрасна оцена те дисперзије је

$$\widehat{D(\bar{x}_{LRC})} = \frac{1}{N^2}\sum_{h}^{L} \frac{1}{n_h} \cdot N_h^2 \left(1 - \frac{n_h}{N_h}\right)\left(s_{n_h}^2(x) - 2b_0 \hat{\rho}_h s_{n_h}(x) s_{n_h}(y) + b_0^2 s_{n_h}^2(y)\right).$$

Доказ: 1) $E(\bar{x}_{LRC}) = E(\bar{x}_{STR}) + b_0\left(E(\bar{y}) - E(\bar{y}_{STR})\right) \overset{STR}{=} \bar{X} + b_0(\bar{Y} - \bar{Y}) = \bar{X}$;

2)
$$D(\bar{x}_{LRC}) = D\left(\bar{x}_{STR} + b_0(\bar{y} - \bar{y}_{STR})\right) = D\left[\frac{1}{N}\sum^{L} N_h \cdot \bar{x}_{n_h} + b_0\left(\bar{y} - \frac{1}{N}\sum^{L} N_h \bar{y}_{n_h}\right)\right]$$

$$= D\left[\frac{1}{N}\sum^{L} N_h \cdot \left(\bar{x}_{n_h} + b_0(\bar{y} - \bar{y}_{n_h})\right)\right] \overset{(*)}{=} D\left[\frac{1}{N}\sum^{L} N_h \cdot \bar{e}_{n_h}\right]$$

независност стратума

$$= \frac{1}{N^2}\sum^{L} N_h^2 \cdot D[\bar{e}_{n_h}] =$$

$$\overset{ПСУ}{=} \boxed{\frac{1}{N^2}\sum^{L} N_h^2 \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot \frac{1}{n_h} \cdot s_h^2(e)}$$

$$= \frac{1}{N^2}\sum^{L} N_h^2 \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot \frac{1}{n_h} \cdot \frac{1}{N_h - 1}\sum^{N_h}(e_{hi} - \bar{e}_h)^2$$

$$= \frac{1}{N^2}\sum^{L} N_h^2 \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot \frac{1}{n_h} \cdot \frac{1}{N_h - 1}\sum^{N_h}\left(x_{hi} + b_0(\bar{y} - y_{hi}) - \left(\bar{x}_h + b_0(\bar{y} - \bar{y}_h)\right)\right)^2$$

$$= \frac{1}{N^2}\sum^{L} N_h^2 \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot \frac{1}{n_h} \cdot \frac{1}{N_h - 1}\sum^{N_h}\left(x_{hi} - \bar{x}_h - b_0(y_{hi} - \bar{y}_h)\right)^2$$

$$= \frac{1}{N^2}\sum^{L} N_h^2 \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot \frac{1}{n_h} \cdot \frac{1}{N_h - 1}\sum^{N_h}\left[(x_{hi} - \bar{x}_h)^2 - 2b_0(x_{hi} - \bar{x}_h)(y_{hi} - \bar{y}_h) + b_0^2(y_{hi} - \bar{y}_h)^2\right]$$

$$= \frac{1}{N^2}\sum^{L} N_h^2 \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot \frac{1}{n_h} \cdot \left(s_h^2(x) - 2b_0 \rho_h s_h(x) s_h(y) + b_0^2 s_h^2(y)\right);$$

(*) $e_{hi} = x_{hi} + b_0(\bar{y} - y_{hi})$

$\bar{e}_h = \bar{x}_h + b_0(\bar{y} - \bar{y}_h)$

$\bar{e}_{n_h} = \bar{x}_{n_h} + b_0(\bar{y} - \bar{y}_{n_h})$

3) $\frac{1}{N^2}\sum^{L} N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{1}{n_h} s_{n_h}^2(e) \underset{(**)}{=} \frac{1}{N^2}\sum^{L} N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{1}{n_h}\cdot\frac{1}{n_h-1}\sum_{i\in S}(e_i-\bar{e}_{n_h})^2$

(**) исти трик као у $\boxed{26}$

$$= \frac{1}{N^2}\sum^{L} N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{1}{n_h}\frac{1}{n_h-1}\sum_{i\in S}\left(x_{hi}+b_0(\bar{y}-y_{hi}) - (\bar{x}_{n_h}+b_0(\bar{y}-\bar{y}_{n_h}))\right)^2$$

$$= \frac{1}{N^2}\sum^{L} N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{1}{n_h}\frac{1}{n_h-1}\sum_{i\in S}\left((x_{hi}-\bar{x}_{n_h}) - b_0(\cancel{\bar{y}}-\bar{y}_{n_h}-\cancel{\bar{y}}+y_{hi})\right)^2$$

$$= \frac{1}{N^2}\sum^{L} N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{1}{n_h}\frac{1}{n_h-1}\sum_{i\in S}\left[(x_{hi}-\bar{x}_{n_h})^2 - 2b_0(x_{hi}-\bar{x}_{n_h})(y_{hi}-\bar{y}_{n_h}) + b_0^2(y_{hi}-\bar{y}_{n_h})^2\right]$$

$$= \frac{1}{N^2}\sum^{L} N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{1}{n_h}\left[\underline{s_{n_h}^2(x)} - 2b_0\underline{\hat{\rho}_h}\,\underline{s_{n_h}(x)}\,\underline{s_{n_h}(y)} + b_0^2\underline{s_{n_h}^2(y)}\right]$$

$$= \widehat{D(\bar{x}_{LRC})}$$

Напомена: Ако b_0 није познато, оцењујемо га са:

$$\hat{b} := \frac{\sum^{L}\left[\frac{N_h^2\left(1-\frac{n_h}{N_h}\right)}{n_h\cdot(n_h-1)}\cdot\sum_{i\in S_h}(x_{hi}-\bar{x}_{n_h})(y_{hi}-\bar{y}_{n_h})\right]}{\sum^{L}\left[\frac{N_h^2\left(1-\frac{n_h}{N_h}\right)}{n_h\cdot(n_h-1)}\cdot\sum_{i\in S_h}(y_{hi}-\bar{y}_{n_h})^2\right]}$$

30. Регресионо оцењивање на основу стратификованог узорка без понављања посебна по стратумима оцена

деф. Посебна линеарно-регресиона оцена:

$$\bar{X}_{LRS} := \frac{1}{N}\sum_{h}^{L} N_h \cdot \underbrace{\left(\bar{x}_{n_h} + b_h(\bar{y}_h - \bar{y}_{n_h})\right)}_{= \bar{x}_{LR}(h)} \quad (\text{стратум} = \text{псу})$$

Теорема 1: Нека се бира стратификован узорак и b_h познати. (претх. искуство)

1) Оцена $\bar{X}_{LRS}$ је непристрасна оцена за $\bar{X}$;

2) За дисперзију те оцене важи:

$$D(\bar{X}_{LRS}) = \frac{1}{N^2}\sum_{h}^{L} N_h^2 \cdot \frac{1}{n_h}\left(1 - \frac{n_h}{N_h}\right)\left(s_h^2(x) + b_h^2 s_h^2(y) - 2b_h \rho_h s_h(x) s_h(y)\right);$$

3) Непристрасна оцена те дисперзије је

$$\widehat{D(\bar{X}_{LRS})} = \frac{1}{N^2}\sum_{h}^{L} N_h^2 \cdot \frac{1}{n_h}\left(1 - \frac{n_h}{N_h}\right)\left(s_{n_h}^2(x) + b_h s_{n_h}^2(y) - 2b_h \hat{\rho}_h s_{n_h}(x) s_{n_h}(y)\right).$$

Доказ: 1) $E(\bar{X}_{LRS}) = \frac{1}{N}\sum^{L} N_h\left[E(\bar{x}_{n_h}) + b_h(\bar{y}_h - E(\bar{y}_{n_h}))\right] \overset{\text{псу}}{=} \frac{1}{N}\sum N_h\left[\bar{X}_h + b_h(\bar{y}_h - \bar{y}_h)\right] = \frac{1}{N}\sum N_h \bar{X}_h = \bar{X};$

2) $D(\bar{X}_{LRS}) \overset{\text{нез. страт.}}{=} \boxed{\frac{1}{N^2}\sum^{L} N_h^2\, D(\bar{x}_{LR}(h))}$

$\overset{26}{=} \frac{1}{N^2}\sum_{h}^{L} N_h^2 \cdot \frac{1}{n_h}\left(1 - \frac{n_h}{N_h}\right)\left(s_h^2(x) + b_h^2 s_h^2(y) - 2b_h \rho_h s_h(x) s_h(y)\right);$

3) $\frac{1}{N^2}\sum^{L} N_h^2\, \widehat{D(\bar{x}_{LR}(h))} \underset{(**)}{=} \frac{1}{N^2}\sum^{L} N_h^2 \cdot \frac{1}{n_h}\left(1 - \frac{n_h}{N_h}\right)\left(s_{n_h}^2(x) + b_h s_{n_h}^2(y) - 2b_h \hat{\rho}_h s_{n_h}(x) s_{n_h}(y)\right)$

$= \widehat{D(\bar{X}_{LRS})}$

(**) исти трик као у 26

Напомена: Ако b_h-ови нису познати, оцењујемо их са: $\hat{b}_h := \hat{\rho}_h \cdot \frac{s_{n_h}(x)}{s_{n_h}(y)}$.

* Када користимо комбиновану оцену, а када посебну по стратумима?

1° b_h-ови слични: комбинована;
2° b_h-ови различити: посебна.

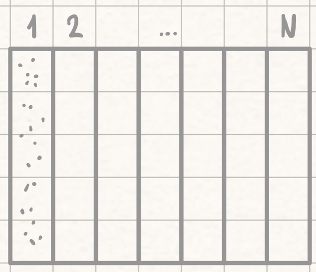
II део

НАСТАВАК

Кластер узорак

Идеја је да узорак делимо на дисјунктне подскупове – кластере, тј. примарне јединке.

Јединке унутар кластера се зову секундарне јединке.



Од N кластера,
бирамо њих n.

Напомена: у пракси, бирамо сек. јединке
и онда узмемо цео кластер ком она припада.

Дакле, узоркује се тако што бирамо целе кластере,
тј. све секундарне јединке из кластера које смо изабрали.

N – број примарних јединки;

M_i – број секундарних јединки у i-том кластеру $(i = \overline{1,N})$

$M_u = \sum^{N} M_i$ – укупан број секундарних јединки

Кластер узорак користимо када немамо добар узорачки оквир.
Такође, знатно је јефтинији. Ипак, оцене ће бити доста слабије.

Пример: број мобилних телефона у једној кући у Аустралији: кластер = град.

Уместо целу државу да обиђемо, бирамо само градове које обилазимо ⇒ јефтино.

Ипак, неки градови су сиромашнији од других ⇒ можда слабије оцене.

Кластери треба да буду што хетерогенији унутар себе, а што хомогенији између себе.
Дакле, супротно од стратума.

16. Кластер узорак

прим. јед. – ПСУ без понављања

$$\begin{array}{ccccc} 1 & 2 & 3 & \dots & N \\ x_{11} & x_{21} & x_{31} & \dots & x_{N1} \\ x_{12} & x_{22} & \vdots & \dots & x_{N2} \\ \vdots & x_{23} & x_{3M_3} & \dots & \vdots \\ x_{1M_1} & \vdots & & \dots & x_{NM_N} \\ & x_{2M_2} & & \dots & \\ & & & \dots & \\ \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ M_1 & M_2 & M_3 & & M_N \end{array}$$

деф. $t_i := \sum_j^{M_i} x_{ij}$ – тотална сума i-тог кластера

То значи: $t = \sum_i^N t_i = \sum_i^N \sum_j^{M_i} x_{ij}$.

деф. $\hat{t} := \frac{N}{n} \cdot \left(\sum_{i \in S} t_i\right)$

Теорема 1: 1) Оцена $\hat{t}$ је непристрасна оцена за t;

2) За дисперзију те оцене важи: $D(\hat{t}) = N^2 \cdot \left(1 - \frac{n}{N}\right) \frac{1}{n} \cdot \frac{1}{N-1} \sum^N \left(t_i - \frac{t}{N}\right)^2$;

3) Непристрасна оцена те дисперзије је: $\widehat{D(\hat{t})} := N^2 \cdot \left(1 - \frac{n}{N}\right) \frac{1}{n} \cdot \frac{1}{n-1} \sum_{i \in S} \left(t_i - \frac{1}{n} \sum_{i \in S} t_i\right)^2$.

Доказ: Следи из теорије за ПСУ без понављања. (само $x_i \mapsto t_i$)

1) $E\left(\frac{N}{n} \sum_{i \in S} t_i\right) = N \cdot E\left(\frac{1}{n} \sum_{i \in S} t_i\right) \overset{\text{ПСУ за } \bar{x}}{=} N \cdot \frac{1}{N} \sum^N t_i = t$;

2) $D(\hat{t}) \overset{\text{ПСУ}}{=} N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} s^2 = N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \cdot \frac{1}{N-1} \sum^N \left(t_i - \frac{t}{N}\right)^2$;

3) $\widehat{D(\hat{t})} \overset{\text{ПСУ}}{=} N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} s_n^2 = N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \cdot \frac{1}{n-1} \sum_{i \in S} \left(t_i - \frac{\sum_{j \in S} t_j}{n}\right)^2$.

деф. $\hat{x} := \frac{1}{M_u} \cdot \hat{t}$

Теорема 2: 1) Оцена $\hat{x}$ је непристрасна оцена за $\bar{x}$;

2) За дисперзију те оцене важи: $D(\hat{x}) = \frac{1}{M_u^2} \cdot D(\hat{t})$;

3) Непристрасна оцена те дисперзије је: $\widehat{D(\hat{x})} := \frac{1}{M_u^2} \cdot \widehat{D(\hat{t})}$.

Доказ: следи из Т1.

* Осим ове оцене, постоји и количничка оцена:

Као и до сада, први кластер посматрамо као прву јединку, други као другу итд.
Од N кластера, бирамо n.

Ако изаберемо i-ти кластер, као главно обележје (x) узимамо t_i.
Као помоћно обележје (y), узимамо M_i.

Ово можемо да урадимо: → јесу у корелацији
→ $y=0 \Rightarrow x=0$ (ако кластер има нула елемената нема шта да се сабере)

Количник популације је: $R = \frac{\sum^N t_i}{\sum^N M_i} = \frac{t}{M_u}$

Његова оцена је: $\hat{R} = \frac{\sum_{i\in S} t_i}{\sum_{i\in S} M_i}$

Количничке оцене су: 1) $\hat{t}_R = \hat{R} \cdot \sum^N M_i = \hat{R} \cdot M_u$

2) $\overline{x}_R = \frac{1}{M_u} \cdot \hat{t}_R = \hat{R}$

јер су кол. оцене асимпт. неприс.

Ове оцене су асимптотски непристрасне, па за велико n важи:

$$D(\hat{t}_R) \approx N^2 \left(1-\frac{n}{N}\right) \frac{1}{n} \cdot \frac{1}{N-1} \sum^N (t_i - RM_i)^2$$

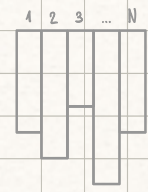
Оцена ове дисперзије је:

$$\widehat{D(\hat{t}_R)} = N^2 \left(1-\frac{n}{N}\right) \frac{1}{n} \cdot \frac{1}{n-1} \sum^N (t_i - \hat{R}M_i)^2$$

17. Кластер узорак

прим. јед. – пропорционално

деф. $p_i = \frac{M_i}{M_u}$ је вероватноћа избора i-тог кластера.



Видимо да су вероватноће избора пропорционалне величинама кластера.

Узорак може бити: са понављањем или без понављања.

I) са понављањем – НН оцене.

деф. $\hat{t}_{НН} := \frac{M_u}{n} \cdot \sum_{i \in S} \frac{t_i}{M_i}$.

Теорема 1: 1) Оцена $\hat{t}_{НН}$ је непристрасна оцена за t;

2) За дисперзију те оцене важи: $D(\hat{t}_{НН}) = \frac{M_u}{n} \cdot \sum_{i=1}^{N} M_i \cdot \left(\frac{t_i}{M_i} - \bar{x}\right)$;

3) Непристрасна оцена те дисперзије је: $\widehat{D(\hat{t}_{НН})} := \frac{1}{n(n-1)} \sum_{i \in S} \left[\frac{M_u}{M_i} \cdot t_i - \hat{t}_{НН}\right]^2$.

Доказ: $p_i = \frac{M_i}{M_u}$ у изразима НН оцена. [8]

деф. $\bar{x}_{НН} := \frac{1}{M_u} \cdot \hat{t}_{НН}$

Теорема 2: 1) Оцена $\bar{x}_{НН}$ је непристрасна оцена за $\bar{x}$;

2) За дисперзију те оцене важи: $D(\bar{x}_{НН}) = \frac{1}{M_u^2} \cdot D(\hat{t}_{НН})$

3) Непристрасна оцена те дисперзије је: $\widehat{D(\bar{x}_{НН})} := \frac{1}{M_u^2} \cdot \widehat{D(\hat{t}_{НН})}$

Доказ: следи из Т1.

II) без понављања – немамо адекватну теорију, па користимо НТ оцене (које увек важе)

деф. $\hat{t}_{HT} := \sum^{\nu} \frac{t_i}{\pi_i}$.

Теорема 3: 1) Оцена $\hat{t}_{HT}$ је непристрасна оцена за t;

2) За дисперзију те оцене важи: $D(\hat{t}_{HT}) = \sum^{N} \frac{1-\pi_i}{\pi_i} \cdot t_i^2 + \sum_i^N \sum_{j \neq i}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} t_i t_j$;

3) Непристрасна оцена те дисперзије је: $\widehat{D(\hat{t}_{HT})} := \sum^{\nu} \frac{1-\pi_i}{\pi_i^2} \cdot t_i^2 + \sum_i^{\nu} \sum_{j \neq i}^{\nu} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \cdot \frac{t_i t_j}{\pi_i \pi_j}$.

Доказ: $p_i = \frac{M_i}{M_u}$ у изразима НТ оцена. ∎

деф. $\bar{x}_{HT} := \frac{1}{M_u} \cdot \hat{t}_{HT}$

Теорема 4: 1) Оцена $\bar{x}_{HT}$ је непристрасна оцена за $\bar{x}$;

2) За дисперзију те оцене важи: $D(\bar{x}_{HT}) = \frac{1}{M_u^2} \cdot D(\hat{t}_{HT})$;

3) Непристрасна оцена те дисперзије је: $\widehat{D(\bar{x}_{HT})} := \frac{1}{M_u^2} \cdot \widehat{D(\hat{t}_{HT})}$.

Доказ: следи из Т3.

Напомена: Нама требају π_i и π_{ij}, а знамо само $p_i = \frac{M_i}{M_u}$

За узорковање са понављањем, важи:

1) $\pi_i = 1 - (1-p_i)^n$; (тривијално)

2) $\pi_{ij} = \pi_i + \pi_j - \left(1 - (1-p_i-p_j)^n\right)$ (формула укљ. и искљ.: $P(AB) = P(A) + P(B) - P(A \cup B)$)

18. Упоређивање оцена
ПСУ без понављања и кластер узорак

Претпоставимо да су сви кластери исте величине $M = M_1 = \ldots = M_N$.

То значи да је обим популације $N \cdot M$, а обим узорка $n \cdot M$.

деф. **Унутаркластерни коеф. корелације** је коеф. корел. између парова јединки у истом кластеру:

$$\rho_{uk} = \frac{\sum_i^N \sum_j^M \sum_{k \neq j}^M (x_{ij} - \bar{x})(x_{ik} - \bar{x})}{(M-1)(N \cdot M - 1) s^2}.$$

Лема 1: 1) $\rho_{uk} > 0$: кластери су хомогени унутар себе;

2) $\rho_{uk} < 0$: кластери су хетерогени унутар себе.

Доказ: 1) $\rho_{uk} > 0 \Rightarrow (x_{ij} - \bar{x})(x_{ik} - \bar{x}) > 0 \Rightarrow$ истог су знака (++ или −−)
$\Rightarrow x_{ij}, x_{ik}$ су са исте стране $\bar{x}$

2) различит знак $\Rightarrow$ са различитих страна $\bar{x}$.

из 16 Т1.2)

Лема 2: $D(\hat{t}) = \frac{N^2(1-\frac{n}{N})}{n} \cdot \frac{NM-1}{N-1} \cdot s^2 \left(1 + (M-1)\rho_{uk}\right).$ (алтернативни облик)

Доказ: $D(\hat{t}) \overset{16\,T1.2)}{=} N^2(1-\frac{n}{N}) \frac{1}{n} \frac{1}{N-1} \sum_i^N (t_i - \frac{t}{N})^2 = N^2(1-\frac{n}{N}) \frac{1}{n} \frac{1}{N-1} \sum_i^N \left(\sum_j^M x_{ij} - M \cdot \frac{t}{N \cdot M}\right)^2$

$= N^2(1-\frac{n}{N}) \frac{1}{n} \frac{1}{N-1} \sum_i^N \left(\sum_j^M x_{ij} - M \cdot \bar{x}\right)^2 = N^2(1-\frac{n}{N}) \frac{1}{n} \frac{1}{N-1} \sum_i^N \left(\sum_j^M (x_{ij} - \bar{x})\right)^2$

$= N^2(1-\frac{n}{N}) \frac{1}{n} \frac{1}{N-1} \sum_i^N \left[\sum_j^M (x_{ij} - \bar{x})^2 + \sum_j^M \sum_{k \neq j}^M (x_{ij} - \bar{x})(x_{ik} - \bar{x})\right]$

$= N^2(1-\frac{n}{N}) \frac{1}{n} \frac{1}{N-1} \left[\sum_i^N \sum_j^M (x_{ij} - \bar{x})^2 + \sum_i^N \sum_j^M \sum_{k \neq j}^M (x_{ij} - \bar{x})(x_{ik} - \bar{x})\right]$

$= N^2(1-\frac{n}{N}) \frac{1}{n} \frac{1}{N-1} \left[(NM-1) \cdot \frac{1}{NM-1} \cdot \sum_i^N \sum_j^M (x_{ij} - \bar{x})^2 + \sum_i^N \sum_j^M \sum_{k \neq j}^M (x_{ij} - \bar{x})(x_{ik} - \bar{x})\right]$

$= N^2(1-\frac{n}{N}) \frac{1}{n} \frac{1}{N-1} \left[(NM-1) \cdot s^2 + (M-1)(NM-1) s^2 \rho_{uk}\right]$

$= N^2(1-\frac{n}{N}) \frac{1}{n} \frac{NM-1}{N-1} \cdot s^2 \left[1 + (M-1)\rho_{uk}\right]$

Дакле, за велико N (самим тим и велико NM)

склонили -1

$$D(\hat{t}_K) \overset{л2}{=} \frac{N^2\left(1-\frac{n}{N}\right)}{n} \cdot \frac{\cancel{N}M}{\cancel{N}} \cdot s^2\left(1+(M-1)\rho_{uk}\right) = N^2\left(1-\frac{n}{N}\right)\frac{1}{n} M s^2 \left(1+(M-1)\rho_{uk}\right)$$

$$D(\hat{t}_{PSU}) \overset{\boxed{1}}{=} N^2 M^{\cancel{2}} \left(1-\frac{n\cancel{M}}{N\cancel{M}}\right) \cdot \frac{1}{n\cancel{M}} \cdot s^2 = N^2\left(1-\frac{n}{N}\right)\frac{1}{n} M s^2$$

Закључак: $D(\hat{t}_{PSU}) < D(\hat{t}_K) \Leftrightarrow \rho_{uk} > 0$. (тј. по ЛМ: ако су $\underbrace{\text{кластери хомогени}}_{\text{лоше}}$ унутар себе)

19. Систематски узорак

деф. Нека је N обим популације, n обим узорка и нека је $k = \frac{N}{n}$.

Од првих k бирамо једну јединку (нпр. i-ту),
а онда бирамо сваку k-ту јединку након ње ($i+k$, $i+2k$, ...).

Овако добијен узорак зове се систематски узорак.

Пример: Ако $\frac{N}{n}$ није цео број, оцене ће бити благо пристрасне

На даље, претпостављамо да јесте цео број, тј. $N = n \cdot k$. Тада су оцене непристрасне.

Приметимо да ово подсећа на стратификован узорак.
Ипак, то није исто - овде имамо неслучајно бирање (остали зависе од првог).

Систематски узорак јесте кластер узорак (имамо k кластера и бирамо један).
↑први елем.

деф. Средина систематског узорка је $\bar{X}_{sis} : \begin{pmatrix} \bar{X}_1 & \bar{X}_2 & \dots & \bar{X}_k \\ \frac{1}{k} & \frac{1}{k} & \dots & \frac{1}{k} \end{pmatrix}$. ($\bar{x}_i$ - ср. вр. i-тог могућег сис. узорка)

деф. $s^2_{sis} := \frac{1}{k(n-1)} \sum_i^k \sum_j^n (x_{ij} - \bar{x}_i)^2$. ($x_{ij}$ је j-ти елемент i-тог узорка)

Теорема 1: 1) Оцена $\bar{X}_{sis}$ је непристрасна оцена за $\bar{X}$;

2) За дисперзију те оцене важи: $D(\bar{X}_{sis}) = \frac{N-1}{N} s^2 - \frac{k(n-1)}{N} s^2_{sis}$.

Доказ: 1) $E(\bar{X}_{sis}) = \frac{1}{k} \cdot \sum_i^k \bar{x}_i = \frac{1}{k} \sum_i^k \frac{1}{n} \sum_j^n x_{ij} = \frac{1}{kn} \sum^k \sum^n x_{ij} = \frac{1}{N} \cdot t = \bar{X}$;

2) $* \; D(\bar{X}_{sis}) \overset{\text{деф.}}{=} E(\bar{X}_{sis} - \bar{X})^2 \overset{\text{деф.}}{=} \frac{1}{k} \cdot \sum^k (\bar{x}_i - \bar{X})^2$

$* \; (N-1) s^2 = \sum_i^k \sum_j^n (x_{ij} - \bar{X})^2 = \sum_i^k \sum_j^n (x_{ij} - \bar{x}_i + \bar{x}_i - \bar{X})^2$

$= \sum_i^k \sum_j^n (x_{ij} - \bar{x}_i)^2 + \sum_i^k \sum_j^n (\bar{x}_i - \bar{X})^2 + 2 \cdot \underbrace{\sum_i^k (\bar{x}_i - \bar{X})} \cdot \sum_j^n (x_{ij} - \bar{x}_i)$

$\overset{(*)}{=} k(n-1) s^2_{sis} + nk \cdot D(\bar{X}_{sis}) + 0 = k(n-1) s^2_{sis} + N \cdot D(\bar{X}_{sis})$

$\Rightarrow D(\bar{X}_{sis}) = \frac{N-1}{N} s^2 - \frac{k(n-1)}{N} s^2_{sis}$

$(*) \; \underbrace{\sum_i^k (\bar{x}_i - \bar{X})} = \sum_i^k \frac{\sum_j^n x_{ij}}{n} - k \cdot \bar{X} = \frac{1}{n} \sum \sum x_{ij} - \frac{N}{n} \cdot \frac{t}{N} = \frac{t}{n} - \frac{t}{n} = 0$

Последица: $\bar{x}_{sis}$ је боља оцена од $\bar{x}_n$ акко $S^2 < S^2_{sis}$.

Другим речима, $\bar{x}_{sis}$ је боља ако је дисперзија унутар систематских узорака већа, што значи да су систематски узорци (њих k) хетерогенији унутар себе.

- што су хомогенији $\Rightarrow$ дају мању информацију о популацији;
- што су хетерогенији $\Rightarrow$ дају бољу информацију.

Доказ: $D(\bar{x}_{sis}) < D(\bar{x}_n) \Leftrightarrow \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S^2_{sis} < \left(1-\frac{n}{N}\right)\frac{1}{n} S^2$

$\Leftrightarrow \left(\frac{N-1}{N} - \frac{N-n}{Nn}\right) S^2 < \frac{k(n-1)}{N} S^2_{sis}$

$\Leftrightarrow \left(1-\cancel{\frac{1}{N}}-\frac{1}{n}+\cancel{\frac{1}{N}}\right) S^2 < \frac{k(n-1)}{N} S^2_{sis}$

$\Leftrightarrow \cancel{\frac{n-1}{n}} S^2 < \cancel{\frac{k(n-1)}{N}} S^2_{sis}$

$k = N \cdot n$

$\Leftrightarrow S^2 < S^2_{sis}$

* Сада повезујемо ову причу са причом о кластерима.

Теорема 2: $D(\bar{x}_{sis}) = \frac{S^2}{n} \cdot \frac{N-1}{N} \cdot (1 + (n-1)\rho_{UK})$. (алтернативни облик)

Доказ: Као што смо рекли: систематски = кластер узорак где од k кластера бирамо 1.

$\Rightarrow \hat{t}_k = \frac{k}{1}\boxed{\sum_{i\in S} t_i}$ (бр. кластера; само један сабирак; бр. изабраних), $\quad \bar{x}_{sis} = \frac{\sum_{i\in S} t_i}{n} = \frac{\hat{t}_k}{nk} = \frac{\hat{t}_k}{N}$.

$D(\bar{x}_{sis}) = D\left(\frac{\hat{t}_k}{N}\right) = \frac{1}{N^2} D(\hat{t}_k) \overset{\text{18 Л2, (**)}}{=} \frac{k^2}{N^2}\left(1-\frac{1}{k}\right)\cdot\frac{1}{1}\cdot\frac{kn-1}{k-1} S^2 (1+(n-1)\rho_{UK})$

$= \frac{1}{N^2}\cdot\frac{k^2(k-1)}{k}\cdot\frac{kn-1}{k-1} S^2 (1+(n-1)\rho_{UK}) = \frac{1}{n^2k^2}\cdot k\cdot (N-1) S^2 (1+(n-1)\rho_{UK})$

$= \frac{S^2}{n}\cdot\frac{N-1}{N}\cdot(1+(n-1)\rho_{UK})$.

(**) $N \to k$
$M \to n$
$n \to 1$

Напомена: $D(\bar{x}_{sis})$ мање акко је ρ_{UK} мање $\Rightarrow$ кластери што хетерогенији

* **Што нисмо тражили $\widehat{D(\bar{x}_{sis})}$?**

Систематско = кластер за $n=1$.

У кластер оцени $\widehat{D}(\hat{x})$ [16 Т2] се појављује члан $\frac{1}{n-1}$ $\Rightarrow$ дељење нулом ↯

20. Квалитет оцене добијене на основу систематског узорка

Први случај

Имамо узорачки списак који је одређен случајно.

(нпр. меримо плату људи, а они су поређани од 1 – 9999 по последње 4 цифре броја телефона).

Тада ће систематски узорак вероватно дати исти квалитет оцене као и ПСУ.

Други случај

Ако је списак направљен и види се да постоји неки линеарни тренд: $x_i = a + b \cdot i$

$$D(\bar{x}_n) = \ldots = \frac{(k-1)(N+1)}{12} \cdot b^2; \quad (*)$$

$$D(\bar{x}_{SIS}) = \ldots = \frac{(k-1)(k+1)}{12} \cdot b^2; \quad (**)$$

$$D(\bar{x}_{STR}) = \ldots = \frac{(k-1)(k+1)}{12n} \cdot b^2. \quad (***)$$

$$\Rightarrow \quad \underbrace{D(\bar{x}_{STR}) \leq}_{\text{због } n \text{ у имениоцу}} D(\bar{x}_{SIS}) \underbrace{\leq}_{k \leq N} D(\bar{x}_n)$$

Дакле, најбољу оцену даје стратификовани узорак.

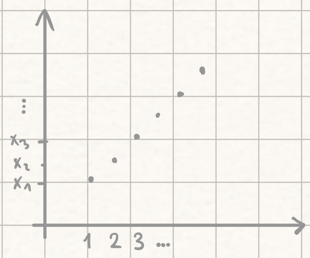
Трећи случај

Имамо списак јединки и уочавамо да се вредности обележја периодично понављају.

Има ли смисла користити систематски узорак? Зависи од k.

нпр. 3 1 2 3 1 2 ... и $k = 3$: 2 2 2 ... → нема смисла;
$k = 4$: 2 3 1 2 3 1 → букв. идеалан узорак.

Дакле, квалитет оцене зависи од k.



$$x_i = a + b\cdot i, \quad \forall i \quad \Rightarrow \quad \bar{x} = a + b\cdot\frac{N+1}{2}$$

$N = n\cdot k$ – укупан број јединки

n – обим узорка

k – корак код систематског узорка

(*) $$D(\bar{x}_n) = \left(1-\frac{n}{N}\right)\frac{S^2}{n} = \left(1-\frac{n}{N}\right)\frac{1}{n}\cdot\frac{1}{N-1}\sum^{N}(x_i-\bar{x})^2 = \left(1-\frac{n}{N}\right)\frac{1}{n}\cdot\frac{1}{N-1}\,b^2\sum^{N}\left(i-\frac{N+1}{2}\right)^2$$

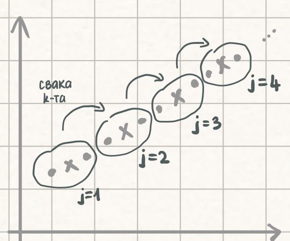
$$= \left(1-\frac{n}{N}\right)\frac{1}{n}\cdot\frac{1}{N-1}\,b^2\left[\sum^{N} i^2 - 2\,\frac{N+1}{2}\sum^{N} i + N\cdot\frac{(N+1)^2}{4}\right]$$

$$= \left(1-\frac{n}{N}\right)\frac{1}{n}\cdot\frac{1}{N-1}\,b^2\left[\frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{2} + \frac{N(N+1)^2}{4}\right]$$

$$= \left(1-\frac{n}{N}\right)\frac{1}{n}\cdot\frac{1}{N-1}\,b^2\cdot\frac{N(N+1)}{12}\left[2(2N+1) - 6(N+1) + 3(N+1)\right] = \frac{N-n}{\cancel{N}}\cdot\frac{1}{n}\cdot\frac{1}{\cancel{N-1}}\,b^2\cdot\frac{\cancel{N}(N+1)}{12}\cdot\cancel{(N-1)}$$

$$= (\cancel{n}k-\cancel{n})\,\frac{1}{\cancel{n}}\,b^2\,\frac{N+1}{12} = \frac{(k-1)(N+1)}{12}\,b^2$$

(**)



$j=1: \quad X_{i+0\cdot k}$

$j=2: \quad X_{i+1\cdot k}$

$j=3: \quad X_{i+2\cdot k}$

$\vdots$

$X_{i+(j-1)\cdot k}$

Подсетимо се из $\boxed{19}$:

$$\bar{x}_{sis}: \begin{pmatrix} \bar{x}_1 & \cdots & \bar{x}_k \\ 1/k & \cdots & 1/k \end{pmatrix}$$

$$D(\bar{x}_{sis}) = E(\bar{x}_{sis} - \bar{x})^2 = \frac{1}{k}\sum_{i}^{k}(\bar{x}_i - \bar{x})^2 = \frac{1}{k}\sum_{i}^{k}\left[\frac{1}{n}\sum_{j}^{n}\underbrace{X_{i+(j-1)\cdot k}}_{=\,a+b(i+(j-1)k)} - \left(a + b\,\frac{N+1}{2}\right)\right]^2$$

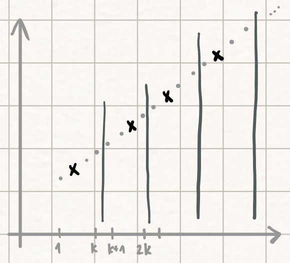
$$= \frac{1}{k}\sum_{i}^{k}\left[\cancel{a} + \frac{bni}{\cancel{n}} + \frac{bk\sum_j^n (j-1)}{n} - \cancel{a} - b\,\frac{N+1}{2}\right]^2$$

$$= \frac{1}{k}\,b^2\sum_{i}^{k}\left[i + \frac{k}{\cancel{n}}\cdot\frac{\cancel{n}(n-1)}{2} - \frac{N+1}{2}\right]^2 = \frac{1}{k}\,b^2\sum_{i}^{k}\left[i + \frac{\cancel{N} - k - \cancel{N} - 1}{2}\right]^2 = \frac{1}{k}\,b^2\sum_{i}^{k}\left(i - \frac{k+1}{2}\right)^2$$

$$= \frac{1}{k}\,b^2\left[\sum_{i}^{k} i^2 - 2\cdot\frac{k+1}{2}\sum_{i}^{k} i + k\,\frac{(k+1)^2}{4}\right] = \frac{1}{k}\,b^2\left[\frac{k(k+1)(2k+1)}{6} - \frac{k(k+1)^2}{2} + \frac{k(k+1)^2}{4}\right]$$

$$= \frac{1}{\cancel{k}}\,b^2\,\cancel{k}(k+1)\cdot\frac{k-1}{12} = \frac{(k-1)(k+1)}{12}\,b^2$$

(***)



Како правимо страт. узорак?

Имамо n стратума величине k
и из сваког од њих бирамо узорак обима 1

$\Rightarrow \quad N_h \mapsto k \quad , \quad L \mapsto n \quad , \quad n_h \mapsto 1$

Ако бирамо из h-тог стратума, $(h-1)\cdot k$ елемената је прошло и ако узимамо j-ти

$\Rightarrow \quad X_{(h-1)\cdot k+j}$

Знамо $D(\bar{x}_{STR}) = \frac{1}{N^2}\sum_h^L N_h^2 \frac{s_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)$ $\Rightarrow$ морамо још да одредимо $s_h^2 \mapsto ?$

(све остало познато)

$$* \; s_h^2 = \frac{1}{k-1}\sum_j^k \Big[X_{(h-1)k+j} - \frac{1}{k}\sum_l^k X_{(h-1)k+l} \Big]^2$$

$$= \frac{1}{k-1}\sum_j^k \Big[a + b((h-1)k+j) - \frac{1}{k}\sum_l^k (a+b((h-1)k+l)) \Big]^2$$

$$= \frac{1}{k-1}\sum_j^k \Big[\cancel{a} + \cancel{b(h-1)k} + bj - \cancel{\tfrac{1}{k}\cdot ka} - \cancel{\tfrac{1}{k}\cdot kb(h-1)k} + \frac{1}{k} b \sum_l^k l \Big]^2 = \frac{1}{k-1} b^2 \sum_j^k \left(j - \frac{k+1}{2}\right)^2$$

$$= \frac{1}{k-1} b^2 \Big[\sum_j^k j^2 - 2\,\frac{k+1}{2}\sum_j^k j + k\,\frac{(k+1)^2}{2} \Big]$$

$$= \frac{b^2}{12} k(k+1)$$

$$* \; D(\bar{x}_{STR}) = \frac{1}{N^2}\sum_h^L N_h^2 \frac{s_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right) = \frac{1}{N^2}\sum_h^n k^2 \cdot \frac{k(k+1)}{12} b^2 \left(1-\frac{1}{k}\right)$$

$$= \frac{1}{n^2k^2}\cdot n \cdot k^2 \cdot \frac{k(k+1)}{12}\cdot b^2 \cdot \frac{k-1}{k}$$

$$= \frac{(k^2-1)}{12n}\cdot b^2$$