УНИВЕРЗИТЕТ У НОВОМ САДУ

ПРИРОДНО-МАТЕМАТИЧКИ ФАКУЛТЕТ

ДЕПАРТМАН ЗА
МАТЕМАТИКУ И ИНФОРМАТИКУ

Радослав Божић

Примене стратификованог узорка

- мастер рад -

Нови Сад, *2012*

Садржај

Предговор

Прва статистичка истраживања вршена су у Кини пре око 4000 година, када су прикупљани подаци о бројном стању становништва, војске и сл., док се први озбиљнији кораци у статистичким истраживањима срећу тек крајем XVIII и почетком XIX века. Данас статистика има широку примену, а због великог броја података које је потребно обрадити, у истраживањима се посматра само један део популације, који се назива узорак.

Како би резултати истраживања били што поузданији, потребно је изабрати репрезентативан узорак. Проучавањем избора узорка и оцењивања одговарајућих параметара се бави теорија узорака. Постоји више планова узорка, као што су прост случајан узорак, узорак са неједнаким вероватноћама, стратификован узорак, узорак скупина, вишеетапни узорак, двофазни узорак,...

Стратификовани узорак је један од најчешће коришћених планова узорка. Његова примена је изузетно велика у истраживањима јавног мњења, али и у многим другим истраживањима, као што је оцењивање различитих демографских параметара. Овај план узорка такође заузима веома значајно место у испитивањима у области привреде.

Међутим, стратификовани узорак је сложен план узорка, што подразумева да се увек примењује у комбинацији са још неким планом. У овом раду су, поред стратификованог, описани и они планови узорка који се са њим најчешће примењују. Дати су и примери примене стратификованог узорка.

Овом приликом се захваљујем свим професорима и асистентима на пренесеном знању током студирања. Посебно се захваљујем свом ментору, Проф. др Загорки Лозанов-Црвенковић, на подршци и разумевању током писања овог рада.

Нови Сад, *јануар 2012.* *Радослав Божић*

1. Увод

1.1 Основни појмови теорије узорака

Да би се испитала нека карактеристика популације, неопходно је анализирати карактеристике елемената те популације, било да се ради о становницима неке области, запосленима у неком предузећу, одређеним производима,... Међутим, често због бројности популације није могуће анализирати карактеристике (обележја) свих јединица, већ се посматра само један део популације, на основу чијих се карактеристика изводи закључак о читавој популацији. Тај део се назива узорак.

Важно је да се резултати испитивања узорка, без већих одступања, могу применити на читаву популацију. Дакле, битно је да узорак буде репрезентативан.

За разлику од статистичке теорије, у теорији узорака се посматра коначна популација. Нека популација има N јединица ($u_1, u_2,..., u_N$). Посматрана карактеристика јединице обележава се са y_i, где је $i = 1,..., n$, и назива се обележје. Када се одабере узорак и региструју обележја, приступа се оцењивању одређених функција обележја.

Поступак којим се јединица из популације бира у узорак назива се план узорка. Планови се деле на стандарсне, који могу бити конвенционални и адаптивни, и нестандардне. Код конвенционалних планова, вероватноћа избора јединице у узорак не зависи ни од једне величине која се испитује, док код адаптивних поступак избора може да зависи од испитиване величине, али само на елементима који су изабрани у узорак. Код нестандардних планова вероватноћа избора зависи од посматране величине.

Нека је θ посматрана карактеристика. Оцена $\hat{\theta}$ је центрирана (непристрасна), ако је њена средња вредност, узета по свим могућим узорцима, једнака θ, односно:

$$E(\hat{\theta}) = \theta$$

Ако оцена није непристрасна, тада се величина:

$$B = E(\hat{\theta}) - \theta$$

назива пристрасност (бијас) у $\hat{\theta}$.

За поређење различитих оцена неког параметра користи се средње квадратна грешка (одступање) оцене:

$$MSE(\hat{\theta}) = V(\hat{\theta}) + (B(\hat{\theta}))^2 ,$$

где је $V(\hat{\theta})$ варијанса оцене $\hat{\theta}$, а $B(\hat{\theta})$ бијас. Од две оцене повољнија је она која има мању средње квадратну грешку.

Уређен узорак величине n је низ $s_0 = (i_1, i_2,\ldots, i_n)$ од n ознака, при чему неке ознаке могу бити исте ако се ради о узорку са понављањем. Уређење је одређено редоследом избора елемената. Редуковани узорак s се сасоји од v различитих ознака из s_0, а уређен је по растућем редоследу индекса.

Минимална довољна статистика за узорке из коначне популације је неуређен скуп различитих вредности обележја за јединица из узорка, и ознака тих јединица. За сваку оцену која није функција минималне довољне статистике може се добити оцена која зависи од минималне довољне статистике.

1.2 Прост случајни узорак

Прост случајни узорак је план узорка у коме се n различитих јединица бира из популације тако да свака могућа комбинација од n јединица има исту вероватноћу да буде изабрана у узорак.

Узорак од n јединица назива се случајан узорак са понављањем ако се свака изабрана јединица након избора враћа у популацију и може поново бити изабрана. Код узорка без понављања изабрана јединица се може одстранити из популације, или вратити у исту, али тако да буде занемарена приликом евентуалног каснијег извлачења.

Основне оцене карактеристика популације су тотал и средина. Њих дефинишемо на следећи начин:

За популацију:

Тотал: $$Y = \sum_{i=1}^{N} y_i = y_1+y_2+\ldots+y_N$$

Средина: $$\overline{Y} = \frac{y_1+y_2+\cdots+y_N}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

За узорак:

Тотал: $$y = \sum_{i=1}^{n} y_i = y_1+\ldots+y_n$$

Средина: $$\overline{y} = \frac{y_1+\cdots+y_n}{n} = \frac{1}{n}\sum_{i=1}^{N} y_i\, I_i$$

где је I_i индикатор функција ($I_i = \begin{cases} 0, & i-\text{та јединица није укључена у узорак} \\ 1, & i-\text{та јединица је изабрана} \end{cases}$).

Вероватноћа избора i-те јединице у узорак је:

$$\pi_i = P(I_i = 1) = \binom{N-1}{n-1} / \binom{N}{n} = \frac{n}{N},$$

јер је број узорака који садрже јединицу i $\binom{N-1}{n-1}$, а вероватноћа да јединица i буде изабрана је $\binom{N}{n}$. Дакле, важи:

$$E(\overline{y}) = \frac{1}{n}\sum_{i=1}^{N} y_i\, P(I_i = 1) = \frac{1}{N}\sum_{i=1}^{N} y_i = \overline{Y},$$

па је $\overline{y}$ центрирана оцена за $\overline{Y}$.

Број различитих елемената узорка (ν) назива се ефективна величина узорка, а n_i је број појављивања i-те јединице у узорку, па важи и:

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{N} y_i n_i$$

$\overline{y}$ је центрирана оцена за $\overline{Y}$, било да се ради о узорку са или без понављања. $\hat{Y} = N\,\overline{y}$ је центрирана оцена тотала обележја популације.

Варијанса обележја је: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{Y})^2$,

а варијанса популације: $S^2 = \frac{N}{N-1}\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})^2$

Варијанса средине $\overline{y}$ је: $V(\overline{y}) = E(\overline{y} - \overline{Y})^2 = \frac{S^2}{n}\frac{N-n}{N} = \frac{S^2}{n}(1-f)$,

где је $f = \frac{n}{N}$ фракција узорка (фактор се $\frac{N}{n}$ назива експанзија узорка).

Стандардна грешка од $\overline{y}$ је:

$$\sigma_{\overline{y}} = \frac{S}{\sqrt{n}}\sqrt{(N-n)/N} = \frac{S}{\sqrt{n}}\sqrt{1-f}$$

Варијанса од $\hat{Y} = N\,\overline{y}$ је:

$$V(\hat{Y}) = E(\hat{Y} - Y)^2 = N^2\,V(\overline{y}) = \frac{N^2S^2}{n}\frac{N-n}{N} = \frac{N^2S^2}{n}(1-f)$$

Стандардна грешка од $\hat{Y}$ је:

$$\sigma_{\hat{Y}} = \frac{N\,S}{\sqrt{n}}\sqrt{(N-n)/N} = \frac{N\,S}{\sqrt{n}}\sqrt{1-f}$$

Варијанса средње вредности за прост случајни узорак из бесконачне популације једнака је σ^2/n , а фактори $1-(N/n)$ за варијансу и $\sqrt{1-(N/n)}$ за стандардну грешку се називају фактори корекције за коначну популацију. Ако је узорак мали у односу на популацију, ти фактори су приближно једнаки 1, па варијанса од $\overline{y}$ тежи ка S^2/n. Аналогно, за јако велики узорак фактори корекције теже нули, па је варијанса занемариво мала.

Средња вредност елемената узорка са различитим ознакама је такође центрирана оцена средине обележја популације:

$$\overline{y}_\nu = \frac{1}{\nu}\sum_{i=1}^{\nu} y_i = \frac{1}{\nu}\sum_{i=1}^{N} y_i I_i$$

Оцена $\overline{y}_\nu$ је ефикаснија од $\overline{y}$: $V(\overline{y}_\nu) < V(\overline{y})$

У формулама стандардне грешке популацијске средине и тотала фигурише варијанса популације S^2, која у пракси обично није позната, па се зато оцењује:

Израз:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

представља центрирану оцену за:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})^2 .$$

Центриране оцене варијансе од $\overline{y}$ и $\hat{Y} = N\,\overline{y}$ су:

$$\hat{V}(\overline{y}) = \frac{s^2}{n}\frac{N-n}{N} = \frac{s^2}{n}(1-f)$$

$$\hat{V}(\hat{Y}) = \frac{N^2 s^2}{n}\frac{N-n}{N} = \frac{N^2 s^2}{n}(1-f)$$

Оцене стандарсних грешака имају веома малу пристрасност:

$$\hat{\sigma}_{\overline{y}} = \frac{s}{\sqrt{n}}\sqrt{1-f} \qquad \hat{\sigma}_{\hat{Y}} = \frac{N\,s}{\sqrt{n}}\sqrt{1-f}$$

Интервали поверења

Када се, након избора узорка, оцени одређени параметар, важно је проценити тачност те оцене, што се постиже налажењем интервала поверења.

Нека је *I* интервал поверења за средину обележја популације $\overline{Y}$. Ако је α вероватноћа грешке, за интервал поверења важи:

$$P(\overline{Y} \in I) = 1-\alpha$$

Интервал *I* се зове *100*(1-α)% интервал поверења, а величина 1-α ниво поверења. Параметар $\overline{Y}$ је фиксиран, а крајеви интервала зависе од узорка. Најчешће се за α бирају вредности 0.01, 0.05, 0.1.

Приближни *100*(1-α)% интервали поверења за неке оцене:

- За средину обележја популације:

$$\overline{y} \pm c\sqrt{\frac{N-n}{N}\frac{s^2}{n}},$$

где је *c* квантил реда $1-\frac{\alpha}{2}$ Студентове *t*–расподеле са *n-1* степени слободе.

- За тотал обележја популације:

$$\hat{Y} \pm c\sqrt{N(N-n)\frac{s^2}{n}}$$

Ако је обим узорка већи од 30, вредност *c* је квантил нормалне $\mathcal{N}(0,1)$ расподелу.

Централна гранична теорема за коначну популацију

Нека је $y_1, y_2, \ldots, y_N$ низ независних случајних променљивих са једнаким расподелама, коначном средином и варијансом, тада је расподела од $\frac{\overline{y}-\overline{Y}}{\sqrt{V(\overline{y})}}$ приближно стандардна нормална за довољно велико n.

Нека је $\overline{Y}_N$ средина обележја, а $\overline{y}_N$ средина узорка популације величине N из које се бира случајни узорак без понављања, тада је расподела од $\frac{\overline{y}_N-\overline{Y}_N}{\sqrt{V(\overline{y}_N)}}$ приближно стандардна нормална за довољно велико n и N-n.

Величина узорка

Нека је $\hat{\theta}$ оцена параметра θ, а d највећа дозвољена разлика између оцене и стварне вредности, и нека је α вероватноћа да је грешка већа од d. Тада се величина узорка бира тако да важи:

$$P(\left|\hat{\theta}-\theta\right| > d) < \alpha$$

За прост случајни узорак, код оцене средине обележја популације, потребна величина узорка се добија решавањем једначине $z\sqrt{\frac{N-n}{N}\,\frac{s^2}{n}} = d$ по n:

$$n = \frac{1}{(\frac{d^2}{z^2\,s^2}+\frac{1}{N})} = \frac{1}{\frac{1}{n_0}+\frac{1}{N}},$$

где је $n_0 = \frac{z^2\,s^2}{d^2}$, а z квантил нормалне $\mathcal{N}(0,1)$ расподеле.

Код оцене тотала, n се добија решавањем једначине $z\sqrt{N(N-n)\,\frac{s^2}{n}} = d$:

$$n = \frac{1}{(\frac{d^2}{N^2\,z^2\,s^2}+\frac{1}{N})} = \frac{1}{\frac{1}{n_0}+\frac{1}{N}},$$

где је $n_0 = \frac{N^2\,z^2\,s^2}{d^2}$.

Како за прост случајни узорак, тако и за било који други план узорка, важи правило да се величина узорка бира тако да се постигне што већа прецизност приликом оцењивања, али и да трошкови целокупног истраживања буду што мањи. Када трошкове истраживања, укључујући и утрошено време, не бисмо узимали у обзир, тада би оптимално било посматрати целу популацију као узорак. Са друге стране, када бисмо гледали само на трошкове, узорак би био сувише мали и добијене оцене би биле неупотребљиве.

1.3 Узорак са неједнаким вероватноћама

1.3.1 Избор узорка

Прост случајни узорак има велики теоријски значај, али се ретко користи у пракси, јер се јединице које се бирају у узорак често разликују по величини, па би применом простог случајног узорка оцене имале велике варијансе. Због тога се чешће примењује избор јединица са вероватноћом пропорционалном величини јединица узорка. Тако ће, на пример, приликом анкете која обухвата више фабрика, фабрика са већим бројем запослених имати веће шансе да буде изабрана у узорак.

Код примене узорка са вероватноћом пропорционалном величини (*PPS*), *i*-та јединица из популације се бира у узорак са вероватноћом $p_i = M_i / M$, где је M_i величина *i*–те јединице, а $M = \sum_i M_i$ величина целе популације. *PPS* узорак може бити са и без понављања, а њихова ефикасност се не разликује много када се употребљавају скупови са фракцијом $f = n / N$.

Постоје два поступка за избор оваквих узорака. У *PPS* узорак се бира *n* јединица из популације од *N* јединица чије су величине $M_1, M_2....,M_n$, где су $M_1, M_2....,M_n$ цели бројеви.

Први поступак се састоји у томе да се бира случајан број између 1 и *M*, па ако изабрани број припада интервалу $[1,M_1]$ у узорак се бира јединица 1. Ако је изабран случајан број из интервала $[M_1+1,M_1+M_2]$ у узорак се бира јединица 2, ако је број из интервала $[M_1+M_2+1,M_1+M_2+M_3]$, бира се трећа јединица, итд. Поступак се понавља све до избора *n*-те јединице у у зорак. Уколико се, код *PPS* узорка без понављања, у једном тренутку изабере јединица која је раније већ изабрана, она се одбацује и поступак се наставља. Међутим овај поступак захтева познавање величине свих јединица популације, што може бити проблематично када је популација велика.

Други поступак, који се назива Лахиријев метод, подразумева избор пара случајних бројева (i,R), таквих да је $1 \le i \le N$ и $1 \le R \le K$, где је $K = \max_i\{M_i\}$. Ако је $R \le M_i$, *i*-та јединица ће бити изабрана у узорак. У супротном, она се одбацује и поступак се понавља.

Код Лахиријевог метода, вероватноћа да *i*-та јединица буде изабрана при избору првог пара случајних бројева је:

$$p_1(i) = \frac{1}{N}\frac{M_i}{K},$$

док је вероватноћа да у првом покушају не буде изабрана ни једна јединица из популације:

$$q = 1 - \sum_{i=1}^{N} \frac{1}{N}\frac{M_i}{K} = 1 - \frac{\overline{M}}{K}$$

$\overline{M}$ је просечна велчина јединице. Вероватноћа избора *i*-те јединице у другом покушају је:

$$p_2(i) = q\,\frac{1}{N}\frac{M_i}{K}$$

Даљим поступком добија се да је вероватноћа избора *i*-те јединице из популације:

$$p_i = p_1(i) + p_2(i) + ... = p_1(i) + q\,p_1(i) + q^2\,p_1(i) + ... = \frac{p_1(i)}{1-q} = \frac{M_i}{M},$$

јер бесконачни ред конвергира, пошто је $q = 1 - \frac{\overline{M}}{K} \leq 1$.

За разлику од досад посматраних, у неким плановима узорка јединице могу имати различиту вероватноћу укључења у узорак. Нека јединица i има вероватноћу укључења π_i, која је функција од p_i. Важи:

$$E(y) = \frac{1}{n}\sum_{i-1}^{N} y_i P(I_i = 1) = \frac{1}{N}\sum_{i-1}^{N} y_i \pi_i$$

добијени израз је, у општем случају, различит од $\overline{Y}$, што значи да $\overline{y}$ није непристрасна оцена за $\overline{Y}$.

Пример 1.3.1

Популација се састоји од $N = 12$ јединица чије су величине M_i дате у табели:

Јединица	1	2	3	4	5	6	7	8	9	10	11	12
M_i	8	17	14	25	6	11	26	36	19	9	21	32

Изабраћемо узорак са неједнаким вероватноћама без понављања обима $n = 3$, и то применом Лахиријевог метода.

Бирамо пар случајних бројева (i,R), такав да је $1 \leq i \leq$ N и $1 \leq R \leq K$, где је $K = \max_i\{M_i\}$. Неко је то пар (6,23). Будући да је $23 > M_6 = 11$, јединицу 6 нећемо изабрати у узорак, него ћемо је одбацити и наставити поступак. Сада бирамо пар (8,19), па пошто је $19 < M_8 = 36$, јединицу 8 узимамо у узорак. Даље бирамо пар (3,11). Видимо да је $11 < M_3 = 14$, па у узорак бирамо и јединицу 3. Нека је следећи пар (11,34). Јединицу 11 не бирамо у узорак јер је $34 > M_{11} = 21$. Поступак настављамо избором пара (12,17) и закључујемо да је $17 < M_{12} = 32$, па ћемо и јединицу 12 изабрати у узорак.

Дакле, применом Лахиријевог метода добили смо узорак $s = (8, 3, 12)$.

1.3.2 Horvitz - Thompson-ова оцена

Нека је π_i вероватноћа укључења за i-ту јединицу, нека је $\pi_i > 0$, за све $i = 1,..., N$. Нека су $y_1, y_2, ...y_v$ вредности обележја за v различитих елемената у узорку. Horvitz - Thompson-ова оцена средине обележја популације је :

$$\widehat{\overline{Y_{HT}}} = \frac{1}{N}\sum_{i=1}^{v} \frac{y_i}{\pi_i}$$

Уколико су вредности π_i различите, оцена зависи од тога који индекси су изабрани у узорак. Такође важи:

$$\widehat{\overline{Y_{HT}}} = \frac{1}{N}\sum_{i=1}^{N} \frac{y_i I_i}{\pi_i} = \frac{1}{N}\sum_{i=1}^{N} y_i \frac{I_i}{E(I_i)}$$

Показујемо да је $\widehat{\overline{Y_{HT}}}$ центрирана оцена:

$$E\left(\widehat{\overline{Y_{HT}}}\right) = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\pi_i}E(I_i) = \overline{Y}$$

Нека је $\pi_{ij} = P(I_iI_j = 1)$ заједничка вероватноћа укључења, $m = N$ и $\pi_{ii} = \pi_i$. Тада је варијанса Horvitz-Thompson - ове оцене:

$$V\left(\widehat{\overline{Y_{HT}}}\right) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\ y_iy_j\ ,$$

док је центрирана оцена варијансе:

$$\hat{V}\left(\widehat{\overline{Y_{HT}}}\right) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j\pi_{ij}}\ y_iy_jI_iI_j = \frac{1}{N^2}\sum_{i=1}^{\nu}\sum_{j=1}^{\nu}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j\pi_{ij}}y_iy_j\ ,$$

где је $\pi_{ij} > 0$ за све i и j таква да је $i \neq j$. Иако се ради о оцени варијансе, овај израз може бити негативан. Због тога су дате и друге оцене, али и оне су имале своје недостатке. Тако је, на пример, једна од њих центрирана само за фиксирану величину узорка.

Непристрасна оцена тотала обележја популације је:

$$\hat{Y}_{HT} = \sum_{i=1}^{\nu}\frac{y_i}{\pi_i} = \sum_{i=1}^{N}\frac{y_iI_i}{\pi_i}$$

Варијанса ове оцене је:

$$V\left(\hat{Y}_{HT}\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\ y_iy_j = \sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}y_i^2 + \sum_{i=1}^{N}\sum_{j\neq i}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\ y_iy_j$$

Ако важи $\pi_{ij} > 0$, за све i, j, тада је центрирана оцена варијансе оцене тотала:

$$\begin{aligned}\hat{V}\left(\hat{Y}_{HT}\right) &= \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j\pi_{ij}}\ y_iy_jI_iI_j \\ &= \sum_{i-1}^{\nu}\ \sum_{i=1}^{\nu}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j\pi_{ij}}y_iy_j = \sum_{i-1}^{\nu}\ \frac{1-\pi_i}{\pi_i^2}y_i^2 + \sum_{i=1}^{\nu}\ \sum_{j\neq i}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j\pi_{ij}}y_iy_j \\ &= \sum_{i=1}^{\nu}\ \left(\frac{1}{\pi_i^2}-\frac{1}{\pi_i}\right)y_i^2 + 2\sum_{i=1}^{\nu}\ \sum_{j>i}\left(\frac{1}{\pi_i\pi_j}-\frac{1}{\pi_{ij}}\right)\ y_iy_j\end{aligned}$$

Приближни $(1-\alpha)100\%$ интервал поверења за тотал обележја популације је једнак:

$$\hat{Y}_{HT} \pm z\sqrt{\hat{V}\left(\widehat{Y_{HT}}\right)}$$

z је, за велике узорке, квантил реда 1 - α стандардне нормалне расподеле $Z: \mathcal{N}(0,1)$. За узорке чији је обим мањи од 30, z представља квантил Студентове t-расподеле са v-1 степеном слободе.

Ако су y_i и π_i пропорционални, оцена $\overline{Y_{HT}}$ има малу варијансу, али ако y_i и π_i нису приближно пропорционални, варијанса ће бити већа и оцена може бити непоуздана. Због тога је Хајек предложио следећу модификацију ове оцене:

$$\overline{Y_{HT}}^{*} = \sum_{i=1}^{v} \frac{y_i}{\pi_i} \; / \sum_{i=1}^{v} \frac{1}{\pi_i}$$

Ова оцена се користи када број јединица популације (N) није познат, односно када се N оцењује непрстрасном оценом:

$$\widehat{N} = \sum_{i=1}^{v} \frac{1}{\pi_i} = \sum_{i=1}^{N} \frac{I_i}{\pi_i}$$

Коришћењем теорије количинског оцењивања добија се апроксимација варијансе Хајекове оцене.

$$V\left(\overline{Y_{HT}}^{*}\right) \approx \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}\right)\left(y_i - \overline{Y}\right)\left(y_j - \overline{Y}\right),$$

као и оцена ове варијансе:

$$\hat{V}\left(\overline{Y_{HT}}^{*}\right) = \frac{1}{N^2} \sum_{i=1}^{v} \sum_{j=1}^{v} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}}\right)\left(y_i - \overline{Y_{HT}}^{*}\right)\left(y_j - \overline{Y_{HT}}^{*}\right)$$

Пример 1.3.2

Из популације величине N = 11 изабрли смо узорак обима n = 3. Вероватноће са којима су елементи изабрани, и y–вредности елемената су: y_1 = 5, p_1 = 0.21, y_2 = 12, p_2 = 0.06, y_3 = 7, p_3 = 0.12.

Сада ћемо оценити тотал и средину обележја популације користећи Horvitz - Thompson-ову оцену, као и варијансе тих оцена.

Најпре за сваку јединицу налазимо вероватноћу укључења у узорак. Ако је вероватноћа избора i–те јединице у једном покушају једнака p_i, онда је вроватноћа да та јединица уопште не буде изабрана: $(1- p_i)^n$. Одатле следи да је вероватноћа укључења i–те јединице у узорак $\pi_i = 1 - (1 - p_i)^n$, па имамо:

$\pi_1 = 1 - (1 - p_1)^3 = 1 - (1 - 0.21)^3 = 1- 0.4930 = 0.5070$
$\pi_2 = 1 - (1 - p_2)^3 = 1 - (1 - 0.06)^3 = 1- 0.8306 = 0.1694$
$\pi_3 = 1 - (1 - p_3)^3 = 1 - (1 - 0.12)^3 = 1- 0.6815 = 0.3185$

Horvitz - Thompson-ова оцена тотала обележја популације је:

$$\hat{Y}_{HT} = \sum_{i=1}^{v} \frac{y_i}{\pi_i} = \frac{5}{0.5070} + \frac{12}{0.1694} + \frac{7}{0.3185} = 102.68$$

Оцена средине обележја популације је:

$$\widehat{\overline{Y_{HT}}} = \frac{1}{N}\sum_{i=1}^{v}\frac{y_i}{\pi_i} = \frac{102.68}{11} = 9.33$$

Заједничке вероватноће укључења су $\pi_{ij} = \pi_i + \pi_j - [1 - (1 - p_i - p_j)^n]$, па добијемо да је $\pi_{12} = 0.0654$, $\pi_{13} = 0.1263$, $\pi_{23} = 0.0393$.

Оцена варијансе добијене оцене тотала је:

$$\begin{aligned}\hat{V}(\hat{Y}_{HT}) &= \sum_{i=1}^{v}\left(\frac{1}{\pi_i^2} - \frac{1}{\pi_i}\right)y_i^2 + 2\sum_{i=1}^{v}\sum_{j>i}\left(\frac{1}{\pi_i\pi_j} - \frac{1}{\pi_{ji}}\right)y_i y_j \\ &= \left(\frac{1}{0.5070^2} - \frac{1}{0.5070}\right)\cdot 5^2 + \left(\frac{1}{0.1694^2} - \frac{1}{0.1694}\right)\cdot 12^2 \\ &+ \left(\frac{1}{0.3185^2} - \frac{1}{0.3185}\right)\cdot 7^2 \\ &+2\left(\frac{1}{0.5070\cdot 0.1694} - \frac{1}{0.0654}\right)\cdot 5\cdot 12 \\ &+2\left(\frac{1}{0.5070\cdot 0.3185} - \frac{1}{0.1263}\right)\cdot 5\cdot 7 \\ &+2\left(\frac{1}{0.1694\cdot 0.3185} - \frac{1}{0.0393}\right)\cdot 12\cdot 7 \\ &= 2825.69\end{aligned}$$

Оцена варијансе оцене средине је:

$$\hat{V}\left(\widehat{\overline{Y_{HT}}}\right) = \frac{1}{N^2}\,\hat{V}\left(\widehat{Y_{HT}}\right) = 23.35$$

1.3.3 Hansen - Hurwitz-ова оцена

Hansen - Hurwitz-ова оцена је предложена код узорка са понављањем. Нека је P_i вероватноћа избора *i*–те јединице у једном кораку, а n_i број избора *i*–те јединице. Тада је:

$$\widehat{\overline{Y_{HH}}} = \frac{1}{nN}\sum_{i=1}^{n}\frac{y_i}{p_i} = \frac{1}{nN}\sum_{i=1}^{N}\frac{y_i n_i}{p_i}$$

будући да n_i има биномну $\mathcal{B}(n_i, P_i)$ расподелу, важи:

$$\widehat{\overline{Y_{HH}}} = \frac{1}{N}\sum_{i=1}^{N} y_i\,\frac{n_i}{E[n_i]}$$

Такође важи:

$$E\left(\widehat{\overline{Y_{HH}}}\right) = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{Y}\ ,$$

па је $\widehat{\overline{Y_{HH}}}$ центрирана оцена.

Знамо да важи: $E(n_i) = n\,p_i$, $V(n_i) = np_i\,(1\text{-}p_i)$, $cov(n_i,n_j) = -\,np_i\,p_j (i\neq j)$ и $\sum_i p_i = 1$.

Одатле добијамо:

$$V\left(\widehat{\overline{Y_{HH}}}\right) = \frac{1}{n^2N^2}\left\{\sum_{i=1}^{N}\frac{y_i^2}{p_i^2}V(n_i) + \sum_{i=1}^{N}\sum_{j\neq i}\frac{y_j y_i}{p_i p_j}cov\left(n_i n_j\right)\right.$$

$$= \frac{1}{nN^2}\left\{\sum_{i-1}^{n}\frac{y_i^2}{p_i}(1-p_i) - \sum_{i-1}^{N}\sum_{j\neq i}y_i y_j\right\}$$

$$= \frac{1}{nN^2}\left\{\sum_{i=1}^{N}\frac{y_i^2}{p_i} - \left(\sum_{i=1}^{N}y_i\right)^2\right\} = \frac{1}{n}\sum_{i=1}^{N}p_i\,(\frac{y_i}{Np_i} - \bar{Y})^2$$

Пошто се на овај начин не добија непристрасна оцена од $V\left(\widehat{\overline{Y_{HH}}}\right)$, потребно је направити општији приступ, у коме се $\widehat{\overline{Y_{HH}}}$ записује као узорачка средина независних случајних променљивих са истом расподелом. Нека је T случајна променљива која узима вредности y_i/Np_i са вероватноћама p_i, i=1,...,N.

Тада важи:

$$\overline{Y_T} = E(T) = \sum_{i=1}^{N}\frac{y_i}{Np_i}P_i = \bar{Y}$$

$$\sigma_T^2 = V(T) = \sum_{i=1}^{N}\left(\frac{y_i}{Np_i} - \bar{Y}\right)^2 P_i$$

Нека је t_i елемент случајног узорка из расподеле T. Пошто је у питању узорак са понављањем, важи:

$$\widehat{\overline{Y_{HH}}} = \frac{1}{n}\sum_{i=1}^{n}t_i = \bar{t}$$

Даље важи:

$$E\left(\widehat{\overline{Y_{HH}}}\right) = E(\bar{t}) = \bar{Y}_T = \bar{Y} \qquad \text{и} \qquad V\left(\widehat{\overline{Y_{HH}}}\right) = \frac{\sigma_T^2}{n} = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{y_i}{Np_i} - \bar{Y}\right)^2 p_i$$

Тражимо центрирану оцену за $V(\widehat{\overline{Y_{HH}}})$:

$$\hat{V}\left(\widehat{\overline{Y_{HH}}}\right) = \frac{S_T^2}{n} = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\,t_i - \bar{t})^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{Np_i} - \widehat{\overline{Y_{HH}}}\right)^2$$

Уколико се се ради о случајном узорку с понављањем, важи: $p_i = \frac{1}{N}$ и $\hat{V}\left(\widehat{\overline{Y_{HH}}}\right) = \bar{y}$, па је:

$$V(\bar{y}) = \frac{1}{nN}\sum_{i=1}^{N}(y_i - \bar{Y})^2$$

Такође важи:

$$\hat{V}(\bar{Y}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Центрирана оцена тотала, њена варијанса и центрирана оцена варијансе су дате следећим изразима:

$$\hat{Y}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$$

$$V(\hat{Y}_{HH}) = \frac{1}{n}\sum_{i=1}^{n} p_i \left(\frac{y_i}{p_i} - Y\right)^2$$

$$\hat{V}(\hat{Y}_{HH}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \widehat{Y_{HH}}\right)^2$$

$(1-\alpha)100\%$ интервал поверења за тотал обележја популације је:

$$\hat{Y}_{HH} \pm z\sqrt{\hat{V}(\widehat{Y_{HH}})}$$

За веће узорке z представља квантил реда 1- α стандардне нормалне расподеле Z: $\mathcal{N}(0,1)$, а за узорке обима мањег од 30 користи се t-расподела са n-1 степени слободе.

Пример 1.3.3

Прво ћемо за узорак из примера 1.2.2 помоћу Hansen - Hurwitz-ове оцене наћи оцену тотала обележја популације, а затим и њену варијансу.

$$\hat{Y}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} = \frac{1}{3}\left(\frac{5}{0.21} + \frac{12}{0.06} + \frac{7}{0.12}\right) = \frac{1}{3}(23.81 + 200 + 58.33) = 94.05$$

$$\begin{aligned}\hat{V}(\hat{Y}_{HH}) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \widehat{Y_{HH}}\right)^2 \\ &= \frac{1}{3\cdot 2}[(23.81 - 94.05)^2 + (200 - 94.05)^2 + (58.33 - 94.05)^2] \\ &= 2905.82\end{aligned}$$

Оцена станадрдне грешке је: $\sqrt{\hat{V}(\hat{Y}_{HH})} = \sqrt{2905.82} = 53.91$

1.4 Систематски узорак

Систематски узорак са кораком k јесте узорак добијен на следећи начин: Из популације од N јединица на случајан начин бирамо једну од првих k јединица,коју означавамо са i. Затим бирамо сваку k–ту јединицу, тако да се узорак формира од јединица са индексима:

$$i,\ i+k,\ i+2k,\ \ldots,\ i+(n-1)k$$

На овај начин се само прва јединица бира случајно, а остале су аутоматски одређене изабраном јединицом. Број могућих узорака код систематског избора је мањи него код, на пример, простог случајног узорка, и до њих се једноставније долази. Такође је, у већини случајева, стандардна грешка код систематског узорка мања него код простог случајног узорка.

Нека је $N=nk$. Тада је број могућих систематских узорака једнак k. Они су дати као колоне у табели 1.3.1, и сваки од њих има n елемената.

Табела 1.4.1:

Број узорка					
1	2	...	i	...	k
y_1	y_2	…	y_i	…	y_k
y_{k+1}	y_{k+2}	…	y_{k+i}	…	y_{2k}
…	…	…	…	…	…
$y_{(n-1)k+1}$	$y_{(n-1)k+2}$	…	$y_{(n-1)k+i}$	…	y_{nk}
$\bar{y}_1$	$\bar{y}_2$		$\bar{y}_i$		$\bar{y}_k$

Вероватноћа избора једног узорка је $1/k$. Такође, постоји могућност да је $N\neq nk$, па различити систематски узорци из исте популације имају различит број елемената. У овом случају, аритметичка средина узорка није непристрасна оцена , али за веће узорке, са више од 50 елемената, пристрасност аритметичке средине није велика.

Варијанса узорачке средине

Нека је y_{ij} j–ти члан у i–том систематском узорку, где j=1, …, n , а i=1, …, k. Нека су:

$\bar{y}_i$ - средина i –тог узорка

$\bar{y}_{sy}$ – средина систематског узорка

$\bar{y}_{sy}$ је случајна променљива чије су вредности $\bar{y}_i$.

Теорема 1.4.1 Нека је:

$$S^2_{wsy} = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2$$

варијанса између јединица које су у истом систематском узорку.

Тада је варијанса средине систематског узорка дата са:

$$V(\bar{y}_{sy}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S^2_{wsy}$$

Доказ: На основу једначине за анализу варијансе, знамо:

$$(N-1)S^2 = \sum_i \sum_j (y_{ij} - \bar{Y})^2 = n \sum_i (\bar{y}_i - \bar{Y})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$$

$\bar{y}_{sy}$ је непристрасна оцена средине популације $\bar{Y}$, па важи: $E(\bar{y}_{sy}) = \bar{Y}$
Одатле следи да је варијанса за $\bar{y}_{sy}$:

$$V(\bar{y}_{sy}) = \frac{1}{k} \sum_{i=1}^{k} (\bar{y}_i - \bar{Y})^2$$

Даље важи:

$$(N-1)S^2 = nkV(\bar{y}_{sy}) + k(n-1)S^2_{wsy}$$

а одатле следи:

$$V(\bar{y}_{sy}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S^2_{wsy}$$

што је и требало доказати. □

Последица 1.4.1.1 Средина систематског узорка је прецизнија од средине простог случајног узорка, ако и само ако је:

$$S^2_{wsy} > S^2$$

Доказ: Нека је $\bar{y}$ средина простог случајног узорка величине *n*. Тада је:

$$V(\bar{y}) = \frac{N-1}{N} \cdot \frac{S^2}{n}$$

Знамо да важи: $V(\bar{y}_{sy}) < V(\bar{y})$, ако и само ако је:

$$\frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S^2_{wsy} < \frac{N-1}{N} \cdot \frac{S^2}{n}$$

односно:

$$k(n-1)S^2_{wsy} > \left(N - 1 - \frac{N-n}{n}\right) S^2 = k(n-1)S^2$$

□

Може се закључити да систематски узорак има већу прецизност у односу на прост случајни узорак уколико је варијанса унутар систематских узорака већа од варијансе целе популације. Дакле, систематски узорак је прецизан када су јединице унутар истог узорка хетерогене, а непрецизан када су хомогене, јер ће узастопне јединице у узорку давати приближно исте информације када су варијације унутар систематског узорка мање од популацијске.

Теорема 1.4.2 Нека је ρ_ω коефицијент корелације између парова јединица које су у истом систематском узорку, дефинисан са:

$$\rho_\omega = \frac{E(y_{ij}-\bar{Y})(y_{in}-\bar{Y})}{E(y_{ij}-\bar{Y})^2}$$

Тада важи:

$$V(\bar{y}_{sy}) = \frac{S^2}{n} \cdot \frac{N-1}{N}(1 + (n-1) \cdot \rho_\omega)$$

Одавде видимо да позитивна корелација између јединица у истом узорку повећава варијансу узорачке средине.

Уколико је списак јединица популације, на основу кога бирамо узорак, уређен случајно, систематски узорак се неће разликовати од простог случајног узорка без враћања. У том случају се могу применити исте формуле за оцену варијансе. Ако је списак јединица унапред формиран, у систематском узорку ћемо имати позитивну корелацију, која повећава варијансу и смањује прецизност. Понекад се, ради повећања прецизности, списак јединица одређује на унапред одређен начин. Тада је корелација унутар класа негативна. Ни у једном од ова два случаја не може се применити прост случајан узорак.

Приликом спровођења анкета, често се примењује систематски узорак са вероватноћом пропорционалном величини (систематски ППС узорак). Овде се свакој јединици придружује цео број x_i, који представља њену величину, а затим се формирају кумулане тих бројева према списку јединица Нека је:

$$T = X = \sum_{i=1}^{N} X_i \; ; \qquad T_i = X_1 + \cdots + X_i$$

$$k = \frac{T}{n} = \frac{X}{n}$$

Бирамо случајан број r између 1 и k. Тако је узорак од n јединица одређен бројевима $r+jk$; j=0, 1, …, n-1. Јединица i је укључена у узорак ако за неко j важи:

$$T_{i-1} < r + jk \leq T_i$$

Вероватноћа укључења i–те јединице у узорак је:

$$\pi_i = \frac{X_i}{k} = \frac{nX_i}{X} = np_i \; ,$$

при чему је $k > X_i$,за свако i.
Оцена тотала у систематском ППС узорку је:

$$\hat{Y} = \sum_i \frac{Y_i}{\pi_i} = k \sum_{i=1}^{n} \frac{y_i}{x_i} = \frac{X}{n} \sum_{i=1}^{n} \frac{y_i}{x_i} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i}$$

Ова оцена је центрирана јер је добијена из Horvitz-Thompson–ове оцене.

1.5 Узорак скупина

У случајевима када је потребно испитати карактеристике јако великих популација, није практично примењивати прост случајан, стратификовани или систематски узорак. Због тога се основни скуп дели на примарне јединице (скупине), од којих се свака састоји од секундарних јединица. Затим се, једном од метода избора узорка, бира одређени број скупина, а у даљој анализи се посматрају сви њихови елементи. На тај начин се формира узорак скупина.

Међутим, овај узорак је мање прецизан од стратификованог и простог случајног узорка. Наиме, потребно је да скупине по својој структури буду што сличније основном скупу, док се стратуми формирају као хомогени скупови. Стратификовани узорак се користи када желимо да што прецизније оценимо параметре, а узорак скупина када је потребно смањити време и трошкове оцењивања.

1.6 Вишеетапни узорак

Двоетапни узорак представља план узорка код кога се популација дели на одређен број примарних јединица, након чега се, из сваке примарне јединице, бира узорак од секундарних јединица. Уколико поступак наставимо, добија се вишеетапни узорак.

Разликују се случајеви када су скупине на које је подељена популације једнаке и када су различите, а подузорци се из примарних јединица могу бирати и неким другим планом узорка, најчешће стратификованим или систематским узорком.

Основна разлика између стратификованог и вишеетапног узорка је у томе што се, код стратификованог узорка, елементи бирају у узорак из сваке групе (стратума), док се код вишеетапног узорка одређене групе (примарне јединице) бирају у узорак, а потом се, из тако одабраних примарних јединица, бирају секундарне јединице.

1.7 Двофазни узорак

Често се избор узорка врши на основу неког познатог параметра. На пример, стратификација популације се врши на основу неке величине чију расподелу познајемо. Међутим, када нам такви параметри нису познати, могуће је наћи најпре њихове оцене, на основу узорка одабраног у првој фази, а затим, у другој фази, одабрати нови узорак и на основу њега оценити жељени параметар. Узорак се у другој фази најчешће бира као подузорак узорка одабраног у првој фази.

Ова техника се назива двофазни (дупли) узорак, а уколико је потребно оценити више непознатих параметара, поступак се понавља у више ваза (онолико колико је потребно) и тада се цео процес назива вишефазни узорак. План узорка се може разликовати у различитим фазама.

2. Стратификовани узорак

Код стратификованог узорка се, ради повећања прецизности оцене, врши стратификација. Она подразумева поделу популације на делове који се називају стратуми. Приликом поделе важно је водити рачуна о томе да стратуми буду релативно хомогени, али међусобно разграничени. Као критеријум поделе се узима бар једна карактеристика популације. Сваки стратум мора имати бар две јединице.

Стратификацијом се може потићи изузетно велика прецизност у оценама целе популације. Хетерогене популације је могуће поделити на потпопулације, од којих је свака хомогена унутар себе, на шта асоцира и сам назив „стратум“, који означава поделу на слојеве. Тако се, на пример, ради постизања што већег степена хомогености, код стратификације привреде велике и мале фирме сврставају у различите стратуме, код стратификације људске популације одвајају се људи који живе у домаћинствима од оних који бораве у институцијама попут затвора, болница, и сл. Ако је сваки стратум хомоген, разлике у мерењима између јединица су мале, па се прецизне оцене могу добити на основу малог узорка у стратуму. Овако добијене оцене могу такође бити прецизне и када се уопште на целу популацију.

Теорија стратификованог узорковања се бави својствима оцена добијених помоћу стратификованог узорка, најбољим избором величине узорка и оптималним распоредом, а све у циљу постизања што веће прецизности приликом оцењивања параметара.

Поступак стратификације се састоји у томе да се популација која садржи N јединица дели на L потпопулација (стратума) који немају заједничких елемената. Нека је број елемената у h-том стратуму N_h, где је $h = 1,...,L$. Тада је $N_1 +...+ N_L = L$. Након формирања стратума, по одређеном плану се бира узорак из сваког стратума, при чему су избори елемената из стратума међу собом независни. Нека је n_h величина узорка из h-тог стратума, тада је $n_1+...+n_L = n$ обим узорка.

Са y_{h_i} означавамо вредност i-те јединице, у h-том стратуму, $W_h = N_h/N$ је релативна фреквенција узорка у стратуму (тежина стратума), $f_h = n_h/N_h$ фракција узорка у стратуму, а Y_h тотал стратума. Средине стратума и узорка и варијанса стратума, дате су редом, следећим изразима:

$$\bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{h_i}$$

$$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{h_i}$$

$$S_h^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(y_{h_i} - \bar{Y}_h)^2$$

Уколико се из сваког стратума бира прост случајни узорак, цео поступак називамо стратификовани случајни узорак. Стратификовани узорак се често комбинује и са другим плановима узорка,као што су узорак са неједнаким вероватноћама (пре свега узорак са вероватноћама пропорционалним величини), систематски узорак, узорак скупина, вишеетапни и вишефазни узорак, у зависности од потреба које одређују врсте и особине популације.

2.1 Варијанса основног скупа

Следећи израз представља варијансу стратификованог основног скупа:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N} (y_i - \overline{Y})^2 = \frac{1}{N-1}\sum_{h=1}^{L}\sum_{i=1}^{N_h} (y_{h_i} - \overline{Y})^2$$

Међутим, варијанса основног скупа се састоји од варијансе унутар стратума (S_u^2) и варијансе између стратума (S_i^2):

$$\begin{aligned} S^2 &= \frac{1}{N-1}\sum_{h=1}^{L} \sum_{i=1}^{N_h} (y_{h_i} - \overline{Y})^2 \\ &= \frac{1}{N-1}\sum_{h=1}^{L} \sum_{i=1}^{N_h} \left[(y_{h_i} - \overline{Y}_h) + (\overline{Y}_h - \overline{Y})\right]^2 \\ &= \frac{1}{N-1}\left[\sum_{h=1}^{L} \sum_{i=1}^{N_h} (y_{h_i} - \overline{Y}_h)^2 + \sum_{h=1}^{L} N_h(\overline{Y}_h - \overline{Y})^2\right] \\ &= \frac{1}{N-1}\sum_{h=1}^{L} (N_h - 1)S_h^2 + \frac{1}{N-1}\sum_{h=1}^{L} N_h (\overline{Y}_h - \overline{Y})^2 \\ &= S_u^2 + S_i^2 \end{aligned}$$

Ефикасност стратификације је већа што је варијанса унутар стратума мања.

2.2 Оцењивање

Средина обележја популације по јединици је:

$$\overline{Y} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h} y_{h_i} = \frac{1}{N}\sum_{h=1}^{L} N_h\overline{Y}_h$$

а њена оцена на основу *L* стратума је:

$$\overline{y}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h\overline{y}_h$$

Средина узорка је:

$$\overline{y} = \frac{1}{n}\sum_{h=1}^{L}\sum_{i=1}^{n_h} y_{h_i} = \frac{1}{n}\sum_{h=1}^{L} n_h\overline{y}_h$$

а разликује се од $\overline{y}_{st}$ по томе што код $\overline{y}_{st}$ оцене из сваког стратума посебно добијају корекцију тежина N_h/N. $\overline{y}$ и $\overline{y}_{st}$ се покалапају ако за сваки стратум важи да је фракција узорка иста у свим стратумима: $\frac{n_h}{n} = \frac{N_h}{N} <=> \frac{n_h}{N_h} = \frac{n}{N} <=> f_h = f$

Оваква стратификација се зове стратификација са пропорционалним распоредом.

Теорема 2.1 Ако је узорачка оцена $\bar{y}_h$ центрирана оцена за $\bar{Y}_h$, тада је $\bar{y}_{st}$ центрирана оцена за $\bar{Y}$.

Доказ:

$$E(\bar{y}_{st}) = E\left(\frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h\right) = \frac{1}{N}\sum_{h=1}^{L} N_h E(\bar{y}_h) = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{Y}_h = \bar{Y}$$

(Важи $E(\bar{y}_h) = \bar{Y}_h$, јер је $\bar{y}_h$ центрирана оцена од $\bar{Y}_h$) □

Последица 2.1.1 $\bar{y}_{st}$ је центрирана оцена од $\bar{Y}$ за стратификован случајни узорак.

Теорема 2.2 Нека је $\bar{y}_h$ центрирана оцена од $\bar{Y}_h$ и нека се узорци бирају независно из различитих стратума. Тада је варијанса оцене $\bar{y}_{st}$:

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 V(\bar{y}_h) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h) \ ,$$

где је, $V(\bar{y}_h) = E(\bar{y}_h - \bar{Y}_h)^2$.

Доказ:

$$\begin{aligned}(\bar{y}_{st} - \bar{Y})^2 &= \left(\frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h - \frac{1}{N}\sum_{h=1}^{L} N_h \bar{Y}_h\right)^2 \\ &= \left(\frac{1}{N}\sum_{h=1}^{L} N_h (\bar{y}_h - \bar{Y}_h)\right)^2 \\ &= \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 (\bar{y}_h - \bar{Y}_h)^2 + \frac{2}{N^2}\sum_{h<j} N_h N_j (\bar{y}_h - \bar{Y}_h)(\bar{y}_j - \bar{Y}_j)\end{aligned}$$

Даље рачунамо средину преко свих узорака. Код међупроизвода, за други члан, фиксирамо узорак из *h*-тог стратума па узимамо средину свих узорака у *j*-том стратуму.

Узорци у *j*-том стратуму имају исте вероватноће без обзира на узорак у стратуму *h*, јер су узорци у различитим стратумима независни. y_j је непристрасна оцена, па важи да је: $E\left(\bar{y}_j - \bar{Y}_j\right) = 0$, одакле следи да су сви међупроизводи једнаки нули. Даље је:

$$V(\bar{y}_{st}) = E(\bar{y}_{st} - \bar{Y})^2 = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 E(\bar{y}_h - \bar{Y}_h)^2 = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 V(\bar{y}_h)$$

□

Варијанса оцене $\bar{y}_{st}$ зависи само од варијансе оцена средина $\bar{Y}_h$ у појединим стратумима.

Теорема 2.3 Варијанса оцене $\bar{y}_{st}$ у стратификованом случајном узорку је:

$$\begin{aligned}V(\bar{y}_{st}) &= \frac{1}{N^2}\sum_{h=1}^{L} N_h (N_h - n_h)\frac{S_h^2}{n_h} \\ &= \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}(1 - f_h) \\ &= \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h}\end{aligned}$$

Доказ:

Знамо да важи: $V(\bar{y}_h) = \frac{S_h^2}{n_h}\frac{N_h - n_h}{N_h}$, где је $\bar{y}_h$ средина обележја простог случајног узорка у h-том стратуму. На основу претходне теореме важи:

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2\, V(\bar{y}_h) = \frac{1}{N^2}\sum_{h=1}^{L} N_h\,(N_h - n_h)\frac{S_h^2}{n_h}$$

□

Ако су узорачке фракције биле занемарене важило би:

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2 S_h^2}{n_h} = \sum_{h=1}^{L}\frac{W_h^2 S_h^2}{n_h}$$

Овде такође занемарујемо и фактор корекције коначне популације. Заменом $n_h = \frac{n\,N_h}{N}$ и израз за варијансу оцене $\bar{y}_{st}$ за прост случајан узорак за пропорционалан распоред добија се:

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h}{N}\frac{S_h^2}{n}\frac{N-1}{n} = \frac{1-f}{n}\sum_{h=1}^{L} W_h\, S_h^2$$

Ако су узорци пропорционални, а сву стратуми имају исту варијансу (S_ω^2), важи :

$$V(\bar{y}_{st}) = \frac{S_\omega^2}{n}\frac{N-n}{n}$$

Центрирана оцена тотала h-тог стратума је $\hat{Y}_h = N_h\bar{y}_h$, а оцена тотала популације $\hat{Y}_{st} = \sum_{h=1}^{L} N_h\bar{y}_h = N\bar{y}_{st}$.

Теорема 2.4 Ако је $\hat{Y}_{st} = N\bar{y}_{st}$ оцена тотала обележја популације, тада је њена варијанса:

$$V\left(\hat{Y}_{st}\right) = \sum_{h=1}^{L} N_h\,(N_h - n_h)\frac{S_h^2}{n_h}$$

Ако је из сваког стратума биран прост случајан узорак, тада је непристрасна оцена варијансе стратума (S_h^2):

$$s_h^2 = \frac{1}{n_h}\sum_{i=1}^{L} N_h(\,y_{h_i} - \bar{y}_h)^2$$

Теорема 2.5 Непристрасна оцена варијансе тотала варијансе од $\bar{y}_{st}$ код стратификованог случајног узорка дате су следећим изразима:

$$\hat{V}\left(\hat{Y}_{st}\right) = \sum_{h=1}^{L} N_h\,(N_h - n_h)\frac{s_h^2}{n_h}$$

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h\,(N_h - n_h)\frac{s_h^2}{n_h} = \sum_{h=1}^{L}\frac{W_h^2 s_h^2}{n_h} - \sum_{h=1}^{L}\frac{W_h^2 s_h^2}{N_h}$$

Приближни $(1-\alpha)100\%$ интервали поверења за тотал обележја популације и за средину су:

$$\hat{Y}_{st} \pm z\sqrt{\hat{V}\left(\hat{Y}_{st}\right)}$$

$$\bar{y}_{st} \pm z\sqrt{\hat{V}\left(\bar{y}_{st}\right)}$$

Ако су сви узорци обима већег или једнаког 30, за z се узима приближна вредност квантила реда 1 - α нормалне расподеле Z: $\mathcal{N}(0,1)$. У супротном се користи t-расподела чији се број степени слободе одређује апроксимацијом:

$$d = \left(\sum_{h=1}^{L} a_h s_h^2\right)^2 / \left(\sum_{h=1}^{L}((a_h s_h^2)^2/(n_h-1))\right)$$

где је $a_h = N_h(N_h - n_h)/n_h$. поред ове апроксимације, могу се користити још неке. Ако су сви стратуми једнаких величина, као и сви узорци, тада је број степени слободе $n - L$.

Пример 2.1

За оцену месечне продаје хране у једном региону, прикупљени су подаци из 20 продавница, од укупно 120, колико их има у том региону. Будући да је већина продавница део одређеног трговинског ланца, формирају се стратуми састављени од продавница из истог ланца. У наредној табели приказан је број продавница у оквиру сваког ланца:

Фирма	Број продавница
А	50
Б	30
Ц	20
Д	10
Е	6
Остале продавнице	4
Укупно	120

Пошто је број продавница у ланцу Е, као и продавница које нису део ниједног ланца мали, оне се спајају у један стратум. Из сваког од 5 тако добијених стратума бирају се узорци чије су величине: n_1=8, n_2=5, n_3=3, n_4=2, n_5=2.

Нека су дате вредности месечне продаје за сваку од изабраних продавница, у хиљадама динара:

Стратум	Вредност продате робе							
1	250	330	210	280	380	190	220	305
2	420	380	470	520	440			
3	200	190	160					
4	100	78						
5	70	58						

Оценићемо просечну вредност продате робе. Да бисмо то учинили, најпре морамо наћи средине узорка по стратумима:

$\bar{y}_1 = \frac{250+330+\cdots+305}{8} = 270.625$

$\bar{y}_2 = 446.000$

$\bar{y}_3 = 183.333$

$\bar{y}_4 = 89.000$

$\bar{y}_5 = 64.000$

Оцена просечне вредности продате робе је:

$$\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h = \frac{50}{120}\,270.625 + \frac{30}{120}\,446.000 + \ldots + \frac{10}{120}\,64.000 = 267.5659$$

Оцене варијанси стратума су:

$$s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{N_h}(y_{1_i} - \bar{y}_1)^2$$

$$= \frac{1}{8-1}\,[(250 - 270.625)^2 + \ldots + (305 - 270.325)^2] = 4274.5536$$

$s_2^2 = 2780.0000$

$s_3^2 = 433.3333$

$s_4^2 = 242.0000$

$s_5^2 = 72.0000$

Оцена варијансе добијене оцене средине је једнака:

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h\,(N_h - n_h)\,\frac{s_h^2}{n_h}$$

$$= \frac{1}{120^2}\,[50\,(50 - 8)\,\frac{4274.5536}{8} + \ldots + 10\,(10 - 2)\,\frac{72.0000}{2}] = 111.1626$$

Оцена укупне вредности продаје је:

$$\hat{Y}_{st} = N\,\overline{y}_t = 120 * 267.5659 = 32108$$

2.3 Оптималан распоред

Одређивање обима узорка n, и обима по стратумима n_h може се вршити тако што се минимизира $V(\bar{y}_{st})$ за фиксне трошкове или тако што се минимизирају трошкови за фиксно $V(\bar{y}_{st})$.

У трошкове спадају издаци за избор узорка, прикупљање и обраду података, и многи други. Деле се на сталне трошкове и трошкове везане за испитивање јединица по стратумима. Функција трошкова дата је следећом једначином:

$$C = c_0 + \sum_{h=1}^{L} c_h n_h \ ,$$

где су C укупни, c_0 стални, а c_h трошкови по јединици узорка у стратуму.

Теорема 2.6 У стратификованом случајном узорку са претходно датом функцијом трошкова, варијанса оцене средине $\bar{y}_{st}$ је минимална када је n_h пропорционално са $\frac{n_h S_h}{\sqrt{c_h}}$.

Доказ:

Уз услов: $\sum_{h=1}^{L} c_h n_h = C - c_0$, тражимо минимум функције:

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}(1 - f_h) = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h}$$

Применом методе Лагранжових мултипликатора за одређивање екстрема функције, добијамо n_h и мултипликатор λ, тако да функција:

$$F(n_h, \lambda) = V(\bar{y}_{st}) + \lambda\left(\sum_{h=1}^{L} c_h n_h + c_0 - C\right)$$

буде минимална. Даље тражимо извод функције $F(n_h, \lambda)$ по n_h:

$$-\frac{W_h^2 S_h^2}{n_h^2} + \lambda c_h = 0, \quad h = 1, \dots, L$$

Очигледно важи: $n_h\sqrt{\lambda} = \frac{W_h S_h}{\sqrt{c_h}}$

Након сабирања по стратумима добија се:

$$n\sqrt{\lambda} = \sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{c_h}}$$

Количник претходних двеју једначина је:

$$\frac{n_h}{n} = \frac{W_h S_h/\sqrt{c_h}}{\sum_{h=1}^{L}(W_h S_h/\sqrt{c_h})} = \frac{N_h S_h/\sqrt{c_h}}{\sum_{h=1}^{L}(N_h S_h/\sqrt{c_h})}$$

Ако је фиксирана варијанса, заменом n_h у формулу за $V(\bar{y}_{st})$ добија се:

$$n = \frac{\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)}{V + \left(\frac{1}{N}\right)\sum_{h=1}^{L} W_h S_h^2} \ ,$$

а ако су фиксирани трошкови, добија се:

$$n = \frac{(C-c_0)\sum_{h=1}^{L}(N_h S_h \sqrt{c_h})}{\Sigma(N_h S_h \sqrt{c_h})}$$

□

У специјалном случају, када су трошкови по јединици исти у свим стратумима ($c_h = c$), важи да је $C = c_0 + cn$, а оптималан распоред за фиксиране трошкове се своди на оптималан распоред за фиксиран обим узорка.

Теорема 2.7 У стратификованом случајном узорку $V(\bar{y}_{st})$ има минималну вредност за фиксно *n* ако важи:

$$n_h = n\frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h} = n\frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}$$

Овај распоред се назива Neyman-ов распоред.

Минимална вредност $V(\bar{y}_{st})$ за фиксно n је:

$$V_{min}(\bar{y}_{st}) = \frac{1}{n}(\sum_{h=1}^{L} W_h S_h)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

Пример 2.2

Нека је популација од $N = 12$ елемената подељена на два стратума, чији су елементи (2, 3, 5, 7, 8, 11) и (12, 14, 17, 19, 25, 33). Сада ћемо, по оптималном распореду, изабрати узорак од $n = 6$ елемената, ако знамо да су трошкови по јединици узорка $c_1 = 4$ и $c_2 = 9$, а стандардне девијације $S_1 = 3.34$ и $S_2 = 7.88$.

Дакле, треба одредити величине узорка по стратумима:

$$n_1 = n\ \frac{N_1 S_1/\sqrt{c_1}}{\sum_{h=1}^{L}(\ N_h S_h/\sqrt{c_h})} = 6\ \frac{6 * 3.34/2}{6 * 3.34/2 + 6 * 7.80/3} = 2.33\ \approx 2$$

$$n_2 = n\ \frac{N_2 S_2/\sqrt{c_2}}{\sum_{h=1}^{L}(\ N_h S_h/\sqrt{c_h})} = 6\ \frac{6 * 7.80/3}{6 * 3.34/2 + 6 * 7.80/3} = 3.66\ \approx 4$$

У изузетно ретким случајевима може се, приликом прављења оптималног распореда, десити да се укаже потреба за тим да узорак унутар једног стратума буде већи од самог стратума ($n_1 > N_1$). У том случају узимамо да је $\tilde{n}_1 = N_1$. Применом формуле:

$$\tilde{n}_h = (n - N_1)\frac{W_h S_h}{\sum_{h=2}^{L} W_h S_h}\ ,\quad h \geq 2$$

обезбеђујемо да, за свако $h \geq 2$, важи: $\tilde{n}_h \leq N_h$. Ако би се десило да је $\tilde{n}_2 > N_2$, тада бирамо:

$$\tilde{n}_1 = N_1\ ;\quad \tilde{n}_2 = N_2\ ;\quad \tilde{n}_h = (n - N_1 - N_2)\frac{W_h S_h}{\sum_{h=3}^{L} W_h S_h}$$

и тако обезбеђујемо да, за свако $h \geq 3$, важи: $\tilde{n}_h \leq N_h$.

Овај поступак настављамо све дотле док не постигнемо да за свако h важи: $\tilde{n}_h \leq N_h$.

У овом случају можемо применити формулу за варијансу оцене средине дату код Neyman–овог распореда, али израз за минималну вредност те варијансе не важи. Тада примењујемо нови израз:

$$V_{min}(\bar{y}_{st}) = \frac{1}{n'}(\sum_{h=1}^{'L} W_h S_h)^2 - \frac{1}{N}\sum_{h=1}^{'L} W_h S_h^2$$

где n' представља величину целокупног узорка.

Међутим, нисмо увек у могућности да применимо оптималан распоред. И у таквим случајевима портебно је оценити одговарајуће параметре, као и величину узорка. Очекивано је да свака оцена има одређену варијансу *V*. Али, у појединим случајевима, позната је граница грешке *d* ($V = (d/t)^2$), где је *t* стандардно одступање.

Ако је *V* варијанса оцене средине, узмимо да је s_h оцена за S_h, и нека је $n_h = w_h n$, за изабрано w_h. Тада је:

$$V = \frac{1}{n}\sum_h \frac{W_h^2 s_h^2}{w_h} - \frac{1}{N}\sum_h W_h s_h^2$$

Одатле добијамо израз за величину узорка:

$$n = \frac{\sum_h \frac{W_h^2 s_h^2}{w_h}}{V + \frac{1}{N}\sum_h W_h s_h^2}$$

Ако је *n* занемариво, тада важи:

$$n = \frac{n_0}{V + \frac{1}{NV}\sum_h W_h s_h^2}$$

Специјално, за узорак са пропорционалним распоредом важи:

$$w_h = W_h = N_h/N\ ; \qquad n_0 = \frac{\sum_h W_h s_h^2}{V}\ ; \qquad n = \frac{n_0}{1+\frac{n_0}{N}}$$

У случају када је *V* варијанса оцене тотала, имамо:

$$n = \frac{\sum_h \frac{N_h^2 s_h^2}{w_h}}{V + \sum_h N_h s_h^2}$$

А за узорак са пропорционалним распоредом важи:

$$n_0 = \frac{N}{V}\sum_h N_h s_h^2\ ; \qquad n = \frac{n_0}{1+\frac{n_0}{N}}$$

Веома је важно да поведемо рачуна о томе колико варијнса опада када број стратума расте, и колико расту трошкови истраживања са порастом броја стратума. Претпоставимо да се први стратум конструише на основу вредности обележја *y*. Посматрајмо један од најједноставнијих случајева, када *y* има униформну расподелу на интервалу (*a*, *a* + *d*). Тада S_y^2, пре стратификације, износи $d^2/12$, тако да са простим случајним узорком величине *n* варијанса средине обележја износи: $V(\bar{y}) = d^2/12n$. Ако су свих L стратума исте величине, варијанса унутар стратума је $S_{yh}^2 = d^2/12L^2$. Дакле, за стратификовани узорак у коме је $W_h = 1/L$ и $n_h = n/L$, важи:

$$V(\bar{y}_{st}) = \frac{1}{n}(\sum_{h=1}^{L} W_h S_{yh})^2 = \frac{1}{n}\left(\sum_{h=1}^{L} \frac{1}{L}\frac{d}{\sqrt{12}L}\right)^2 = \frac{d^2}{12nL^2} = \frac{V(\bar{y})}{L^2}$$

Дакле, када имамо униформну расподелу, варијанса оцене средине обележја узорка опада обрнуто пропорционално расту квадрата броја стратума.

Што се тиче трошкова истраживања, најчешће коришћена трошковна функција, коју је дао Даленијус и која приказује зависност трошкова од величине стратума, дата је следећим изразом:

$$C = LC_s + nC_n$$

Индекс трошкова, C_s/C_n, зависи од врсте истраживања. Већи број стратума изискује веће ангажовање око прикупљања података, планирања и спровођења анализе и приказивања резултата, па самим тим и веће трошкове. У неким случајевима раст трошкова условљен повећањем броја стратума је велики, док је у ређим случајевима незнатан. Када је раст трошкова исувише велики, не исплати се повећавати број стратума, упркос мањој варијанси, односно већој прецизности која се на тај начин постиже.

2.4 Избор стратификованог узорка за пропорције

У случају да желимо да оценимо удео јединица из одређене класе *K* у читавој популацији, идеална стратификација би била она код које постоје само два стратума, при чему би у првом стратуму биле смештене јединице које припадају класи *K*, а у другом све остале јединице. Међутим, то није могуће извести, па се праве стратуми у којима се разликује удео јединица из класе *K*.

Нека је $P_h = \frac{A_h}{N_h}$ удео јединица из класе *K* у *h*–том стратуму, а $p_h = \frac{a_h}{n_h}$ удео таквих јединица у узорку изабраном из *h*–тог стратума. Оцена за удео јединица из класе *K* у читавој популацији, која одговара стратификованом узорку, дата је следећим изразом:

$$p_{st} = \sum_h \frac{N_h p_h}{N}$$

Теорема 2.8 Код стратификованог случајног узорка, варијанса за p_{st} износи:

$$V(p_{st}) = \frac{1}{N^2}\sum_h \frac{N_h^2(N_h - n_h)}{N_h - 1}\frac{P_h Q_h}{n_h}$$

Доказ: Ово је специјалан случај теореме о варијанси оцене средине, тако да важи:

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_h N_h^2(N_h - n_h)\frac{S_h^2}{n_h}$$

Знамо да важи:

$$S_h^2 = \frac{N_h}{N_h - 1}P_h Q_h \ ,$$

па одатле дирекно следи тврђење. □

У већини случајева вредност $1/N_h$ је занемарива, па важи:

$$V(p_{st}) = \frac{1}{N^2}\sum_h N_h(N_h - n_h)\frac{P_h Q_h}{n_h} = \sum_h \frac{W_h^2 P_h Q_h}{n_h}(1 - f_h)$$

Последица 2.8.1 Ако је могуће занемарити корективни фактор популације, тада је :

$$V(p_{st}) = \sum_h \frac{W_h^2 P_h Q_h}{n_h}$$

Последица 2.8.2 За стратификовани узорак са пропорционалним распоредом важи:

$$V(p_{st}) = \frac{N-n}{N}\frac{1}{nN}\sum_h \frac{N_h^2 P_h Q_h}{N_h - 1} = \frac{1-f}{n}\sum_h W_h P_h Q_h$$

Што се тиче одређивања величине узорка по стратумима, уколико примењујемо принцип минималне варијансе за фиксирану величину целог узорка, важи:

$$n_h \approx N_h\sqrt{N_h/(N_h - 1)}\sqrt{P_h Q_h} = N_h\sqrt{P_h Q_h}$$

Дакле:

$$n_h = n\frac{N_h\sqrt{P_h Q_h}}{\sum_h N_h\sqrt{P_h Q_h}}$$

Уколико примењујемо принцип минималне варијансе за фиксне трошкове, важи:

$$n_h = n\frac{N_h\sqrt{P_h Q_h/c_h}}{\sum_h N_h\sqrt{P_h Q_h/c_h}}$$

Ако су трошкови по јединици једнаки у сваком стратуму, предност стратификованог случајног узорка у односу на прост случајни узорак је минимална, осим у случају када се удео јединица из класе *K* значајно разликује од стратума до стратума. Оптималан распоред има малу предност у однасу на пропорционалан распоред када је удео јединица класе *K* у стратумима између 0.1 и 0.9.

2.5 Последице одступања од оптималног распореда

Није увек могуће постићи оптималан распоред, и одступња се не могу избећи. Зато је неопходно знати какве ефекте може проузроковати одступање од оптималног распореда, како би се избегло значајније умањење прецизности.

Нека је величина узорка у *h*–том стратуму:

$$n_h' = n\frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}, \qquad (1)$$

а минимална варијанса:

$$V_{min}(\bar{y}_{st}) = \frac{1}{n}(\sum_{h=1}^{L} W_h S_h)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 \qquad (2)$$

У пракси нам n_h није познато, тако да можемо само да апроксимирамо овакав распоред. Ако је $\hat{n}_h$ апроксимирана величина узорка из *h*–тог стратума, малопређашњи израз за варијансу добија следећи облик:

$$V(\bar{y}_{st}) = \frac{1}{\hat{n}_h}(\sum_{h=1}^{L} W_h S_h)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

Повећање варијансе, узроковано одступањем од оптималног распореда, износи:

$$V(\bar{y}_{st}) - V_{min}(\bar{y}_{st}) = \frac{1}{\hat{n}_h}(\sum_{h=1}^{L} W_h S_h)^2 - \frac{1}{N}(\sum_{h=1}^{L} W_h S_h)^2$$

Увршатавањем израза (1) у претходни израз, добија се:

$$V(\bar{y}_{st}) - V_{min}(\bar{y}_{st}) = \frac{(\sum_{h=1}^{L} W_h S_h)^2}{n^2}\left(\sum_{h=1}^{L}\frac{n_h'^2}{\hat{n}_h} - n\right) = \frac{(\sum_{h=1}^{L} W_h S_h)^2}{n^2}\sum_{h=1}^{L}\frac{(\hat{n}_h - n_h'^2)^2}{\hat{n}_h}$$

Ако занемаримо корективни фактор популације, на основу (2) имамо:

$$\frac{V_{min}(\bar{y}_{st})}{n} = \frac{(\sum_{h=1}^{L} W_h S_h)^2}{n^2}$$

Због тога што је пропорционални раст варијансе настао као последица одступања од оптималног распореда, важи:

$$\frac{V(\bar{y}_{st}) - V_{min}(\bar{y}_{st})}{V_{min}(\bar{y}_{st})} = \frac{1}{n}\sum_{h=1}^{L}\frac{(\hat{n}_h - n_h')^2}{\hat{n}_h}, \qquad (3)$$

где је $\hat{n}_h$ актуелна, а n_h' оптимална величине узорка у *h*–том стратуму. Ако занемаримо корективни фактор популације, важиће:

$$\frac{V(\bar{y}_{st}) - V_{min}(\bar{y}_{st})}{V_{min}(\bar{y}_{st})} \geq \frac{1}{n}\sum_{h=1}^{L}\frac{(\hat{n}_h - n_h')^2}{\hat{n}_h}$$

Нека је $g_h = |\hat{n}_h - n'_h|/\hat{n}_h$ апсолутна разлика величина узорака у h–том стратуму. Тада израз (3) добија облик:

$$\frac{V-V_{min}}{V_{min}} = \sum_{h=1}^{L} \frac{\hat{n}_h}{n} g_h^2$$

Претходни израз представља тежинску средину за g_h^2. Према томе, горња граница за $(V - V_{min})/ V_{min}$ је g^2, где је g највећа пропорционална разлика у било ком стратуму. Дакле, ако је g = 0.2=20%, пропорционални раст варијансе не може прекорачити $(0.2)^2$, односно 4%. Ако је g = 30%, он не може бити већи од 9%.

За жељени тип стратификације, величина стратума, као ни други параметри, није увек прецизно одређена, па се користе разне оцене. Тако, уместо стварне тежине стратума W_h, користимо њену оцену w_h. Узорачке оцене су пристрасне. Због пристрасности, тачност оцене рачунамо помоћу њене средње квадратне грешке, што је бољи начин него да тачност рачунамо помоћу варијансе оцене средине. Пристрасност је константна, без обзира на раст величине узорка. У таквим случајевима, често се достиже она величина узорка за коју је оцена мање прецизна него за прост случајни узорак. На тај начин се губи прецизност постигнута стратификацијом. Уобичајена оцена $s(\bar{y}_{st})$ потцењује стварну грешку за $\bar{y}_{st}$, јер не садржи пристрасност. Због свега овога, оцена популацијске средине је: $\bar{Y} = \sum w_h \bar{Y}_h$, док пристрасност износи $\sum(w_h - W_h)\bar{Y}_h$. Средња квадратна грешка, помоћу које рачунамо тачност оцене, дата је следећим изразом:

$$MSE(\bar{y}_{st}) = \sum \frac{w_h^2 S_h^2}{n_h}(1 - f_h) + [\sum(w_h - W_h)\bar{Y}_h]^2$$

2.6 Проблем распореда приликом оцењивања више параметара истовремено

Оптималан распоред за оцењивање једног параметра у већини случајева није оптималан за оцењивање других параметара. Због тога је неопходно постићи компромис у прављењу распореда када се оцењује више параметара. Најпре се мора редуковати број параметара које желимо да оцењујемо, при чему одбацујемо све параметре мањег значаја, а у разматрање узимамо само најважније. На основу познатих података, можемо направити оптималне распореде за сваки параметар посебно, а затим одредити колики је степен неслагања. Код одређених посматрања постоји могућност велике корелације између параметара, а у том случају су разлике између њихових оптималних распореда релативно мале.

Посматрајмо случај када су дати пропорционални распоред, оптимални распоред за сваки параметар посебно и компромисни распоред за све параметре. Нека је m_h средња вредност величине узорака по одговарајућем стратуму за све параметре и нека је дата варијанса оцене средине обележја популације по јединици за сва три распореда, редом:

$$v_{prop} = \frac{\sum W_h s_h^2}{n}, \qquad v_{opt} = \frac{(\sum W_h s_h)^2}{n}, \qquad v_{comp} = \frac{(\sum W_h S_h)^2}{m_h}$$

Не тако ретко, применом компромисног распореда добијамо резултате готово исте прецизности као и када смо у могућности да применимо оптималан распоред за све параметре. Оно што је још занимљивије јесто то што, у појединим случајевима, ни прецизност постигнута применом пропорционалног распореда не заостаје значајно у односу на индивидуалне оптималне распореде и компромисни распоред.

Постоје и друге методе распоређивања када оцењујемо више параметара. Једна од њих јесте да изаберемо величину узорка по стратуму, n_h, тако да минимизирамо аритметичку средину пропорционалног раста варијансе оцене средине обележја популације по параметру:

$$n_h = n\sqrt{\sum_j n_{jh}'^2} \,/\, \sum_h \sqrt{n_{jh}'^2}\,,$$

где је $n_{jh}'^2$ оптимална величина узорка у стратуму h за параметар j.

У одређеним случајевима оптимални распореди индивидуалних параметара се толико разликују да не постоји могућност проналажења компромисног распореда. Тада је потребно применити одређена правила за одређивање распореда. Један од начина је посматрање у одређеном смеру, где се губитак до кога је дошло услед погрешног распореда може изразити у новцу, односно где се посматра директан утицај грешака на трошкове. У овом случају, укупни очекивани губитак дат је следећим изразом:

$$L = \sum_j^k a_j V(\bar{y}_{jst}) = \sum_j^k a_j \sum_h^L W_h^2 S_{jh}^2 (\frac{1}{n_h} - \frac{1}{N_h})$$

где је S_{jh}^2 варијанса *j*–тог параметра у *h*–том стратуму. Заменом редоследа сума, претходни израз добија облик:

$$L = \sum_h^L \frac{W_h^2}{n_h} (\sum_j^k a_j S_{jh}^2) - \frac{1}{N} \sum_h^L W_h (\sum_j^k a_j S_{jh}^2)$$

Минимизацијом производа (C - c_0) из функције трошкова и првог израза у функцији L (оног који зависи од n_h), и применом Коши-Шварцове неједнакости, добијамо:

$$n_h \approx \frac{W_h}{\sqrt{c_h}} \sqrt{\sum_j a_j S_{jh}^2}$$

Посматрајмо случај када је вредност функције L одређена, и када можемо занемарити корективни фактор популације. Тада је величина узорка по стратуму:

$$n_h = \frac{n(W_h A_h/\sqrt{c_h})}{\sum(W_h A_h/\sqrt{c_h})}$$

где је . Величина целог узорка је, у овом случају:

$$n = \frac{1}{L}\left(\sum_h \frac{W_h A_h}{\sqrt{c_h}}\right)(\sum_h W_h A_{h\sqrt{c_h}})$$

2.7 Поређење стратификованог узорка и простог случајног узорка

Стратификација је сложен поступак, а да би се видело колико је оправдан, неопходно је упоредити прецизност простог случајног узорка са прецизношћу стратификованог узорка (са пропорционалним и оптималним распоредом). Иако се применом стратификованог узорка постижу изузетно прецизне оцене, не може се рећи да су варијансе код сратификованог узорка увек мање него код простог случајног узорка. Када је величина узорка унутар стратума далеко од оптималне, оне могу бити и знатно веће. У изузетно ретким случајевима се дешава да, чак и онда када је оптималан распоред постигнут, варијансе оцена буду велике. Ипак, у општем случају није тако.

Нека су:

V_{ran} - варијанса оцене средина за прост случајни узорак

V_{prop} - варијанса оцене средине за стратификовани узорак са пропорционалним распоредом

V_{opt} - варијанса оцене средине за стратификовани узорак са оптималним распоредом

Теорема 2.9 Нека је распоред оптималан за фиксирано n ($n_h \approx N_h S_h$). Ако претпоставимо да је $n_h/N_h \approx 0$, тада важи:

$$V_{opt} \leq V_{prop} \leq V_{ran}$$

Доказ: Занемарићемо 1- n/N (корективни фактор основног скупа. Тада је:

$$V_{ran} = \frac{S^2}{n}; \qquad V_{prop} = \frac{\sum_{h=1}^{L} N_h S_h^2}{nN}; \qquad V_{opt} = \frac{(\sum_{h=1}^{L} N_h S_h)^2}{nN^2}$$

Важи:

$$\begin{aligned}(N-1)S^2 &= \sum_h \sum_i (y_{hi} - \bar{Y})^2 \\ &= \sum_h \sum_i (y_{hi} - \bar{Y}_h)^2 + \sum_h N_h (\bar{Y}_h - \bar{Y})^2 \\ &= \sum_h (N_h - 1) S_h^2 + \sum_h N_h (\bar{Y}_h - \bar{Y})^2\end{aligned}$$

Претходни израз се може записати и као:

$$N\left(1 - \frac{1}{N}\right) S^2 = \sum_h N_h (1 - \frac{1}{N_h}) S_h^2 + \sum_h N_h (\bar{Y}_h - \bar{Y})^2$$

Знамо да је $1/N_h \approx 0$, па важи:

$$NS^2 = \sum_h N_h S_h^2 + \sum_h (\bar{Y}_h - \bar{Y})^2$$

Одатле следи да је:

$$\frac{S^2}{n} = \frac{\sum_{h=1}^{L} N_h S_h^2}{nN} + \frac{\sum_{h=1}^{L} N_h (\bar{Y}_h - \bar{Y})^2}{nN},$$

односно:

$$V_{ran} = V_{prop} + \frac{\sum_{h=1}^{L} N_h (\bar{Y}_h - \bar{Y})^2}{nN}$$

По дефиницији варијансе оцене средине за стратификовани узорак са оптималним распоредом важи да је $V_{opt} \leq V_{prop}$. Такође знамо да је:

$$V_{prop} - V_{opt} = \frac{1}{nN}\left[\sum_h N_h S_h^2 - \frac{\left(\sum_{h=1}^{L} N_h S_h\right)^2}{N}\right] = \frac{1}{nN}\sum_h N_h \left(S_h - \bar{S}\right)^2,$$

где је: $\bar{S} = \frac{\sum_{h=1}^{L} N_h S_h}{N}$.

На основу последњих двеју једначина, имамо да је:

$$V_{ran} = V_{opt} + \underbrace{\frac{\sum_h N_h \left(S_h - \bar{S}\right)^2}{nN}}_{\text{A}} + \underbrace{\frac{\sum_{h=1}^{L} N_h (\bar{Y}_h - \bar{Y})^2}{nN}}_{\text{B}}$$

Компоненте А и В опадају у варијанси кад прелазимо са простог случајног узорка на оптималан распоред. Компонента А долази од елиминација ефекта разлика између стандардних девијација стратума и представља разлику у варијанси између пропорционалног и оптималног распореда, а компонента В од елиминација између средина стратума.

Уколико није могуће занемарити корективни фактор основног скупа, сличним поступком долазимо до следећег израза:

$$V_{ran} = V_{prop} + \frac{N-n}{nN(N-1)}\left[\sum_h N_h (\bar{Y}_h - \bar{Y})^2 - \frac{1}{N}\sum_h (N - N_h) S_h^2\right]$$

Одатле следи да пропорционална стратификација даје већу варијансу ако је:

$$\sum_h N_h (\bar{Y}_h - \bar{Y})^2 < \frac{1}{N}\sum_h (N - N_h) S_h^2 \qquad \square$$

Најбоље би било када бисмо могли вршити стратификацију на основу оне величине коју желимо да оценимо. Тада не би било преклапања међу стратумима, а варијанса унутар стратума би била знатно мања. У пракси то, у већини случајева, није изводљиво, али могуће је пажљиво одабрати критеријум стратификације, тако да се оствари значајан напредак у прецизности. Да би се то постигло, неопходно је узети у обзир чињеницу да се јединице од којих се састоји популација значајно разликују по величини, те да су најважније променљиве, чије се вредности одређују, тесно повезане са величинама јединица. Такође је неопходно прецизно одредити величине јединица и стратума. Тако су, на пример, разлике у пребукираности у болницама прилично велике између болница са великим капацитетима и оних са малим бројем кревета. Али могуће је и да пребукираност буде већа у болницама већег капацитета, јер се оне обично налазе у већим местима, док

мале болнице, у мањим местима, нису толико оптерећене, јер је мањи и број пацијената. Дакле, овде на посматрани параметар посредно утиче и величина града, односно број пацијената, а не само величина (капацитет) болнице.

У неким случајевима величина није толико битна, али тада је неопходно прецизно одредити неки други параметар, кључан у датој ситуацији. Пребукираност у болницама често зависи и од њихове намене, што нема везе са величином. Ако величина јединица не варира у времену, што је чест случај када се посматрају краћи временски периоди, најбоље је посматрати ону величину која је актуелна у време вршења мерења.

Када се јединице значајно разликују по величини, стратификација са пропорционалним распоредом не даје добре резултате, јер је варијанса већа када су јединице веће. Такође је неопходно узети у обзир и трошкове спровођења истраживања, јер понекад су трошкови испитивања одређених елемената узорка сувише велики да би били оправдани прецизношћу која се постигне.

2.8 Стратификација са малим узорцима

Претпоставимо да имамо два критеријума стратификације, на основу којих је популација подељена на R редова и C колона, као што је приказано у табели 2.1 на следећој страни. На тај начин добијамо RC ћелија. Ако је $n \geq RC$, свака ћелија може бити заступљена у узорку. Проблем настаје када је $n < RC$. Будући да тада не могу све ћелије бити заступљене, морамо водити рачуна да оба критеријума стратификације буду подједнако заступљена.

Један од једноставнијих начина распоређивања којим се подједнака заступљеност може постићи захтева да је $n > R$ и $n > C$. Овај метод ћемо илустровати на примеру мале популације од 165 школа, која се стратификује на основу величине места у коме се налази школа, и на основу просечног трошка по ученику. На основу првог критеријума популација се дели на 5 стратума, а на основу другог на 4 стратума, што је такође приказано у табели 2.1. Овде је, дакле, $R = 5$, а $C = 4$. Нека је m_{ij} број школа у ћелији, а $P_{ij} = m_{ij}/165$ удео школа у свакој од 20 ћелија. Најбоље би било дати свакој школи једнаке шансе да буде изабрана у узорак, при том дајући једнаке шансе и обема класама.

Табела 2.1

Величина града	Трошак по ученику					Укупно		n_i
		A	B	C	D			
I	m_{1j} P_{1j}	15 0.091	21 0.127	17 0.103	9 0.055	m_{ij} P_{ij}	62 0.376	4
II	m_{2j} P_{2ij}	10 0.061	8 0.049	13 0.079	7 0.042	m_{ij} P_{ij}	38 0.231	2
III	m_{3j} P_{3j}	6 0.036	9 0.055	5 0.030	8 0.049	m_{ij} P_{ij}	28 0.170	2
IV	m_{4j} P_{4j}	4 0.024	3 0.018	6 0.036	6 0.036	m_{ij} P_{ij}	19 0.114	1
V	m_{5j} P_{5j}	3 0.018	2 0.012	5 0.030	8 0.049	m_{ij} P_{ij}	18 0.109	1

Укупно	$m_{.j}$ $P_{.j}$ $n_{.j}$	38 0.230 2	43 0.261 3	46 0.278 3	38 0.231 2	165 1.000

Узмимо да је n = 10. Сада рачунамо и , и добијене вредности на одговарајући начин заокружујемо на најближи цео број, као што је приказано у табели. Након тога правимо нову табелу (2.2), помоћу које, на случајан начин, бирамо ћелије у узорак.

Табела 2.2

		1 2 *A*	3 4 5 *B*	6 7 8 *C*	9 10 *D*
1 2 3 4	I	X 	 X 	 X 	
5 6	II	 	 	X 	 X
7 8	III	 	X 	 	 X
9	IV		X		
10	V				X

Најбољи начин је да у сваком реду изаберемо по једну ћелију на случајан начин. При томе морамо водити рачуна да се не понављају колоне. Дакле, избор можемо вршити користећи прост случајни узорак без понављања. Вероватноћа избора ћелије ij је $n_i n_j / P_{ij}$. Ове вероватноће нису међусобно једнаке, осим у случају када је $P_{ij} = n_i n_j / n^2$. Изабрану ћелију означавамо са X. Непристрасна оцена средине обележја по школи је:

$$\bar{y}_U = \frac{1}{n} \sum \frac{n^2 P_{ij}}{n_i n_j} y_{ij}$$

2.9 Формирање стратума

Приликом формирања стратума, треба водити рачуна да стратумске варијансе оцењиваног обележја буду што мање. Зато стратуми треба да буду што хомогенији унутар себе. Размотримо одређивање граница међу стратумима код узорка са понављањем, применом пропорционалног и Neyman-овог распореда.

Нека су y_0 и y_L најмања,односно највећа вредност промењљиве y на нивоу популације. Тражимо границе стратума, $y_1, \dots, y_{L-1}$, тако да $V(\bar{y}_{st})$ буде минимална. Нека је $f(y)$ функција фреквенције обележја популације у дискретном, односно функција густине расподеле у непрекидном случају. Тада је релативна фреквенција *h*–тог стратума:

$$W_h = \int_{y_{h-1}}^{y_h} f(t)dt$$

Функција густине расподеле обележја *h*–тог стратума је:

$$f_h(y) = \begin{cases} \frac{f(y)}{W_h}, y \in [y_{h-1}, y_h] \\ 0, \quad y \notin [y_{h-1}, y_h] \end{cases}, \quad h = 1, \dots, L$$

Тада су средња вредност и варијанса *h*–тог стратума редом:

$$\bar{Y}_h = \int_{y_{h-1}}^{y_h} t f_h(t) dt = \frac{1}{W_h} \int_{y_{h-1}}^{y_h} t f(t) dt$$

$$S_h^2 = \int_{y_{h-1}}^{y_h} (t - \bar{Y}_h)^2 f_h(t) dt = \frac{1}{W_h} \int_{y_{h-1}}^{y_h} (t - \bar{Y}_h)^2 f(t) dt = \frac{1}{W_h} \int_{y_{h-1}}^{y_h} t^2 f(t) dt - \bar{Y}_h^2$$

Оцена средине и њена варијанса код стратификованог случајног узорка дате су са:

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$$

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h}$$

У случају пропорционалног распореда имамо да је $n_h = n\frac{N_h}{N} = nW_h$, тако да важи:

$$V(\bar{y}_{st}) = \frac{1}{n} \sum_{h=1}^{L} W_h S_h^2 - \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2$$

Будући да користимо узорак са понављањем,можемо занемарити корективни фактор коначне популације, па је довољно минимизирати следећу суму:

$$\sum_{h=1}^{L} W_h S_h^2 = \sum_{h=1}^{L} \left[\int_{y_{h-1}}^{y_h} t^2 f(t) dt - W_h \bar{Y}_h^2 \right]$$

Фактор y_h се појављује само у једном делу претходног израза:

$$G = -W_h \bar{Y}_h^2 - W_{h+1} \bar{Y}_{h+1}^2 = \frac{(W_h \bar{Y}_h)^2}{W_h} - \frac{(W_{h+1} \bar{Y}_{h+1})^2}{W_{h+1}}$$

Знамо да важи:

$$W_h = \int_{y_{h-1}}^{y_h} f(t) dt; \qquad \frac{\partial W_h}{\partial y_h} = f(y_h)$$

$$W_{h+1} = \int_{y_h}^{y_{h+1}} f(t) dt; \qquad \frac{\partial W_{h+1}}{\partial y_h} = -f(y_h)$$

Такође важи:

$$W_h\bar{Y}_h = \int_{y_{h-1}}^{y_h} tf(t)dt; \qquad \frac{\partial(W_h\bar{Y}_h)}{\partial y_h} = y_h f(y_h)$$

$$W_{h+1}\bar{Y}_{h+1} = \int_{y_h}^{y_{h+1}} tf(t)dt; \qquad \frac{\partial(W_{h+1}\bar{Y}_{h+1})}{\partial y_h} = -y_h f(y_h)$$

За $y_h = \frac{\bar{Y}_{h+1}+\bar{Y}_h}{2}$ важи:

$$\frac{\partial G}{\partial y_h} = \frac{\partial\left(-W_h\bar{Y}_h^2 - W_{h+1}\bar{Y}_{h+1}^2\right)}{\partial y_h} = -\frac{\partial\left(\frac{\left(W_h\bar{Y}_h^2\right)}{W_h}\right)}{\partial y_h} - \frac{\partial\left(\frac{\left(W_{h+1}\bar{Y}_{h+1}^2\right)}{W_{h+1}}\right)}{\partial y_h}$$

$$= \frac{\frac{\partial\left(W_h\bar{Y}_h^2\right)}{\partial y_h}W_h - \left(W_h\bar{Y}_h\right)^2\frac{\partial W_h}{\partial y_h}}{W_h^2} - \frac{\frac{\partial\left(W_{h+1}\bar{Y}_{h+1}^2\right)}{\partial y_h}W_{h+1} - \left(W_{h+1}\bar{Y}_{h+1}\right)^2\frac{\partial(W_{h+1})}{\partial y_h}}{W_{h+1}^2}$$

$$= -\frac{2W_h\bar{Y}_h y_h f(y_h)W_h - f(y_h)W_h^2\bar{Y}_h^2}{W_h^2} - \frac{2W_{h+1}\bar{Y}_{h+1}(-y_h)f(y_h)W_{h+1} + f(y_h)W_{h+1}^2\bar{Y}_{h+1}^2}{W_{h+1}^2}$$

$$= -2\bar{Y}_h y_h f(y_h) - f(y_h)\bar{Y}_h^2 + 2y_h f(y_h)\bar{Y}_{h+1} - f(y_h)\bar{Y}_{h+1}^2$$

$$= 2y_h f(y_h)\left(\bar{Y}_{h+1} - \bar{Y}_h\right) - f(y_h)(\bar{Y}_{h+1}^2 - \bar{Y}_h^2)$$

$$= 2\frac{\left(\bar{Y}_{h+1}+\bar{Y}_h\right)}{2}f(y_h)\left(\bar{Y}_{h+1} - \bar{Y}_h\right) - f(y_h)(\bar{Y}_{h+1}^2 - \bar{Y}_h^2)$$

$$= f(y_h)(\bar{Y}_{h+1}^2 - \bar{Y}_h^2) - f(y_h)(\bar{Y}_{h+1}^2 - \bar{Y}_h^2) = 0$$

Пошто $\bar{Y}_h$ зависи од y_h, тачке y_h, где је h = 1,2,...,L-1, се одређују итеративним поступком.

У случају Neyman-овог распореда имамо да је:

$$n_h = n\frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h},$$

као и:

$$V(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

Пошто и овде користимо узорак са понављањем, након занемаривања корективног фактора, преостаје да нађемо минимум суме $\sum_{h=1}^{L} W_h S_h$. Вредност y_h се овде појављује само у члановима $W_h S_h + W_{h+1} S_{h+1}$, тако да важи:

$$\frac{\partial\left(\sum_{h=1}^{L} W_h S_h\right)}{\partial y_h} = \frac{\partial W_h}{\partial y_h}S_h + W_h\frac{\partial S_h}{\partial y_h} + \frac{\partial W_{h+1}}{\partial y_h}S_{h+1} + W_{h+1}\frac{\partial S_{h+1}}{\partial y_h}$$

Знамо да важи:

$$S_h^2 = \int_{y_{h-1}}^{y_h} (t - \overline{Y}_h)^2 \, f_h\,(t)dt = \frac{1}{W_h} \int_{y_{h-1}}^{y_h} (t - \bar{Y}_h)^2 \, f(t)dt \ ,$$

па следи:

$$W_h S_h^2 = \int_{y_{h-1}}^{y_h} (t - \overline{Y}_h)^2 \, f(t)dt$$

$$\frac{\partial (W_h S_h^2)}{\partial y_h} = (y_h - \bar{Y}_h)^2 \, f(y_h) \qquad (1)$$

Такође важи:

$$\frac{\partial (W_h S_h^2)}{\partial y_h} = S_h^2 \frac{\partial W_h}{\partial y_h} + 2 W_h S_h \frac{\partial S_h}{\partial y_h} \ ,$$

односно :

$$\frac{\partial (W_h S_h^2)}{\partial y_h} = S_h^2 f(y_h) + 2 W_h S_h \frac{\partial S_h}{\partial y_h} \qquad (2)$$

Изједначавањем десних страна једначина (1) и (2) добија се:

$$(y_h - \bar{Y}_h)^2 f(y_h) = S_h^2 f(y_h) + 2 W_h S_h \frac{2 S_h}{2 y_h} \,,$$

односно:

$$W_h \frac{\partial S_h}{\partial y_h} = \frac{1}{2} f(y_h) \frac{(y_h - \bar{Y}_h)^2 - S_h^2}{S_h}$$

Аналогно долазимо до израза:

$$W_{h+1} \frac{\partial S_{h+1}}{\partial y_h} = -\frac{1}{2} \, f(y_h) \frac{(y_h - \bar{Y}_{h+1})^2 - S_{h+1}^2}{S_{h+1}}$$

Заменом добијених вредности у једначину:

$$\frac{\partial (\sum_{h=1}^{L} W_h S_h)}{\partial y_h} = \frac{\partial W_h}{\partial y_h} S_h + W_h \frac{\partial S_h}{\partial y_h} + \frac{\partial W_{h+1}}{\partial y_h} S_{h+1} + W_{h+1} \frac{\partial S_{h+1}}{\partial y_h} = 0,$$

добија се израз:

$$\frac{(y_h - \bar{Y}_h)^2 + S_h^2}{S_h} = \frac{(y_h - \bar{Y}_{h+1})^2 + S_{h+1}^2}{S_{h+1}}, \qquad h = 1, \ldots L - 1$$

Будући да $\bar{Y}_h$ и S_h^2 зависе од y_h , ове једначине није могуће решити алгебарски, па се траже апроксимативна решења.

Постоји више начина за проналажење апроксимативних решења, а један од тих начина је Dalenius - Hodges-ов поступак. Нека је:

$$Z_h = \int_{y_0}^{y_h} \sqrt{f(t)}\ dt$$

У случају да је број стратума велики, $f(t)$ је приближно једнако константи f_h унутар датог сратума. Даље важи:

$$W_h\ = \int_{y_{h-1}}^{y_h} f(t)dt = f_h\,(y_h - y_{h-1})\ ,$$

$$S_h\ = \frac{1}{\sqrt{12}}(y_h - y_{h-1})\ ,$$

$$Z_h - Z_{h-1} = \int_{y_{h-1}}^{y_h} \sqrt{f(t)}\,dt = \sqrt{f_h}\,(y_h - y_{h-1})$$

Заменом ових вредности добија се:

$$\sqrt{12}\ \sum_{h=1}^{L} W_h S_h = \sum_{h=1}^{L} f_h\ (y_h\ - y_{\,h-1})^2 = \sum_{h=1}^{L} (\,Z_h - Z_{h-1})^2$$

Сума $\sum_{h=1}^{L}(\,Z_h - Z_{h-1})^2$ има минималну вредност када је $Z_h - Z_{h-1}$ константно, јер је $Z_L - Z_0$ фиксирано. За дато $f(y)$ се формирају кумулативне функције за функцију $\sqrt{f(y)}$, па се тачке y_h бирају тако да чине константне интервале на кумуланти $cum\sqrt{f(t)}$.

Пример 2.3

Ако је функција густине расподеле обележја популације *Y* дата са:

$$f(y) = \begin{cases} \frac{y}{32}, & y \in [0,8] \\ 0, & y \notin [0,8] \end{cases},$$

одредићемо границе стратума за $L = 2$, према Neyman-овом распореду.
Границе стратума се одређују тако да важи:

$$\frac{(y_h - \bar{Y}_h)^2 + S_h^2}{S_h} = \frac{(y_h - \bar{Y}_{h+1})^2 + S_{h+1}^2}{S_{h+1}}, \qquad h = 1, \ldots L-1$$

Будући да $\bar{Y}_h$ и S_h^2 зависе од y_h, ове једначине се не могу алгебарски решити. Због тога се границе стратума y_h налазе тако што се полази од произвољних граница, које се коригују до оптималних вредности.

Узмимо да је $y_h = 3$. Израчунаћемо средњу вредност $\bar{Y}_h$ и варијансу S_h^2 у оба стратума.

Из формула:

$$\bar{Y}_h = \frac{1}{W_h} \int_{y_{h-1}}^{y_h} y\, f(y) dy$$

и

$$S_h^2 = \frac{1}{W_h} \int_{y_{h-1}}^{y_h} y^2 f(y) dy - \bar{Y}_h^2 ,$$

где је $W_h = \int_{y_{h-1}}^{y_h} f(y) dy$, добијамо:

$$W_1 = \int_0^3 \frac{y}{32} dy = \frac{9}{64}, \; W_2 = 1 - W_1 = \frac{55}{64}$$

$$\bar{Y}_1 = \frac{64}{9} \int_0^3 y \, \frac{y}{32} dy = \frac{2}{9} \int_0^3 y^2 \, dy = 2$$

$$\bar{Y}_2 = \frac{64}{55} \int_3^8 y \, \frac{y}{32} dy = \frac{2}{55} \int_3^8 y^2 \, dy = 5.8787$$

$$S_1^2 = \frac{64}{9} \int_0^3 y^2 \frac{y}{32} dy - 2^2 = 0.5008$$

$$S_2^2 = \frac{64}{55} \int_3^8 y^2 \frac{y}{32} dy - 5.8787^2 = 1.9398$$

Важи:

$$\frac{(y_h - \bar{Y}_1)^2 + S_1^2}{S_1} = \frac{(3-2)^2 + 0.5008}{\sqrt{0.5008}} = 2.1209$$

$$\frac{(y_h - \bar{Y}_2)^2 + S_2^2}{S_2} = \frac{(3 - 5.8787)^2 + 1.9398}{\sqrt{1.9398}} = 7.3430$$

Ове вредности се, очигледно, значајно разликују, па $y_h = 3$ није добар избор за границу стратума. Зато бирамо $y_h = 4$ и $y_h = 5$ и понављамо поступак. Добијамо резултате дате у наредној табели.

y_h	$\bar{Y}_1$	S_1^2	$\bar{Y}_2$	S_2^2	$\frac{(y_h - \bar{Y}_1)^2 + S_1^2}{S_1}$	$\frac{(y_h - \bar{Y}_2)^2 + S_2^2}{S_2}$
3	2	0.5008	5.8787	1.9398	2.1209	7.3430
4	2.67	0.887	6.222	1.2842	2.8482	5.49
5	3.33	1.3888	6.6154	0.7637	3.5450	3.8989

На основу података из табеле можемо закључити да је за границу стратума најбоље узети вредност $y_h = 5$. Дакле, стратуми су (0,5) и (5,8).

2.10 Постстратификација

Постстратификација представља стратификацију после узимања узорка, односно поступак поделе по стратумима узорка који је већ изабран по неком другом плану. На тако добијене стратуме примењују се оцене за стратификован узорак. Постстратификација се примењује код анализирања оних параметара код којих се величине стратума и распоред могу одредити тек након избора узорка. Неки од таквих параметара су старосно доба, пол, расна припадност, итд.

Величине узорака по стратумима код постстратификације су, за разлику од стратификованог узорка, случајне променљиве. Важно је да и релативна фреквенција сваког стратума (N_h/N) буде позната, или оцењена на одговарајући начин.

Ако се ради о узорку са пропорционалним распоредом, величина узорка у сваком стратуму је фиксна и једнака: $n_h = n\,N_h / N$, док је варијанса оцене средине једнака:

$$V(\bar{y}_{st}) = \frac{N-n}{Nn}\sum_{h=1}^{L}\frac{N_h}{N}S_h^2$$

Узорак који се постстратификацијом добије из простог случајног узорка има приближно пропорционалан распоред, јер је величина узорка у h-том стратуму:

$$n_h = nN_h/N$$

Нека су дате стратификоване оцене средине и тотала:

$$\bar{y}_{st} = \sum_{h=1}^{L}\frac{N_h}{N}\bar{y}_h$$

$$\hat{Y}_{st} = N\bar{y}_{st}$$

Тада су њихове варијансе:

$$V(\bar{y}_{st}) \approx \frac{N-n}{Nn}\sum_{h=1}^{L}\frac{N_h}{N}S_h^2 + \frac{1}{n^2}\frac{N-n}{N-1}\sum_{h=1}^{L}\frac{N-N_h}{N}S_h^2$$

$$V\left(\hat{Y}_{st}\right) = N^2V(\bar{y}_{st})$$

Први члан на десној страни израза за варијансу оцене средине представља варијансу оцене $\bar{y}_{st}$ за пропорционалан распоред, док се, због случајних величина узорка, други члан додаје варијанси постстратификацијом.

Када је реч о стратификацији елемената одабраних простим случајним узорком, долази у обзир и следећа оцена средине обележја узорка:

$$\bar{y}_w = \sum W_h\bar{y}_h\ ,$$

где је $W_h = N_h/N$. Нека је m_h број јединица из h-тог стратума, које припадају узорку.

Тада израз за варијансу средине обележја узорка добија облик:

$$V(\bar{y}_W) = \sum \frac{W_h^2 S_h^2}{m_h} - \frac{1}{N} \sum W_h S_h^2$$

Неопходно је израчунати средњу вредност варијансе $V(\bar{y}_W)$ за различите величине узорка. Понекад се може десити да је $m_h = 0$. У таквим случајевима, пре оцењивања морамо укомбиновати два или више стратума, што за последицу може имати мање прецизну оцену. Ипак, када је узорак већи, шансе да из неког стратума ниједна јединица не буде изабрана у узорак су минималне. Ако такве случајеве занемаримо, важиће:

$$E\left(\frac{1}{m_h}\right) = \frac{1}{nW_h} + \frac{1-W_h}{n^2 W_h^2}$$

Дакле, важи:

$$E\big(V(\bar{y}_W)\big) = \frac{1-f}{n} \sum W_h S_h^2 + \frac{1}{n^2} \sum (1 - W_h) S_h^2$$

Први израз одговара вредности $V(\bar{y}_{st})$ за пропорционалну стратификацију. Други представља раст варијансе услед тога што елементи узорка нису пропорционално распоређени по стратумима.

Уопштење претходно наведеног дато је следећим изразом:

$$\frac{1}{n^2} \sum (1 - W_h) S_h^2 = \frac{1}{n} \left(\frac{L}{n}\right) \bar{S}_h^2 - \frac{1}{n^2} \sum W_h S_h^2 = \frac{1}{n\bar{n}_h} \bar{S}_h^2 - \frac{1}{n^2} \sum W_h S_h^2 \ ,$$

где $\bar{S}_h^2$ представља средњу вредност за S_h^2, а $\bar{n}_h = n/L$ средину величине стратума. Ако се S_h^2 не разликују значајно, варијанса оцене средине код пропорционалне стратификације расте $(L-1)/L\bar{n}_h$ пута, уз занемаривање корективног фактора популације. Раст варијансе ће бити мали уколико је $\bar{n}_h$ довољно велико.

2.11 Узорковање по принципу квоте

Метода која се веома често користи у истраживањима тржишта и јавног мњења подразумева да је величина узорка по стратумима унапред израчуната, тако да је стратификација пропорционална. У том случају, истраживање се врши дотле док се не достигне одређена квота. Најпогоднији параметри за стратификацију су географско подручје, животно доба, пол, расна или национална припадност и поједини економски параметри. Ако бисмо, на пример, на случајан начин бирали особе унутар географских подручја, па их онда разврставали у одговарајуће стратуме, та метода би била идентична стратификованом случајном узорку. Важно је да теренски рад буде доста заступљен приликом испуњавања квота.

У већини случајева, ипак, узорковање по принципу квоте може се сматрати стратификованим узорковањем са мање или више „неслучајним“ избором јединица из стратума. Због тога се код узорка изабраног по принципу квота не могу поуздано применити оцене коришћене код стратификованог случајног узорка.

2.12 Оцена побољшања прецизности

Понекад је од значаја да се, након избора стратификованог случајног узорка, оцени побољшање прецизности у односу на прост случајни узорак. Подаци који су нам познати из узорка јесу $N_h, n_h, \bar{y}_h$ и s_h^2. Знамо да је оцена варијансе тежинске средине стратификованог узорка дата са:

$$v(\bar{y}_{st}) = \sum \frac{W_h^2 s_h^2}{n_h} - \sum \frac{W_h s_h^2}{N}$$

Проблем је упоредити ову варијансу са варијансом оцене средине добијеном из простог случајног узорка. Једна од процедура које се за то користе, подразумева рачунање сродне средње квадратне грешке на основу узорачке средине:

$$s^2 = \frac{\sum(y_{hi} - \bar{y})^2}{n-1}$$

при чему игноришемо стратуме. Ово се узима као оцена за s^2, тако да је $\hat{V}_{ran} = (N-n)/Nn$ оцена средине простог случајног узорка. Ова метода се може успешно применити за пропорционалан распоред, ако је и прост случајни узорак расподељен приближно пропорционално између стратума. Али, уколико се не ради о пропорционалном распореду, ова метода није примењива. У таквим случајевима користимо следећу теорему.

Теорема 2.9 За дате резултате за стратификовани случајни узорак, непристрасна оцена варијансе оцене средине за прост случајни узорак је:

$$V_{ran} = \frac{(N-n)}{n(N-1)} \left[\frac{1}{N} \sum_{h=1}^{L} \frac{N_h}{n_n} \sum_{j=1}^{n_h} y_{hj}^2 - \bar{y}_{st}^2 + v(\bar{y}_{st})\right]$$

при чему је $v(\bar{y}_{st})$ уобичајена непристрасна оцена за $V(\bar{y}_{st})$.

Доказ Познато нам је да важи:

$$V_{ran} = \frac{(N-n)}{nN} S^2 = \frac{(N-n)}{n(N-1)} \left[\frac{1}{N} \sum_{h=1}^{L} \sum_{j=1}^{N_h} y_{hj}^2 - \bar{Y}^2\right]$$

Одатле је:

$$\frac{1}{N} E\left(\sum_{h=1}^{L} \frac{N_h}{n_n} \sum_{j=1}^{n_h} y_{hj}^2\right) = \frac{1}{N} \sum_{h=1}^{L} \sum_{j=1}^{N_h} y_{hj}^2$$

С обзиром на то да су $v(\bar{y}_{st})$ и $\bar{y}_{st}$ непристрасне оцене за $V(\bar{y}_{st})$ и $\bar{Y}$, важи:

$$Ev(\bar{y}_{st}) = V(\bar{y}_{st}) = E(\bar{y}_{st}^2) - \bar{Y}^2 \ ,$$

а одатле је $\bar{y}_{st}^2 - \ v(\bar{y}_{st})$ непристрасна оцена за $\bar{Y}^2$.

Из претходних једначина следи да је непристрасна оцена за V_{ran} управо:

$$V_{ran} = \frac{(N-n)}{n(N-1)}\left[\frac{1}{N}\sum_{h=1}^{L}\frac{N_h}{n_n}\sum_{j=1}^{n_h} y_{hj}^2 - \bar{y}_{st}^2 + v(\bar{y}_{st})\right]$$

што је и требало доказати. □

За пропорционалан распоред $N_h/n_h = N/n$, израз из претходног тврђења добија облик:

$$v_{ran} = \frac{(N-n)}{n(N-1)}\left[\frac{(n-1)}{n}s^2 + v(\bar{y}_{st})\right]$$

За велике вредности , можемо сматрати да је и , па у том случају претходни израз добија облик:

$$v_{ran} = \frac{(N-n)}{nN}s^2$$

Ово важи за пропорционалан распоред, док је у општем случају:

$$v_{ran} = \frac{(N-n)}{nN}\left[\frac{1}{N}\sum_{h=1}^{L}\frac{N_h}{n_n}\sum_{j=1}^{n_h} y_{hj}^2 - \bar{y}_{st}^2\right]$$

2.13 Оцена варијансе код једноелементних стратума

Ако је популација таква да постоји већи број ефективних критеријума који би се могли применити за стратификацију, може се извршити таква стратификација да сваки стратум садржи само једну јединицу. У том случају није могуће применити раније коришћене оцене за за тотал и средину обележја. Уколико је број стратума паран, оцењивање се може вршити тако што се стратуми благовремено упарују како бисмо добили што уједначеније тотале стратума.

Посматрајмо парове узорака са обележјима y_{j1} и y_{j2}, где је $j = 1, \dots L/2$. Оцењиваћемо тотале стратума, $\hat{Y}_{j1} = N_{j1}y_{j1}$ и $\hat{Y}_{j2} = N_{j2}y_{j2}$. Сада је:

$$\hat{Y}_{j1} - \hat{Y}_{j2} = \left(Y_{j1} - Y_{j2}\right) + \left(\hat{Y}_{j1} - Y_{j1}\right) - (\hat{Y}_{j2} - Y_{j2})$$

Средња вредност на нивоу оба узорка *j*-тог пара је:

$$E(\hat{Y}_{j1} - \hat{Y}_{j2})^2 = \left(Y_{j1} - Y_{j2}\right)^2 + N_{j1}(N_{j1} - 1)S_{j1}^2 + N_{j2}(N_{j2} - 1)S_{j2}^2$$

За варијансу оцене тотала обележја популације користимо оцену:

$$v_1(\hat{Y}_{st}) = \sum_{j=1}^{L/2}(\hat{Y}_{j1} - \hat{Y}_{j2})^2 \qquad (1)$$

Очекивана вредност ове величине је:

$$Ev_1(\hat{Y}_{st}) = \sum_{h=1}^{L} N_h(N_h - 1)S_h^2 + \sum_{j=1}^{L/2}(\hat{Y}_{j1} - \hat{Y}_{j2})^2$$

У претходној једначини, први израз на десној страни представља тачну варијансу, док други израз представља позитивнну пристрасност, која зависи од успешности избора парова стратума, чији се тотали незнатно разликују.

За непарно L, најмање једна група мора бити различита од 2. Тада израз (1) добија облик:

$$v_2(\hat{Y}_{st}) = \sum_{j=1}^{G} \frac{L_j}{L_j - 1} \sum_{j=1}^{L_j} (\hat{Y}_{jk} - \hat{Y}_j / L_j)^2 \ ,$$

где је број група било које величине , а оцена тотала за групу j. За , када је , претходни израз је идентичан изразу (1).

Када је позната помоћна променљива A_h, за сваки стратум за који је потребно израчунати тотал Y_h, користимо следећу оцену за варијансу:

$$v_2(\hat{Y}_{st}) = \sum_{j=1}^{G} \frac{L_j}{L_j - 1} \sum_{j=1}^{L_j} (\hat{Y}_{jk} - A_{jk} \hat{Y}_j / A_j)^2$$

Ако је помоћу променљиве A_h добро процењен тотал стратума, позитивна пристрасност, у претходној једначини за v_2 изражена преко $(\hat{Y}_{jk} - A_{jk} \hat{Y}_j / A_j)^2$, биће мања у односу на одговарајући израз у v_1.

Постоји и метода конструкције стратума помоћу које се гарантовано добија непристрасна оцена варијансе оцене тотала, при чему су сви стратуми једноелементни. Посматрајмо најједноставнији случај, када је $N/n = N/L = k$, где је k цео број. На случајан начин бирамо цео број r, који се налази између 1 и k. Први стратум садржи јединице под редним бројевима од (r + 1) до (r + k), други садржи једнице под редним бројевима од (r + k + 1) до (r + 2k), и тако даље, до последњег стратума који асдржи јединице под редним бројевима од r + (n - 1)k +1 до N = nk. На први поглед, чини се да за последњи стратум нема великог избора. Међутим, ова метода даје одличне резултате када се врши стратификација у разним географским истраживањима.

2.14 Стратуми као предмет проучавања

У овом одељку ћемо се укратко позабавити методама чији је примарни циљ поређење различитих стратума. Као што смо већ видели, правила по којима се одређује величина стратума и распоред јединица се разликују у зависности од врсте истраживања. Такође нам је познато да се оцене код стратификованог узорка значајно разликују у односу на све остале методе које подразумевају оцењивање на нивоу целе популације.

Ако посматрамо само два стратума, из њих бирамо узорке величине n_1 и n_2, тако да варијанса разлике $(\bar{y}_1 - \bar{y}_2)$ буде минимална. Ако занемаримо коректив фактор популације, биће:

$$V(\bar{y}_1 - \bar{y}_2) = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

Нека је функција трошкова дата са:

$$C = c_0 + c_1 n_1 + c_2 n_2$$

Тада је варијанса минимална ако су испуњени следећи услови:

$$n_1 = \frac{\frac{nS_1}{\sqrt{c_1}}}{S_1/\sqrt{c_1}+S_2/\sqrt{c_2}} \qquad n_2 = \frac{\frac{nS_2}{\sqrt{c_2}}}{S_1/\sqrt{c_1}+S_2/\sqrt{c_2}}$$

Када посматрамо L стратума ($L > 2$), оптималан распоред зависи од захтеване прецизности за различита поређења. На пример, трошкови могу бити минимални ако је задовољен сет од $L(L - 1)/2$ услова да је $V(\bar{y}_1 - \bar{y}_2) \leq V_{hi}$, при чему вредности бирамо на основу тога колика прецизност нам је потребна у поређењу h–тог и i–тог стратума.

Посматраћемо једну од једноставнијих метода распоређивања, која се може успешно применити када се вредности S_h и c_h не разликују значајно (за различито h). Ова метода подразумева минимизацију просечне варијансе разлике између $L(L - 1)/2$ парова стратума. Дакле, минимизирамо вредност следећег израза:

$$\bar{V} = \frac{2}{L}\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} + \cdots + \frac{S_L^2}{n_L}\right)$$

Вредност $\bar{V}$ је минимална када је, за константну вредност трошкова, задовољен следећи услов:

$$n_h \approx \frac{S_h}{\sqrt{c_h}}$$

Испуњавањем овог услова може се постићи изузетно висока прецизност приликом поређења парова стратума.

Други начин јесте да изаберемо n_h тако да варијанса разлике има исту вредност за сваки пар стратума, и износи $\sqrt{V}$. У том случају, за сваки стратум важи да је $S_h^2/n_h = V/2$. За фиксне трошкове, ова метода је мање прецизна у односу на претходну. За две врсте оптималног распореда добијамо:

$$\bar{V} = \frac{2(\sum S_h\sqrt{c_h})^2}{L(C-c_0)}, \qquad V = \frac{2(\sum S_h^2 c_h)}{(C-c_0)}$$

На основу Коши-Шварцове неједнакости важи да је $V > \bar{V}$, осим у случају када је $S_h\sqrt{c_h}$ константно. Уколико је V много веће од $\bar{V}$, у већини случајева можемо, из неколико покушаја, успешно одредити просечну варијансу, која се не разликује много од $\bar{V}$, а да притом $V(\bar{y}_h - \bar{y}_i)$ остане приближно константно за сваки пар стратума.

Понекад је пожељно наћи одговарајуће оцене за сваки стратум, а не само за читаву популацију. Тада се постављају додатна два услова:

$$V(\bar{y}_h) = \frac{S_h^2}{n_h}(1 - f_h) \leq V_h, \qquad V(\bar{y}_{st}) = \sum \frac{N_h^2 S_h^2}{n_h}(1 - f_h) \leq V$$

У ове изразе је укључен и корективни фактор популације, зато што је циљ поступка да се одреди прецизност са којом ће се одредити оцене за читаву популацију.

3. Примене стратификованог узорка

Као што смо раније споменули, стратификовани узорак има широку примену у разним истраживањима. Често се користи, на пример, приликом испитивања јавног мњења, али и у разним привредним, географским, демографским и и другим истраживањима. Тако се, приликом спровођења анкета пред изборе, ради процене изборних резултата, Србија дели у шест стратума. То су: Београд, Централна Србија, Војводина, Рашка област, Косово и Метохија и општине Прешево, Медвеђа и Бујановац, које чине један стратум. Можемо закључити да су, у овом случају, коришћени различити критеријуми за стратификацију. Прва три стратума формирана су на основу броја становника, односно броја бирача. Покрајина Косово и Метохија се посматра као посебан стратум, упркос малом броју бирача, због тренутних политичких прилика. Преостала два стратума формирана су на основу етничке структуре становништва.

Овде су критеријуми стратификације комбиновани на тај начи како процена изборних резултата не би битније одступала од остварених резултата. Ако би критеријуми по којима се врши стратификација били погрешно одабрани, десило би се да узорак изабран из стратума не буде довољно репрезентативан, што би свакако имало за последицу велика одступања и нетачне резултате истраживања. У даљем тексту ћемо, на неколико примера показати како се врши стратификација, како се бира узорак и оцењују параметри, али и како избор критеријума за поделу популације на стратуме утиче на касније оцењивање параметара.

3.1 Оцена просечне нето зараде у Републици Србији на основу стратификованог узорка

Посматраћемо Републику Србију као једну популацију, чији су елементи општине у њеном саставу. Циљ нам је да, на основу репрезентативног узорка, добијеног помоћу стратификације, оценимо колика је просечна нето зарада на нивоу целе популације, односно целе Србије, на месечном нивоу. Дакле, посматрано обележје (карактеристика) је нето зарада. Познате су нам нето зараде за појединачне општине.

Сада желимо да одредимо репрезентативан узорак, који ће се састојати од одређеног броја општина, на основу чијих нето зарада ћемо оценити просечну месечну нето зараду у Србији у 2010-ој години. Располажемо следећим подацима: број становника за сваку општину, просечне нето зараде за 2008. и 2009. годину за сваку општину и за целу Србију и просечне нето зараде за сваку општину за 2010. годину. Најпре морамо одредити критеријум на основу кога ћемо популацију поделити на стратуме. Поједини критеријуми су добри за нека истраживања, а за нека нису. Прво ћемо посматрати случај када је као критеријум поделе одабран број становника општина, а затим случај када се популација дели на стратуме на основу просечне нето зараде из претходне године. Онда ћемо упоредити та два случаја и установити који критеријум је боље применити у датој ситуацији, тј. када оцењујемо просечну нето зараду.

Посматрамо, дакле, популацију од 159 општина, подељену у 8 стратума, као што је приказано у табели на следећој страни. Важно је напоменути да се у популацији не налазе све општине Републике Србије. Наиме, тачни подаци везани за број становника и просечне зараде нису познати за општине у саставу Аутономне покрајине Косово и Метохија, па су ове општине изузете из истраживања.

Табела 3.1

Редни број стратума (h)	Број становника	Број општина (N_h)
1	до 12000	16
2	12000 до 15000	21
3	15000 до 20000	24
4	20000 до 30000	24
5	30000 до 40000	18
6	40000 до 60000	21
7	60000 до 100000	15
8	Преко 100000	20
		N=159

Након што смо популацију изделили на стратуме, потребно је да одредимо величину узорка, као и број елемената који ће бити изабрани у узорак из сваког стратума. Будући да нам трошкови истраживања нису познати, применићемо Neyman–ов распоред. Изабраћемо узорак од n = 31 елемената. За одређивање броја елемената изабраних у узорак из h–тог стратума, користимо раније дефинисану формулу:

$$n_h = n\frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h} = n\frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}$$

Једина непозната величина у претходном изразу S_h, односно варијанса стратума (S_h^2). Њу рачунамо по формули:

$$S_h^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(y_{h_i} - \bar{Y}_h)^2 ,$$

при чему знамо да је $\bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{h_i}$. Резултати су приказани у следећој табели:

Табела 3.2

h	$\bar{Y}_h$	S_h^2	S_h	n_h
1	24268.50	8513503.467	2917.790854	2
2	25898.86	20419495.03	4518.793537	4
3	27181.54	36689787.30	6057.209531	5
4	27291.46	23604343.74	4858.430172	4
5	26802.00	13607178.00	3688.790859	3
6	31390.48	42904242.86	6550.133042	5
7	31576.47	33146925.70	5757.336684	3
8	34064.75	43507037.46	6595.986466	5

Сада, пошто смо одредили број елемената који се бирају у узорак из сваког стратума, применом методе простог случајног узорка без враћања, изабраћемо елементе узорка. Изабрани узорак приказан је у табели 3.3 на следећој страни.

Табела 3.3

Општина	Стратум (h)	Број становника	Просечна нето зарада у 2010-ој години (у динарима)
Димитровград	1	10587	25019
Ражањ	1	9655	25721
Сечањ	2	14652	26562
Осечина	2	12921	24377
Мало Црниће	2	12836	21828
Бојник	2	12133	25150
Тител	3	16577	23940
Чајетина	3	15350	25332
Мионица	3	15533	25978
Ада	3	18032	27068
Брус	3	17248	23193
Куршумлија	4	25607	16503
Мајданпек	4	20986	32104
Нови Бечеј	4	25095	26586
Сјеница	4	28301	28359
Петровац на Млави	5	32447	25563
Бачка Топола	5	36213	27909
Власотинце	5	31527	22061
Бор	6	52083	34301
Рума	6	58963	28663
Врачар	6	56380	45419
Вршац	6	53556	41617
Трстеник	6	46180	24092
Јагодина	7	75759	27634
Гроцка	7	85764	31347
Лозница	7	83915	25215
Краљево	8	139411	29095
Лесковац	8	155682	26733
Смедерево	8	116941	37524
Нови Сад	8	335381	30554
Нови Београд	8	225917	52654

Сада, на основу изабраног узорка, рачунамо оцену средине обележја популације по јединици ($\bar{y}_{st}$), односно оцену просечне месечне нето зараде у Републици Србији. За то рачунање користимо следећу формулу:

$$\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h$$

при чему је:

$$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{h_i}$$

Средине узорка за сваки стратум дате су у следећој табели.

Табела 3.4

h	1	2	3	4	5	6	7	8
$\bar{y}_h$	25370	24479	25102	25888	25178	34818	28065	35312

Одатле добијамо да је средина обележја популације по јединици $\bar{y}_{st} = 28020.98$. Дакле, према нашој оцени, просечна месечна нето зарада у Републици Србији у 2010-ој години је износила приближно 28021 динара.

Применом израза:

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h\,(N_h - n_h)\frac{S_h^2}{n_h}$$

добијамо да варијанса оцене средине обележја популације износи $V(\bar{y}_{st})$ = 698612.8357. Сада ћемо одредити 95% интервал поверења за средину обележја популације. За то користимо следећи израз:

$$\bar{y}_{st} \pm z\sqrt{V(\bar{y}_{st})}$$

С обзиром на то да је наш узорак обима већег од 30, узимамо да је z приближна вредност квантила реда 1 - α нормалне $\mathcal{N}(0,1)$ расподеле, односно z = 1.6449. Даљим рачунањем добијамо да је тражени интервал поверења (26646.14 , 29395.86).

С обзиром на чињеницу да располажемо потребним подацима за читаву популацију, можемо одредити и реализовану вредност просечне нето зараде у Републици Србији. Она износи 28529 динара и, као што видимо, та вредност припада 95% интервалу поверења који смо одредили. Дакле, оцену средине обележја популације добијену на овај начин можемо сматрати задовољавајућом.

Сада ћемо оценити просечну месечну зараду на основу новог стратификованог узорка, добијеног тако што је као критеријум стратификације одабрана просечна месечна нето зарада у претходној (2009-ој) години. Поступак је сличан претходном, посматрамо популацију од 159 општина и делимо је у 5 стратума, као што је приказано у табели 3.5.

Табела 3.5

Редни број стратума (h)	Просечна месечна нето зарада у 2009-ој години	Број општина (N_h)
1	до 20000	11
2	20000 до 25000	60
3	25000 до 30000	59
4	30000 до 40000	21
5	преко 40000	8
		N=159

У следећој табели дати су подаци за варијансу стратума у број елемената стратума који се бирају у узорак. Величина узорка је n = 31.

Табела 3.6

h	$\overline{Y_h}$	S_h^2	S_h	n_h
1	21329.09	11798870.09	3434.948399	3
2	24793.88	3524919.715	1877.476688	10
3	28636.78	3336238.968	1826.537426	9
4	36173.86	11993739.03	3454.524429	6
5	45578.50	11128755.14	3335.981886	2

Сада, применом методе простог случајног узорка без враћања, бирамо елементе узорка. Изабрани узорак је приказан у наредној табели.

Табела 3.7

Општина	Стратум (h)	Просечна нето зарада у 2009-ој години (у динарима)	Просечна нето зарада у 2010-ој години (у динарима)
Гаџин Хан	1	17069	18894
Бела Паланка	1	13357	19677
Рача	1	17996	21486
Младеновац	2	23639	26010
Крупањ	2	22649	23882
Жабари	2	24552	26597
Мало Црниће	2	20532	21828
Баточина	2	23709	24930
Ариље	2	20025	21886
Прибој	2	21187	24990
Александровац	2	20888	22723
Житорађа	2	22510	24891
Алибунар	2	24365	23007
Шид	3	27187	28695
Жабаљ	3	26647	28894
Врање	3	25245	27214
Ниш	3	27007	29132
Тутин	3	28095	26327
Лучани	3	27441	29651
Мајданпек	3	29672	32104
Свилајнац	3	25537	28709
Велика Плана	3	26958	28832
Шабац	3	28369	32422
Земун	4	39123	42902
Смедерево	4	33115	37524
Пожаревац	4	35907	37822
Кладово	4	31339	33397
Ужице	4	30542	32746

Вршац	4	37416	41617
Косјерић	5	41238	43103
Лазаревац	5	45257	47093

У табели 3.8 су дате средине узорка за сваки стратум.

Табела 3.8

h	1	2	3	4	5
$\bar{y}_h$	20019	24074	29198	37668	45098

На основу добијених података имамо да оцена средине обележја популације по јединици износи $\bar{y}_{st} = 28548.07$, односно да је просечна месечна нето зарада у Републици Србији у 2010-ој години износила приближно 28548 динара.

Аналогно претходном примеру, добијамо да варијанса ове оцене износи 120017.62, а одговарајући 95% интервал поверења (27957.17 , 29138.83). И у овом случају добијена оцена припада интервалу поверења и можемо је сматрати задовољавајућом.

Али, ако обратимо пажњу, приметићемо да је оцена коју смо добили из другог узорка приближнија реализованој вредности, него оцена добијена из првог узорка. Одатле можемо закључити да је бољи критеријум за стратификацију, у овом случају, просечна месечна нето зарада из претходне године у односу на посматрану, па овај критеријум и треба примењивати онда када је то могуће, односно када су нам познати одговарајући подаци. Међутим, када то није могуће, стратификовање на основу броја становника даје довољно прецизне резултате, највероватније због тога што број становника општина у одређеној мери зависи од развијености привреде, тј. код нас је чест случај да људи напуштају мање развијена места у потрази за запослењем. Овакве миграције, нарочито у последње време, у великој мери утичу на пораст броја становника у развијенијим градским центрима, односно пад броја становника у местима у којима је привреда замрла.

3.2 Оцене приноса појединих пољопривредних култура на основу стратификованог узорка

Наредним примером ћемо показати како се, помоћу узорка добијеног стратификацијом, може оценити принос једне пољопривредне културе на подручју Аутономне Покрајине Војводине. Познати су нам подаци о приносу ове културе за све општине Покрајине за 2009. и 2010. годину. Ми желимо да оценимо укупан годишњи принос на нивоу Покрајине. Дакле, у овом случају ћемо оцењивати тотал обележја популације. Популацију пердставља Аутономна Покрајина Војводина, а њени елементи су општине на територији Покрајине. Обележје популације представља годишњи принос посматране културе у 2010-ој години.

Овога пута ћемо извршити стратификацију популације на три начина, односно применом три различита критеријума, а потом ћемо одредити који критеријум је бољи када је у питању истраживање у области пољопривреде. Најпре ћемо популацију стратификовати према броју становника општина, као и у претходном примеру. Други критеријум стратификације ће нам бити просечна месечна нето зарада одговарајуће општине у 2010-ој години. Као трећи критеријум, узећемо принос посматране пољопривредне културе у претходној (2009-ој) години. Оцењиваћемо приносе пшенице.

Сада ћемо извршити стратификацију општина АП Војводине на основу броја становника. Популација се дели на 5 стратума, што је приказано у наредној табели.

Табела 3.9

Редни број стратума (h)	Број становника	Број општина (N_h)
1	до 15000	10
2	15000 до 25000	9
3	25000 до 35000	8
4	35000 до 55000	8
5	преко 55000	9
		N=45

Нека је узорак који бирамо величине n = 15. Сада морамо за сваки стратум израчунати колико елемената ће из тог стратума бити изабрано у узорак, након чега, методом простог случајног узорка бирамо елементе у узорак. Резултати су дати у табелама 3.10 и 3.11.

Табела 3.10

h	$\bar{Y}_h$	S_h^2	S_h	n_h
1	13579.90	103153546.10	10156.453	4
2	12214.78	59042307.440	7683.8992	3
3	15834.33	15963170.000	3995.3935	1
4	21405.67	48299033.500	6949.7506	2
5	37665.50	305625838.90	17482.158	5

Табела 3.11

Општина	Стратум (h)	Број становника	Годишњи принос пшенице (t)
Ириг	1	11818	6437
Опово	1	10690	6971
Нова Црња	1	11138	10636
Бачки Петровац	1	14082	9118
Беочин	2	16266	1960
Ада	2	18032	6110
Тител	2	16577	10051
Оџаци	3	38201	18971
Кикинда	4	63018	34367
Инђија	4	50928	19828
Панчево	5	127736	27275
Нови Сад	5	335381	25713
Зрењанин	5	128386	54249
Сомбор	5	92729	69481
Сремска Митровица	5	83982	29744

Оцена тотала обележја популације се одређује по формули:

$$\hat{Y}_{st} = \sum_{h=1}^{L} N_h \bar{y}_h = N\bar{y}_{st} \ ,$$

док се оцена средине обележја популације одређује раније описаним поступком. Средине узорка по стратуму дате су у табели 3.12.

Табела 3.12

h	1	2	3	4	5
$\bar{y}_h$	8290.50	6040.33	18971.00	27097.50	41292.40

Према овим подацима, оцена средине обележја популације износи $\bar{y}_{st} = 19604.97$, а оцена тотала обележја популације $\hat{Y}_{st} = 882223.70$. Дакле, можемо закључити да је годишњи принос пшенице на територији АП Војводине у 2010-ој години износио приближно 882223 тоне.

Применом израза:

$$V(\widehat{Y_{st}}) = \sum_{h=1}^{L} N_h \, (N_h - n_h) \frac{S_h^2}{n_h} ,$$

добијамо да је варијанса оцене тотала обележја популације изузетно велика и износи 6747836547. Сада тражимо одговарајући 90% интервал поверења, помоћу следећег израза:

$$\hat{Y}_{st} \pm z\sqrt{V\left(\hat{Y}_{st}\right)}$$

С обзиром на чињеницу да смо овога пута анализу вршили на основу узорка мањег од 30, z ће бити квантил t-расподеле са d степени слободе, где се d рачуна помоћу следећег израза:

$$d = \left(\sum_{h=1}^{L} a_h S_h^2\right)^2 / \left(\sum_{h=1}^{L} ((a_h S_h^2)^2/(n_h - 1))\right) \ ,$$

где је $a_h = N_h(N_h - n_h)/n_h$. Дакле, добијамо да је број степени слободе приближно $d = 8$. У том случају је $z = 1.397$, па је интервал поверења (767466.13 , 996979.87). Стварни годишњи принос пшенице у 2010-ој години био је 904037 тона. Видимо да ова вредност припада формираном интервалу поверења, па можемо бити задовољни оценом годишњег приноса коју смо добили из узорка добијеног стратификовањем АП Војводине на основу броја становника општина.

Сада ћемо принос пшенице оценити на основу узорка добијеног тако што је као критеријум стратификације одабрана просечна нето зарада у општинама за 2010. годину. Популација се дели на 3 стратума, а величина узорка ће такође бити $n = 15$. Поступак је идентичан претходном, а одговарајући резултати су преиказани у табелама 3.13, 3.14, 3.15, и 3.16.

Табела 3.13

Редни број стратума (h)	Просечна месечна нето зарада (у динарима)	Број општина (N_h)
1	до 25000	6
2	од 25000 до 30000	23
3	преко 30000	16
		N=45

Табела 3.14

h	$\bar{Y}_h$	S_h^2	S_h	n_h
1	15190.00	49387685.20	7027.6372	1
2	17966.61	86592995.43	9305.5357	6
3	23615.25	339157400.6	18416.227	8

Табела 3.15

Општина	Стратум (h)	Просечна нето зарада (у динарима)	Годишњи принос пшенице (t)
Алибунар	1	23007	16740
Мали Иђош	2	27909	12574
Опово	2	27384	6971
Бечеј	2	26316	18939
Пећинци	2	28663	14443
Бач	2	27070	5663
Оџаци	2	25310	20150
Врбас	3	31844	20790
Рума	3	31236	21540
Кањижа	3	31726	12273
Сомбор	3	30159	69481
Сремски Карловци	3	30554	258
Ириг	3	38271	6437
Беочин	3	40316	1960
Кикинда	3	30018	34367

Табела 3.16

h	1	2	3
$\bar{y}_h$	16740.00	13123.33	20888.25

На исти начин као у претходном примеру добијамо да је оцена средине обележја популације по јединици $\bar{y}_{st} = 16366.41$, а оцена тотала обележја популације $\hat{Y}_{st} = 736488.67$. Дакле, према нашој оцени, годишњи принос пшенице у АП Војводини је 736489 тоне. Варијанса је у овом случају такође изузетно велика и износи 12551125835.

На раније описан начин формирамо 90% интервал поверења. У овом случају, z ће бити квантил t-расподеле са 12 степени слободе, односно z = 1.356. Тако добијамо да је одговарајући интервал поверења (584573.87 , 888404.13). Видимо да, и поред изузетно великог интервала поверења, стварни принос пшенице не припада том интервалу. Дакле, оцена добијена оваквом стратификацијом је у незадовољавајућа и непримењива.

Сада ћемо стратификацију извршити тако што ћемо као критеријум узети принос пшенице у претходној (2009-ој) години. Популацију делимо на 5 стратума, бирамо узорак од 15 елемената, поступак је исти као у претходним примерима, резултати су дати у табелама 3.17, 3.18, 3.19 и 3.20.

Табела 3.17

Редни број стратума (h)	Принос пшенице у 2009-ој години (t)	Број општина (N_h)
1	до 10000	6
2	од 10000 до 20000	14
3	од 20000 до 30000	9
4	од 30000 до 40000	11
5	преко 40000	5
		N=45

Табела 3.18

h	$\overline{Y_h}$	S_h^2	S_h	n_h
1	5403.8330	14054710.97	3748.9613	1
2	12455.143	11937577.05	3455.0798	3
3	18219.889	4896721.361	2212.8536	1
4	26337.091	43335154.49	6582.9442	4
5	44346.800	368764406.2	19203.239	6

Видимо да се указала потреба да узорак изабран из петог стратума буде већи од самог стратума. О овој појави је било речи у поглављу 2.3. Због тога ћемо узети да је величина узорка из петог стратума $\tilde{n}_5 = N_5 = 5$. С обзиром на то да је пети стратум уједно и последњи, не постоји могућност да промена величине узорка изабраног из њега утиче на величине узорака изабраних из других стратума, тако да ће сада величина целог узорка бити $\tilde{n} = 14$.

Табела 3.19

Општина	Стратум (h)	Годишњи принос пшенице у 2009-ој годину (t)	Годишњи принос пшенице у 2010-ој години (t)
Ада	1	9745	6110
Ковачица	2	18684	16489
Стара Пазова	2	18486	15797
Мали Иђош	2	14481	12574
Жабаљ	3	28408	15139

Рума	4	32032	21540
Вршац	4	38428	38619
Бачка Паланка	4	30573	14461
Сечањ	4	35354	33228
Зрењанин	5	76472	54249
Бечеј	5	40325	18939
Сомбор	5	79424	69481
Суботица	5	58091	44698
Кикинда	5	40472	34367

Табела 3.20

h	1	2	3	4	5
$\bar{y}_h$	6110.00	14953.33	15139.00	26962.00	44346.80

Добијамо да оцена средине обележја популације износи $\bar{y}_{st} = 20012.75$, а оцена тотала обележја популације $\hat{Y}_{st} = 900573.67$. Дакле, можемо закључити да је годишњи принос пшенице на територији АП Војводине у 2010-ој години износио приближно 900574 тона. Варијанса ове оцене износи 1913898941. Тражимо 90% интервал поверења. z је квантил t–расподеле са приближно 5 степени слободе, и износи 1.476. Одговарајући интервал поверења је (836001.76 , 965146.24). Видимо да стварни принос пшенице припада интервалу поверења, па је и наша оцена задовољавајућа.

Ако упоредимо све три добијене оцене, видећемо да је стварној вредности приноса пшенице најприближнија она коју смо добили помоћу трећег узорка, код којег је као критеријумстратификације одабран принос пшенице у претходној години у односу на посматрану. То је логично, јер људи који се баве производњом пшенице обично дуже време не мењају културу. Али да смо посматрали индустријско биље, могло би се десити да ни овај критеријум стратификације не буде добар, јер је чест случај да пољопривредници који су једне године садили, на пример, шећерну репу, наредне године засаде соју, и сл. Због тога стратификација на основу приноса из претходне године вероватно не би дала репрезентативан узорак за оцењивање приноса репе у посматраној години.

Прва оцена добијена из узорка одабраног стратификацијом на основу броја становника је задовољавајуће, али ипак мање прецизна у односу на трећу оцену. С обзиром да је као популација посматрана регија у којој се становништво претежно бави пољопривредом, логично је да се у већим местима произведе више пшенице, јер велики градови у Војводини нису нужно и индустријски центри, у којима је пољопривреда слабије заступљена.

„Најлошију“ оцену дала је стратификација на основу висине просечних нето зарада. То је такође логично, јер је познато да су код нас зараде у пољопривреди знатно ниже него у индустрији, тако да је, у већини општина са већом просечном зарадом, знатно мање пољопривредног становништва, те је и принос пшенице мањи.

Литература

1. *William Gemmell Cochran*, Sampling Techniques, Wiley Series in Probability and Mathematical Statistics

2. Steven K. Thompson, Sampling, Wiley Series in Probability and Statistics

3. Steven K. Thompson, Stratified adaptive cluster sampling, Biometrika, *Biometrika (1991) 78 (2): 389-397. doi: 10.1093/biomet/78.2.389*

4. *Др Љиљана Петровић*, Теорија узорака и планирање експеримената, Универзитет у Београду, Економски факултет, 2009

5. Monroe G. Sirken, Stratified Sample Surveys with Multiplicity, Journal of the American Statistical Association, Vol 67, 1972, 224-227

6. *www.stat.gov.rs*, Републички завод за статистику, одељење у Новом Саду

Кратка биографија

Радослав Божић је рођен 14.11.1987. у Новом Саду. Основну школу "Јован Грчић Миленко" је завршио у Беочину, а гимназију "Јован Јовановић Змај" у Новом Саду. Основне студије је завршио на Природно математичком факултету у Новом Саду, смер Математичар математике финансија, 2010-те године.

Нови Сад, *јануар 2012.* *Радослав Божић*

УНИВЕРЗИТЕТ У НОВОМ САДУ
ПРИРОДНО - МАТЕМАТИЧКИ ФАКУЛТЕТ
КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА

Редни број:

РБР

Идентификациони број:

ИБР

Тип документације: Монографска документација

ТД

Тип записа: Текстуални штампани материјал

ТЗ

Врста рада: Мастер рад

ВР

Аутор: Радослав Божић

АУ

Ментор: Проф. Др Загорка Лозанов-Црвенковић

МН

Наслов рада: Стратификовани узорак и примене

МР

Језик публикације: Српски (ћирилица)

ЈП

Језик извода: с / е

ЈИ

Земља публиковања: Србија

ЗП

Уже географско подручје: Војводина

УГП

Година: 2012

ГО

Издавач: Ауторски репринт

ИЗ

Место и адреса: Нови Сад, Департман за математику и информатику, ПМФ, Трг Доситеја Обрадовића 4

МА

Физички опис рада: (3,65,0,27,0,0,0); бр. поглавља – 3, бр. страна – 65, бр. литерарних цитата – 0, бр. табела – 27, бр. слика – 0, бр. графика – 0, бр. прилога – 0

ФО

Научна област: Математика

НО

Научна дисциплина: Статистика

НД

Кључне речи: Стратификовани узорак

ПО

УДК:

Чува се: У библиотеци департмана за математику и информатику

ЧУ

Важна напомена:

ВН

Извод:

ИЗ

У овом раду је описан стратификовани узорак, као један од најчешће коришћених планова узорка, а такође је приказано и неколико примера његове примене.

Датум прихватања од стране НН већа: 07.12.2011

ДП

Датум одбране:

ДО

Чланови комисије:

КО

Председник: Др Љиљана Гајић, редовни професор Природно–математичког факултета у Новом Саду

Члан: Др Загорка Лозанов-Црвенковић, редовни професор Природно–математичког факултета у Новом Саду

Члан: Др Данијела Рајтер-Ћирић, ванредни професор Природно – математичког факултета у Новом Саду

UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCE
KEY WORDS DOCUMENTATION

Accession number:

ANO

Identification number:

INO

Document type: Monograph type

DT

Type of record: Printed text

TR

Contents Code:

CC

Author: Radoslav Božić

AU

Mentor: Prof. Dr Zagorka Lozanov-Crvenković

MN

Language of text: Serbian

LT

Language of abstract: English

LA

Contry of publication: Serbia

CP

Locality of publication: Vojvodina

LP

Publication year: 2012

PY

Publisher: Author's reprint

PU

Publ. place: Novi Sad, Department of mathematics and informatics, Faculty of Science, Squer Dositeja Obradovića 4

PP

Physical description: (3,65,0,27,0,0,0)

PD

Scientific field: Mathematics

SF

Scientific discipline: Statistics

SD

Key words: Stratified sample

SKW

UC:

Holding data: In library of Department of mathematics

HD

Note:

N

Abstract:

AB

In this paper we study stratified sample, as one of the most recently used sampling methods. There are also exposed some examples of it's application.

Accepted by the Scientific Board on: 07.12.2011

ASB

Defended:

DE

Thesis defend board:

DB

President: Dr Ljiljana Gajić, Full Professor, Faculty of Science and Mathematics, University of Novi Sad

Member: Dr Zagorka Lozanov-Crvenković, Full Professor, Faculty of Science and Mathematics, University of Novi Sad

Member: Dr Danijela Rajter-Ćirić, Associate Professor, Faculty of Science and Mathematics, University of Novi Sad