ТЕОРИЈА УЗОРАКА час 7

16. април ’14.

СТРАТИФИКОВАН УЗОРАК (*Stratified Sampling*)

- Стратификован узорак примењује се, пре свега, онда када је потребно повећати прецизност оцена параметара, односно смањити грешке узорка. Други разлози могу бити: економичност, једноставност истраживања и сл. У неким ситуацијама истраживање је могуће спроводити једино по стратумима.
- Стратификација (раслојавање) је подела популације на потпопулације, слојеве – стратуме (*strata*), при чему би требало формирати релативно хомогене, међу собом разграничене стратуме, што значи да вредности обележја, које је предмет истраживања, буду приближне на јединицама у сваком стратуму, а да се вредности обележја јединица из различитих стратума међусобно битно разликују.

- Читава популација се класификује у стратуме на основу неких додатних, претходно прикупљених информација. Као критеријум за стратификацију користи се нека (једна или више њих) карактеристика популације, за коју се сматра да је са посматраним обележјем у корелацији.
- Заправо, повећање прецизности оцене зависи од хомогености јединица у оквиру стратума и на њега, у великој мери, утиче начин стратификације. Зато се, природно, постављају питања: како формирати стратуме; како одредити број стратума; како распоредити узорак по појединим стратумима и сл.
- Након извршене стратификације узорци, унапред одређеног обима, бирају се унутар сваког стратума. При томе, узорци се бирају међусобно независно из различитих стратума и није неопходно користити исти план узорковања за све стратуме.

- Примери ситуација, код којих би било погодно користити стратификацију:
  - јединице популације: пољопривредна газдинства
    обележје: принос пшенице
    стратификација: укупна површина обрадивог земљишта по фарми
  - јединице популације: области у географским регионима
    обележје: број домаћинстава
    стратификација: густина насељености; рељеф
  - јединице популације: људи
    обележје: разна
    стратификација: пол; старост; образовање; верска припадност; етничка припадност; област живљења; социјално-економски фактори и сл.

Уопште, онда када је приметно да обележје од интереса има различите средње вредности у различитим подгрупама јединица популације.

- Захтева се:
  - стратуми морају бити међусобно дисјунктни, тј. свака јединица популације мора припадати тачно једном стратуму
  - стратуми морају "покривати" целу популацију, тј. не сме се појавити јединица која није укључена ни у један стратум
  - требало би да стратуми буду интерно хомогени, а да се међусобно значајно разликују
  - број стратума може бити већи или мањи, с тим што се због мерења прецизности оцене, захтева да број јединица у сваком стратуму не сме бити мањи од две јединице

Такође, инсистира се на томе да начин поделе популације на стратуме буде, што је више могуће, "природан".

- Предности:
  - могућност да се не само оцене параметри на целој популацији, него и да се донесу закључци на нивоу, тј. унутар самих стратума, и да се изврши поређење по стратумима
  - могућност да истраживач сам контролише величине узорка унутар сваког стратума
  - повећање прецизности оцене у смислу смањења дисперзије оцена (нпр. у односу на узорак *SRSWOR* истог обима)
  - повећање репрезентативности узорка, јер омогућава да елементи сваког стратума буду заступљени у финалном узорку
  - могућност да истраживач користи различите планове узорковања на различитима стратумима, у зависности од сопствених потреба и доступности информација
  - јефтиније је

- Недостаци:
  - врши се у складу са конкретним проблемом, уз претходно проучавање појаве и њене структуре. Стога, захтева велику количину претходних знања о популацији. Долазак до тих сазнања може представљати дуготрајан и скуп процес.
  - избор фактора по којима се врши стратификација може бити тежак, ако је истраживање комплексно и укључује велики број параметара
  - анализа података је комплексна, нарочито коректно оцењивање дисперзија оцена

What Are the Steps in Selecting a Stratified Sample?

There are eight major steps in selecting a stratified random sample:

1. Define the target population.
2. Identify stratification variable(s) and determine the number of strata to be used. The stratification variables should relate to the purposes of the study. If the purpose of the study is to make subgroup estimates, the stratification variables should be related to those subgroups. The availability of auxiliary information often determines the stratification variables that are used. More than one stratification variable may be used. However, in order to provide expected benefits, they should relate to the variables of interest in the study and be independent of each other. Considering that as the number of stratification variables increases, the likelihood increases that some of the variables will cancel the effects of other variables, not more than four to six stratification variables and not more than six strata for a particular variable should be used.
3. Identify an existing sampling frame or develop a sampling frame that includes information on the stratification variable(s) for each element in the target population. If the sampling frame does not include information on the stratification variables, stratification would not be possible.
4. Evaluate the sampling frame for undercoverage, overcoverage, multiple coverage, and clustering, and make adjustments where necessary.
5. Divide the sampling frame into strata, categories of the stratification variable(s), creating a sampling frame for each stratum. Within-stratum differences should be minimized, and between-strata differences should be maximized. The strata should not be overlapping, and altogether, should constitute the entire population. The strata should be independent and mutually exclusive subsets of the population. Every element of the population must be in one and only one stratum.
6. Assign a unique number to each element.
7. Determine the sample size for each stratum. The numerical distribution of the sampled elements across the various strata determines the type of stratified sampling that is implemented. It may be a proportionate stratified sampling or one of the various types of disproportionate stratified sampling.
8. Randomly select the targeted number of elements from each stratum. At least one element must be selected from each stratum for representation in the sample; and at least two elements must be chosen from each stratum for the calculation of the margin of error of estimates computed from the data collected.

- Ознаке:
  - L - број стратума у популацији
  - N_h - број јединица у h-том стратуму, $h = 1, 2, \dots, L$
  - Y_{hj} - вредност обележја y j-те јединице h-тог стратума, $h = 1, 2, \dots, L$, $j = 1, 2, \dots, N_h$
  - n_h - величина узорка који се бира из h -тог стратума
  - Y_h - тотал обележја h-тог стратума
  - $\bar{Y}_h$ - средина обележја h-тог стратума
  - y_{hj} - вредност обележја y j-те јединице одабране у узорак у h-том стратуму
  - $\bar{y}_h$ - узорачка средина обележја за h-ти стратум
- Важи:
  - $N = \sum_{h=1}^{L} N_h \; ; \; n = \sum_{h=1}^{L} n_h$
  - $S_h^2 = \frac{1}{N_h - 1} \sum_{j=1}^{N_h} [Y_{hj} - \bar{Y}_h]^2 \quad s_h^2 = \frac{1}{n_h - 1} \sum_{j=1}^{n_h} [y_{hj} - \bar{y}_h]^2$

- Оцена тотала:
  Ако је $\hat{Y}_h$, $h = \overline{1, L}$, непристрасна оцена тотала обележја Y_h h-тог стратума, тада је непристрасна оцена тотала обележја популације $\hat{Y}_{st} = \sum_{h=1}^{L} \hat{Y}_h$, а оцена њене дисперзије је $v[\hat{Y}_{st}] = \sum_{h=1}^{L} v[\hat{Y}_h]$, где су $v[\hat{Y}_h]$ непристрасне оцене дисперзија $V[\hat{Y}_h]$.
- Оцена средине:
  Ако је $\hat{\bar{Y}}_h$, $h = \overline{1, L}$, непристрасна оцена средине обележја $\bar{Y}_h$ h-тог стратума, тада је непристрасна оцена средине обележја популације $\hat{\bar{Y}}_{st} = \frac{1}{N} \sum_{h=1}^{L} N_h \hat{\bar{Y}}_h$

СТРАТИФИКОВАН СЛУЧАЈАН УЗОРАК

- Према томе, ако је коришћен случајан узорак (са или без понављања) у свих L стратума, оцена тотала обележја популације је

$$\hat{Y}_{st} = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{j=1}^{n_h} y_{hj}$$

а непристрасна оцена њене дисперзије је

$v[\hat{Y}_{st}] = \sum_{h=1}^{L} \frac{N_h^2}{n_h} s_h^2$, односно $v[\hat{Y}_{st}] = \sum_{h=1}^{L} \frac{N_h^2 (N_h - n_h)}{N_h n_h} s_h^2$.

- Фракција узорка у h-том стратуму: $f_h = \frac{n_h}{N_h}$.

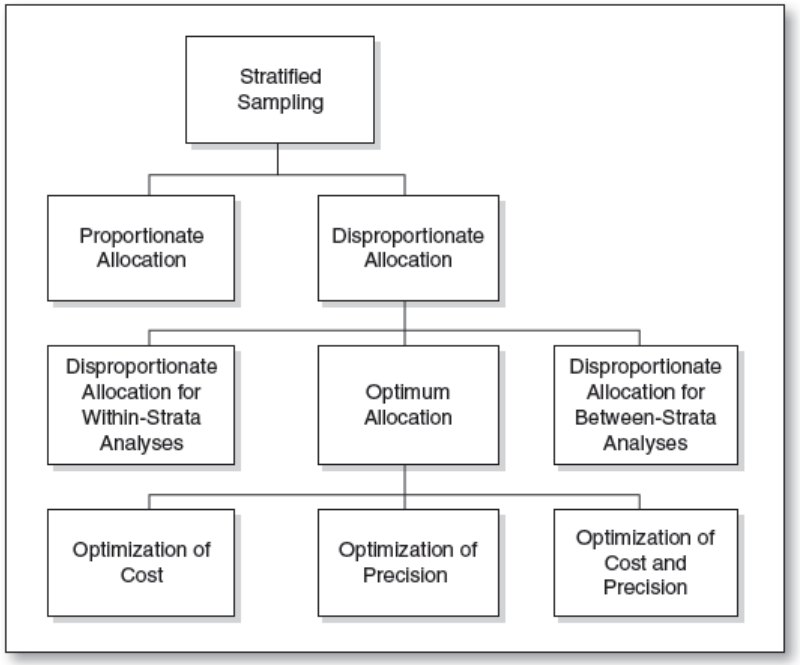
РАСПОДЕЛА/РАСПОРЕД ОБИМА УЗОРКА
(*Sample Size Allocation*)

- Када је већ одређен и фиксиран обим узорка, треба приступити одлучивању о обиму узорка n_h за сваки стратум појединачно, $h = 1, 2, \ldots, L$.
- У пракси се за решавање овог проблема обично користи нека од две популарне технике:
  - пропорционални распоред
  - *Neyman*-ов оптималан распоред

Одређени распоред обима узорка по стратумима примењује се, пре свега, у циљу смањења дисперзије. Међутим, и други чиниоци условљавају размештај обима узорка.

Главни принципи распореда приказани су на следећој шеми:

Figure 5.2 Subtypes of Stratified Sampling Based on Stratum Allocation



ПРОПОРЦИОНАЛАН РАСПОРЕД

- Код пропорционалног распореда, број јединица које се бирају у узорак из појединог стратума, пропорционалан је броју јединица у том стратуму, тј. $n_h = \frac{n}{N} N_h \ (f_h = f\ ), h = 1, 2, \dots, L$.
- Оцена тотала:

  Код пропорционалног распореда и стратификованог СУ, непристрасна оцена тотала обележја популације дата је са $\hat{Y}_{st} = \frac{N}{n} \sum_{h=1}^{L} \sum_{j=1}^{n_h} y_{hj}$.

- Ова техника базирана је на оригиналној идеји о репрезентативном узорку. Као резултат добија се тзв. *self-weighting sample*.
- Она, дакле, даје обиме узорака по стратумима онда када је унапред познат обим целог узорка и не узима у обзир трошкове. Међутим, трошкови су увек значајно ограничење при организовању било каквог истраживања. Зато је од интереса размотрити пропорционални распоред за задати укупан трошак.
- Нека је c_h, $h = 1, 2, \ldots, L$, трошак прикупљања информације од једне јединице из h-тог стратума. Ови трошкови могу се значајно разликовати међу стратумима.

- Укупан трошак истраживања је:

$$C = C_0 + \sum_{h=1}^{L} c_h n_h$$

где је C_0 општи (сталан) трошак.

- Пропорционални распоред за дати трошак дат је са

$$n_h = \frac{C - C_0}{\sum_{h=1}^{L} c_h N_h} N_h$$

а укупан обим узорка је, тада, једнак

$$n = \frac{C - C_0}{\sum_{h=1}^{L} c_h N_h} N$$

НЕПРОПОРЦИОНАЛНИ РАСПОРЕДИ

- Претходно описана техника пропорционалног распореда не узима у разматрање ниједан други аспект предмета истраживања, осим величине стратума (тј. броја јединица у стратуму). Она у потпуности игнорише унутрашњу структуру стратума у смислу инхерентног одступања вредности обележја унутар стратума и сл.
- Зато су предложене и шеме распореда, које воде рачуна о поменутом.
- *Disproportionate allocation for within strata analyses*

- *Disproportionate allocation for between-strata analyses*

  Овде се може користити најједноставнији метод размештаја обима узорка, који се састоји у избору подједнаког броја јединица из сваког стратума, тј. тако да

$$n_h = \frac{n}{L} \quad h = 1, 2, \ldots, L$$

  Уколико је код неког стратума $\frac{n}{L} > N_h$ узима се $n_h = N_h$, а остатак узорка $n - N_h$ се распореди равномерно на $(L-1)$ стратума. Реч је о равномерном распореду.

- *Optimum allocation*

  У пракси се користе две шеме распореда које минимизирају дисперзију оцена. Како је минимална дисперзија оптимално својство оцене, овакви распореди се називају оптималним.

- *Neyman Optimum Allocation*

  “Given a fixed sample size, how should sample be allocated to get the most precision from a stratified sample?”

Neyman-ов распоред минимизира дисперзију оцене, за познат и фиксиран обим целог узорка.

Код стратификованог случајног узорка без понављања, дисперзија оцене тотала обележја $\hat{Y}_{st}$ износи

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

Циљ је одредити $n_1, n_2, \ldots, n_L$, који минимизирају наведену дисперзију, под условом да важи $\sum_{h=1}^{L} n_h = n$.

Рачуницом се добија: $n_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} n$

- *Cost Optimum Allocation*
  "Given a fixed budget, how should sample be allocated to get the most precision from a stratified sample?"

Овај распоред минимизира дисперзију оцене, за познат и фиксиран укупан трошак истраживања.

Рачуницом се добија: $n_h = \frac{\frac{N_h S_h}{\sqrt{c_h}}(C - C_0)}{\sum_{h=1}^{L} N_h S_h \sqrt{c_h}}$

а укупан обим узорка је, тада, једнак $n = \frac{(C - C_0)\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^{L} N_h S_h \sqrt{c_h}}$

Што је већа варијабилност посматраног обележја у одређеном стратуму / што је већи обим стратума / што је јефтиније узорковање унутар стратума, требало би узети већи узорак из стратума.

ОДРЕЂИВАЊЕ ОБИМА УЗОРКА ЗА ЗАДАТУ ТАЧНОСТ

- Нека је d апсолутна грешка оцене непознатог параметра и α праг значајности.
- Код стратификованог случајног узорка без понављања потребан обим узорка за оцену тотала обележја популације је:
  - код равномерног распореда (ту је $n_h = \frac{n}{L}$)

$$n = \frac{n_0}{1 + \frac{z^2}{d^2}\sum_{h=1}^{L} N_h S_h^2} \qquad n_0 = \frac{Lz^2}{d^2}\sum_{h=1}^{L} N_h^2 S_h^2$$

  - код пропорционалног распореда (ту је $n_h = \frac{nN_h}{N}$)

$$n = \frac{1}{\frac{1}{n_0} + \frac{1}{N}} \qquad n_0 = \frac{Nz^2}{d^2}\sum_{h=1}^{L} N_h S_h^2$$

  - код *Neyman*-овог распореда (ту је $n_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} n$)

$$n = \frac{n_0}{1 + \frac{z^2}{d^2}\sum_{h=1}^{L} N_h S_h^2} \qquad n_0 = \frac{z^2}{d^2}\left(\sum_{h=1}^{L} N_h S_h\right)^2$$

Table 5.6 Strengths and Weaknesses of Stratified Sampling Compared to Simple Random Sampling

Strengths	Weaknesses
Unlike simple random sampling, stratified sampling:	Unlike simple random sampling, stratified sampling:
Has greater ability to make inferences within a stratum and comparisons across strata.	Requires information on the proportion of the total population that belongs to each stratum.
Has slightly smaller random sampling errors for samples of same sample size, thereby requiring smaller sample sizes for the same margin of error.	Information on stratification variables is required for each element in the population. If such information is not readily available, they may be costly to compile.
Obtains a more representative sample because it ensures that elements from each stratum are represented in the sample.	More expensive, time-consuming, and complicated than simple random sampling.
Takes greater advantage of knowledge the researcher has about the population.	Selection of stratification variables may be difficult if a study involves a large number of variables.
Data collection costs may be lower if the stratification variable breaks up the population into homogeneous geographical areas, or so as to facilitate data collection.	In order to calculate sampling estimates, at least two elements must be taken in each stratum.
Permits different research methods and procedures to be used in different strata.	The analysis of the data collected is more complex than the analysis of data collected via simple random sampling.
Permits analyses of within-stratum patterns and separate reporting of the results for each stratum.	If disproportionate allocation is used, weighting is required to make accurate estimates of population parameters.

Sampling Design Evaluation

- Главни разлог због кога би предност требало дати стратификованом узорковању јесте што овај план узорковања даје прецизније оцене него (прост) СУ.
- Генерално, план узорковања може се вредновати поређењем дисперзије одговарајуће оцене непознатог параметра, добијене тим планом узорковања, са дисперзијом исте оцене добијене код СУ без понављања (ради се са узорком истог обима). Количник те две дисперзије је ефекат дизајна (*design effect*) − *DEFF*.

$$DEFF\left[\hat{Y}_{st}\right] = \frac{V[\hat{Y}_{st}]}{V[\hat{Y}_{srs}]}$$, односно $$DEFF\left[\hat{\bar{Y}}_{st}\right] = \frac{V[\hat{\bar{Y}}_{st}]}{V[\hat{\bar{Y}}_{srs}]}$$.

Ако *DEFF* има вредност мању од 1 то указује да је стратификован случајни узорак ефикаснији од СУ без понављања истог обима. Ако има вредност већу од 1, ефикаснији је СУ без понављања истог обима.

Нпр. при важењу два (не посебно строга) услова, а то су: $N_h \approx N_h - 1$ и $N \approx N - 1$, *DEFF* за оцену тотала (он је, истовремено, једнак и *DEFF* за оцену средине) обележја популације, код пропорционалног распореда, може се апроксимирати формулом:

$$DEFF\left[\hat{Y}_{st}\right] \approx \frac{V\left[\hat{Y}_{st}\right]}{V\left[\hat{Y}_{st}\right] + \left(\frac{1-f}{f}\right)\sum_{h=1}^{L} N_h(\bar{Y}_h - \bar{Y})^2}$$

ДИСПЕРЗИЈА ОБЕЛЕЖЈА ПОПУЛАЦИЈЕ

- Поправљена дисперзија обележја популације

дата је са $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}[Y_i - \bar{Y}]^2 = \frac{1}{N-1}\sum_{h=1}^{L}\sum_{j=1}^{N_h}[Y_{hj} - \bar{Y}]^2$.

Даље је: $S_y^2 = \frac{1}{N-1}\sum_{h=1}^{L}\sum_{j=1}^{N_h}[(Y_{hj} - \bar{Y}_h) - (\bar{Y} - \bar{Y}_h)]^2$

$$= \frac{1}{N-1}\left[\sum_{h=1}^{L}\sum_{j=1}^{N_h}(Y_{hj} - \bar{Y}_h)^2 + \sum_{h=1}^{L} N_h(\bar{Y} - \bar{Y}_h)^2\right]$$

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

- Последњи израз може се записати и у облику

$$S_y^2 = S_W^2 + S_B^2$$

  где је S_W^2 дисперзија унутар (*within*) стратума, а S_B^2 дисперзија између (*between*) стратума.
- Ако је дисперзија унутар стратума мала, ефикасност стратификације је већа.
- Поређење квалитета оцена код стратификованог случајног узорка без понављања са оценама код СУ без понављања, за исти, унапред одређен и фиксиран, обим узорка n:

  Уведу се ознаке за дисперзије оцена – $V_{srs}, V_{prop}, V_{opt}$ и претпостави се да се фактор корекције популације и фактор корекције за стратуме могу занемарити.

  Тада важи: $V_{opt} \leq V_{prop} \leq V_{srs}$

- Нпр. ако су $V_{srs}, V_{prop}, V_{opt}$ ознаке дисперзија одговарајућих оцена тотала обележја популације, под наведеним претпоставкама и за фиксирани обим узорка, важи:

$$V_{srs} = \frac{S_y^2}{n} \qquad V_{prop} = \frac{1}{nN}\sum_{h=1}^{L} N_h S_h^2 \qquad V_{opt} = \frac{1}{nN^2}\left(\sum_{h=1}^{L} N_h S_h\right)^2$$

код *Neyman*-овог распореда је $n_h \approx N_h S_h$.

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

$$V_{prop} = V_{opt} + \frac{1}{nN}\sum_{h=1}^{L} N_h \left(S_h - \frac{1}{N}\sum_{h=1}^{L} N_h S_h\right)^2$$

Може се закључити да постоје две компоненте дисперзије које опадају при преласку са СУ (без понављања) на оптималан распоред. Прва компонента потиче од елиминације разлика између средина стратума; друга компонента потиче од елеминације ефекта разлика између стандардних одступања стратума.

ИЗБОР И ФОРМИРАЊЕ СТРАТУМА

- Један од основних проблема, који се јављају код стратификованог узорка, тиче се питања броја стратума (а, посредно, и величине стратума, тј. броја јединица унутар стратума).
- Мали број стратума може довести до значајне варијабилности, тј. инхерентног одступања у вредностима обележја јединица унутар истог стратума.
- Велики број стратума отежава рад и знатно повећава трошкове истраживања.

- Јасно је (између осталог, и из формула за дисперзију) да би стратуме требало формирати тако да имају што већу "хомогеност", тј. тако да је S_h^2 у сваком стратуму, $h = 1, 2, \ldots, L$, што мање. Самим тим и мали обим узорка по стратуму обезбеђује довољну прецизност оцена.
- Најбоље би било да се стратификација врши директно на основу вредности самог обележја које се испитује.
- Међутим, стратификација (раслојавање) по вредностима обележја које се изучава је ретко изводљива, или је чак и бесмислена, јер захтева познавање свих вредности обележја популације.
- Ипак, стратификација по самом обележју ("одокативно") је понекад прилично једноставна.
- Стратификација се најчешће врши према неком обележју за које постоји основана индиција да је у корелацији са испитиваним обележјем. При томе, очекује се да хомогеност у стратумима у односу на "помоћно" обележје/обележја значи и хомогеност вредности посматраног обележја.