ТЕОРИЈА УЗОРАКА 3

19. 04. ’13.

ВЕЖБАЊА

- Код PPSWOR, када постоји основано веровање да су вредности величине (X_i) јединица популације "довољно" пропорционалне вредностима посматраног обележја (Y_i), вероватноће укључења првог реда могу се израчунавати по формули

$$\pi_i = np_i = n\frac{X_i}{X}$$

Ако се, у претходној формули, добије $\pi_i > 1$, одговарајућа i-та јединица се бира у узорак (са вероватноћом укључења једнаком 1), а онда се прерачунају остали π_i модификовањем наведене формуле.

- Systematic Sampling
- Sunter's Method

Познате су вредности величине свих 10 јединица дате популације. Оне износе: 10, 10, 8, 6, 6, 4, 2, 2, 1, 1. Одабрати узорак са неједнаким вероватноћама, пропорционалним наведеним вредностима, обима $n = 4$, коришћењем поменутог метода.

Sunter's Method се састоји у томе да се иде редом, по уређеном (на основу вредности величине) списку свих јединица и за свако k (k иде од 1 до N) поступа на следећи начин:

- генериши случајан број u_k из $U[0, 1]$ расподеле
- ако је $k = 1$, задржи јединицу k у узорку акко је (КОРАК 1) $u_1 \leq \pi_1$
- ако је $k \geq 2$, задржи јединицу k у узорку акко је (КОРАК k)

$$u_k \leq \frac{n - n_{k-1}}{n - \sum_{i=1}^{k-1} \pi_i} \pi_k$$

где n_{k-1} представља број јединица већ одабраних у узорак на крају $k-1$-ог корака.

СТРАТИФИКОВАН УЗОРАК
Stratified Sampling

- Стратификован узорак примењује се онда када је потребно повећати прецизност оцена параметара, односно смањити грешке узорка.
- Стратификација је подела популације на потпопулације − стратуме (strata), при чему треба 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 сакупљених информација. Као критеријум за стратификацију користи се нека (или више) 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, се бирају 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
  - јединице популације: људи
    обележје: разна
    стратификација: пол; старост; образовање; верска припадност; етничка припадност; област живљења; социјално-економски фактори и сл.
- 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 за мерење прецизности оцене, број јединица за сваки стратум не сме бити мањи од две

- 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 различитим стратумима, у зависности од његових потреба и доступности информација
  - јефтиније је

- Мане:
  - 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 знања 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 $\hat{Y}_h$, $h = 1, 2, \dots, L$, 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 непристрасна оцене 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 $\hat{\bar{Y}}_h$, $h = 1, 2, \dots, L$, 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 прост случајан узорак 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

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

- 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 распоред

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 за стратификован случајан узорак:

  Код пропорционалног распореда, непристрасна оцена тотала обележја популације је

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

  а оцена њене дисперзије је

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

- Оцена средине:

  Ако је $\hat{\bar{Y}} = \frac{\hat{Y}_{st}}{N}$, код пропорционалног распореда, онда је $\hat{\bar{Y}}$ непристрасна оцена средине обележја популације.
- Ова техника, 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, $h = 1, 2, \dots, L$, трошак прикупљања информације од једне јединице из h-тог стратума. Ови трошкови се могу битно 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

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

ОПТИМАЛАН РАСПОРЕД

- Претходно описана техника пропорционалног распореда не узима у разматрање ниједан други аспекат 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.
- У пракси се користе две шеме распореда које минимизирају дисперзију оцена. Како је минимална дисперзија оптимално својство оцене, овакви распореди се називају оптимални.

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, дисперзија оцене тотала обележја $\hat{Y}_{st}$ износи

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

- Циљ је одредити $n_1, n_2, \dots, n_L$ 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?”
- Cost optimum 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

$$n = \frac{(C - C_0)\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^{L} N_h S_h \sqrt{c_h}}$$