ТЕОРИЈА УЗОРАКА - ЗАДАЦИ СА ВЕЖБИ

1. Показати да је $MSE(\hat{\theta}) = D(\hat{\theta}) + b^2(\hat{\theta})$.

2. Нека су $X_1, X_2, ..., X_n$ независне случајне величине из исте расподеле са коначним очекивањем μ и коначном дисперзијом $\sigma^2, \sigma > 0$.

а) Одредити пристрасност оцене (за σ^2) $S_n^2 = \sum_{k=1}^{n}(X_k - \overline{X}_n)^2$.

б) Одредити непристрасну оцену за σ^2.

3. Колико има простих случајних узорака без понављања, а колико са понављањем, ако се вади узорак обима 7 из популације која садржи 20 јединики?

4. Марко има 7 јабука, Петар 2, Јован 2 и Саша 6. Да ли је оцена укупног броја јабука непристрасна, ако се узорак обима 2 бира тако да је вероватноћа да су Марко и Петар у узорку $\frac{1}{3}$, вероватноћа да су Марко и Јован у узорку $\frac{1}{2}$, вероватноћа да су Петар и Јован $\frac{1}{6}$, док је вероватноћа избора свих осталих узорака једнака 0. (Као оцена укупног броја јабука користи се статистика $\hat{t} = N\overline{X}_n$).

5. Испитати да ли је боља (у средње квадратном смислу) оцена укупне суме обележја ($\hat{t} = N\overline{X}_n$) на основу узорка обима 2, или на основу узорка обима 3, из популације $\{1, 2, 3, 4\}$ ако се користи прост случајан узорак без понављања.

6. Из скупа $\{1, 2, 3, 4, 5, 6, 7, 8\}$ ваде се узорци без понављања обима 2, тако да сваки који садржи непаран број има вероватноћу 0. Они узорци који садрже 2 имају вероватноћу обрнуто пропорционалну другом елементу узорка, а они који садрже 4, а не садрже 2, имају вероватноћу $\frac{1}{8}$. Испитати непристрасност оцене средње вредности и одредити њену средње квадратну грешку.

7. Из популације са вредностима обелезяа $\{1, 2, ..., 100\}$ извадити 15 простих случајних узорака без понављања обима 20. На сваком од њих наћи оцену укупне суме обележја и испитати који од узорака је најрепрезентативнији, тј. где је реализована вредност статистике $\hat{t} = N\overline{X}_n$ најближа стварној вредности.

8. Нека је $Z = \sum_{i=1}^{N} a_i I_i$, где су $a_1, a_2, ..., a_N$ константе, а I_i индикатори укључења i-те јединке у узорак, $i \in \{1, 2, ..., N\}$. Одредити очекивање и дисперзију случајне величине Z (у зависности од вероватноћа укључења првог и другог реда).

9. Дата је популација $\{1, 2, 3, 4, 5, 6, 7, 8\}$ и размотрен је следећи план узорковања:

S	$\{1,3,5,6\}$	$\{2,3,7,8\}$	$\{1,4,6,8\}$	$\{2,4,6,8\}$	$\{4,5,7,8\}$
$P(S)$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

а) Наћи вероватноће укључења π_i за сваки елемент i.

б) Испитати непристрасност оцене $\hat{t} = N\overline{X}_n$.

1 ПРОСТ СЛУЧАЈАН УЗОРАК

1. У ресторану се служе четири врсте колача - баклаве, тулумбе, еклери и шампите. Десеторо људи је купило баклаву, двадесеторо тулумбу, двадесет и петоро еклер, а петнаесторо шампиту. На случајан начин се одаберу 2 различите врсте колача и бележи се број људи који су купили ту врсту.

а) Оценити укупан број особа које су купиле колач, а затим испитати непристрасност оцене која се користи.

б) Испитати да ли су узорачка дисперзија и узорачка стандардна грешка непристрасне оцене дисперзије и стандардне грешке на читавој популацији.

2. Узет је прост случајан узорак без понављања од 10 ученика од 100 ученика трећег разреда и бележене су њихове оцене из математике. Забележени су резултати: $(4, 5, 5, 2, 3, 1, 3, 4, 4, 5)$. Оценити просечну оцену из математике, а затим израчунати оцену дисперзије те оцене.

3. Одредити расподелу случајне величине μ_i која представља број појављивања i-те јединке популације, $(i \in \{1, 2, ..., N\})$, у простом случајном узорку са понављањем обима n.

4. У кутији се налази 100 папирића на којима се налазе каро, пик, херц и треф. Десет пута се извлачи папирић, забележи се знак, а затим се папирић врати у кутију. Добијен је узорак (К,П,П,Х,К,Т,Т,П,Х,Х).

а) Ако каро има вредност 1 динар, пик 2, херц 3, а треф 4, оценити укупну суму новца која се налази у кутији.

б) Одредити оцену броја пикова.

в) Одредити непристрасне оцене дисперзија ових оцена.

5. За који од следећих планова простог случајног узорковања без понављања ће бити добијена најпрецизнија оцена средње вредности обележја на популацији? Претпоставимо да обележје на свакој од популација има дисперзију 100.

а) Узорак обима 400, добијен из популације обима 4000.

б) Узорак обима 30, добијен из популације обима 200.

в) Узорак обима 3000, добијен из популације обима 300000000.

6. Истраживач жели да оцени пропорцију деце са плавим очима у вртићу од 4000 деце. Прихватљиво одступање му је 5% са ризиком 8%. Одредити обим простог случајног узорка без понављања који је потребан истраживачу. Ако је истраживач већ радио истраживање у другом вртићу и добио пропорцију $\frac{1}{3}$, који му је онда обим узорка потребан?

7. Узет је прост случајан узорак од 10 различитих кућа од 100 кућа које се налазе у једном насељу. Број становника у кућама из узорка је 2,5,1,4,4,3,2,5,2,3.

а) Оценити укупан број становника у том насељу и оценити дисперзију те оцене.

б) Оценити просечан број становника по кући и оценити дисперзију те оцене.

в) Наћи приближни 90%-ни интервал поверења за укупан број становника, као и за просечан број становника по кући.

8. Оцењује се број дивљих животиња у неком региону који је подељен на 286 области. Изабран је прост случајан узорак без понављања од 15 области и дат је број животиња у њима:

1, 50, 21, 98, 2, 36, 4, 29, 7, 15, 86, 10, 21, 5, 4

а) Наћи узорачку дисперзију, оцену просечног броја животиња по области, оцену дисперзије те оцене, као и оцену стандардне грешке.

б) Наћи 90% интервал поверења за укупан број животиња.

в) Коју величину узорка треба узети за оцењивање укупног броја животиња тако да одступање не буде веће од 2000 животиња са вероватноћом 0.9.

9. Ботаничар жели да оцени број стабала брезе у некој области. Област је подељена на 1000 делова. Познато је из претходних испитивања да је дисперзија броја стабала по области приближно 45. Одредити величину простог случајног узорка без понављања, потребну да са вероватноћом 0.95 одступање не буде веће од 500 стабала.

10. Посматрамо популацију обима 12 са вредностима обележја 3, 4, 12, 9, 8, 2, 6, 15, 14, 17, 1, 7, 12. Размотримо принцип простог случајног узорковања без понављања за узорак обима 4.

а) Одредити све могуће узорке и вероватноће да сваки од тих узорака буде изабран.

б) За сваки узорак израчунати узорачку средњу вредност, узорачку дисперзију, оцену дисперзије узорачке средине и медијану.

в) Показати да је средња вредност обележја на узорку непристрасна оцена средње вредности обележја популације и да је узорачка дисперзија непристрасна оцена дисперзије популације.

11. У датотеци *deca* дати су подаци о броју деце у свакој од 512 улица у неком граду. Наћи 95%-ни интервал поверења за укупан број деце у том граду користећи прост случајан узорак без понављања обима 200.

12. У бази података *Brand Laptops* налазе се подаци о 991 лаптопу са сајта *Smartprix*.

а) Изабрати прост случајан узорак без понављања обима 200 лаптопова. Оценити колико лаптопова сваког бренда се налази на сајту *Smartprix*.

б) Одредити 95% интервал поверења за просечну цену лаптопова (у индијским рупијама).

в) Поновити узорковање 5000 пута и нацртати хистограме узорачких медијана и узорачких средњих вредности просечних цена.

2 УЗОРКОВАЊЕ СА НЕЈЕДНАКИМ ВЕРОВАТНОЋАМА

1. Одредити $Hansen-Hurwitz$-ову оцену укупне суме обележја популације ако се користи прост случајан узорак са понављањем, а затим и $Hansen-Hurwitz$-ову оцену дисперзије те оцене.

2. Из популације се вади узорак са понављањем обима n и претпоставља се да је p_i, вероватноћа избора i-те јединке у узорак у једном покушају, позната за свако $i \in \{1,2,...,N\}$. Одредити π_i - вероватноћу укључења i-те јединке у узорак и π_{ij} - заједничку вероватноћу укључења јединки i и j $(i,j \in \{1,2,...,N\})$.

3. Популација се састоји од 15 јединки чије су величине 23, 30, 41, 26, 53, 60, 28, 52, 113, 72, 80, 35, 42, 38 и 52. Изабрати узорак са понављањем и узорак без понављања обима 5 са вероватноћама пропорционалним величинама користећи Лахиријев метод.

4. Изабран је узорак са вероватноћама пропорционалним величини, са понављањем, обима 3 из популације величине 10. Дате су вредности обележја изабраних елемената и вероватноће избора:

i	1	2	3
x_i	3	10	7
p_i	0.06	0.2	0.1

а) Наћи оцену укупне суме обележја популације користећи $Hansen-Hurwitz$-ову оцену.

б) Оценити дисперзију те оцене.

в) Наћи оцену укупне суме обележја популације користећи $Horvitz-Thompson$-ову оцену.

д) Оценити дисперзију те оцене.

5. Ако се бира узорак обима n без понављања, показати да важи:

а) $\sum_{i=1}^{N} \pi_i = n.$

б) $\sum_{j=1, j\neq i}^{N} \pi_{ij} = (n-1)\pi_i$, за свако фиксирано i, $i = \overline{1,n}$.

6. За испитивање загађености 320 језера укупне површине $80km^2$ изабран је узорак са понављањем обима 4. Прво језеро из узорка бирано је два пута, а остала два по једном. Концентрација загађености у та три језера у узорку је редом 2, 5 и 10, а површине тих језера (у km^2) су редом 1.2, 0.2 и 0.5. Наћи $Hansen-Hurwitz$-ову оцену средњег загађења по језеру у посматраној популацији, као и оцену дисперзије добијене оцене.

7. У датотеци *populacija* налазе се величине 120 јединки из популације. Извадити узорак са понављањем и узорак без понављања обима 90 користећи метод кумуланти.

8. Из популације коју чине три поља на којима се узгаја пшеница бира се узорак обима 2 са вероватноћама пропорционалним величинама, са понављањем. У следећој табели су дати подаци о количини произведене пшенице на сваком пољу и вероватноће избора сваког поља.

i	1	2	3
x_i	11	6	25
p_i	0.3	0.2	0.5

Налажењем свих могућих узорака одредити вероватноће укључења сваког елемента $\pi_i, i = 1,2,3$, као и Хансен-Хурвицову и Хорвиц-Томпсонову оцену за укупну производњу пшенице за сваки узорак.

9. У датотеци *radnici* дати су подаци о броју радника и производњи у 10 фабрика у индустријској зони. Избрати узорак обима 3 са понављањем са вероватноћама избора пропорционалним броју радника у фабрици. Користећи добијени узорак одредити $Hansen-Hurwitz$-ову оцену укупне производње.

10. Дата је популација од четири прашуме, њихове површине и бројеви тигрова који живе у њима:

Прашума	1	2	3	4
Површина (у km^2)	100	200	300	500
Број тигрова	11	20	23	54

Вади се узорак од 2 прашуме. Оценити укупан број тигрова, одредити дисперзију те оцене и наћи непристрасну оцену те дисперзије ако је узорковање вршено:

а) Са понављањем (Користити Хансен-Хурвицову оцену),

б) Без понављања (Користити Хорвиц-Томпсонову оцену),

и извучен је узорак (1,2).

11. Популацију чини база *trees* која садржи податке о 31 дрвету.

а) Изабрати прост случајан узорак без понављања обима 10 и оценити средњу вредност обележја $Volume$ користећи добијени узорак. Затим поновити поступак 1000 пута и за тако добијене вредности нацртати хистограм. Упоредити оцену средње вредности добијену на основу једног узорка и ону добијену симулацијом.

б) Изабрати узорак са понављањем обима 10 са вероватноћама пропорционалним обиму стабла (променљива $Girth$) и оценити средњу вредност обележја $Volume$ користећи $Hansen-Hurwitz$-ову оцену.

3 СТРАТИФИКОВАН УЗОРАК

1. Популација од 6 јединки подељена је на 2 стратума, тако да се у првом налазе јединке са вредностима обележја 0, 1 и 2, а у другом јединке са вредностима обележја 4, 6 и 11. У узорак без понављања су методом простог случајног узорковања изабране јединке са вредностима обележја 0 и 2 из првог стратума, а 6 и 11 из другог.

а) Одредити оцену средње вредности обележја популације, дисперзију те оцене и оцену те дисперзије.

б) Упоредити добијену дисперзију са дисперзијом оцене средње вредности предложене код простог случајног узорковања без понављања.

в) Оценити оценом укупну суму обележја популације, дисперзију те оцене и оцену те дисперзије на основу добијеног стратификованог узорка.

2. Испитује се број претрчаних метара ученика једног разреда на часу физичког васпитања. У разреду има 112 ученика, од чега 59 девојчица и 53 дечака. Посматрано је 7 дечака и 8 девојчица.
Дечаци: 879, 810, 789, 567, 900, 870, 777.
Девојчице: 450, 234, 679, 456, 239, 555, 560, 467.

а) Оценити укупан број метара које су претрчала деца тог разреда и оценити дисперзију те оцене. Затим урадити исто коришћењем оцена предложених код простог случајног узорковања без понављања.

б) Оценити просечан број претрчаних метара и оценити дисперзију те оцене.

3. У датотеци *stratumi* налазе се подаци о вредностима обележја на популацији која се састоји из 2 стратума (стратуми су различитих величина, али да бисмо могли да их сместимо у две колоне базе, мањи стратум је допуњен NA вредностима до величине већег стратума. Избацити их при рачунању.). Вади се стратификован узорак обима 500, тако што се из мањег стратума вади прост случајан узорак са понављањем обима 200, а из већег обима 300.

а) Оценити укупну суму обележја популације, израчунати дисперзију те оцене и оцену те дисперзије.

б) Оценити средњу вредност обележја популације, израчунати дисперзију те оцене и оцену те дисперзије.

4. а) Популација је подељена на три стратума чије су величине 123, 102 и 180, а одговарајуће дисперзије обележја су 116, 143 и 170, редом. Ако се бира стратификован случајан узорак обима 100, одредити величину узорка који се бира из сваког стратума користећи пропорционални избор.

б) Извести формулу за одређивање обима узорака који се узимају из сваког стратума ако се користи Нејманова метода, а затим одредити величину узорка који се узима из сваког стратума за популацију и дати обим узорка из дела а).

5. Популација од 2000 елемената подељена је на три стратума величина $N_1 = 600, N_2 = 800$ и $N_3 = 600$. Познато је да је $s_1 = 2, s_2 = 4s_3$ и $s_1 = 100$. Ако се бира стратификован узорак обима 300, одредити величину узорка који се вади из сваког стратума користећи:

а) пропорционални избор;

б) Нејманов избор.

6. За испитивање просечне недељне потрошње бензина град је подељен на 4 дела, који се посматрају као стратуми. Извађен је стратификован случајан узорак (без понављања) и забележена је потрошња бензина за протеклу недељу код сваког возача из узорка. Добијени су следећи подаци:
Стратум 1: $N_1 = 3750, n_1 = 50, \overline{x}_{n1} = 12.6, s^2_{n1} = 2.8$
Стратум 2: $N_2 = 3272, n_2 = 45, \overline{x}_{n2} = 14.5, s^2_{n2} = 2.9$
Стратум 3: $N_3 = 1387, n_3 = 30, \overline{x}_{n3} = 18.6, s^2_{n3} = 4.8$
Стратум 4: $N_4 = 2475, n_4 = 30, \overline{x}_{n4} = 13.8, s^2_{n4} = 3.2$.

а) Оценити просечну недељну потрошњу за цео град.

б) Наћи 95%-ни интервал поверења добијене оцене.

б) Ако треба изабрати стратификовани случајни узорак обима 1000, одредити величину узорка који се вади из сваког стратума користећи пропорционални распоред.

7. Средњошколци су подељени у три групе по успеху у школи са циљем да се испита колико су заинтересовани за позориште. У фајлу *posete* дати су подаци о броју посета позоришту за ученике током годину дана, као и којој групи по успеху припадају. Изабрати стратификован узорак без понављања обима 200, користећи пропорционални избор, а затим оценити просечан број посета позоришту током годину дана за средњошколце, дисперзију те оцене и оцену те дисперзије. Затим урадити то исто користећи Нејманов избор и упоредити добијене дисперзије.

8. У датотеци *netflix_titles* налазе се подаци о филмовима и серијама додатим на платформу *Netflix* у периоду од 2008. до 2021. године.

а) Поделити наслове у стратуме по годинама када су додати, али тако да се сви наслови додати пре 2015. године налазе у једном стратуму. Изабрати стратификован случајан узорак без понављања обима 500 користећи Нејманов избор, па оценити удео серија у насловима.

б) Поделити серије у стратуме по рејтингу (постоји само једна серија са рејтингом *TV-Y7-FV*, њу сместити у стратум са серијама са рејтингом *R*). Изабрати стратификован случајан узорак без понављања обима 200 серија користећи Нејманов избор, па оценити просечан број сезона.

9. Поделити лаптопове из базе *BrandLaptops* у стратуме по бренду процесора. Користећи пропорционални распоред, одредити колико би лаптопова из сваког стратума требало одабрати у стратификован узорак обима 300. Из сваког стратума бирати прост случајан узорак без понављања, осим код лаптопова са процесором *Intel* - ту бирати прост случајан узорак са понављањем. Оценити број лаптопова са оперативним системом *Windows* и одредити оцену дисперзије те оцене.

4 КОЛИЧНИЧКО ОЦЕЊИВАЊЕ

1. Узет је прост случајан узорак без понављања обима 30 из велике популације и добијени су подаци $\overline{y}_n = 5$ и $\overline{x}_n = 18$, где је x обележје од интереса, а y помоћно обележје. Такође, познато је да је $t_y = 1891$. Наћи оцену количника R и количничку оцену суме обележја x.

2. Агент за некретнине жели да процени релативну промену у просечној цени кућа за период од две године. Посматрао је 20 кућа од укупно 1000 за које је задужен. Дате су цене кућа у хиљадама евра ове године и одговарајуће цене пре две године за сваку од 20 посматраних кућа:
Пре две године: 290, 53, 300, 60, 110, 311, 150, 150, 240, 190, 32, 180, 90, 351, 241, 300, 230, 84, 140, 155;
Сада: 255, 50, 275, 66, 109, 321, 113, 126, 215, 177, 24, 179, 65, 339, 314, 271, 235, 74, 135, 154.
Оценити релативну промену у просечној цени. Да ли је могуће добити количничку оцену просечне цене кућа данас?

3. Дата је популација обима 4 и вредности обележја које се испитује (x) и помоћног обележја (y):
x: 3, 5, 7, 9
y: 1, 2, 2, 3
На основу простих случајних узорака обима 2 и 3 испитати да ли је количничка оцена количника популације непристрасна.

4. Дате су следеће вредности:

Школа	1	2	3	4
Број деце која имају 5 из математике	200	300	100	250
Број посматраних разреда	4	5	3	4

Истраживач жели посматрањем прве и треће школе да донесе закључак о укупном броју деце која имају пет из математике. Испитати да ли је количничка оцена која се користи непристрасна (ако се користи просто случајно узорковање без понављања), одредити њену вредност на датом узорку, а затим упоредити са оценом добијеном на основу простог случајног узорка без понављања. Која је боља?

5. У датотеци *kol* дати су подаци за прост случајан узорак без понављања обима 100 из популације обима 530 у коме је x обележје које се испитује, а y помоћно обележје. Познато је да је популацијска средња вредност обележја y једнака $\overline{y} = 3.33$.

а) Наћи количничку оцену средње вредности за обележје x.

б) Оценити дисперзију добијене количничке оцене.

в) Наћи 95%-ни интервал поверења за средњу вредност обележја популације.

6. У граду који има 28753 домаћинстава, циљ је да се испита просечан рачун за струју. Посматрана су четири домаћинства и забележен је број чланова, као и рачуни за струју за свако од њих.
Прво домаћинство: 1 члан, рачун од 1282 динара;
Друго домаћинство: 3 члана, рачун од 4375 динара;
Треће домаћинство: 2 члана, рачун од 2333 динара;
Четврто домаћинство: 4 члана, рачун од 5789 динара.
Укупан број становника града је 70351.
Наћи количничку оцену просечног рачуна за струју, као и оцену дисперзије те оцене. Колико је укупно новца потрошено на струју?

7. У популацији обима 15 познате су вредности помоћног обележја y: 2.3, 3.0, 4.1, 2.6, 5.3, 6.0, 2.8, 5.2, 11.3, 7.2, 8.0, 3.5, 4.2, 3.8, 5.2. Из популације је одабран прост случајан узорак обима 5 и добијени су следећи резултати:

x_i	40	40	70	82	52
y_i	4,1	5.2	7.2	8.0	5.2

Наћи количничку оцену средње вредности обележја x и оцену дисперзије те оцене.

8. У датотеци *prodavnica* дати су подаци о дневној нето заради за одређени дан у години, подаци о дневној нето заради за исти тај дан претходне године, као и подаци о броју запослених у продавници, за 1534 продавнице у једном насељу. Извадити узорак са понављањем обима 400, са вероватноћама пропорционалним броју запослених, па количничком оценом оценити укупну нето зараду за све продавнице за дан који се посматра, ако се као помоћно обележје користе зараде од претходне године за исти дан. Израчунати дисперзију те оцене.

9. Дати су подаци за два стратума исте величине из популације величине 12.

h	x_{h_i}	y_{h_i}
1	2,2,3,4,6,7	6,6,7,7,8,8
2	10,11,12,14,15,16	10,11,16,17,18,18

Из сваког стратума бирају се у узорак по 3 јединке. Вредности јединки укључених у узорак дате су у табели:

h	x_{h_i}	y_{h_i}
1	2,3,7	6,7,8
2	10,14,15	10,17,18

Наћи комбиновану и посебну количничку оцену средње вредности обележја популације, а затим одредити дисперзије тих оцена.

10. У бази података *Tweets* налазе се подаци о одређеном броју твитова објављених у току неког дана. Твитови су сврстани у 3 категорије према емоционалном тону - негативни, неутрални и позитивни. Желимо да проценимо просечан број речи које садржи твит (речју сматрамо низ карактера између размака). Познато је да је тог дана објављено 500 000 758 твитова и да је просечан број карактера по твиту 28.

а) Оценити просечан број речи по твиту количничком оценом, ако се као помоћно обележје користи број карактера у твиту и подаци су добијени методом простог случајног узорковања без понављања. Одредити оцену дисперзије те оцене.

б) Оценити просечан број речи по твиту комбинованом оценом, за исто помоћно обележје, ако су подаци добијени стратификованим случајним узорковањем без понављања, где су стратуми категорије и коришћен је пропорционални распоред.