**Пояснювальна записка для користувачів до файлів мікроданих "Анонімні мікродані за основними показниками щодо доходів, витрат та умов життя домогосподарств" за 2018 рік**

Відповідно до плану державних статистичнихспостережень на 2019 рік, затвердженого розпорядженням Кабінету Міністрів України від 10.04.2019 № 283-р, за результатами вибіркового обстеження умов життя домогосподарств (ОУЖД) вперше підготовлено анонімні мікродані за основними показниками щодо доходів, витрат та умов життя для публічного використання: "Анонімні мікродані по домогосподарствах" (далі − файл по домогосподарствах) та "Анонімні мікродані по особах, які входять до складу домогосподарства" (далі − файл по особах) за 2018 рік.

Метою зазначеного обстеження є отримання даних щодо соціально-демографічних характеристик домогосподарств, структури їх витрат та ресурсів, а також для комплексного дослідження рівня матеріального добробуту різних соціальних груп населення.

Детальну інформацію про організацію, методологію проведення ОУЖД, поширення його результатів на генеральну сукупність, якість даних тощо розміщено на офіційному сайті Держстату в розділі "Методологія та класифікатори"/"Статистична методологія"/"Демографічна та соціальна статистика"/"Доходи та умови життя":

* <http://www.ukrstat.gov.ua/metod_polog/metod_doc/2018/32/mp_vond.pdf>,
* <http://www.ukrstat.gov.ua/metod_polog/metod_doc/2019/168/mp_ouzh_dg.pdf>,
* <http://www.ukrstat.gov.ua/metod_polog/metod_doc/2014/246/Mordvo.pdf>,
* <http://www.ukrstat.gov.ua/suya/st_zvit/2018/st_zvit_obum_domogosp_2017.pdf> .

Узагальнена інформація за результатами ОУЖД оприлюднюється щоквартально у статистичній інформації, щорічно − у статистичному збірнику "Витрати і ресурси домогосподарств України" та тематичних збірниках, які розміщуються на вищезазначеному сайті в розділі "Статистична інформація"/"Публікації"/  
"Демографічна та соціальна статистика"/"Доходи та умови життя" (<http://www.ukrstat.gov.ua/druk/publicat/kat_u/2019/zb/06/zb_vrdu2018.pdf> ).

Обстеження не охоплює інституційне населення (військовослужбовців строкової служби, осіб, що знаходяться в місцях позбавлення волі, осіб, що постійно проживають у будинках-інтернатах, будинках для осіб похилого віку), маргінальні прошарки (безпритульні тощо), а також населення, яке проживає на тимчасово окупованій території Автономної республіки Крим, м. Севастополя та частині тимчасово окупованих територій у Донецькій та Луганській областях.

Файли представлено у форматі "sav" (програмне забезпечення SPSS) та "xlsx".

Інформація файлів надає можливість здійснити аналіз рівня та умов життя домогосподарств залежно від рівня матеріального добробуту, наявності в домогосподарстві дітей та за іншими соціально-економічних ознаками.

Файл по домогосподарствах сформовано на підставі опитування 8051 домогосподарства (17,1 тис. осіб). У файл по домогосподарствах включено змінні щодо:

* доходів та ресурсів,
* окремих соціально-демографічних характеристик домогосподарств,
* витрат домогосподарств (дані згруповані відповідно до Класифікації індивідуального споживання за цілями (КІСЦ)[[1]](#footnote-1), яка розміщена на вищезазначеному сайті в розділі "Методологія та класифікатори" /"Класифікатори"/"Класифікація індивідуального споживання за цілями"),
* споживання домогосподарствами продуктів харчування,
* окремих характеристик житла,
* наявності у домогосподарствах товарів тривалого користування,
* самооцінки респондентами рівня своїх доходів тощо.

У файл по особах включено змінні щодо:

* віку,
* статі,
* рівня освіти,
* соціально-економічного статусу особи,
* індивідуальних доходів,
* окремих характеристик рівня доступності медичних послуг,
* доступу до інтернету тощо.

Файли є набором анонімних знеособлених мікроданих, до яких були застосовані методи захисту знеособлених мікроданих (глобального перекодування, агрегування та маскування) з метою мінімізації ризику опосередкованого встановлення конкретної статистичної одиниці. Файли містять тільки певну кількість відібраних змінних, оскільки не всі змінні, які містяться в первинній базі даних, можуть бути надані без порушення правил захисту конфіденційності.

Підготовка файлів проводилась із застосуванням до окремих змінних методів маскування, зокрема некоригувальних методів захисту знеособлених мікроданих, які не змінюють початкових даних. У файлі по домогосподарствах зменшено рівень деталізації шляхом глобального перекодування й агрегування в інтервальні ряди по змінних: hsize "кількість осіб, які проживають у домогосподарстві", SLIV "житлова площа" та ROOMS "кількість кімнат", а у файлі по кожній особі, яка входить до складу домогосподарства, − по змінних: AGE "вік", L\_EDUC\_M "Рівень освіти (для осіб у віці 6 років і старше)" та SES\_MEM "Соціально-економічний статус".

Наприклад, у файлі по особах первинні дані містять інформацію щодо віку кожної особи. Для зменшення ризику розкриття конфіденційної інформації сформовано п’ять вікових груп інтервального розподілу: до 18 років, 18-35 років, 36-55 років, 56-59 років, 60 років і старші.

До певних змінних по витратах домогосподарств були застосовані коригувальні методи маскування, а саме локального закриття (процедура верхнього перекодування top-coding − заміна найбільших значень на середнє значення цієї змінної по зазначеній групі домогосподарств).

Наприклад, по змінній "витрати на транспорт" було визначено дванадцять домогосподарств, у яких значення цього виду витрат значно перевищувало середнє значення по цій змінній по всій сукупності домогосподарств (аутлаєри). Щоб захистити ці домогосподарства від можливого розкриття конфіденційної інформації про них, такі значення-аутлаєри були замінені на середнє значення по цих дванадцяти домогосподарствах. При такому підході слід ураховувати, що використання процедури "top-coding" може вплинути на основні показники диференціації по доходах і витратах. Процедура "top-coding" була застосована до таких змінних:

* h041 "Фактична плата за житло",
* h043 "Утримання та ремонт житла",
* h05 "предмети домашнього вжитку, побутова техніка та поточне утримання житла",
* h06 "охорона здоров'я",
* h07 "транспорт",
* h08 "зв’язок",
* h09 "відпочинок і культура",
* h11 "ресторани та готелі",
* h12 "інші товари та послуги",
* n00 "неспоживчі грошові витрати".

Звертаємо увагу, що при аналізі витрат домогосподарств на оплату житла та комунальних послуг слід із загальних сум коштів, наведених у змінній h04 "житло, вода, електроенергія, газ та інші види палива", виключати грошові кошти, зазначені у змінній h043 "Утримання та ремонт житла".

Також до файла по домогосподарствах була застосована процедура маскування шляхом локального закриття окремих значень для забезпечення граничного рівня   
k-анонімності[[2]](#footnote-2) з параметром 3. У результаті застосування цієї процедури було приховано 5 записів по змінній TP\_NS\_P "тип населеного пункту", 10 − по COD\_OBL "регіон", 208 − по HSIZE "розмір домогосподарства", 7 − по TYPE\_DOM "тип домогосподарства", 749 − по SLIV "житлова площа", 305 − по ROOMS "кількість кімнат", 322 − по LANDPLOT "наявність у користуванні земельних ділянок".

Із файла по особах вилучена інформація щодо осіб, які входять до складу великих за розміром (6 і більше осіб) домогосподарств (у 2018 році − 134 домогосподарства), зважаючи на міркування мінімізації ризику їх розкриття, який зумовлений унікальним (нечастим) характером поширення домогосподарств такого чисельного складу.

Після подальшого об’єднання даних по домогосподарствах і по особах, які входять до їх складу, проведено оцінку ризику розкриття конфіденційної інформації по кожній особі з урахуванням даних по домогосподарству. Для випадків суттєвої ймовірності розкриття до змінних по особах, які входять до складу домогосподарства, застосована процедура локального закриття. Мета цієї процедури − досягнення встановленого граничного рівня k-анонімності при одночасній мінімізації кількості випадків необхідного приховування значення змінної. До об’єднаного файла по особах та по домогосподарствах були застосовані умови: k- анонімність дорівнює параметру 3. Таким чином, було приховано 32 записи по змінній "age" (вік), 713 − по змінній "L\_EDUC\_M" (рівень освіти) та 524 − по змінній "ses\_mem" (соціально-економічний статус особи).

До змінних ppinc1 "Дохід від зайнятості" та ppinc2 "Інші доходи" зазначеного файла також була застосована процедура top-coding .

Після застосування процедур анонімізації значення деяких змінних були змінені. У зв'язку з цим відповідні похідні змінні, які розраховуються на основі агрегування змінених первинних змінних, також приймають інші значення порівняно з показниками, наведеними в офіційних публікаціях. До таких показників належать: грошові, сукупні витрати, загальні доходи та сукупні ресурси, а також показники диференціації та нерівності матеріального добробуту населення.

Застосовані методи забезпечують високий рівень захисту від можливого розкриття конфіденційних даних із одночасним збереженням оцінок основних показників ОУЖД на рівні, прийнятному для аналізу. Відхилення середніх показників, розрахованих за даними файлів анонімних мікроданих, від відповідних оприлюднених рівнів не перевищує 3 відсоткових пунктів.

1. Гармонізована з Класифікацією індивідуального споживання за цілями Європейського Союзу   
   (COICOP-HBS, 2003 року) [↑](#footnote-ref-1)
2. K-анонімність - показник ризику, який ґрунтується на тому принципі, що кількість осіб (домогосподарств), які мають однаковий ключ (набір змінних, які в поєднанні з іншими можуть призвести до ідентифікації респондентів), становить не менш ніж *k*. [↑](#footnote-ref-2)