

А. П. Подаваленко, Т. Д. Нессонова, Н. В. Білера

ЕПІДЕМІОЛОГІЧНА ДІАГНОСТИКА. СТАТИСТИЧНІ МЕТОДИ ДОСЛІДЖЕННЯ

Навчальний посібник

*для слухачів курсів тематичного удосконалення,
слухачів циклів спеціалізації та лікарів-інтернів*

МІНІСТЕРСТВО ОХОРОНИ ЗДОРОВ'Я УКРАЇНИ
Харківський національний медичний університет

А. П. Подаваленко, Т. Д. Нессонова, Н. В. Білера

ЕПІДЕМІОЛОГІЧНА ДІАГНОСТИКА.
СТАТИСТИЧНІ МЕТОДИ ДОСЛІДЖЕННЯ

Навчальний посібник
для слухачів курсів тематичного удосконалення,
слухачів циклів спеціалізації та лікарів-інтернів

Харків
ХНМУ
2025

УДК 616-036.22-022.7:519.23(075.8)
П44

*Затверджено Вченою радою ХНМУ.
Протокол № 11 від 27.06.2025.*

Рецензенти

Т. В. Пересипкіна – д-р мед. наук, проф. (ХНУ ім. В.Н. Каразіна).

О. В. Арсен'єв – канд. фіз.-мат. наук, доц. (ПВНЗ «Харківський міжнародний медичний університет»).

Подаваленко А. П., Нессонова Т. Д., Білера Н. В.

П44 Епідеміологічна діагностика. Статистичні методи дослідження : навч. посіб. для слухачів курсів тематичного удосконалення, слухачів циклів спеціалізації та лікарів-інтернів. Харків: ХНМУ, 2025. 136 с.

У навчальному посібнику представлені теоретичні питання епідеміологічної діагностики інфекційної та неінфекційної захворюваності й практичні заняття щодо використання статистичних методів дослідження. До кожного практичного заняття надано файли Excel для виконання поставлених завдань у вправах, алгоритм їх виконання та запропоновані авторами висновки за отриманими результатами. Висвітлено елементи ретроспективного та оперативного аналізу, а також визначення та структуру епідеміологічного методу, який пропонується використовувати в практичній та науковій діяльності лікарів.

Для самоконтролю засвоєння навчального матеріалу складено запитання первинного контролю та тести підсумкового контролю рівня знань.

УДК 616-036.22-022.7:519.23(075.8)

© Харківський національний
медичний університет, 2025

© А. П. Подаваленко, Т. Д. Нессонова,
Н. В. Білера, 2025

ЗМІСТ

Список умовних скорочень	4
Запитання для первинного контролю рівня знань	5
Вступ	6
1. Епідеміологічна діагностика – методична основа епідеміологічного нагляду	7
1.1. Ретроспективний епідеміологічний аналіз	13
1.2. Оперативний епідеміологічний аналіз	14
2. Епідеміологічний метод: визначення та структура	15
2.1. Епідеміологічне спостереження	15
2.2. Епідеміологічний експеримент	15
2.3. Математичне моделювання	15
3. Статистичний аналіз в епідеміології	19
3.1. Використання статистичних методів	24
3.1.1. Практичне заняття 1. Розрахунок основних епідеміологічних показників	24
3.1.2. Практичне заняття 2. Систематизація даних епідеміологічного моніторингу	35
3.1.3. Практичне заняття 3. Визначення функції розподілу та основних числових статистичних характеристик показника	51
3.1.4. Практичне заняття 4. Перевірка гіпотез щодо нормального розподілу у вибірці. Статистичні методи порівняння середніх значень	64
3.1.5. Практичне заняття 5. Дослідження статистичних зв'язків	78
3.1.6. Практичне заняття 6. Числові характеристики та методи прогнозування часових рядів	98
Тести для підсумкового контролю рівня знань	114
Додатки	121
Література	134
Електронні ресурси	135

СПИСОК УМОВНИХ СКОРОЧЕНЬ

ВІ	– випадки інфекцій
ВООЗ	– Всесвітня організація охорони здоров'я
Г	– госпіталізовано
Держстат	– Державна служба статистики України
ЕН	– епідеміологічний нагляд
ЕПК	– екстенційний показник кількості інфікованих
З	– захворюваність
ІЕП	– інтегрований епідеміологічний показник
ІЗ	– визначення випадку інфекційного захворювання
МОЗ	– Міністерство охорони здоров'я
Н	– населення
ПГ	– показник госпіталізації
С	– смертність
СВ	– смертельні випадки
М	– середня арифметична
Me	– медіана
Mo	– мода

ЗАПИТАННЯ ДЛЯ ПЕРВИННОГО КОНТРОЛЮ РІВНЯ ЗНАНЬ

1. Дайте визначення поняттю «епідеміологічний нагляд». [3, 6, 18, 21, 24]
2. Дайте визначення поняттю «епідеміологічна діагностика». [3, 6, 18, 21, 24]
3. Назвіть розділи епідеміологічної діагностики. [3, 6, 18]
4. Назвіть етапи епідеміологічної діагностики. [3, 6, 18]
5. Ретроспективний епідеміологічний аналіз: мета та завдання. [3, 6, 14, 18, 22]
6. Оперативний епідеміологічний аналіз: мета та завдання. [3, 6, 14, 18, 22].
7. Епідеміологічне спостереження в структурі епідеміологічного методу. [7, 23]
8. Епідеміологічний експеримент в структурі епідеміологічного методу. [7, 23]
9. Математичне моделювання в структурі епідеміологічного методу. [28, 29, 30]
10. В чому полягає основна мета епідеміологічної діагностики? [18, 21]
11. Які основні принципи накопичення інформації щодо проявів епідемічного процесу? На що впливає кількість накопиченої інформації? [7, 17]
12. Які основні етапи проведення епідеміологічного нагляду за станом здоров'я населення? [14, 21]
13. Яка різниця між рутинним та дозорним епідеміологічним наглядом? [21]
14. Які кількісні показники використовуються під час аналізу проявів епідемічного процесу? [7, 14, 16]
15. За якими принципами формуються групи ризику інфекційних захворювань? [19, 22, 26]
16. На які групи поділяються інфекційні хвороби? За якими принципами складена класифікація інфекційних хвороб? [3, 6, 24]
17. В яких випадках потрібне статистичне оцінювання різниці між епідеміологічними показниками? [22, 26, 27]
18. Які групи факторів можуть впливати на перебіг епідемічного процесу? [15, 19, 26, 27]
19. За допомогою яких процедур досліджується зв'язок між факторами та кількісними характеристиками епідемічного процесу? [15, 17, 19, 22, 26, 27]
20. На які групи поділяються прогнози перебігу інфекційних хвороб? [1, 17]

ВСТУП

Навчальний посібник підготовлено згідно з вимогами до видання навчальної літератури для вищої школи та уніфікованих програм підготовки лікарів-інтернів, слухачів післядипломної освіти, лікарів-епідеміологів та фахівців центру громадського здоров'я.

Наразі відбувається реформування галузей охорони здоров'я, освіти та науки України, що потребує реорганізації Центру громадського здоров'я й обласних Центрів контролю та профілактики хвороб. Це спонукає до перегляду основних положень системи епідеміологічного нагляду (ЕН) за інфекційними та неінфекційними хворобами, осмислення проведення епідеміологічної діагностики з використанням статистичних методів, які є методичною основою доказової медицини. Водночас вдосконалення діагностичної підсистеми ЕН шляхом застосування сучасних інформаційних технологій у практичній діяльності є однією з найбільш важливих і стійких тенденцій розвитку доказової медицини і стимулює до підготовки компетентних фахівців. На сьогодні здобуття та обробка інформації, обмін нею та її захист, набуття навичок користування джерелами інформації є пріоритетними напрямками розвитку медичної галузі. В умовах тотальної інформатизації всіх сфер життя одним з важливих питань стає використання комп'ютера як засобу та інструмента якісного виконання професійних обов'язків. Сучасні методи збору медичних та епідеміологічних даних (наприклад, через електронні медичні картки, онлайн-реєстрацію захворювання, мобільні додатки) призводять до значного збільшення обсягів інформації, що потребує ефективного аналізу. Комп'ютерні програми можуть обробляти великі обсяги даних, що практично неможливо виконати вручну.

Отже, актуальність запропонованого навчального посібника підтверджується стрімкою цифровізацією всіх сфер діяльності та необхідністю підготовки компетентних лікарів з доказової медицини в епідеміології, а також розробленим у 2023 р. документом «Рамка цифрової компетентності працівника охорони здоров'я України».

1. ЕПІДЕМІОЛОГІЧНА ДІАГНОСТИКА – МЕТОДИЧНА ОСНОВА ЕПІДЕМІОЛОГІЧНОГО НАГЛЯДУ

Дослідження спалахів інфекційних хвороб, епідемій та пандемій сприяло формуванню в медицині нового методу дослідження – епідеміологічного, який з часом все більше вдосконалювався та утверджувався. Епідеміологічний метод давав змогу вивчати внутрішню, невидиму сутність епідемій та пандемій, яка проявлялася захворюваністю в умовах дії на них соціальних, природних та біологічних чинників. За допомогою епідеміологічних досліджень можна встановити причини розвитку будь-яких патологічних процесів серед населення.

Наразі в усьому світі широкого визнання отримало поняття «клінічна епідеміологія», яка є методологічною основою «доказової медицини». Клінічна епідеміологія досліджує закономірності поширення, лікування та профілактику будь-яких захворювань, використовуючи епідеміологічні методи, а доказова медицина – це технологія пошуку, аналізу та оцінки достовірної наукової інформації про найефективніші, найнебезпечніші, найновіші та економічно вигідні медичні технології діагностики, лікування та профілактики захворювань. Таким чином, методичну основу клінічної епідеміології складають епідеміологічні дослідження (випадок-контроль, когортні дослідження, рандомізовані контрольовані дослідження тощо), які є важливим джерелом отримання доказів, необхідних для практичної діяльності.

Аналітичний вид діяльності лікаря-епідеміолога передбачає епідеміологічний аналіз інфекційної захворюваності з використанням статистичних методів, зокрема аналіз кількісних характеристик епідемічного процесу в динаміці, виявлення закономірностей на основі вивчення якісних ознак епідемічного процесу при різних нозологічних формах, а також визначення провідного типу епідемічного процесу за допомогою прийомів епідеміологічної діагностики тощо. Отже, предметом вивчення інфекційної епідеміології залишається епідемічний процес, який є безперервним процесом взаємодії макроорганізму та мікроорганізму (збудника-паразита) на популяційному рівні та проявляється за певних умов поодинокими, множинними та безсимптомними формами захворювання, супроводжується поширенням інфекційної захворюваності серед людей і забезпечує збереження у природі збудника як біологічного виду (Громашевський Л. В. Вибрані праці. Теоретичні питання імунології. Т. 2. Київ : Здоров'я, 1987).

ЕН за інфекційними та неінфекційними захворюваннями можна представити як систему взаємопов'язаних та взаємообумовлених елементів або підсистем (інформаційна, діагностична та управлінська). Функції інформаційної підсистеми – це збір даних про інфекційну захворюваність, стан здоров'я населення та фактори середовища життєдіяльності. Функції діагностичної підсистеми: встановити причинно-наслідковий зв'язок між інфекційними хворобами та факторами середовища життєдіяльності на підставі аналізу отриманої інформації в результаті проведення клініко-діагностичного, мікробіологічного, імунологічного, соціально-екологічного та інших моніторингів і поставити епідеміологічний діагноз. Управлінський моніторинг передбачає розробку, планування та впровадження програм профілактичних та протиепідемічних заходів, адекватних епідемічній ситуації, проведення контролю ефективності профілактичних та протиепідемічних заходів і при необхідності їх корекцію. Діяльність цього блоку

повинна бути направлена на всі ланки епідемічного процесу: джерело збудника інфекції, шляхи і фактори передачі та сприйнятливий організм.

Збір даних при проведенні ЕН є важливим етапом для виконання зазначених завдань, які направлені передусім на розпізнавання захворюваності та епідеміологічного стану населення, що за своєю сутністю є епідеміологічною діагностикою. Результатом є формулювання епідеміологічного діагнозу щодо причин (факторів ризику) виникнення та поширення патологічних станів. Якість інформації про випадки захворювання, фактори середовища життєдіяльності, про стан та чисельність населення є запорукою будь-якого епідеміологічного дослідження. Помилки, які допускають на цьому етапі ЕН, практично неможливо потім усунути.

Таким чином, для потреб ЕН в контексті програми «Єдине здоров'я» на міжнародному, а також на національному, регіональному та локальному рівнях введено поняття «стандартне епідеміологічне визначення випадку», яке ґрунтується на клінічних, епідеміологічних та лабораторних критеріях певної нозологічної форми. Таке визначення випадку забезпечує однакову діагностику кожної нозологічної форми, дозволяє порівняти кількість випадків, які виникли в певний час та в певному місці, з кількістю випадків, які виникли в іншому місці та в інший час. Епідеміологічне визначення далеко не завжди співпадає з клінічним визначенням випадку, оскільки застосовується не для клінічної діагностики та вибору лікування, а для того щоб чітко визначити випадок інфекційного захворювання. В епідеміологічній практиці застосування визначення випадку дозволяє забезпечити активне виявлення випадків при епідеміологічному розслідуванні навіть тоді, коли клінічний діагноз не сформульований.

Аналіз сучасних тенденцій розвитку епідеміології, як загальномедичної науки, свідчить про все ширше використання аналітичних методів. Це значною мірою стимулювало еволюцію епідеміологічної науки в цілому і визначило спрямованість її розвитку: епідеміологія досить чітко трансформувалася з описової науки в аналітичну, що дало можливість сформулювати основні принципи епідеміологічної діагностики, її мету, завдання та сутність. Епідеміологічну діагностику правомірно можна розглядати як розпізнавання епідеміологічних ризиків інфекційної захворюваності та епідеміологічного стану населення, враховуючи основне завдання реформи охорони здоров'я – переорієнтацію від політики лікування до політики профілактики хвороб. Тож, у системі ЕН та контролю за інфекційними хворобами назріла необхідність переходу від стратегії втручання в епідемічний процес на підставі захворюваності (тобто факту, що стався) до стратегії оцінки ризику з розробкою та впровадженням системи забезпечення епідеміологічної безпеки, що базується на цьому підході. Епідеміологічний ризик – це вірогідність негативного впливу на захворюваність (та/або його наслідків) окремих груп населення, обумовлених зовнішніми та/або внутрішніми факторами, які діють в певний час на конкретній території. Епідеміологічний ризик включає визначення потенційних і реальних небезпек ускладнення епідемічної ситуації, оцінку ймовірності (ступеня) її ускладнення, а також розрахунок можливих наслідків та інтегрального збитку під впливом окремих чи сукупних чинників ризику.

Отже, з позицій ЕН, епідеміологічна діагностика – це процес постановки епідеміологічного діагнозу, що є логічною формулою, яка містить оцінку епі-

демічної ситуації та причин її виникнення на конкретній території серед певних груп населення в досліджуваній відрізок часу з метою раціоналізації профілактичних і протиепідемічних заходів та розробки епідеміологічного прогнозу.

Розділи епідеміологічної діагностики: семіотика, діагностична техніка, епідеміолого-діагностичне мислення.

Семіотика епідеміологічної діагностики – розділ вчення про закономірності проявів епідемічного процесу (інтенсивність, тенденція, динаміка, розподіл захворюваності в різних групах населення, по території, за часом).

Кожна із зазначених характеристик епідемічного процесу має своє діагностичне значення, методи оцінки та одиниці виміру, ступінь інформативності та прийоми графічного зображення.

Діагностична цінність тенденції полягає в тому, що вона відображає силу дії причин виникнення та розвитку епідемічного процесу та умов, за яких вони діють. Інтенсивність епідемічного процесу оцінюється у різних вікових, соціальних, професійних та інших групах населення, може бути представлена маніфестними та латентними формами захворювання, а також носійством збудників та хронічним перебігом інфекції.

Динаміка епідемічного процесу характеризує зміни його тенденцій у часі. Ця характеристика відображає посилення або послаблення у часі (багаторічні, місячні чи інші інтервали) сили дії причин та умов розвитку епідемічного процесу. Динаміку епідемічного процесу можна вивчати за нозологічними формами, на окремих територіях та у різних групах населення. У цьому випадку кожна з розробок може представляти свою цінність для постановки діагнозу.

Вивчаючи тенденцію епідемічного процесу на різних територіях, можна виявити відмінності в силі впливу причин та умов в окремих регіонах, районах, містах, населених пунктах тощо. Під час аналізу території необхідно розділити їх залежно від впливу природних, екологічних, соціальних та інших факторів. Застосування картографічного методу істотно розширює можливості та інформативність цього аналізу, а виконання картодіаграм забезпечує наочність отриманих результатів.

Діагностична цінність структурних показників (наприклад, розподіл захворюваності в різних групах) полягає в тому, що вони вказують на співвідношення окремих частин проявів епідемічного процесу. Не висвітлюючи інтенсивності процесу, представлені у поступовій динаміці, вони дозволяють оцінити зміни у співвідношенні елементів. У процесі аналізу структурні показники використовуються для вивчення вікової, соціальної, професійної, етіологічної та інших характеристик епідемічного процесу.

Діагностична техніка – комплексний статистичний аналіз даних обліку інфекційної захворюваності порівняно з даними вивчення біолого-екологічних характеристик збудників інфекцій, імунобіологічних особливостей груп населення, а також природних і соціальних умов, в яких розвивається епідемічний процес.

Діагностична техніка – це сукупність технічних прийомів, що забезпечують розпізнавання проявів епідемічного процесу. Основні технічні прийоми, які використовуються для цього, такі:

- організація збору епідеміологічної інформації;

- використання обліково-звітних форм, що містять відомості про інфекційні хвороби;
- статистична обробка даних про захворюваність;
- облік впливу соціальних та природних умов на рівні захворюваності;
- облік даних лабораторних досліджень (біологічні властивості збудників, імуноструктура населення тощо);
- побудова таблиць, рисунків, графіків, картограм.

Крім цього, як технічні прийоми можуть бути використані гігієнічні, хімічні, бактеріологічні, клінічні, інструментальні, санітарно-технічні, метеорологічні та інші методи.

Якість та повнота епідеміологічної інформації є важливою умовою для проведення епідеміологічної діагностики. До інформації, яка необхідна для епідеміологічних досліджень, відноситься: частота зареєстрованої захворюваності; частота безсимптомних форм інфекції; властивості циркулюючих збудників; структура населення щодо сприйнятливості та імунітету в динаміці; ефективність проведення протиепідемічних заходів; результати спостереження за епідемічно значущими об'єктами та продукцією; епідемічна та епізоотична обстановка на територіях, у тому числі суміжних, звідки можливе занесення збудників інфекції; природні, екологічні та соціально-економічні умови території; демографічні процеси тощо.

Епідеміолого-діагностичне мислення – логічний метод, який використовують в процесі діагностики епідемічної ситуації для обґрунтування гіпотези про причинно-наслідкові зв'язки захворюваності з факторами та умовами виникнення цієї захворюваності. Методологічною основою є професійна підготовка фахівців, методи і прийоми наукового пізнання, зокрема прийоми формальної логіки.

Приєм диференціювання (різниці). Якщо частота захворюваності у двох групах різна і ці групи відрізняються за інтенсивністю впливу факторів (або є фактор тільки в одній групі), то можна припустити, що даний фактор або його відсутність є причиною чи необхідною умовою виникнення захворюваності.

Приєм подібності (схожості). Якщо один і той же фактор є загальним для багатьох різних умов, при яких реєструється захворюваність, то цей фактор розглядається як причина виникнення захворювання.

Приєм об'єднання різниці та схожості. Якщо при порівнянні декількох груп за допомогою методу схожості можна буде виявити причину або умови захворюваності, а потім за допомогою цілеспрямованих заходів у ряді груп цю причину усунути, тобто створити різні групи за цим фактором, то за зміною захворюваності можна було б оцінити роль даного фактора у виникненні захворюваності.

Приєм супутніх змін. Якщо при виявленні фактора його частота та сила змінюються паралельно зі зміною рівня захворюваності, то можна припустити, що існують причинно-наслідкові зв'язки між цим фактором і захворюваністю.

Приєм залишків. Якщо для диференційованої оцінки ролі в епідемічному процесі декілька факторів ризику, припущення про які обґрунтовані іншими методами побудови гіпотез, то можна використовувати послідовне виключення окремих факторів шляхом проведення цілеспрямованих заходів. За величиною

«залишків» захворюваності на кожному етапі і оцінюється значущість кожного із факторів.

Прийом узгодження. Чим більша кількість ознак узгоджується одна з одною і з дією фактора ризику, який ми припускаємо на підставі раніше розглянутих методів, тим вище вірогідність існування причинно-наслідкових зв'язків даного фактора із захворюваністю. Цей метод використовують для підтвердження гіпотез, які побудовані за допомогою інших методів.

Прийом аналогій. Гіпотеза будується на підставі застосування встановлених епідеміологічних принципів і положень, які схожі з природними ситуаціями. Якщо виявлено схожість проявів епідемічного процесу двох або декількох інфекцій, що викликані збудниками близьких еколого-біологічних груп, то закономірності, які встановлені для однієї із цих інфекцій, можуть бути використані для пояснення причин поширення при схожих умовах інших нозологічних форм.

При формулюванні гіпотез щодо причин та умов розвитку епідемічного процесу використовують два взаємопов'язані підходи: 1) встановлюють прояви епідемічного процесу, висувають гіпотезу, яка їх пояснює, а потім зв'язують з науковими даними; 2) на основі наукових даних формують гіпотезу, що пояснює прояви епідемічного процесу, потім її зв'язують з конкретними (фактичними) проявами епідемічного процесу.

Формулювання гіпотез найважливіше та водночас найскладніше завдання епідеміологічних досліджень, що забезпечує виявлення проблем профілактики, а також висування, перевірку та доказ гіпотез щодо умов та механізму розвитку епідемічного процесу при конкретній нозологічній формі інфекційної хвороби. Підвищення рівня захворюваності у певних групах населення дозволяє припустити можливе існування провідних факторів ризику чи індивідуальних чинників. Гіпотезу можна перевірити під час проведення додаткових досліджень, спланованих для того, щоб продемонструвати, чи є виявлена залежність стійкою і чи збільшується рівень захворюваності пропорційно ступеню впливу досліджуваного чинника.

Епідеміолого-діагностичне мислення базується на загальній та спеціальній професійній підготовці епідеміолога (компетентності), вихідних відомостях, які зазначені в інструктивно-методичних документах МОЗ України, та нових наукових даних щодо конкретної інфекційної хвороби.

У ряді випадків сформульована гіпотеза може бути підтверджена контрольованим експериментом, в якому роблять спробу вплинути на частоту виникнення захворювання шляхом зменшення ступеня впливу причинного фактора. Зрештою, тільки таким чином можна встановити причинно-наслідковий зв'язок і продемонструвати ефективність профілактичних та протиепідемічних заходів.

Етапи епідеміологічної діагностики:

1. Визначення конкретної мети та завдань епідеміологічної діагностики.
2. Розробка програми: визначення розділів, обсягу матеріалу, необхідних методів та залучених фахівців з огляду на гіпотезу щодо причин та умов виникнення й розвитку епідемічного процесу.
3. Розробка макета (карти) для збирання вихідної інформації.
4. Збір інформації, перевірка її достовірності.

5. Підготовка макетів аналітичних таблиць.

6. Попереднє угруповання, обробка та осмислення результатів, корекція програм, макетів, таблиць.

7. Статистична обробка, графічне зображення результатів.

8. Осмислення результатів: оцінка, зіставлення, виявлення ознак, визначення території, часу, груп, колективів та факторів ризику (встановлення причинно-наслідкових зв'язків).

9. Формулювання епідеміологічного діагнозу.

10. Прогнозування захворюваності.

За результатами епідеміологічної діагностики передбачено вирішення таких завдань:

1) оцінити прояви епідемічного процесу серед різних груп населення, по території та в часі (виявлення груп, територій та часу ризику);

2) виявити конкретні фактори середовища життєдіяльності, що впливають на прояви епідемічного процесу (формулювання гіпотез про фактори ризику);

3) перевірити сформульовані гіпотези, виявити причинно-наслідкові зв'язки між факторами ризику та захворюваністю;

4) скласти найближчий та (або) віддалений прогноз захворюваності, оцінити достовірність гіпотез про фактори ризику за ефектом;

5) визначити пріоритетні хвороби за епідеміологічною, соціальною та економічною значущістю (ефективністю) заходів профілактики (оцінити витрати і користь);

б) обґрунтувати проведення ефективних профілактичних та протиепідемічних заходів.

Таким чином, в умовах реформування охорони здоров'я, реорганізації санітарно-епідеміологічної служби в центри контролю та профілактики хвороб і створення системи громадського здоров'я важливо та необхідно при впровадженні нових підходів щодо визначення випадку інфекційного захворювання та проведення епідеміологічного нагляду за ними знати розділи епідеміологічної діагностики та використовувати статистичні методи епідеміологічного аналізу для постановки епідеміологічного діагнозу й прогнозування інфекційної захворюваності. За допомогою прийомів епідеміологічної діагностики можна на національному, регіональному та локальному рівнях виявляти епідеміологічні ризики (групи, території, час та фактори) та розробляти ефективні профілактичні заходи щодо запобігання виникненню захворювання. Такий підхід до проведення епідеміологічного нагляду за інфекційними хворобами дозволить здійснити переорієнтацію суспільства від політики спостереження за інфекційними хворобами (в клінічній практиці лікування) до політики профілактики їх виникнення, як зазначено в концепції системи громадського здоров'я.

Епідеміологічна діагностика має два види: ретроспективний та оперативний епідеміологічний аналіз захворюваності на інфекційні хвороби.

1.1. РЕТРОСПЕКТИВНИЙ ЕПІДЕМІОЛОГІЧНИЙ АНАЛІЗ

Ретроспективний аналіз являє собою вивчення сформованих у минулому тенденцій. Ретроспективний епідеміологічний аналіз – це основа епідеміологічної діагностики.

Основні напрямки ретроспективного епідеміологічного аналізу щодо окремих груп чи нозологічних форм такі: санітарно-епідеміологічна експертиза території; аналіз багаторічної динаміки захворюваності населення; аналіз річної (помісячної) динаміки захворюваності за середніми багаторічними даними в різні за інтенсивністю періоди перебігу епідемічного процесу; аналіз рівня та структури захворюваності за нозологічними формами в соціально-вікових групах; аналіз захворюваності за факторами ризику; аналіз якості та ефективності профілактичних заходів.

Всі напрямки взаємопов'язані, а сукупність розрахунків уточнюється під час аналізу. Водночас розрахунки повинні бути спрямовані на виявлення груп, територій та часу ризику в пріоритетних групах і нозологічних формах інфекційних хвороб, а також на встановлення причинно-наслідкових зв'язків захворюваності з факторами ризику, формулювання та обґрунтування гіпотези, яка пояснює прояви епідемічного процесу.

Ретроспективний епідеміологічний аналіз передбачає порівняння отриманих даних щодо певної нозологічної форми з іншими захворюваннями, а також з даними тієї ж інфекції, але на іншій території.

Мета санітарно-епідеміологічної експертизи території – визначення можливого впливу природних та соціальних умов на епідеміологію інфекційних хвороб та рівень захворюваності на певній території.

Мета аналізу рівня та структури захворюваності за нозологічними формами – визначення епідеміологічної, соціальної та економічної значущості хвороб.

Мета аналізу багаторічної динаміки захворюваності – визначення тенденцій розвитку захворюваності, періодичності циклів її коливань; виявлення причин зміни рівня захворюваності за багаторічний період; прогнозування захворюваності.

Мета аналізу річної (помісячної) динаміки захворюваності – встановлення закономірностей помісячної динаміки; виділення основних форм прояву епідемічного процесу; оцінка можливих причин зміни її інтенсивності та структури.

Мета аналізу захворюваності за факторами ризику – виявлення факторів середовища життєдіяльності, які визначають ризик розвитку та поширення епідемічного процесу, щоб скласти план раціональних профілактичних та проти-епідемічних заходів.

Мета аналізу якості та ефективності профілактичних заходів – оцінка своєчасності, повноти та адекватності проведення заходів, якості та правильності використання протиепідемічних засобів, оцінка здатності протиепідемічних засобів і заходів запобігати захворюванням.

1.2. ОПЕРАТИВНИЙ ЕПІДЕМІОЛОГІЧНИЙ АНАЛІЗ

Оперативний (поточний) епідеміологічний аналіз є логічним продовженням ретроспективного аналізу і проводиться безпосередньо в ході епідеміологічного спостереження або одразу після підбиття його підсумків з метою оперативного впливу на досліджуваний епідемічний процес.

Це постійний безперервний моніторинг епідемічного процесу, який дозволяє швидко та своєчасно виявити причини та особливості розвитку епідемічного процесу за період спостереження, забезпечити оперативну епідеміологічну діагностику та обґрунтувати поточні управлінські рішення.

Основні напрямки оперативного епідеміологічного аналізу: оцінка та аналіз результатів лабораторних досліджень; безперервне спостереження за захворюваністю і оцінка її динаміки; епідеміологічне обстеження осередків з поодинокими та множинними (спалахи, епідемії) випадками захворювання.

2. ЕПІДЕМІОЛОГІЧНИЙ МЕТОД: ВИЗНАЧЕННЯ ТА СТРУКТУРА

Епідеміологічний метод – це сукупність методичних прийомів, які використовуються для вивчення захворюваності населення. Об'єднує їх мета дослідження, а саме: виявлення та оцінювання причинно-наслідкових зв'язків між факторами ризику і захворюваністю, тих самих зв'язків, які обумовлюють зростання, зниження або припинення захворюваності для обґрунтування заходів з профілактики захворювань та оцінки їх ефективності.

Структура епідеміологічного методу включає три сукупності методичних прийомів: епідеміологічне спостереження, епідеміологічний експеримент та математичне моделювання.

2.1. ЕПІДЕМІОЛОГІЧНЕ СПОСТЕРЕЖЕННЯ

Методичні прийоми: дескриптивний, або описово-оціночний (скринінг, епідеміологічне обстеження осередку, статистичне спостереження) – кількісна оцінка розподілу захворюваності по територіях, в групах населення та за часом, формування первинних гіпотез про фактори ризику, що дозволяє виявити: хто, коли та де захворів, якщо мова йде про інфекційні захворювання або інші порушення здоров'я; аналітичний (дослідження типу «випадок–контроль», когортне дослідження, статистичне вимірювання зв'язку) – перевірка та оцінка гіпотез про причинно-наслідкові зв'язки проявів захворюваності, спрямована на визначення, чому і як виникають інфекційні хвороби та інші порушення здоров'я, що дозволяє визначити напрямки профілактики відповідно до факторів ризику.

2.2. ЕПІДЕМІОЛОГІЧНИЙ ЕКСПЕРИМЕНТ

Методичні прийоми: природний експеримент, неконтрольований експеримент, контрольований експеримент, фізичне і біологічне моделювання, епізоотологічний експеримент, що дозволяє підтвердити гіпотези, тобто кількісно оцінити ефективність профілактичних заходів та побудувати і оцінити нові наукові гіпотези.

2.3. МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ

Методичні прийоми: аналітичне дослідження, дослідження з використанням інформаційних технологій.

Математичне моделювання передбачає складання та вирішення математичних рівнянь, побудову дерева рішень, стохастичне моделювання. За допомогою цього дослідження прогнозується захворюваність або фактори ризику.

Моделлю (від лат. «*modulus*» – зразок) системи, об'єкта чи явища називають штучну систему або об'єкт, що в певних умовах може замінити систему-оригінал або об'єкт шляхом відтворення найбільш характерних параметрів та властивостей об'єкта, які нас цікавлять. Математична модель – це система математичних залежностей між величинами, що характеризують модельований об'єкт.

Розрізняють теоретико-аналітичні та прикладні медико-соціальні моделі. Перші використовуються для дослідження загальних властивостей і закономірностей епідемічних процесів, другі – для розв'язування конкретних задач аналізу, прогнозування, управління процесами та об'єктами. Залежно від визначеності використовуваної інформації розрізняють детерміновані та ймовірнісні моделі.

Якщо залежність в моделі належить до одного й того ж моменту або періоду часу, то така модель називається статичною. Модель, яка відображає зміни епідемічних процесів у часі, є динамічною моделлю. За формою математичних залежностей моделі поділяються також на лінійні та нелінійні.

Процес побудови та дослідження моделі називається моделюванням. Для побудови моделі необхідно:

- 1) мати достатню сукупність спостережень вхідних даних;
- 2) забезпечити однорідність сукупності спостережень;
- 3) забезпечити точність вхідних даних.

Математичні моделі застосовуються для прогнозу перебігу епідемічного процесу. Основним методом прогнозування еволюційного розвитку є метод прогнозу екстраполяції, який базується на використанні принципу (закону) інерції, тобто перенесенні тенденції (тренду) розвитку інфекційного захворювання, що спостерігалася в минулому і теперішньому періодах, на майбутній період.

Розрізняють базу екстраполяції та строки прогнозування. База екстраполяції – це кількість інтервалів часу (10–15), за які вивчається тенденція, а строки прогнозування – це кількість періодів часу, на які розраховують прогноз. Залежно від строків прогнозу класифікують на короткострокові (1–2 періоди), середньострокові (3–5 періодів) та довгострокові (більше 5 періодів). Прогноз, що характеризується одним значенням, називається точковим, трьома – інтервальним (нижня та верхня границі прогнозу і його точкове значення). Чим довша база та коротші строки прогнозу, тим він точніший.

Математичні моделі поширення інфекцій можна розділити за метою їх використання на такі категорії [29]:

- моделі для ідентифікації спалахів епідемій за даними у режимі реального часу;
- методи машинного навчання для прогнозування поширення інфекції;
- моделі для аналізу та прогнозування поширення інфекції при різних протиепідемічних заходах.

Для першої категорії використовують статистичні моделі, що покликані за поточними даними про досліджувану інфекцію (кількість звернень за медичною допомогою, кількість госпіталізацій, смертність тощо) констатувати факт початку епідемії.

У цій категорії найбільш популярним методом моделювання є тест CUSUM [27], який можна описати таким чином: нехай $y(ti)$, $i = 1, 2, 3, \dots, n$ – число випадків інфікування. Величина кумулятивної суми визначається як:

$$CUSUM(i) = \sum_{j=1}^i (y(t_j) - k_j),$$

де k_j – референтне значення випадків інфікування в момент часу t_j .

При перевищенні CUSUM порогового значення фіксується факт початку епідемії.

Останніми роками набули поширення методи машинного навчання, аналізу та прогнозування характеристик епідемії на основі пошукових запитів і постів у соціальних мережах з визначенням місцеположення хворого. Однак отримані моделі машинного навчання є складними для інтерпретації.

Моделі третьої категорії, так звані компартментні, можуть бути використані для передбачення динаміки поширення вірусних інфекцій та аналізу ефективності протиепідемічних заходів.

Компартментні моделі характеризуються наявністю простору станів (компартментів) та описом правил переходу від одного стану до іншого. Кількість та вид станів можуть бути різними. У найпростішому випадку використовуються три стани: сприйнятливий; інфікований; що одужав (класична SIR-модель, названа за першими літерами станів: S – susceptible, I – infected, R – recovered), які запропоновані шотландськими епідеміологами Кермаком та Маккендріком ще у 1920-х роках (рис. 1).

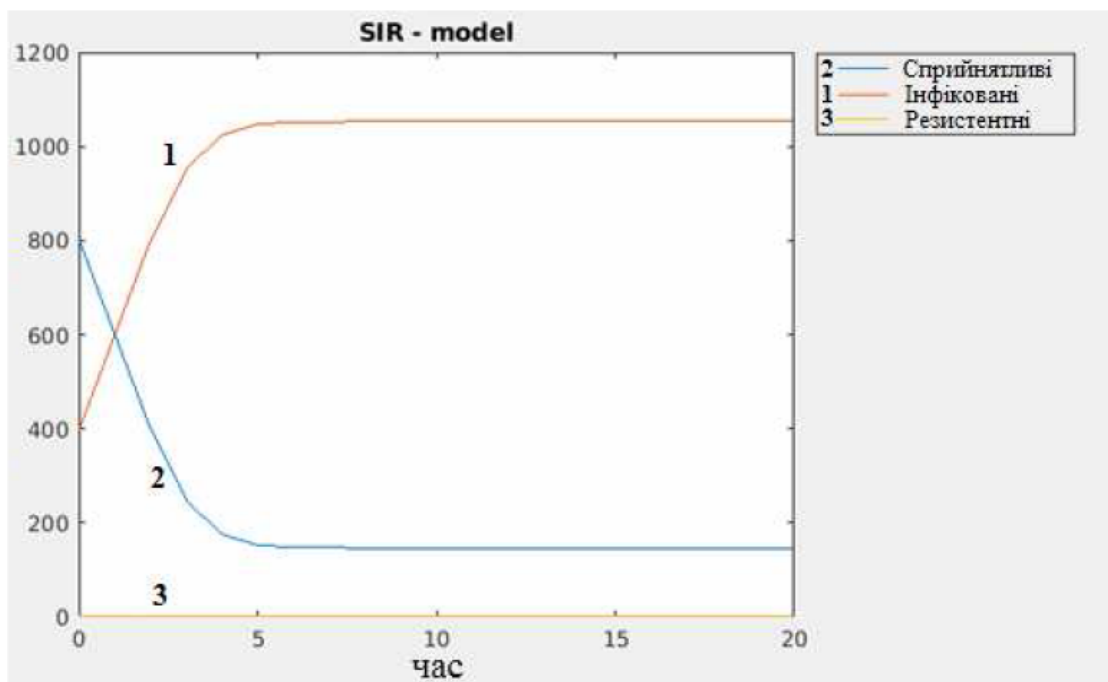


Рис. 1. Компартментна модель прогнозування за сприйнятливістю, інфікованістю та резистентністю

В основі сучасного математичного моделювання епідемій закладений принцип експоненціального зростання числа хворих. Це означає, що зростання числа хворих з плином часу поступово збільшується пропорційно певному коефіцієнту, який є константою і розраховується для кожного захворювання окремо.

Для прогнозування використовуються системи диференціальних рівнянь, де як функції виступають кількість хворих і здорових людей на певній території. Розв'язанням такої системи рівнянь є рівень інфекційної захворюваності в кожен момент модельного часу. Ця методологія була детально розроблена в 1970-ті роки академіком О. В. Барояном і професором Л. О. Рвачовим. Для її створення використовувався метод наукової аналогії у відображенні епідемічного процесу «перенесення» збудника інфекції від хворих до здорових з процесом «перенесення» матерії (енергії, імпульсу тощо) в рівняннях математичної фізики. Побудована модель відповідає загальному закону розвитку епідемії, який базується на принципах пропорційності, незалежності та післядії.

Принцип пропорційності: добовий приріст числа носіїв вірусу в інфікованому суспільстві пропорційний сумарному числу носіїв вірусу, що виникли за попередній час розвитку епідемії.

Принцип незалежності: відносний добовий приріст числа носіїв вірусу не залежить від сумарного числа носіїв вірусу, що виникли за попередній час розвитку епідемії.

Принцип післядії: відносний добовий приріст числа носіїв вірусу із запізненням відображає заходи щодо управління епідемією або їх саботаж. За допомогою моделі Барояна-Рвачова обчислювалися прогнози захворюваності на грип для окремих територій.

На динаміку сучасних вірусних епідемій впливають зміни, які відбуваються в паразитарній системі внаслідок нетривалого специфічного імунітету, можливість повторних інфікувань, високий рівень мутацій вірусу. Для урахування цих особливостей вірусів були розроблені більш складні моделі, до яких входять 4 і більше диференціальних рівнянь. Це такі моделі:

- SIRS – модель перебігу епідемій з тимчасовим специфічним імунітетом;
- SEIR (E – exposed, група людей, які контактували з джерелом інфекції) – модель поширення захворювання з урахуванням інкубаційного періоду;
- SIS – модель поширення захворювання, до якого не формується специфічний імунітет;
- MSEIR (M – maternally derived immunity, мають природний пасивний імунітет) – модель, яка враховує природний пасивний імунітет дітей завдяки материнським антитілам;
- SEIR-HCD (H – hospitalized – госпіталізовані, C – critical – в критичному стані, D – dead – померлі) – модель враховує динаміку госпіталізації пацієнтів, критичний стан пацієнтів, померлих від захворювання;
- інші.

3. СТАТИСТИЧНИЙ АНАЛІЗ В ЕПІДЕМІОЛОГІЇ

Епідеміологія на сучасному етапі одночасно розглядається як:

- особливий підхід до вивчення закономірності перебігу хвороб із застосуванням теорії ймовірностей, статистики та специфічних методів проведення наукових експериментів;

- метод, за допомогою якого можна встановити причинно-наслідкові зв'язки шляхом перевірки гіпотез про причини виникнення хвороб та способи їх попередження;

- інструмент для прийняття управлінських рішень у сфері охорони здоров'я, заснований на наукових даних та розкритих причинно-наслідкових зв'язках.

Епідеміологічний нагляд – це систематичний та безперервно діючий збір, облік, складання та аналіз, тлумачення, поширення даних медико-санітарного призначення, епідемічного благополуччя населення і показників середовища життєдіяльності для оцінки і прийняття необхідних відповідних заходів у сфері громадського здоров'я (наказ МОЗ України № 1726 від 30.07.2020).

Статистичний аналіз епідеміологічних даних є одним із методів епідеміологічного аналізу та одночасно одним із розділів прикладної статистики, що вивчає процеси збирання, аналізу, інтерпретації та подання даних.

У *табл. 1* наведені різні підходи до епідеміологічного аналізу залежно від поставленого завдання.

Таблиця 1

Підходи до епідеміологічного аналізу залежно від поставленого завдання

Ознака класифікації	Підходи до аналізу	Завдання
За групами інфекцій	Тематичний (дослідження однієї нозологічної форми)	Дескриптивний та порівняльний аналіз показників захворюваності населення за нозологічними формами в окремій групі інфекцій
	Комплексний (дослідження декількох нозологічних форм)	Дескриптивний та порівняльний аналіз показників захворюваності населення за групами інфекцій
За визначеннями нозологічних форм	Тематичний (дослідження однієї нозологічної форми)	Дескриптивний та порівняльний аналіз показників захворюваності населення із середнім багаторічним рівнем захворюваності для окремої нозологічної форми
	Комплексний (дослідження декількох нозологічних форм)	Дескриптивний та порівняльний аналіз показників захворюваності населення із середнім багаторічним рівнем захворюваності для всіх зареєстрованих нозологічних форм
За територіями	Тематичний (дослідження однієї нозологічної форми)	Дескриптивний та порівняльний аналіз показників захворюваності населення із середнім багаторічним рівнем захворюваності для окремої території
	Комплексний (дослідження декількох нозологічних форм)	Дескриптивний та порівняльний аналіз показників захворюваності населення за територіями
За групами ризику	Тематичний (дослідження однієї нозологічної форми)	Дескриптивний та порівняльний аналіз показників захворюваності населення із середнім багаторічним рівнем захворюваності для окремої групи ризику. Визначення груп ризику
	Комплексний (дослідження декількох нозологічних форм)	Дескриптивний та порівняльний аналіз показників захворюваності населення по групам ризику. Визначення груп ризику

Ознака класифікації	Підходи до аналізу	Завдання
За охопленням населення	Суцільний	Аналіз даних захворюваності для всього населення
	Вибірковий	Вибіркова оцінка захворюваності населення
За користувачами аналізу	Внутрішній	Облік показників захворюваності населення і профілактичних та протиепідемічних заходів. Складання плану профілактичних заходів на наступний період.
	Зовнішній	Зведення даних захворюваності населення для представлення зацікавленим структурам
За часом	Ретроспективний	Визначення характеристик багаторічної динаміки захворюваності населення (тенденція, циклічність, нерегулярні коливання). Складання прогнозу захворюваності населення на початку часового періоду
	Оперативний	Аналіз даних захворюваності населення за результатами часового періоду

Специфіка статистичного аналізу епідеміологічних даних визначається насамперед їхньою природою і структурою, тобто інформаційною базою. Особливістю епідеміологічної інформації є те, що вона розглядається для певної території, певного контингенту та певного періоду часу. Статистичний аналіз виконується за допомогою методів математичної статистики, які відповідають типу досліджуваної інформації та розподілу числових даних.

Аналіз даних включає в себе ряд процесів, спрямованих на те, щоб надати зібраній інформації осмислений характер. До таких процесів відносяться збір, очищення, аналіз, інтерпретація та візуалізація даних.

Головна мета статистики – знайти закономірності та зв'язки даних і використати їх для прогнозування майбутніх подій або прийняття рішень. Особливістю статистичного дослідження є те, що зв'язки між досліджуваними ознаками через дії випадкових факторів-причин проявляються не в кожному окремому спостереженні, а тільки їх в сукупності.

Класична статистика ділиться на дві основні сфери (рис. 2):

1. **Дескриптивна (описова) статистика** – це статистичний метод, який дозволяє описувати та підсумовувати дані. Вона використовується для представлення даних у зручній та зрозумілій формі, наприклад, у вигляді графіків або таблиць. Дескриптивна статистика допомагає зрозуміти основні характеристики даних, такі як середнє значення, медіана та розкид. До дескриптивної статистики також відноситься візуалізація (унаочнення або графічне зображення даних), яка використовується для перетворення складних і великих обсягів інформації у зрозумілу та наочну форму, щоб швидше її аналізувати, інтерпретувати та робити висновки. Візуалізація передбачає створення таких зображень, які передають зв'язки між представленими даними, і тим самим роблять їх очевидними для оглядача. Така статистика допомагає попередньо визначити пріоритетність інфекційних хвороб (проблем) за характеристиками центральних тенденцій (середня арифметична, медіана, мода) та розподілом за віком, територією, часом.

2. Інференційна статистика – це статистичний метод, який дозволяє робити висновки про генеральну сукупність на основі вибірки даних. Вона використовується для узагальнення результатів дослідження на популяцію. Цей метод допомагає визначити, наскільки достовірними є висновки і які висновки можна зробити про популяцію на основі вибірки.

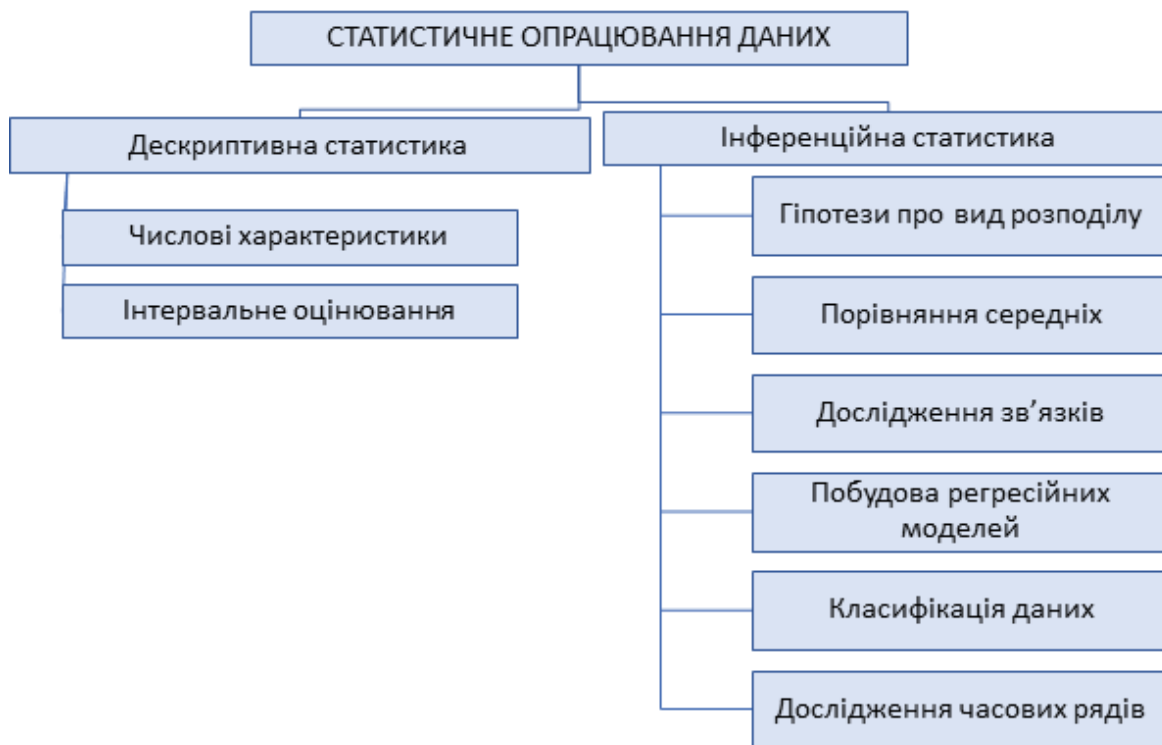


Рис. 2. Класифікація методів статистичного опрацювання даних

Існує ще декілька класифікацій статистичних методів (рис. 3–6).

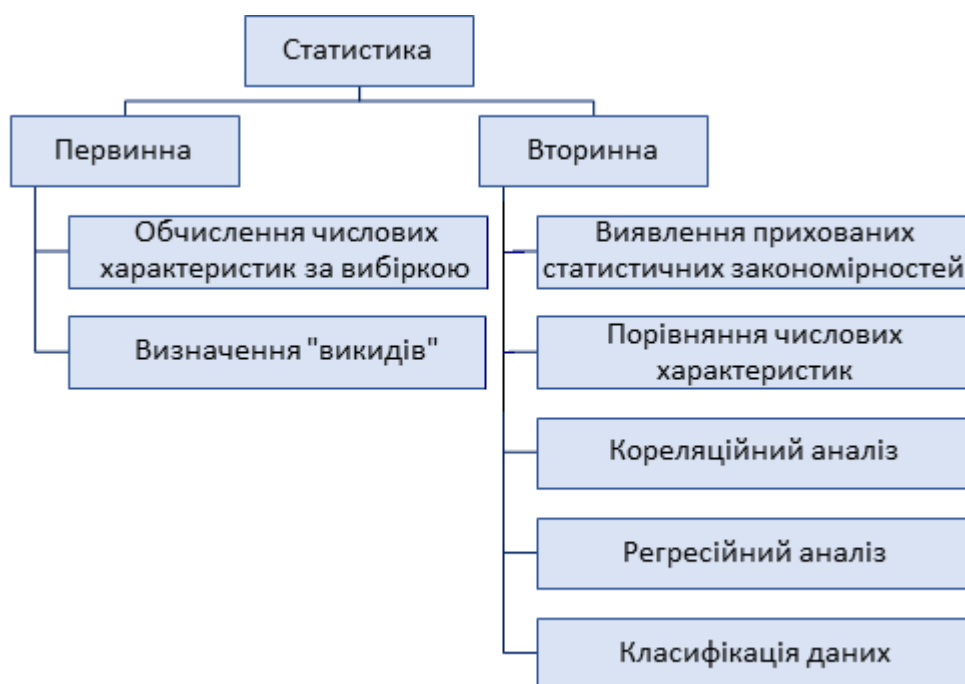


Рис. 3. Класифікація статистичних методів залежно від етапу дослідження



Рис. 4. Класифікація статистичних методів залежно від типу розподілу числових даних

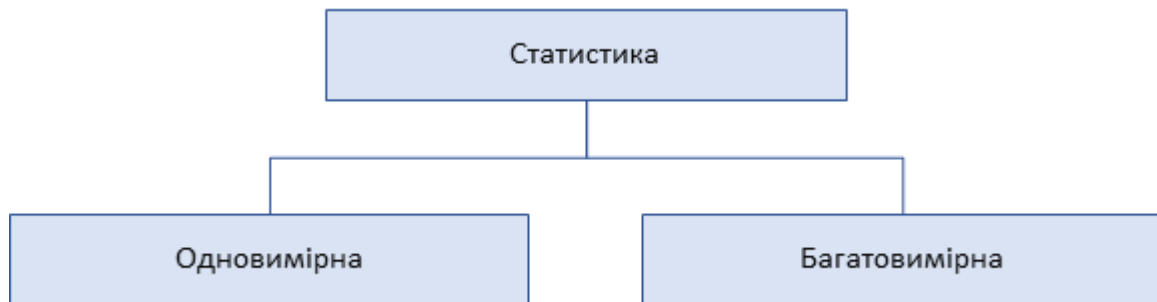


Рис. 5. Класифікація статистичних методів залежно від кількості показників, що досліджуються

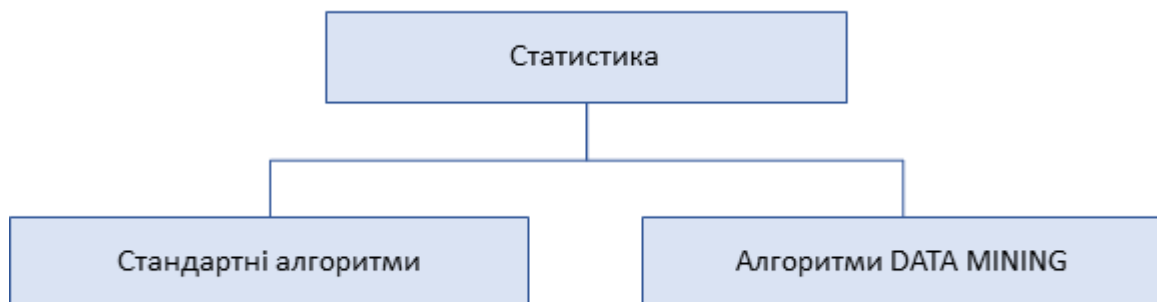


Рис. 6. Класифікація статистичних методів залежно від типу алгоритмів опрацювання даних

Стандартні алгоритми відповідають базовим формулам обчислення числових характеристик та статистичних критеріїв. Алгоритми інтелектуального аналізу даних Data Mining – це унікальні сучасні обчислювальні алгоритми, які працюють з великими обсягами інформації з метою побудови нетривіальних моделей і класифікацій, таких як дерева рішень, узагальнені моделі зв'язку та прогнозу даних, кластеризації, технології нейронних мереж. Для використання алгоритмів Data Mining потрібні потужні комп'ютери та програмне забезпечення.

Статистичний аналіз даних може проводитися вручну або із застосуванням комп'ютерної техніки (рис. 7). В умовах цифровізації суспільства та широкого впровадження комп'ютерної техніки в усі сфери людської діяльності перевага звичайно за комп'ютерним статистичним аналізом. Сучасний ринок програмних продуктів пропонує різноманітні пакети програм для статистичної обробки даних. Всесвітньо відомі статистичні пакети для комплексної обробки даних: BMDP, SPSS, SAS, Systat, Minitab, S-Plus, Statgraphics, Statistica, EpiInfo, EpiToos та ін. Достатня кількість засобів для проведення статистичного аналізу є у популярному табличному процесорі Microsoft Excel.

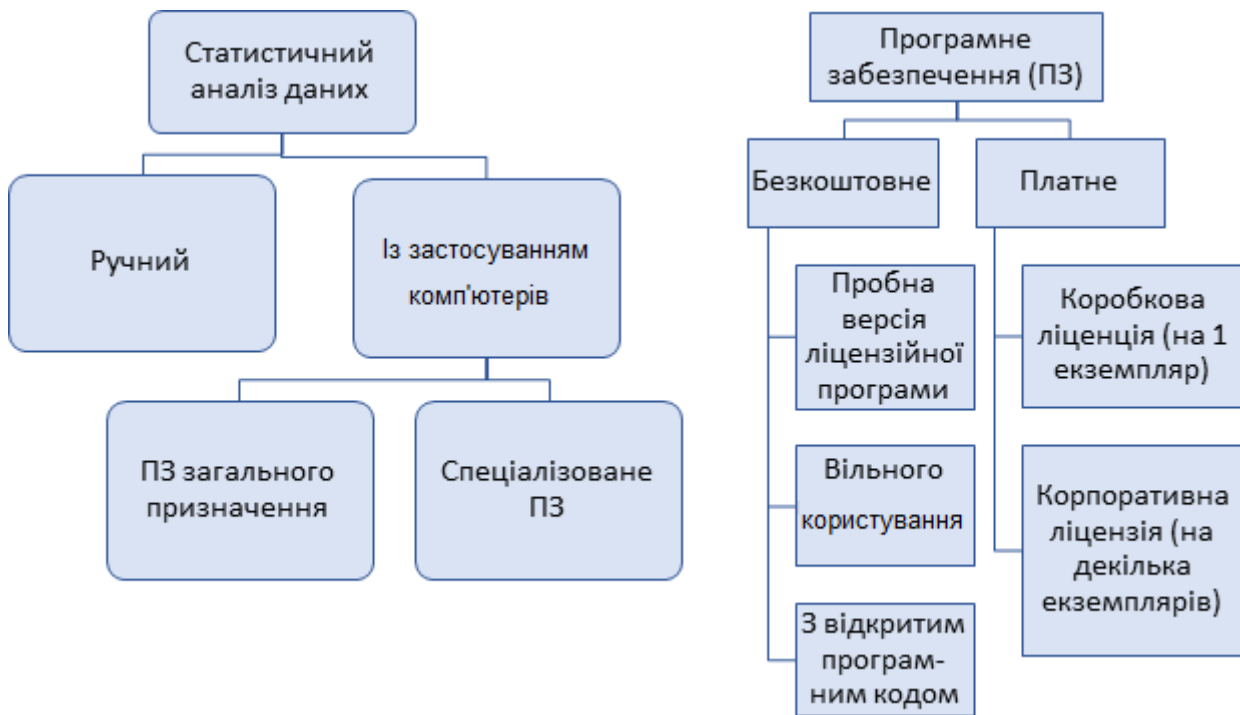


Рис. 7. Класифікація програмних засобів статистичного аналізу даних

Інформаційна база епідеміологічного аналізу

Під інформаційною базою передбачається сукупність впорядкованих даних, які використовуються в процесі проведення аналізу. Якість інформаційної бази визначає об'єктивність та достовірність аналізу.

Перелік показників повинен визначатися за принципами вимірюваності, прогнозованості та значущості.

Суть принципу вимірюваності полягає в тому, що запропонований показник повинен бути вимірюваним, тобто мати конкретне кількісне значення. Цифрові дані для обчислення цього показника повинні бути доступними для дослідника.

Принцип прогнозованості передбачає змогу передбачити майбутню динаміку того чи іншого показника з метою застосування прогнозу в управлінні процесом.

Принцип значущості передбачає врахування лише тих показників, які є суттєвими для конкретного об'єкта дослідження, та дозволяє відокремити основні проблеми чи тенденції розвитку об'єкта дослідження.

Інформаційна база епідеміологічного аналізу складається з нормативних (медичні та державні класифікатори, що використовуються для кодування інформації) та фактографічних даних (медична, соціальна, економічна, екологічна інформація, яка стосується захворюваності населення або факторів, що на неї впливають, і підготовлена за спеціальними затвердженими шаблонами).

Інформацію, що стосується епідеміологічного дослідження, можна поділити на первинну (яка збирається) і вторинну (яка є результатом розрахунків або будь-якого опрацювання: як то сортування, фільтрація даних тощо).

До первинної інформації відносяться такі документи:

- 1) медична форма № 058 «Інформація про інфекційне захворювання» або витяг з державної медичної бази даних з інформацією про інфекційне захворювання;
- 2) інформація про перебіг, результати та тривалість лікування;

- 3) інформація про щеплення та інші профілактичні заходи;
 - 4) архів щорічних, щомісячних, щотижневих звітів про інфекційну захворюваність;
 - 5) державні класифікатори територіальних одиниць, вулиць;
 - 6) класифікатор професій та довідник установ і промислових закладів;
 - 7) дані Державної служби статистики щодо територій та населення;
 - 8) дані Державної служби статистики щодо умов життя та статків населення;
 - 9) дані Державної служби статистики щодо кліматичних умов;
 - 10) дані Державної служби статистики щодо екологічного стану територій.
- До вторинної інформації відносяться результати розрахунків, які отримані під час аналізу первинної документації, а саме:

- 1) захворюваність на певній території за період, темпи росту захворюваності;
- 2) госпіталізація на певній території за період, темпи росту госпіталізації;
- 3) смертність на певній території за період, темпи росту смертності;
- 4) летальність на певній території за період, темпи росту летальності;
- 5) середня тривалість лікування захворювань на певній території за період;
- 6) коефіцієнт тяжкості захворювань на певній території за період;
- 7) вплив факторів ризику на епідеміологічні показники на певній території за період;
- 8) порівняння інфекційної захворюваності на певній території за період з нормами;
- 9) прогноз захворюваності на певну нозологічну форму на певній території на період;
- 10) прогноз смертності внаслідок певного захворювання на певній території на період;
- 11) порівняння довгострокових та короткострокових прогнозів певного захворювання на певній території на період;
- 12) інші.

3.1. ВИКОРИСТАННЯ СТАТИСТИЧНИХ МЕТОДІВ

3.1.1. Практичне заняття 1.

Розрахунок основних епідеміологічних показників

Тема: Дескриптивна статистика епідеміологічних даних. Розрахунок основних епідеміологічних показників. Графічне зображення результатів епідеміологічних досліджень.

Мета: удосконалення навичок з автоматизації розрахунків інтенсивних та екстенсивних епідеміологічних показників; знайомство з методами створення інтегрованих показників.

Теоретичні відомості

Епідемічний процес проявляється інфекційною захворюваністю, яка характеризується різною інтенсивністю захворюваності загалом в популяції, а також в різних групах, на територіях та в різний період часу.

Захворюваність – поширення хвороби серед сукупного населення та в окремих його групах, яке виражається в інтенсивних (на 1 тис., 10 тис., 100 тис. населення) і екстенсивних показниках (питома вага випадків однієї хвороби

серед всіх). Одиниця спостереження – зареєстрований протягом певного часу випадок захворювання.

Захворюваність або інцидентність (від англ. «incidence») – частота вперше у житті діагностованих (вперше або знову виявлених) захворювань протягом досліджуваного періоду, розрахована в інтенсивних (на 1 тис., 10 тис., 100 тис. населення) показниках.

Поширеність або превалентність (від англ. «prevalence») – частота усіх випадків захворювання (як первинних, так і хронічних) протягом певного часу, розрахована в інтенсивних (на 1 тис., 10 тис., 100 тис. населення) показниках.

Показник госпіталізації – частота всіх випадків захворювання, зареєстрована у реєстрі вибулих зі стаціонару протягом досліджуваного періоду, розрахована в інтенсивних (на 1 тис., 10 тис., 100 тис. населення) показниках.

Смертність – кількість померлих (загальна або за певним діагнозом) протягом досліджуваного періоду, розрахована в інтенсивних (на 1 тис., 10 тис., 100 тис. населення) показниках.

Летальність – кількість померлих серед хворих на певні захворювання, виражена у відсотках.

Перед вивченням епідеміологічних показників завжди має бути процес виявлення та реєстрації окремих випадків захворювання. Після накопичення інформації щодо кількості захворілих за певний період часу на певній території з'являється можливість вимірювання захворюваності та відображення її рівня у вигляді статистичних показників. Показники захворюваності виражаються в абсолютних та відносних величинах. Відносні величини є результатом ділення абсолютного показника (кількості випадків захворювання, госпіталізації, померлих) на базисний показник, яким у більшості випадків є чисельність населення певної територіальної одиниці. Отримане таким чином значення виражає ймовірність того, що представник даної групи населення захворів на певну хворобу. В сучасній епідеміології цю ймовірність називають ризиком.

Для унаочнення результатів епідеміологічних досліджень застосовуються діаграми. В епідеміології використовують різні їх види: лінійні, стовпчикові, кругові, картограми.

Лінійні діаграми (графіки) відображають кількісні дані в системі двох координат – осі X (горизонтальна вісь) та осі Y (вертикальна вісь). В епідеміології застосовуються для відображення динаміки захворюваності.

Стовпчикові діаграми служать для зображення кількісної ознаки у вигляді плоского або об'ємного стовпця, висота якого відповідає значенню ознаки, а ширина усіх стовпців має бути однаковою. Розташування може бути вертикальним або горизонтальним. В епідеміології використовуються під час дослідження структури захворюваності для відображення різниці величин захворюваності груп населення, які виділені за індивідуальними або територіальними ознаками. При цьому діаграми можуть відображати середні або прогностичні показники окремих років або інших часових інтервалів.

Кругові, або секторні, діаграми використовують для вивчення структури розподілу захворюваності. Такі діаграми представлені у вигляді кола, розбитого на сектори, величина яких відповідає питомій вазі кожної структурної частини.

Картограма – це географічна карта, на якій за допомогою графічних символів відображають прояви захворюваності, виявлені на різних територіях.

Задача-еталон

Дано: Інформація про випадки захворювань на COVID-19 на території України за період 2020–2022 рр. надана у таблиці Excel (джерела даних – <https://index.minfin.com.ua/ua/reference/coronavirus/ukraine/> та

<https://edata.e-health.gov.ua/e-data/dashboard/covid-package-stat>) (табл. 2).

Таблиця 2

Стовпці (показники) у таблиці Excel та підходи до їх розрахунку

№ з/п	Назва показника	Примітка
1	Регіон	
2	Область	
3	Населення (Н)	Кількісний показник
4	Випадки інфекції (ВІ), абс. показник	Кількісний показник
5	Смертельні випадки (СВ), абс. показник	Кількісний показник
6	Госпіталізовано (Г), абс. показник	Кількісний показник
7	Захворюваність (З) на 100 000 населення	Кількісний показник. Обчислюється за формулою: $Z = (VI, \text{ абс. показник} \times 100000) / N$
8	Смертність (С) на 100 000 населення	Кількісний показник. Обчислюється за формулою: $C = (CB, \text{ абс. показник} \times 100000) / N$
9	Летальність (Л)	Кількісний показник. Обчислюється за формулою: $L = (CB, \text{ абс. показник} \times 100) / VI, \text{ абс. показник}$
10	Показник госпіталізації (ПГ)	Кількісний показник. Обчислюється за формулою: $PG = (G, \text{ абс. показник} \times 100000) / VI, \text{ абс. показник}$
11	Екстенційний показник кількості інфікованих (ЕПКІ)	Кількісний показник. Обчислюється за формулою: $EPKI = VI, \text{ абс. показник} / \text{Загальна кількість I по всім областям. Виразити у відсотках}$
12	Випадки інфекції, рейтинг	Ціле число. Обчислюється за формулою: $VI, \text{ рейтинг} = \text{Ранг у стовпці VI, екстенційний показник, впорядкований за спаданням}$
13	Смертність, рейтинг	Ціле число. Обчислюється за формулою: $C, \text{ рейтинг} = \text{Ранг у стовпці C, впорядкованому за спаданням}$
14	Летальність, рейтинг	Ціле число. Обчислюється за формулою: $L, \text{ рейтинг} = \text{Ранг у стовпці L, впорядкованому за спаданням}$
15	Показник госпіталізації, рейтинг	Ціле число. Обчислюється за формулою: $PG, \text{ рейтинг} = \text{Ранг у стовпці PG, впорядкованому за спаданням}$
16	Інтегрований епідеміологічний показник (ІЕП)	Ціле число. Обчислюється за формулою: $IEP = VI, \text{ рейтинг} + C, \text{ рейтинг} + L, \text{ рейтинг} + PG, \text{ рейтинг (сума показників 12–15)}$

При проведенні поглибленого епідеміологічного аналізу з оцінкою захворюваності за середньою величиною доцільно додатково використовувати інтегрований показник, що суттєво розширює можливості середньої величини захворюваності та точніше відображає прояви епідемічного процесу

Запропонований підхід аналізу окрім захворюваності може бути використаний і для порівняння інших основних епідеміологічних показників, що формують предметну область епідеміології (наприклад, поширеність).

Інтегрований епідеміологічний показник є комбінацією 4 основних епідеміологічних показників, у подальшому аналізі розглядається як узагальнення цього комплексу.

Існує декілька способів обчислення ІЕП, при цьому часто застосовують метод нормалізації даних вибірки з урахуванням мінімального та максимального значення показника.

Дані для аналізу надаються у файлі Lesson 01D.xlsx. Період спостереження – з 03.2020 по 02.2022 р.

Для обчислення показників 8, 9, 10, 11, 12, 13, 14, 15, 16 треба ввести формули у відповідних комірках.

Вправа 1. Обчислити показники захворюваності, смертності, летальності, госпіталізації та інтегрований епідеміологічний показник. Для відображення інтенсивних показників та інтегрованого епідеміологічного показника в таблиці застосувати засіб Microsoft Excel «Умове форматування», «Гістограми».

Алгоритм виконання:

1. Відкрити файл Lesson 01D.xlsx.
2. Для обчислення інтенсивного показника «захворюваність на 100 000 населення» в комірку таблиці G3 ввести формулу $=D3*100\ 000/C3$.
3. Для обчислення інтенсивного показника «смертність на 100 000 населення» в комірку таблиці H3 ввести формулу $=E3*100\ 000/C3$.
4. Для обчислення летальності в комірку I3 ввести формулу $=E3/D3$. Встановити для комірки I3 формат «відсотковий».
5. Для обчислення інтенсивного показника «госпіталізація на 100 000 населення» в комірку таблиці J3 ввести формулу $=F3*100\ 000/C3$.
6. Скопіювати формули з комірок G3:J3 в діапазон G4:J28.
7. Для визначення структури кількості інфікованих по областях розрахувати екстенсивний показник. У комірку K4 ввести формулу $=D4/(\$D\$3)$. Скопіювати формулу в діапазон K5:K28.
8. Для графічного зображення екстенсивного показника можна використовувати секторні, звичайні стовпчикові діаграми або гістограми. У випадку, коли рядків понад 10 більш наочні гістограми або звичайні стовпчикові діаграми. За даними комірок K4:K28 побудувати секторну діаграму. За допомогою інструментів «Змінити тип діаграми» та «Налаштування макету» надати графічному зображенню наступного вигляду (рис. 8):

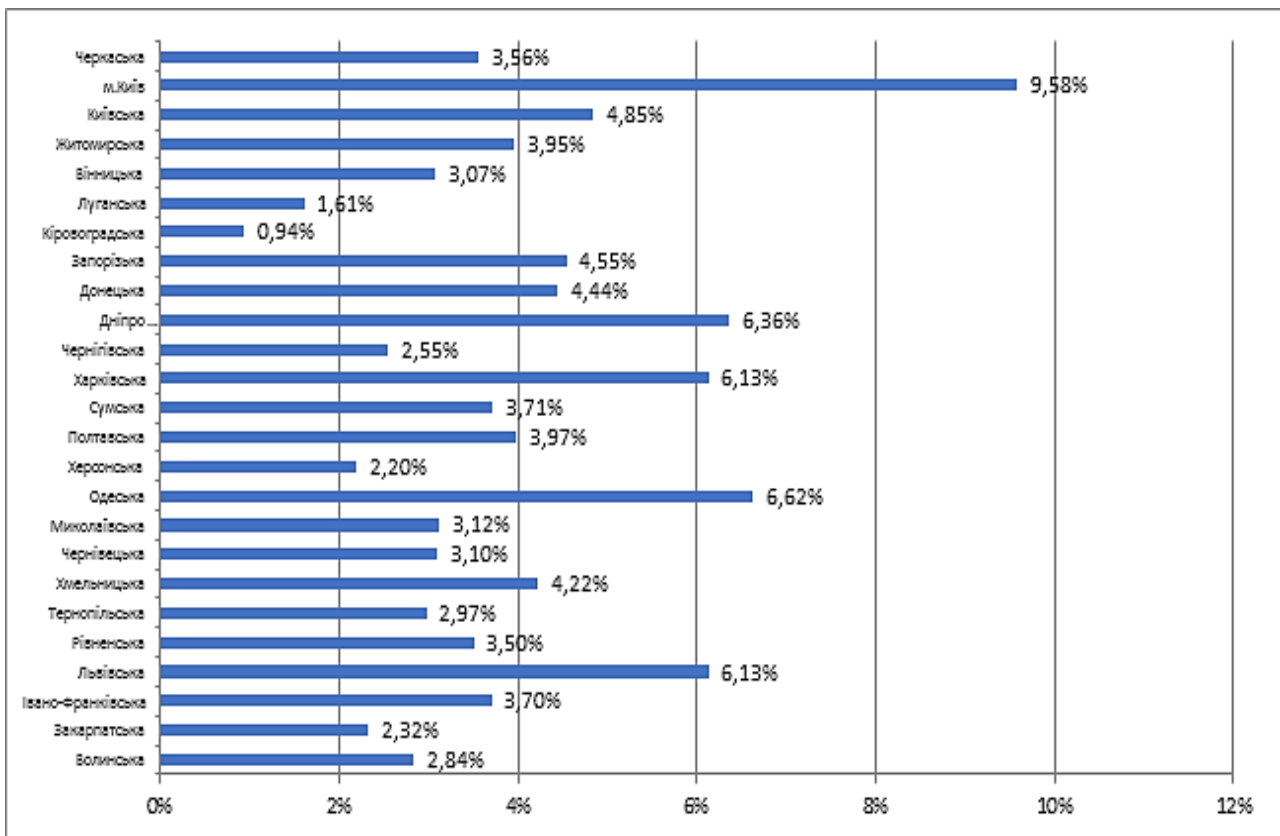


Рис. 8. Гістограма для унаочнення та порівняння екстенсивного показника захворюваності

9. Для визначення рангу інтенсивного показника захворюваності (місця у списку) в комірку L4 ввести формулу =RANK.AVG(G4;\$G\$4:\$G\$28;1).

10. Для визначення рангу інтенсивного показника смертності (місця у списку) в комірку M4 ввести формулу =RANK.AVG(H4;\$H\$4:\$H\$28;1).

11. Для визначення рангу показника летальності (місця у списку) в комірку N4 ввести формулу =RANK.AVG(I4;\$I\$4:\$I\$28;1).

12. Для визначення рангу показника госпіталізації (місця у списку) в комірку O4 ввести формулу =RANK.AVG(J4;\$J\$4:\$J\$28;1).

13. Для визначення інтегрованого епідеміологічного показника в комірку P4 ввести формулу =SUM(L4:O4).

14. Скопіювати формули з комірок L4:P4 в діапазон L5:P28.

15. Виділити діапазон комірок L4:O28. Встановити умовне форматування за допомогою інструмента «Гістограми. Градієнтна заливка».

16. Для графічного зображення рейтингових показників можна використовувати стовпчикові діаграми кільцеві, звичайні та з накопиченням або гістограми. У випадку, коли рядків понад 10 більш наочні гістограми або стовпчикові діаграми з накопиченням. За даними комірок L4:O28 побудувати кільцеву діаграму. За допомогою інструментів «Змінити тип діаграми» та «Налаштування макету» надати графічному зображенню вигляду, який представлений на *рис. 9*.

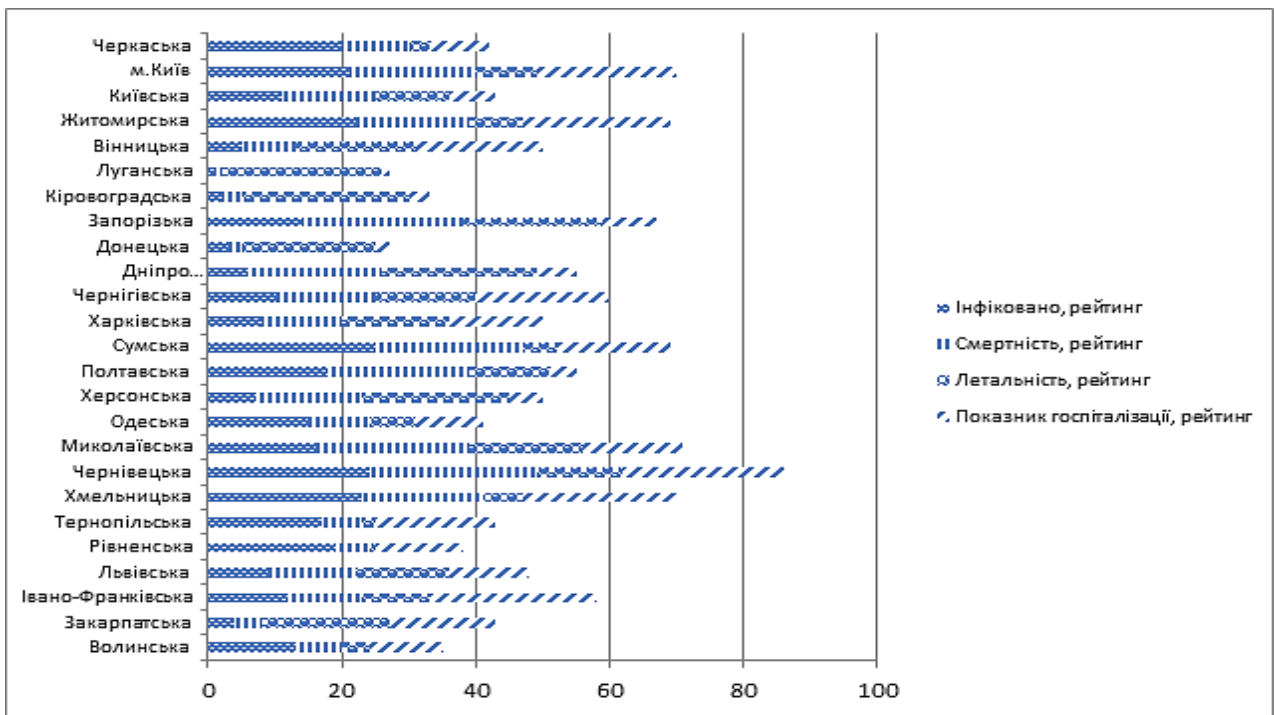


Рис. 9. Діаграма з накопиченням для відображення комплексу показників

17. За даними комірок Р4:Р28 побудувати гістограму і порівняти з рис. 9.

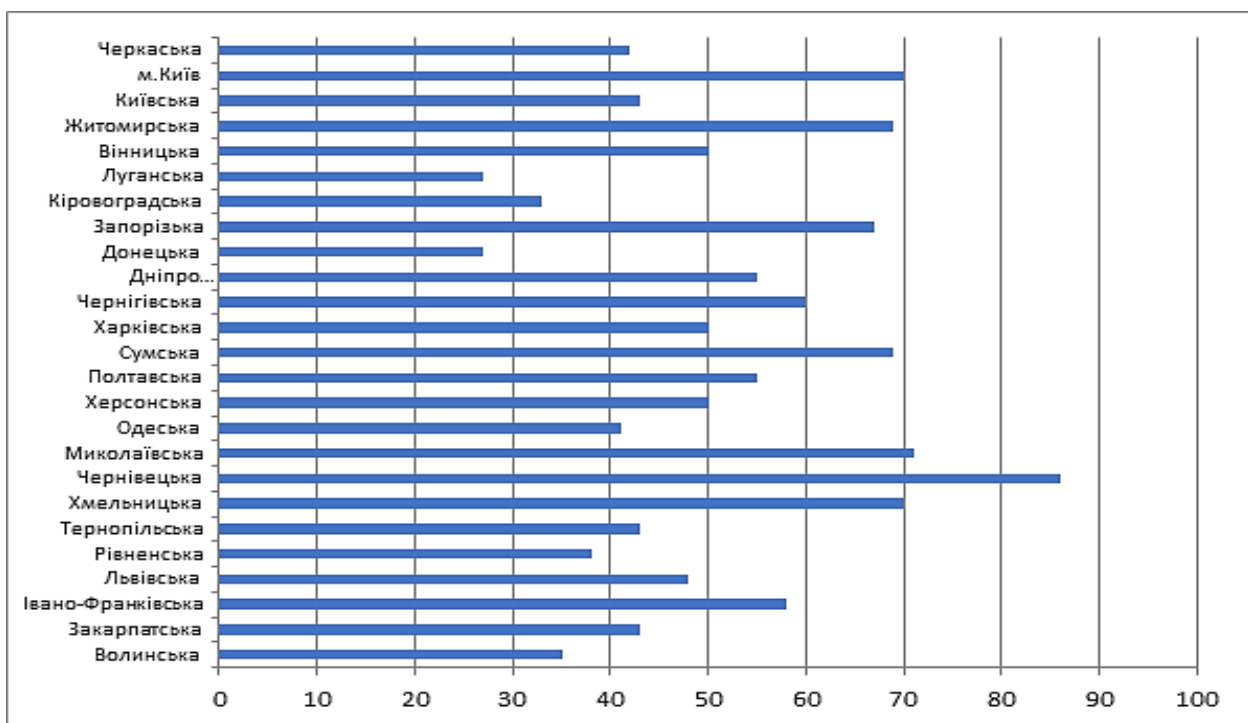


Рис. 10. Гістограма відображення інтегрованого показника захворюваності

18. Виділити діапазон комірок Р4:Р28, встановити умовне форматування за допомогою інструмента «Гістограми. Градієнтна заливка».

Зразок результату розрахунків і умовного форматування представлений на рис. 11.

	A	B	G	H	I	J	K	L	M	N	O	P
1	Коронавірус COVID-19 в Україні 2020-2022											
2	Регіон	Область	Захворюваність на 100000 населення	Смертність на 100000 населення	Летальність	Показник госпіталізації на 100000 населення	Екстенційний показник кількості інфікованих	Інфіковано, рейтинг	Смертність, рейтинг	Летальність, рейтинг	Показник госпіталізації, рейтинг	Інтегрований епідеміологічний показник
3	Україна	Україна	12192,67	256,51	2,10%	2323,22						
4	Західний	Волинська	13930,84	234,52	1,68%	2378,33	2,84%	13	7	4	11	35
5	Західний	Закарпатська	9368,42	219,02	2,34%	2716,33	2,32%	4	4	19	16	43
6	Західний	Івано-Франківська	13740,04	257,37	1,87%	3859,17	3,70%	12	11	10	25	58
7	Західний	Львівська	12420,62	285,30	2,14%	2383,59	6,13%	9	13	14	12	48
8	Західний	Рівненська	15396,88	221,75	1,44%	2544,13	3,50%	19	5	1	13	38
9	Західний	Тернопільська	14602,35	234,18	1,60%	2775,91	2,97%	17	6	2	18	43
10	Західний	Хмельницька	17227,37	301,67	1,75%	3414,91	4,22%	23	18	6	23	70
11	Західний	Чернівецька	17463,31	369,70	2,12%	3590,63	3,10%	24	25	13	24	86
12	Південний	Миколаївська	14347,36	323,70	2,26%	2694,73	3,12%	16	23	17	15	71
13	Південний	Одеська	14126,44	251,82	1,78%	2342,98	6,62%	15	9	7	10	41
14	Південний	Херсонська	11014,19	287,98	2,61%	2027,49	2,20%	7	16	22	5	50
15	Північний	Полтавська	14758,33	309,57	2,10%	1878,50	3,97%	18	21	12	4	55
16	Північний	Сумська	17998,45	310,13	1,72%	2771,85	3,71%	25	22	5	17	69
17	Північний	Харківська	11840,93	284,15	2,23%	2571,93	6,13%	8	12	16	14	50
18	Північний	Чернігівська	13350,01	287,25	2,15%	2808,19	2,55%	10	15	15	20	60
19	Східний	Дніпропетровська	10318,47	305,64	2,96%	2156,63	6,36%	6	20	23	6	55
20	Східний	Донецька	5492,41	138,23	2,52%	982,23	4,44%	3	2	20	2	27
21	Східний	Запорізька	13932,90	358,49	2,57%	2221,84	4,55%	14	24	21	8	67
22	Східний	Кіровоградська	5240,39	192,95	3,68%	1193,39	0,94%	2	3	25	3	33
23	Східний	Луганська	3838,61	113,81	2,96%	620,83	1,61%	1	1	24	1	27
24	Центральний	Вінницька	10211,91	235,19	2,30%	2802,08	3,07%	5	8	18	19	50
25	Центральний	Житомирська	16836,02	300,44	1,78%	3311,65	3,95%	22	17	8	22	69
26	Центральний	Київська	13535,23	279,53	2,07%	2218,71	4,85%	11	14	11	7	43
27	Центральний	м.Київ	16277,56	303,86	1,87%	2974,62	9,58%	21	19	9	21	70
28	Центральний	Черкаська	15413,56	256,47	1,66%	2338,16	3,56%	20	10	3	9	42

Рис. 11. Скріншот таблиці з умовним форматуванням даних в стовпцях

Висновки:

1. Застосування табличного процесора Microsoft Excel значно прискорило обчислення і графічну візуалізацію основних епідеміологічних показників.

2. Епідеміологічні показники в областях західного регіону перевищують показники в інших регіонах.

3. Максимальне значення інтегрованого показника дорівнює 86 (Чернівецька область), що може свідчити про негативний вплив факторів середовища життєдіяльності на поширення COVID-19 у західних областях України, та зокрема у Чернівецькій області, або недоліки в організації профілактичних й проти-епідемічних заходів.

Вправа 2. Обчислити загальні числові характеристики інтенсивних показників: середнє значення; мінімальне значення; максимальне значення; розмах значень; кількість областей, де показник менше показника загалом по Україні; кількість областей, де показник більше показника загалом по Україні. Оформити результати у рядках 31–36. Зробити висновки.

Алгоритм виконання:

1. Ввести підписи в діапазон комірок B31:B36, G30:J30 (як на рис. 12).

2. У комірки G31:J31 скопіювати значення комірок G3:J3, оскільки середнє значення показника – його значення по Україні.

3. Для визначення мінімального значення показника «захворюваність» в комірку G32 ввести формулу =MIN(G4:G28).

4. Для визначення максимального значення показника «захворюваність» в комірку G33 ввести формулу =MAX(G4:G28).

5. Для визначення розмаху значень в комірку G34 ввести формулу =G33-G32.

6. Для визначення «кількість областей, де показник менше показника по Україні», в комірку G35 ввести формулу =COUNTIF(G4:G28;"<"&G3).

7. Для визначення «кількість областей, де показник більше показника по Україні», в комірку G36 ввести формулу =COUNTIF(G4:G28;">"&G3).

8. Виділити діапазон комірок G32:G36 та скопіювати його в діапазон H32:J36.

Зразок результату розрахунків представлений на *рис. 12*.

	A	B	G	H	I	J
			Захворюваність на 100000 населення	Смертність на 100000 населення	Летальність	Показник госпіталізації на 100000 населення
30						
31		Середнє значення	12192,67	256,51	2,10%	2323,22
32		Мінімум	3838,61	113,81	1,44%	620,83
33		Максимум	17998,45	369,70	3,68%	3859,17
34		Розмах	14159,85	255,89	0,02	3238,34
		Кількість областей, де показник менше показника по Україні	8	10	12	8
35		Кількість областей, де показник більше показника по Україні	17	15	13	17
36						

Рис. 12. Результати обчислень загальних числових характеристик інтенсивних показників

Висновки:

1. Інтенсивні показники захворюваності на COVID-19 (на 100 тис. населення) по областях України значно відрізняються. Так, максимальне значення майже в 6 разів перевищує мінімальне, кількість областей, де значення показника перевищує середній показник по Україні, майже у 2 рази більше кількості областей з відносно невисоким рівнем захворюваності.

2. Показник летальності в середньому становить 2,1 %, але у половини областей цей показник перевищує показник загалом по Україні.

3. Максимальне значення показника смертності у 3 рази перевищує мінімальне. При цьому в 15 із 25 областей значення показника перевищує середній показник по Україні. Високий рівень смертності може свідчити про несвоєчасно надану медичну допомогу або про неефективне лікування.

4. Велика різниця між мінімальним і максимальним значенням показника госпіталізації хворих на COVID-19. Так, кількість областей, де показник госпіталізації перевищує середній показник по Україні, майже у 2 рази більше порівняно з кількістю областей з меншим рівнем госпіталізації.

Вправа 3. Побудувати картограми показника захворюваності на 100 тис. населення: а) для всіх областей України; б) для областей Центрального регіону; в) для областей Північного регіону.

Алгоритм виконання:

1. Для побудови картограм скористатися Microsoft Excel версії 2021 р.

2. У файлі Lesson 01D.xlsx створити новий аркуш, назвати його Картограми.

3. На аркуші Картограми сформувати таблицю з даними для побудови картограм наступними чином:

- a) на аркуші Дані виділити діапазон комірок A2:B28, скопіювати вміст комірок до аркушу Картограми, діапазон A1:B27;
- b) до комірки C1 ввести текст «Область»;
- c) видалити з таблиці 2-й рядок, де містяться дані по всій Україні;
- d) до комірки C2 ввести формулу =B2&" область"; скопіювати цю формулу до діапазону C3:C24 та до комірки C26; до комірки C25 скопіювати значення з комірки B25;
- e) до комірки D1 ввести текст «Захворюваність на 100 тис. населення»;
- f) на аркуші Дані виділити діапазон комірок G4:G28, скориставшись функцією Спеціальне вставлення скопіювати значення комірок до діапазону D2:D26;
- g) для діапазону D2:D26 встановити формат даних: числовий, кількість десяткових розрядів – 0.

Таким чином таблиця з даними для побудови картограми має наступний вигляд (рис. 13):

	A	B	C	D	E
1	Регіон	Область	Область	Захворюваність на 100000 населення	
2	Західний	Волинська	Волинська область	13931	
3	Західний	Закарпатська	Закарпатська область	9368	
4	Західний	Івано-Франківська	Івано-Франківська область	13740	
5	Західний	Львівська	Львівська область	12421	
6	Західний	Рівненська	Рівненська область	15397	
7	Західний	Тернопільська	Тернопільська область	14602	
8	Західний	Хмельницька	Хмельницька область	17227	
9	Західний	Чернівецька	Чернівецька область	17463	
10	Південний	Миколаївська	Миколаївська область	14347	
11	Південний	Одеська	Одеська область	14126	
12	Південний	Херсонська	Херсонська область	11014	
13	Північний	Полтавська	Полтавська область	14756	
14	Північний	Сумська	Сумська область	17998	
15	Північний	Харківська	Харківська область	11841	
16	Північний	Чернігівська	Чернігівська область	13350	
17	Східний	Дніпропетровська	Дніпропетровська область	10318	
18	Східний	Донецька	Донецька область	5492	
19	Східний	Запорізька	Запорізька область	13933	
20	Східний	Кіровоградська	Кіровоградська область	5240	
21	Східний	Луганська	Луганська область	3839	
22	Центральний	Вінницька	Вінницька область	10212	
23	Центральний	Житомирська	Житомирська область	16836	
24	Центральний	Київська	Київська область	13535	
25	Центральний	м.Київ	м.Київ	16278	
26	Центральний	Черкаська	Черкаська область	15414	
27					

Рис. 13. Таблиця з даними для побудови картограми

4. Виділити діапазон C1: D26 і виконати команду вкладки «Вставлення → Діаграми → Картти» (рис. 14).

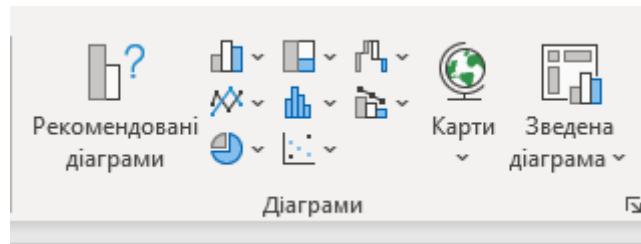


Рис. 14. Перелік видів діаграм, які можна побудувати в середовищі Microsoft Excel 2021

5. Результат побудови карти (рис. 15):

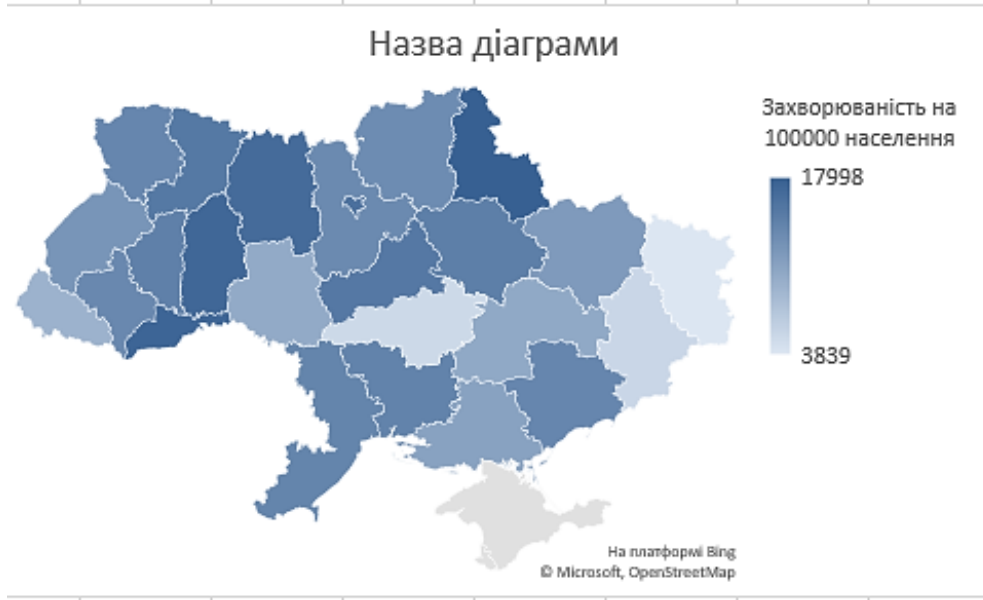


Рис. 15. Результат побудови карти України по областях, що вказані у таблиці діапазону C1: D26

6. Налаштувати карту наступним чином:

- а) змінити назву картограми на «Коронавірус COVID-19 в Україні 2020–2022»;
- б) виділити ряд даних (кляцнути на будь-якій області) і за допомогою контекстного меню налаштувати формат ряду даних (рис. 16–18);

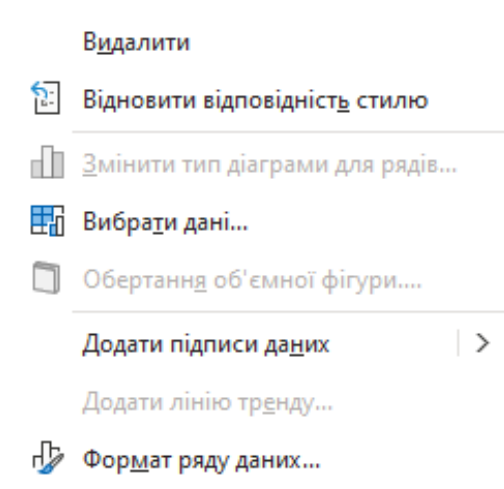


Рис. 16. Контекстне меню ряду даних картограми

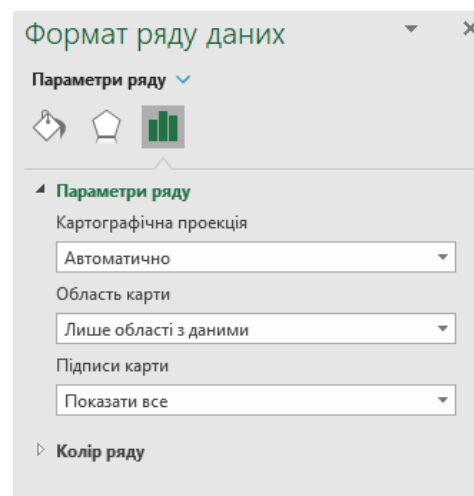


Рис. 17. Панель для встановлення параметрів ряду



Рис. 18. Результат налаштування параметрів ряду даних картограми

с) за допомогою контекстного меню додати підписи даних (значення захворюваності на 100 тис. населення) (рис. 19).



Рис. 19. Результат додавання підписів до картограми

7. Для діапазону комірок A1: D26 аркушу «Картограма» встановити фільтрацію даних.

8. За допомогою фільтра відібрати дані Центрального регіону. Відповідні зміни відбудуться на картограмі (рис. 20).

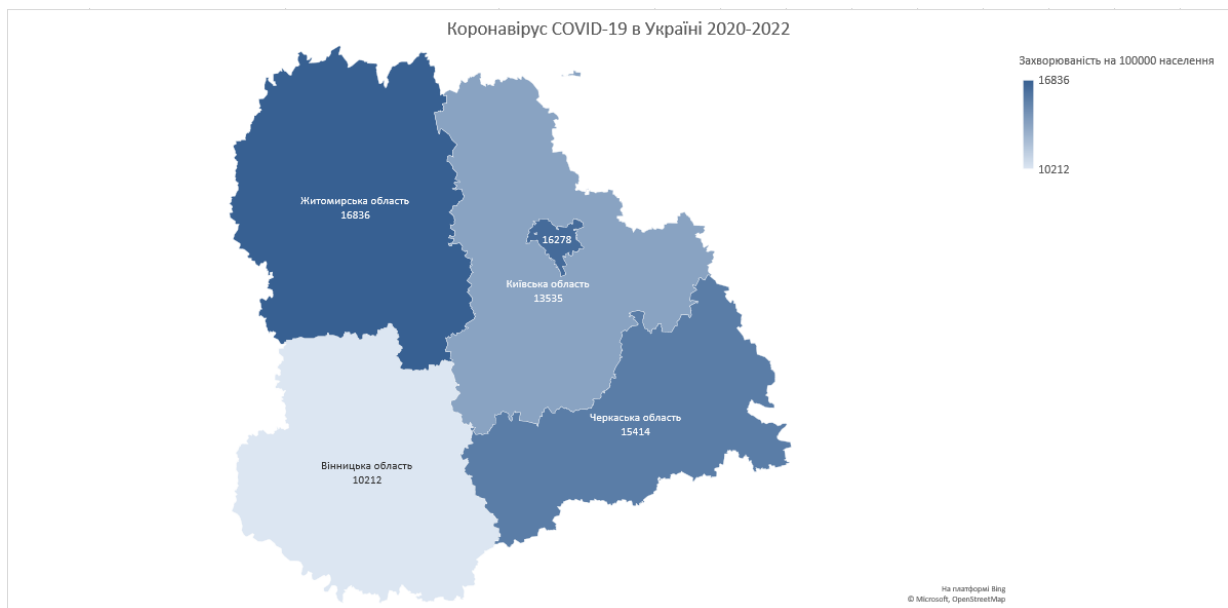


Рис. 20. Картограма захворюваності на COVID-19 в Центральному регіоні України у 2020–2022 рр.

9. Самостійно побудувати картограму захворюваності на COVID-19 в Північному регіоні України.

3.1.2. Практичне заняття 2.

Систематизація даних епідеміологічного моніторингу

Тема: Дескриптивна статистика епідеміологічних даних. Систематизація даних епідеміологічного моніторингу.

Мета: удосконалення навичок з автоматизації створення підсумкових статистичних таблиць.

Теоретичні відомості

Моніторинг (безперервне, динамічне спостереження) – система постійного спостереження за явищами і процесами, що відбуваються в навколишньому середовищі, суспільстві, результати якого слугують для обґрунтування управлінських рішень щодо забезпечення безпеки людей та об'єктів або суб'єктів.

Епідеміологічний нагляд (ЕН) (за визначенням ВООЗ) – це система збору, аналізу та інтерпретації даних про інфекційні захворювання, що включає періодичну звітність про зібрану інформацію перед зацікавленими особами та групами.

Епідеміологічний нагляд (спостереження) в Україні проводиться згідно з наказом Міністерства охорони здоров'я України від 30.07.2020 № 1726 «Про затвердження Порядку ведення обліку, звітності та епідеміологічного нагляду (спостереження) за інфекційними хворобами та Переліку інфекційних хвороб, що підлягають реєстрації» для оцінки та вжиття необхідних відповідних заходів у сфері громадського здоров'я.

Основні завдання епідеміологічного нагляду такі:

- моніторинг і оцінка епідемічної ситуації та виявлення тенденцій і оцінка темпів динаміки епідемічних процесів інфекційних хвороб на відповідній території у певний проміжок часу;

- прогнозування розвитку епідемічного процесу на майбутній період та визначення обсягу основних профілактичних та протиепідемічних заходів; виявлення та ідентифікація будь-яких спалахів інфекційних хвороб з визначенням джерела, часу, місця їх виникнення, постраждалого населення, факторів та умов передачі збудника; надання пропозицій та запровадження обґрунтованих заходів реагування відповідно до епідемічної ситуації;

- моніторинг і оцінка профілактичних та протиепідемічних програмних заходів з метою їх раціонального коректування; визначення груп ризику та потреби у цільових профілактичних заходах; оцінка тягаря інфекційних хвороб у структурі загальної захворюваності та смертності населення та на систему охорони здоров'я.

Аналіз даних епідеміологічного моніторингу проводиться за групами інфекційних хвороб і нозологічними формами за проявами епідемічного процесу: інтенсивність, розподіл захворювань за часом, за територією та серед різних груп населення.

Інтенсивність епідемічного процесу характеризується абсолютними або інтенсивними показниками захворюваності, а також смертності, інвалідності тощо. Розподіл інфекційної захворюваності за часом є важливим показником багаторічної динаміки та річної захворюваності.

Багаторічна динаміка захворюваності визначає епідемічну тенденцію, циклічність і нерегулярні коливання. Розподіл інфекційних хвороб серед різних груп, зокрема, за віком (дорослі, діти 0–17 років, у тому числі до року, 1–4 роки, 5–9, 10–14 та 15–17 років), місцем проживання (місто, село), статтю (жінки, чоловіки) та професією, дозволяє визначити групи ризику.

Отже, розрахунок показників проявів епідемічного процесу дозволить встановити епідеміологічні ризики:

- за територією (для встановлення території «ризик» – де хворіють);
- за соціально-віковими групами (для встановлення груп «ризик» – хто хворіє);
- за конкретними колективами (для встановлення колективів «ризик» – хто хворіє);
- за часом (в багаторічній і річній динаміці – для встановлення часу «ризик» – коли хворіють).

Задача-еталон

Дано: Інформація про окремі випадки захворювань дорослого населення накопичується у спеціальному реєстрі – таблиці Excel (табл. 3).

Таблиця 3

Стовпці (показники) у таблиці Excel та підходи до їх визначення і розрахунку

№ з/п	Назва показника	Примітка
1	Реєстраційний №	
2	Дата реєстрації	
3	Пацієнт, прізвище	
4	Пацієнт, ім'я	
5	Пацієнт, по батькові	
6	Стать	

№ з/п	Назва показника	Примітка
7	Дата народження	
8	Вік на момент реєстрації	Обчислюється за формулою: $\text{Вік} = \frac{\text{Дата реєстрації} - \text{Дата народження}}{365}$ Результат округлити до 1 знака після коми
9	Вікова група на момент реєстрації	Обчислюється за формулою: $\text{Вікова група} = \begin{cases} 29 \text{ і менше} & \text{Вік} < 30 \\ 30 - 39 & 30 \leq \text{Вік} < 40 \\ 40 - 49 & 40 \leq \text{Вік} < 50 \\ 50 - 65 & 50 \leq \text{Вік} < 66 \\ 66 \text{ і більше} & 66 \leq \text{Вік} \end{cases}$
10	Місце проживання, область	
11	Місце проживання, район	
12	Місце проживання, населений пункт	
13	Місце проживання, тип населеного пункту	
14	Місце проживання, вулиця	
15	Місце проживання, будинок	
16	Місце проживання, квартира	
17	Зайнятість	Можливі 2 варіанти: працює, не працює
18	Професія за класифікацією	Група професій за державним класифікатором професій
19	Професійна група ризику	Можливі 2 варіанти: є, немає
20	Визначення випадку інфекційного захворювання (ІЗ)	Використовувати систему «Єдине здоров'я»
21	Групи інфекцій	Інфекції: 1) кишкові; 2) дихальних шляхів; 3) кров'яні; 4) зовнішніх покривів
22	Дата визначення випадку ІЗ	
23	Початок захворювання, рік	Обчислюється за формулою: Початок захворювання, Рік = Рік від дати встановлення діагнозу
24	Початок захворювання, місяць	Обчислюється за формулою: Початок захворювання, Місяць = Місяць від дати встановлення діагнозу
25	Початок захворювання, рік + місяць	Обчислюється за формулою: Початок захворювання, Рік + Місяць = Початок захворювання, Рік+ «_»+Початок захворювання, Місяць. Таким чином, створений допоміжний показник можна використати під час визначення підсумків та аналізу
26	Місце лікування	Можливі 2 варіанти: амбулаторно, в стаціонарі
27	Завершення лікування, дата	Обов'язкове для заповнення поле
28	Результат лікування	Можливі 2 варіанти: одужання, смерть
29	Термін лікування в днях	Обчислюється за формулою: Термін лікування = Завершення лікування - Початок захворювання
30	Ознака тривалого лікування	Обчислюється за формулою: $\text{Тривале лікування} = \begin{cases} \text{ні} & \text{Термін лікування} < 30 \\ \text{так} & \text{Термін лікування} \geq 30 \end{cases}$

Дані для аналізу надаються у файлі Lesson 02D.xlsx. Період спостереження з 31.08.2020 по 11.08.2021 р.

Для обчислення показників 8, 9, 23, 24, 25, 29, 30 треба ввести формули у відповідних комірках.

Алгоритм обчислення показників таблиці-джерела даних:

1. Відкрити файл Lesson 02D.xlsx. Зберегти його з ім'ям Lesson 02R.xlsx.
2. Для обчислення показника 8 – «вік» (на момент реєстрації) в комірку таблиці H2 ввести формулу $= (B2 - G2) / 365$.
3. Для обчислення показника 9 – «вікова група» (на момент реєстрації) в комірку таблиці I2 ввести формулу $= IF(H2 < 30; "29 і менше"; IF(H2 < 40; "30-39"; IF(H2 < 50; "40-49"; IF(H2 < 60; "50-59"; IF(H2 < 66; "60-65"; "66 і більше"))))$.
4. Виділити діапазон комірок H2:I2. Скопіювати формули з комірок H2:I2 в діапазон H101:I101.
5. Для обчислення показника «початок захворювання», «рік» в комірку таблиці X2 ввести формулу $= YEAR(W2)$.
6. Для обчислення показника «початок захворювання», «місяць» в комірку таблиці Y2 ввести формулу $= MONTH(W2)$.
7. Для обчислення показника «початок захворювання», «рік+місяць» в комірку таблиці Z2 ввести формулу $= TEXT(YEAR(W2); "0000") & "-" & TEXT(MONTH(W2); "00")$.
8. Виділити діапазон комірок X2:Z2. Скопіювати формули з комірок X2:Z2 в діапазон X101:Z101.
9. Для обчислення показника «термін лікування в днях» в комірку таблиці AD2 ввести формулу $= AB2 - W2$.
10. Для обчислення показника «тривале лікування» в комірку таблиці AE2 ввести формулу $= IF(AD2 < 30; "ні"; "так")$.
11. Виділити діапазон комірок AD2:AE2. Скопіювати формули з комірок AD2:AE2 в діапазон AD101:AE101.
12. Зберегти файл Lesson 02R.xlsx зі змінами.

На підставі отриманих результатів можна узагальнити показники і оцінити епідемічну ситуацію в певних групах населення на певних територіях за визначений проміжок часу. Для цього можна застосувати засіб Microsoft Excel «Зведені таблиці».

Зведені таблиці Microsoft Excel – це інструмент, який за даними таблиці-джерела формує таблицю-звіт згідно із заданим алгоритмом. Зведена таблиця дозволяє розбити табличну базу даних на групи і вивести проміжні підсумки з будь-яким рівнем деталізації. За допомогою зведених таблиць можна швидко опрацьовувати великі обсяги інформації: відфільтровувати, порівнювати, групувати дані.

Джерело (база) даних для зведеної (статистичної) таблиці – це таблиця з детальною первинною інформацією. У першому рядку таблиці повинні бути назви показників, що досліджуються. У першому стовпці має бути унікальний реєстраційний номер одиниці спостереження. Починаючи зі стовпця 2 розташовується група категорійних показників, за якими відбуватиметься групування даних. Це можуть бути територіальні ознаки, стать, вік або вікова група, місце

роботи або навчання, професія, визначення інфекційного захворювання. Далі розташовується група кількісних інтенсивних епідеміологічних показників.

Алгоритм побудови зведеної таблиці:

1. Виділити будь яку комірку в середині таблиці-джерела даних. Вибрати команду «Вставлення»/Зведена таблиця.

2. На першому етапі створення зведеної таблиці задаються адреси місць розташування таблиці-джерела та звіту зведеної таблиці (рис. 21).

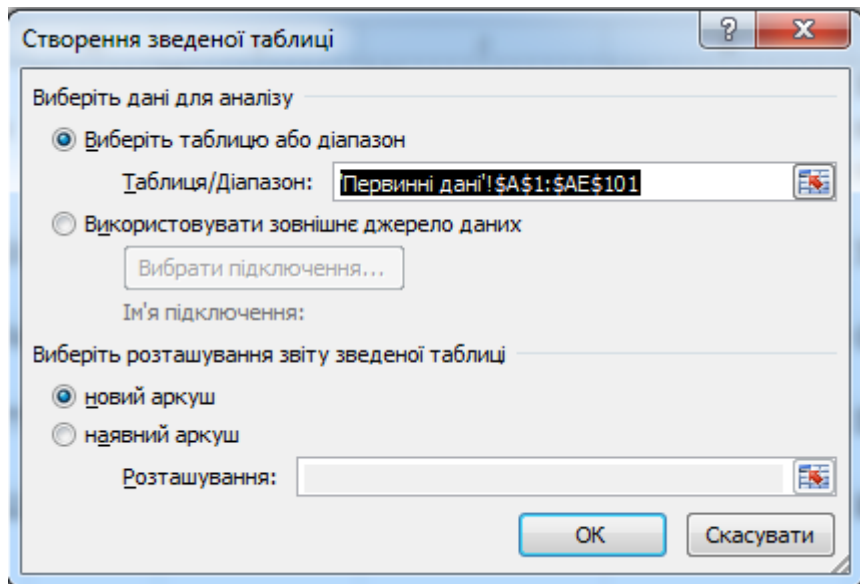


Рис. 21. Вікно майстра створення зведених таблиць

3. У разі вибору варіанта розташування звіту на новому аркуші до книги (файлу Lesson 02R.xlsx) буде доданий новий аркуш та відкрито знаряддя для роботи зі зведеною таблицею (рис. 22).

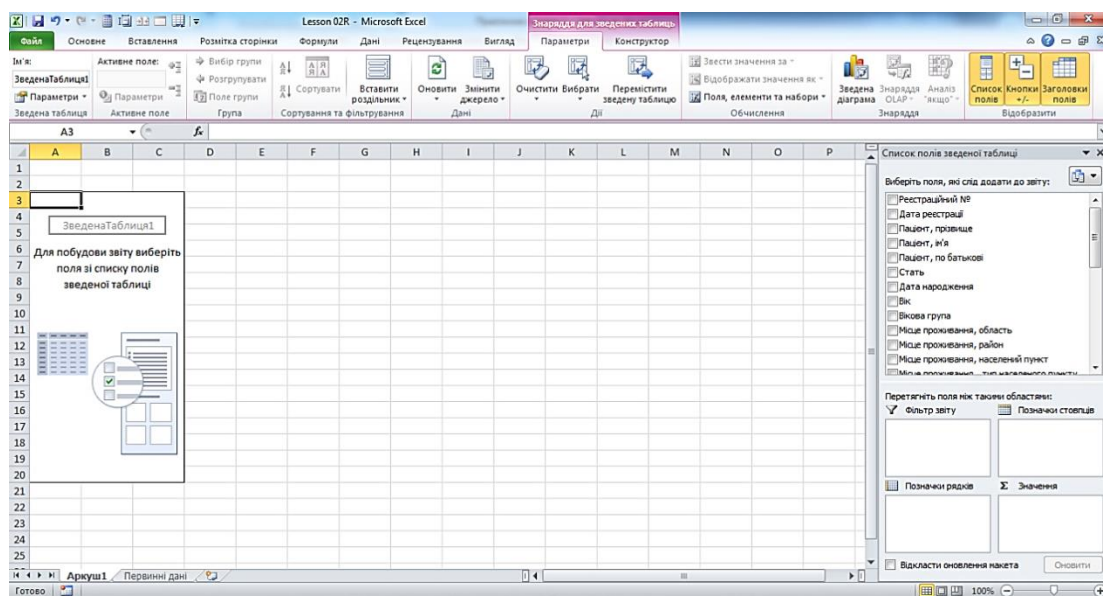


Рис. 22. Знаряддя для роботи зі зведеною таблицею (додаткова вкладка в меню програми Microsoft Excel)

4. Макет створеної зведеної таблиці визначається вмістом 4 областей. Основні елементи макета цієї таблиці представлені на *рис. 23* та *24*.

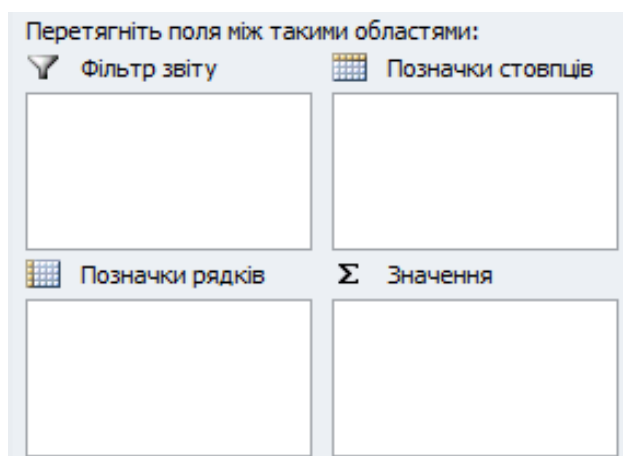


Рис. 23. Основні елементи макета зведеної таблиці, до яких перетягуються поля зі списку полів зведеної таблиці

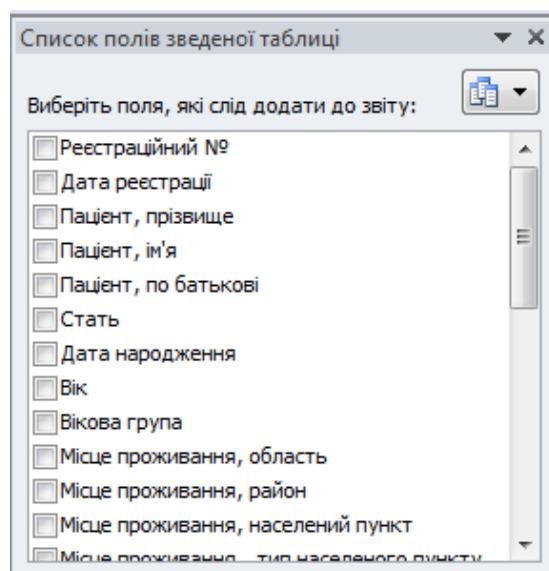


Рис. 24. Список полів зведеної таблиці, який формується із заголовків стовпців джерела даних

Після перетягування поля в область стовпців макета в області стовпців звіту з'являються унікальні значення цього поля. Після перетягування поля в область рядків макета в області рядків звіту з'являються унікальні значення цього поля. Після перетягування поля в область значень макета аналізуються його значення з таблиці-джерела, обирається відповідна типу підсумкова операція і в області значень звіту з'являються результати обчислення.

Наприклад, для обчислення кількості пацієнтів різної статі по областях треба виконати наступне: 1) перетягнути поле «стать» в область стовпців зведеної таблиці; 2) перетягнути поле «місце проживання», «область» в область рядків зведеної таблиці; 3) перетягнути поле «пацієнт», «прізвище» в область даних « Σ значень» зведеної таблиці (*рис. 25*).

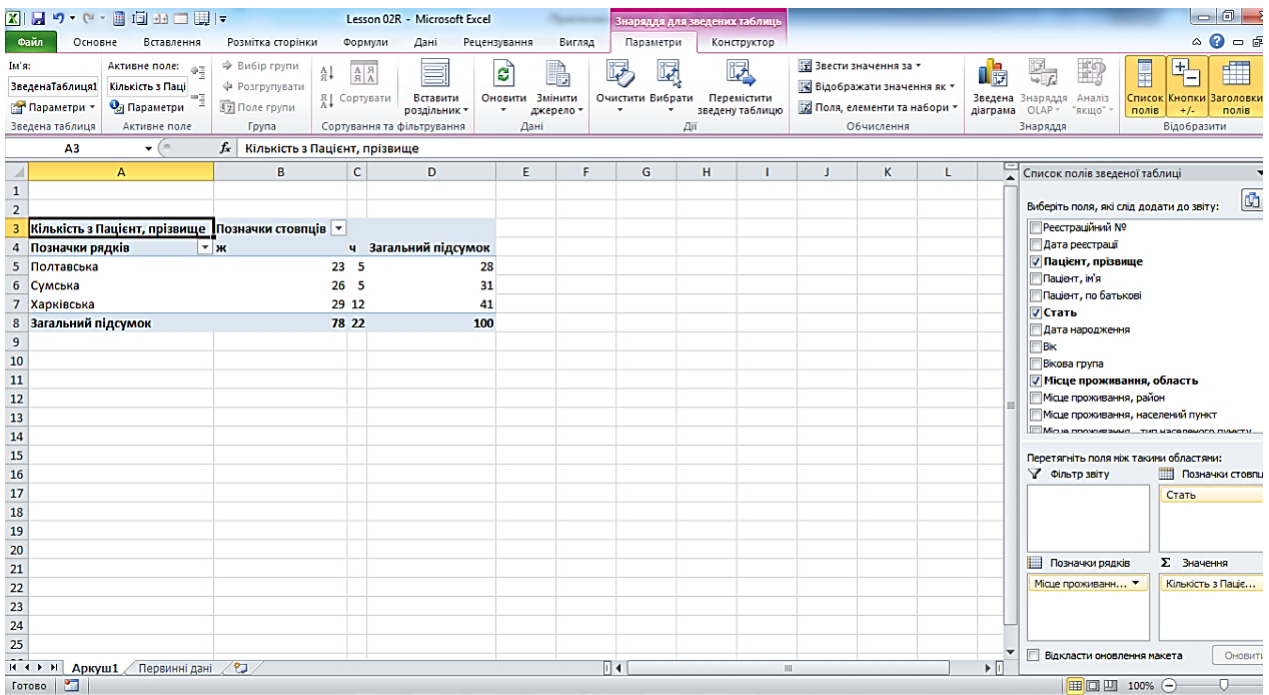


Рис. 25. Аркуш зі зведеною таблицею в Microsoft Excel

Тип підсумкової операції можна налаштувати. Для цього у зведеній таблиці виділяють поле відповідного заголовку, за допомогою контекстного меню обирають режим налаштування параметрів значення поля та у вікні налаштування параметрів обирають потрібну операцію (рис. 26).

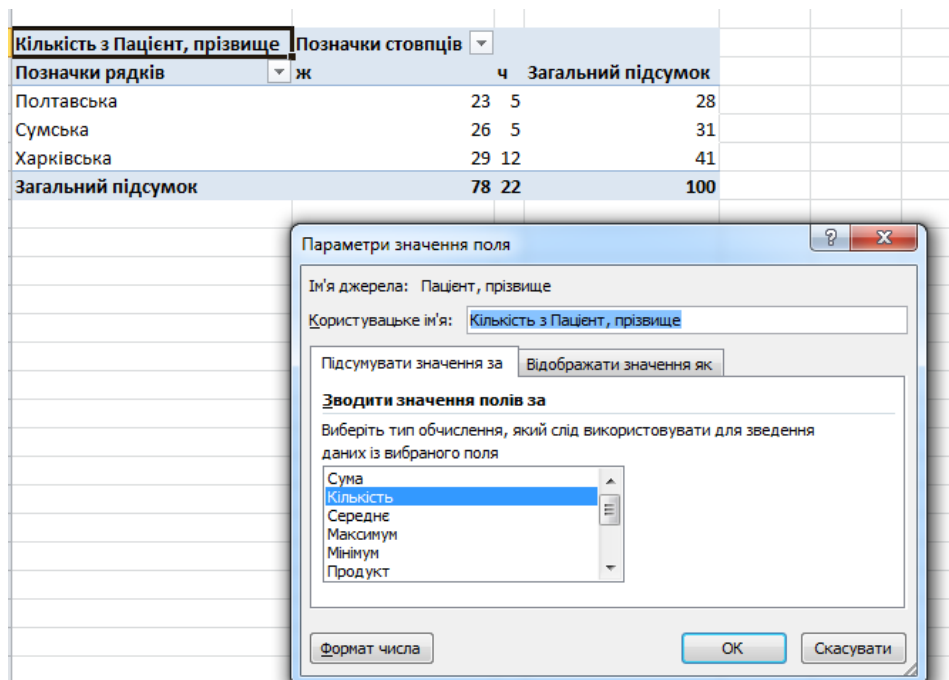
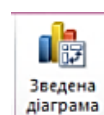


Рис. 26. Вікно налаштування параметрів підсумкового поля зведеної таблиці

Візуалізація зведеної таблиці – зведена діаграма. Для її побудови на вкладці «Знаряддя для зведених таблиць» є спеціальна кнопка.



Вправа 1. Обчислити розподіл хворих на інфекційні захворювання за статтю та областю і районом проживання за 31.08.2020–11.08.2021 р. Результати розмістити у «Зведеній таблиці 1». Зробити висновки.

Алгоритм виконання:

1. Створити зведену таблицю на новому аркуші за інформацією з джерела даних, що розташоване на аркуші «Первинні дані».

2. Макет зведеної таблиці налаштувати таким чином: в області стовпців розташувати поле «стать», в області рядків – поля «місце проживання область» та «місце проживання район», в області підсумкових значень – поле «реєстраційний №» (змінити для нього підсумкову операцію з обчислення суми на обчислення кількості).

3. Змінити текст в комірках-заголовках зведеної таблиці.

Зразок. «Зведена таблиця 1». Розподіл хворих на інфекційні захворювання за статтю та областю і районом проживання за 31.08.2020–11.08.2021 р. (рис. 27).

Область, район	Стать кількість пацієнтів		Загальний підсумок
	ж	ч	
Полтавська	23	5	28
Кременчуцький	6		6
Лубенський	4	2	6
Миргородський	4	3	7
Полтавський	9		9
Сумська	26	5	31
Конотопський	5	1	6
Охтирський	6	2	8
Роменський	6		6
Сумський	7	2	9
Шосткінський	2		2
Харківська	29	12	41
Богодухівський	2	1	3
Ізюмський	6	2	8
Красноградський	2	3	5
Куп'янський	2		2
Лозівський		2	2
Харківський	9	4	13
Чугуївський	8		8
Загальний підсумок	78	22	100

Рис. 27. «Зведена таблиця 1». Розподіл хворих на інфекційні захворювання за статтю та областю і районом проживання

Побудувати зведену діаграму за зразком, який представлений на рис. 28.

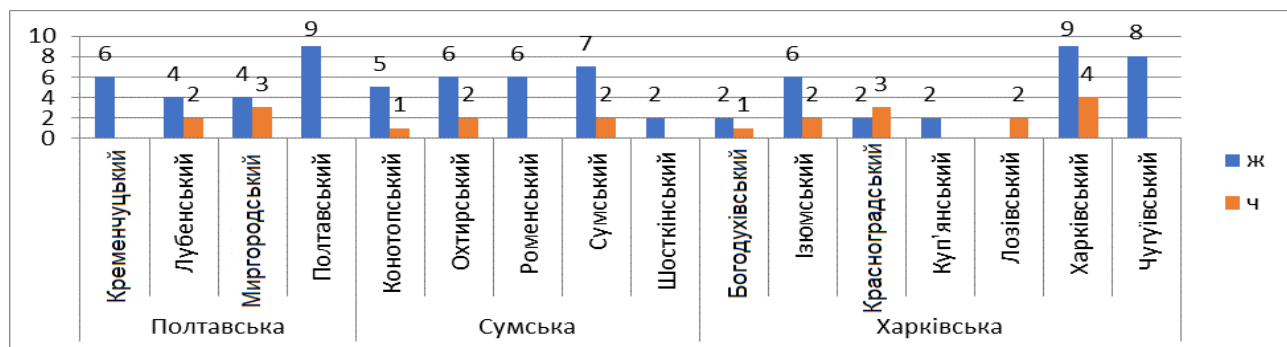


Рис. 28. Зведена діаграма, яка побудована за даними «Зведеної таблиці 1»

Висновки. Незважаючи на те, що жінок, хворих на інфекційні хвороби, було більше, ніж чоловіків, остаточний епідеміологічний висновок щодо порівняння за абсолютними показниками зробити неможливо. Необхідно розрахувати інтенсивні показники за статтю (захворюваність: на 100 тис. чоловіків, на 100 тис. жінок) та кожною окремою територією (область, район) і за їх результатами порівнювати.

Вправа 2. Обчислити розподіл хворих на інфекційні захворювання за статтю і віковою групою та областями і районами проживання за 31.08.2020–11.08.2021 р. Результати розмістити у «Зведеній таблиці 2». Зробити висновки.

Алгоритм виконання:

1. Створити зведену таблицю на новому аркуші за інформацією з джерела даних, що розташоване на аркуші «Первинні дані».

2. Макет зведеної таблиці налаштувати таким чином: в області стовпців розташувати поля «стать» і «вікова група», в області рядків – поля «місце проживання область» та «місце проживання район», в області підсумкових значень – поле «реєстраційний №» (змінити для нього підсумкову операцію з обчислення суми на обчислення кількості). Змінити текст в комірках-заголовках «Зведеної таблиці 1».

Зразок. «Зведена таблиця 2». Розподіл хворих на інфекційні захворювання за статтю і віковою групою та місцем проживання (область і район) за 31.08.2020–11.08.2021 р. представлено на *рис. 29*.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1															
2															
3	Кількість хворих	Стать. Вікова група													
4		ж Підсумок ч Підсумок Загальний підсумок													
5	Область, район	29 і менше	30-39	40-49	50-59	60-65	66 і більше		29 і менше	30-39	40-49	50-59	66 і більше		
6	Полтавська	1	2	6	6	3	5	23	2	1	1	1		5	28
7	Кременчуцький		1	2	1	1	1	6							6
8	Лубенський				2	1	1	4	1			1		2	6
9	Миргородський	1			1		2	4	1	1	1			3	7
10	Полтавський		1	4	2	1	1	9							9
11	Сумська	2	2	5	7	2	8	26	1			1		3	31
12	Конотопський	1		2	1		1	5					1	1	6
13	Охтирський	1	1		1	1	2	6	1				1	2	8
14	Роменський		1		2		3	6							6
15	Сумський			3	2	1	1	7				1		2	9
16	Шосткінський				1		1	2					1		2
17	Харківська	4	1	11	7	3	3	29	1	2	5	3	1	12	41
18	Богодухівський	1				1		2			1			1	3
19	Ізюмський			4	2			6		1			1	2	8
20	Красноградський					1	1	2			1	2		3	5
21	Куп'янський				1	1		2							2
22	Лозівський								1	1				2	2
23	Харківський	2	1	3	2		1	9			3	1		4	13
24	Чугувський	1		4	2		1	8							8
25	Загальний підсумок	7	5	22	20	8	16	78	4	3	6	5	4	22	100
26															

Рис. 29. Складна зведена таблиця з подвійним угрупованням в стовпцях та рядках

Побудувати зведені діаграми розподілу хворих на інфекційні захворювання за віковою групою та областю і районом місця проживання окремо для жінок і чоловіків. Скористатись засобом фільтрації, що вбудований до діаграми.

Зразок. «Зведена діаграма 2.1». Розподіл жінок, хворих на інфекційні захворювання за віковою групою та місцем проживання (область, район) за 31.08.2020–11.08.2021 р. представлено на *рис. 30*.

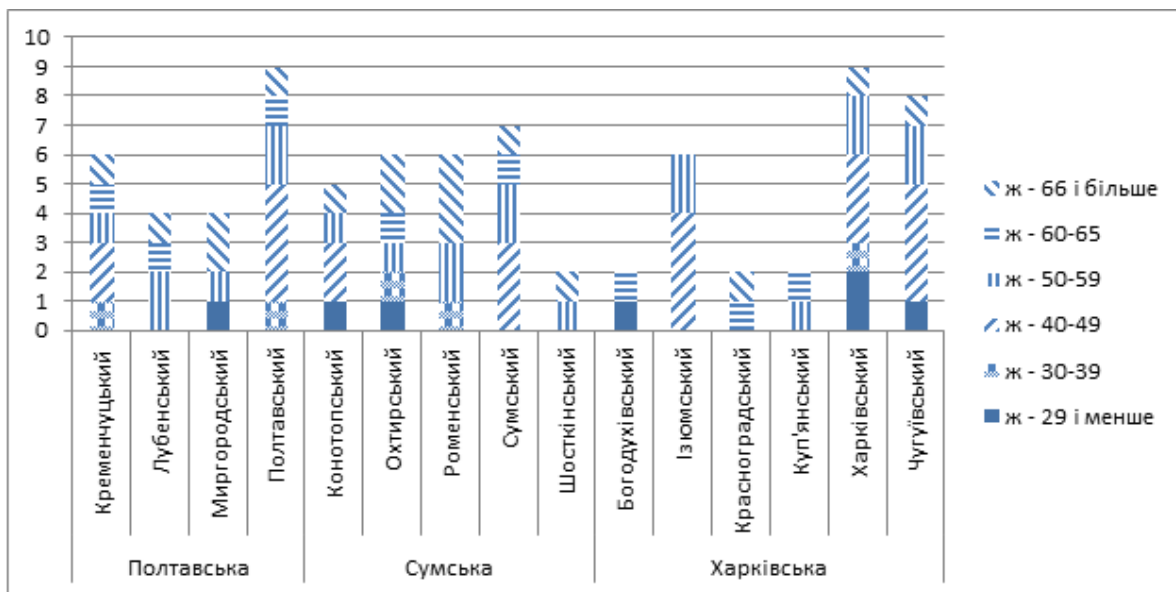


Рис. 30. Зразок зведеної діаграми з накопичуванням (для кожної вікової групи унікальний візерунок заливки; фільтрація даних за статтю – жінки)

Висновки. Для оцінки епідемічної ситуації та порівняння показників інфекційної захворюваності за віковими групами та різними територіями необхідно розрахувати інтенсивні показники з огляду на чисельність населення кожної вікової групи та окремо кожної території (область, район). Щодо отриманих результатів можна тільки констатувати факт реєстрації випадків захворювання.

Зразок. «Зведена діаграма 2.2». Розподіл чоловіків, хворих на інфекційні захворювання за віковою групою та місцем проживання (область, район) за 31.08.2020–11.08.2021 р. представлено на *рис. 31*.

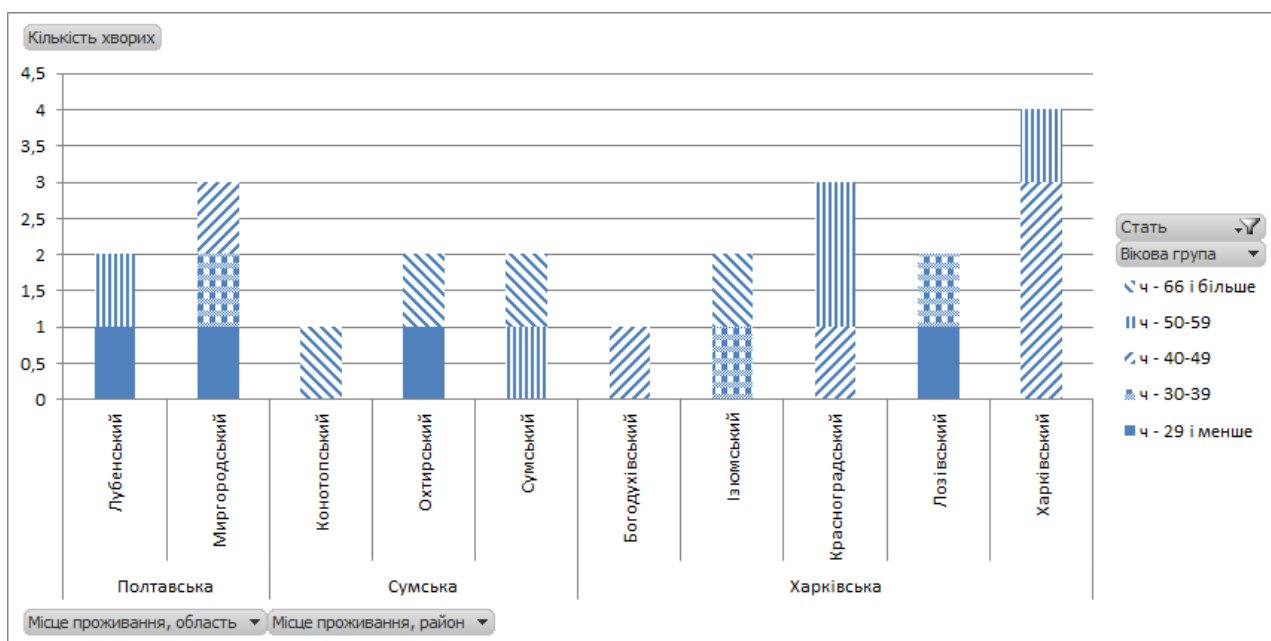


Рис. 31. Зразок зведеної діаграми з накопичуванням (фільтрація даних за статтю – чоловіки)

Висновки. Для оцінки епідемічної ситуації та порівняння показників інфекційної захворюваності між віковими групами та різними територіями необхідно розрахувати інтенсивні показники з огляду на чисельність населення кожної вікової групи та окремо кожної території (область, район). Щодо отриманих результатів можна тільки констатувати факт реєстрації випадків захворювання.

Вправа 3. Обчислити розподіл хворих на інфекційні захворювання за професійною групою та областю проживання за 31.08.2020–11.08.2021 р. Результати розмістити у «Зведеній таблиці 3». Зробити висновки.

Алгоритм виконання:

1. Створити зведену таблицю на новому аркуші за інформацією з джерела даних, що розташоване на аркуші «Первинні дані».

2. Макет зведеної таблиці налаштувати таким чином: в області стовпців розташувати поле «місце проживання область», в області рядків – поле «професія за кваліфікацією», в області підсумкових значень – поле «реєстраційний №» (змінити для нього підсумкову операцію з обчислення суми на обчислення кількості). Змінити текст в комірках-заголовках зведеної таблиці.

Зразок. «Зведена таблиця 3». Розподіл хворих на інфекційні захворювання за професійною групою та областю проживання за 31.08.2020–11.08.2021 р. представлено на *рис. 32*.

	A	B	C	D	E	
1						
2						
3	Кількість хворих	Область проживання				
4	Професія	Полтавська	Сумська	Харківська	Загальний підсумок	
5	Медичні працівники та допоміжний персонал		5	2	4	11
6	Не працює		3	2	5	10
7	Освітні професії		3	4	8	15
8	Пенсіонер		3	6	1	10
9	Працівники в сфері комп'ютерно-математичних, фізичних та соціальних наук			1		1
10	Працівники ділової та фінансової сфер		5	4	7	16
11	Промисловість		3	2	3	8
12	Робітники в сфері громадських та соціальних послуг			1	1	2
13	Робітники в сфері «Догляд за собою та обслуговування»				1	1
14	Робітники в сфері транспорту		1	1	1	3
15	Робітники з офісно-адміністративної підтримки		1	3	1	5
16	Робітники сфери «Монтаж, обслуговування та ремонт»				1	1
17	Робітники сфери продаж		2	1	2	5
18	Співробітники в сфері архітектури та інженерії		2	1	4	7
19	Співробітники з вирішення юридичних питань				2	2
20	Співробітники поліції та охоронних служб			3		3
21	Загальний підсумок		28	31	41	100

Рис. 32. Вікно зведеної таблиці з угрупованням хворих за професією і місцем проживання

Зразок. «Зведена діаграма 3». Розподіл хворих на інфекційні захворювання за професійною групою та областю проживання за 31.08.2020–11.08.2021 р. представлено на *рис. 33*.

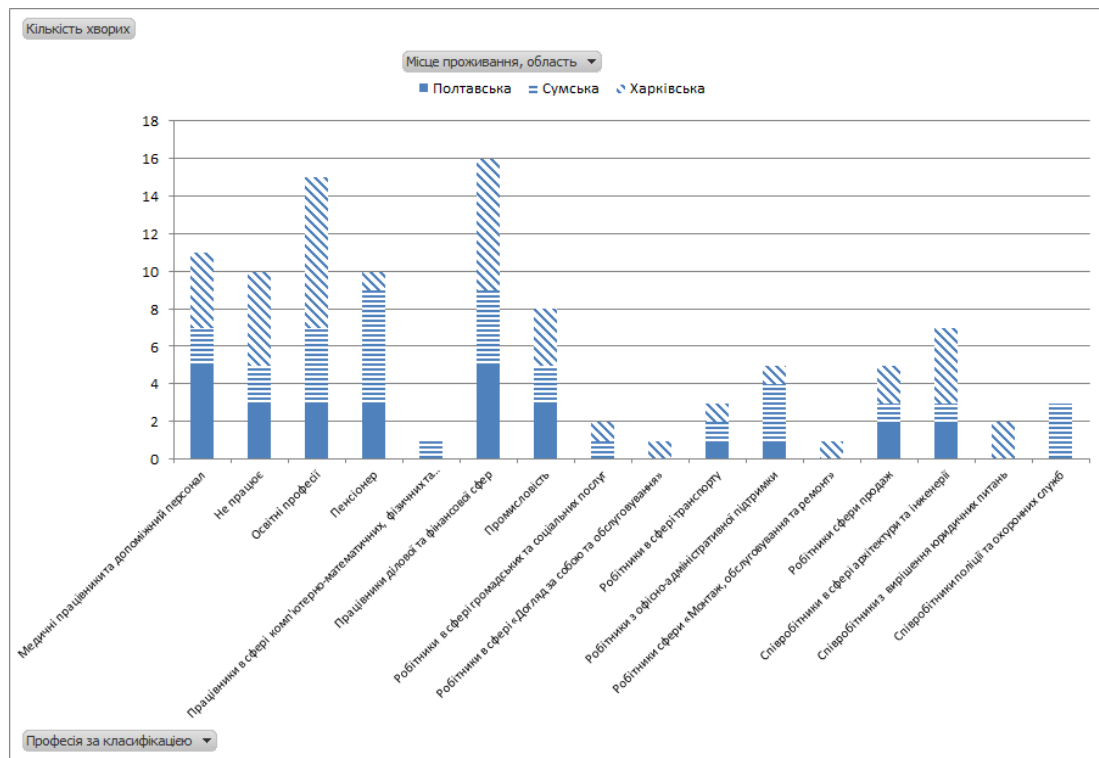


Рис. 33. Зведена діаграма з накопичуванням для аналізу кількості хворих за професією по областях

Висновки. Для оцінки епідемічної ситуації та порівняння показників інфекційної захворюваності між професійними групами та різними територіями необхідно розрахувати інтенсивні показники з огляду на чисельність населення кожної професійної групи окремо на кожній території (область, район) або екстенсивні показники і порівняти співвідношення однієї групи з іншою, у тому числі на різних територіях. Щодо отриманих результатів можна тільки констатувати факт реєстрації випадків захворювання.

Вправа 4. Обчислити розподіл хворих за групами інфекційних захворювань і визначенням випадку нозологічної форми та областю проживання за 31.08.2020–11.08.2021 р. з можливістю фільтрації за місяцем захворювання.

4.1. Відібрати всі результати і розмістити у зведеній таблиці. Зробити висновки.

Алгоритм виконання:

1. Створити зведену таблицю на новому аркуші за інформацією з джерела даних, що розташоване на аркуші «Первинні дані».

2. Макет зведеної таблиці налаштувати таким чином: в області стовпців розташувати поле «місце проживання область», в області рядків – поля «група інфекцій» та «визначення інфекційного захворювання», в області підсумкових значень – поле «реєстраційний №» (змінити для нього підсумкову операцію з обчислення суми на обчислення кількості). Змінити текст в комірках-заголовках зведеної таблиці.

Зразок. «Зведена таблиця 4.1». Розподіл хворих на інфекційні захворювання за групою інфекційних захворювань і визначенням випадку нозологічної форми та областю проживання за 31.08.2020–11.08.2021 р. представлено на *рис. 34*.

	A	B	C	D	E	F
3	Кількість хворих	Область проживання				
4	Інфекції	Полтавська	Сумська	Харківська	Загальний підсумок	
5	▣ Інфекції дихальних шляхів	11	11	13	35	
6	Дифтерія		4	1	5	
7	Кашлюк	3	1	4	8	
8	Кір	1	1	5	7	
9	Менінгококова інфекція	3	1	1	5	
10	Паротит	2		2	4	
11	Туберкульоз органів дихання	2	4		6	
12	▣ Інфекції зовнішніх покривів	3	3	3	9	
13	Короста	3	3	3	9	
14	▣ Кишкові інфекції	11	11	14	36	
15	Бруцельоз	2	2	1	5	
16	Вірусний гепатит А	3	3	3	9	
17	Сальмонельоз	1	2	4	7	
18	Черевний тиф	1	1	2	4	
19	Шигельоз	4	3	4	11	
20	▣ Кров'яні інфекції	3	6	11	20	
21	Кліщовий енцефаліт	1	3	3	7	
22	Рикетсіоз	1	2	4	7	
23	Туляремія	1	1	4	6	
24	Загальний підсумок	28	31	41	100	

Рис. 34. Вікно зведеної таблиці з угрупованням хворих за групою інфекційних захворювань і визначенням випадку нозологічної форми та областю проживання

Побудувати зведену діаграму розподілу за групою інфекційних захворювань і визначенням нозологічної форми та областю проживання.

Зразок. «Зведена діаграма 4.1». Розподіл хворих за групою інфекцій і визначенням випадку нозологічної форми та областю проживання за 31.08.2020–11.08.2021 р. представлено на рис. 35.

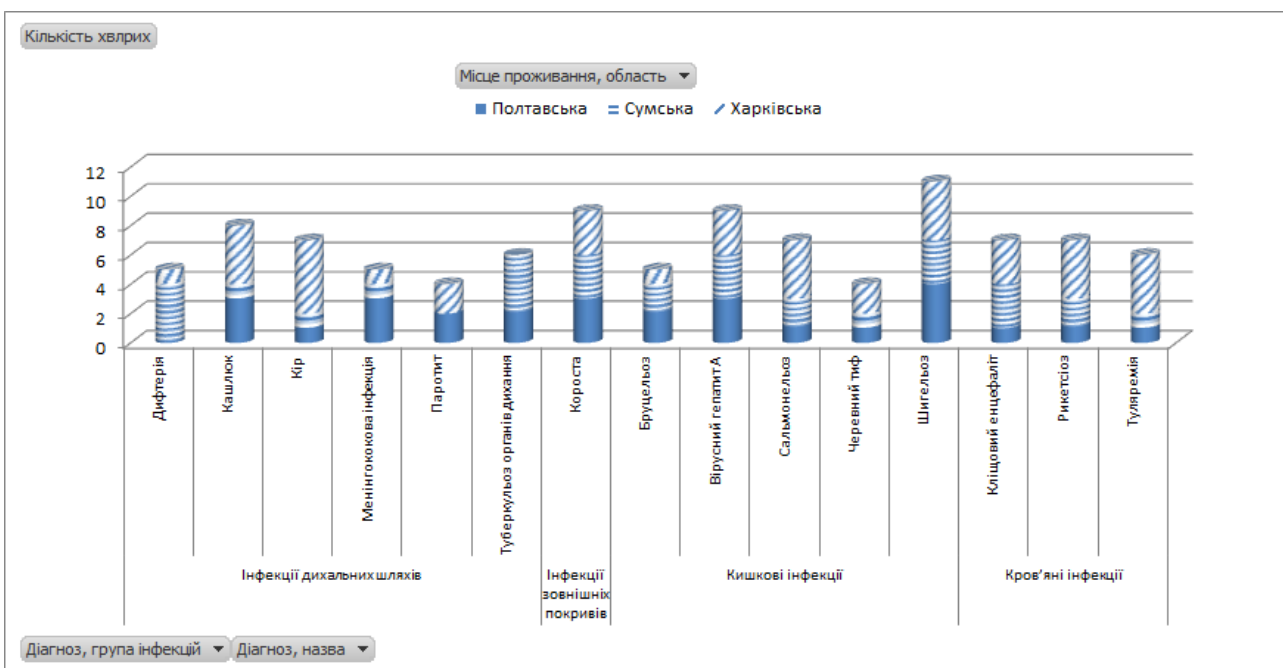


Рис. 35. Зведена діаграма з накопичуванням для аналізу кількості хворих за групами інфекційних захворювань і визначенням випадку нозологічної форми по областях

Висновки. Для оцінки епідемічної ситуації та порівняння показників між групами інфекційних захворювань та окремими нозологічними формами необхідно розрахувати інтенсивні показники з огляду на чисельність населення певної території (область, район) або екстенсивні показники і проаналізувати співвідношення однієї групи інфекцій з іншою або одного інфекційного захворювання з іншим в межах певної території. Щодо отриманих результатів можна тільки констатувати факт реєстрації випадків захворювання.

4.2. Відібрати результати за 2020 р. і розмістити у «Зведеній таблиці 4.2».

Алгоритм виконання:

1. В області фільтрації зведеної таблиці розташувати поле «початок захворювання рік+місяць».

2. Встановити режим фільтрації (рис. 36).

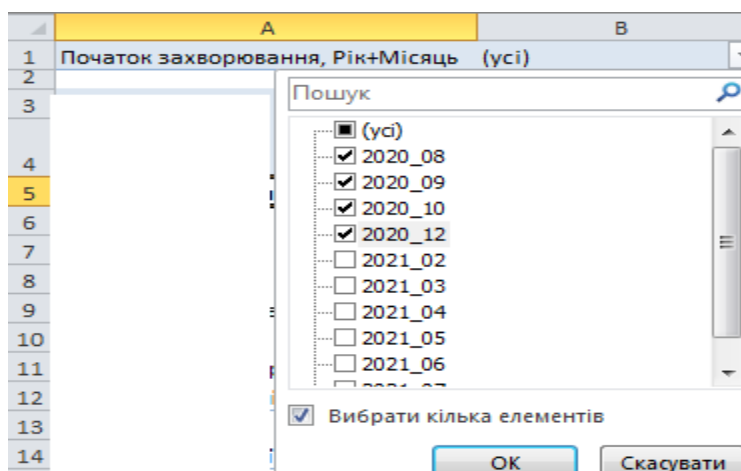


Рис. 36. Вибір параметрів фільтрації зведеної таблиці

Зразок. «Зведена таблиця 4.2». Розподіл хворих за групами інфекційних захворювань і визначенням випадку нозологічної форми та областю проживання за 31.08.2020–31.12.2020 р. представлено на рис. 37.

A	B	C	D	E	
Початок захворювання, Рік+Місяць	(кілька елементів)				
Кількість хворих	Область проживання				
Інфекції	Полтавська	Сумська	Харківська	Загальний підсумок	
☑ Інфекції дихальних шляхів		1	1	2	
Кашлюк			1	1	
Туберкульоз органів дихання		1		1	
☑ Інфекції зовнішніх покривів			1	1	
Короста			1	1	
☑ Кишкові інфекції		1		1	
Шигельоз		1		1	
Загальний підсумок		1	1	2	4

Рис. 37. Вікно зведеної таблиці з угрупованням хворих за групами інфекційних захворювань і визначенням випадку нозологічної форми та областю проживання і фільтрацією за роком початку захворювання

Зразок. «Зведена діаграма з накопичуванням» представлена на рис. 38.

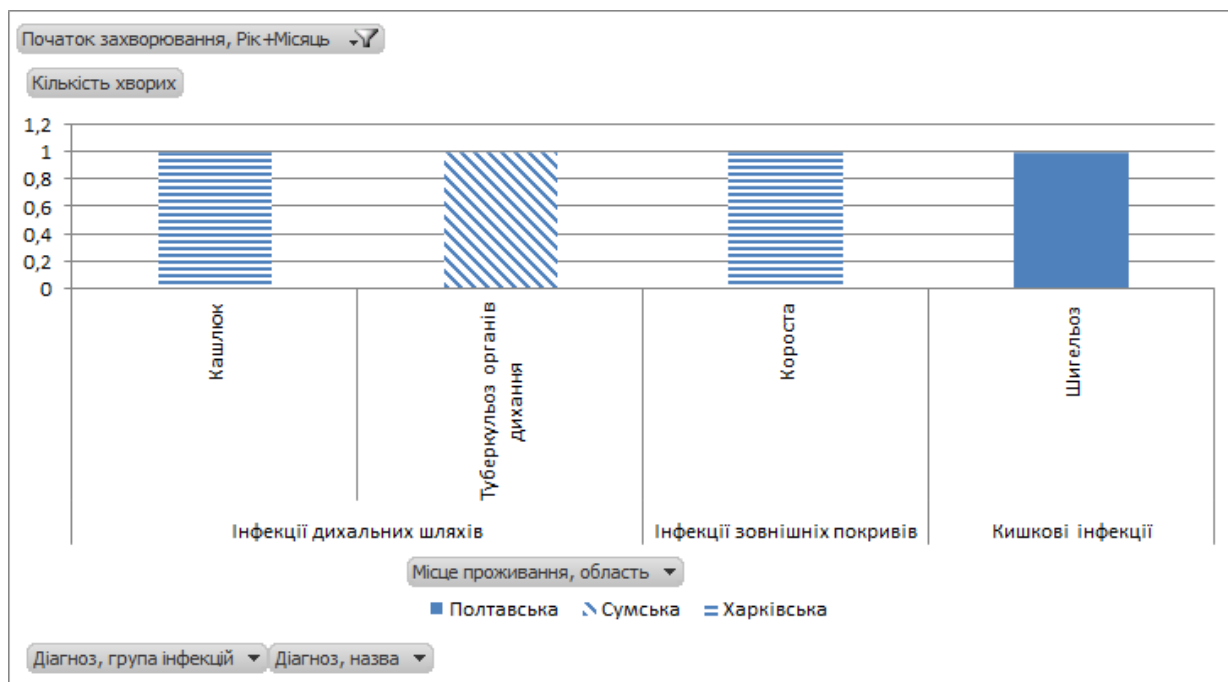


Рис. 38. Зведена діаграма з накопичуванням для аналізу кількості хворих за групами інфекційних захворювань і визначенням випадку нозологічної форми по областях за 2020 р.

Висновки. Для оцінки епідемічної ситуації та порівняння показників між групами інфекційних захворювань та між окремими нозологічними формами необхідно розрахувати інтенсивні показники з огляду на чисельність населення певної території (область, район) або екстенсивні показники і проаналізувати співвідношення однієї групи інфекцій з іншою або одного інфекційного захворювання з іншим у межах певної території. Щодо отриманих результатів можна тільки констатувати факт реєстрації випадків захворювання.

Вправа 5. Обчислити розподіл хворих за групами інфекційних захворювань і визначенням випадку нозологічної форми та місцем проживання (область, район) з можливістю фільтрації за ознакою тривалого лікування за 31.08.2020–11.08.2021 р.

5.1. Відібрати всі результати і розмістити у «Зведеній таблиці 5.1».

5.2. Відібрати результати за ознакою тривалого лікування і розмістити у «Зведеній таблиці 5.2».

Алгоритм виконання:

1. Створити зведену таблицю на новому аркуші за інформацією з джерела даних, що розташоване на аркуші «Первинні дані».

2. Макет зведеної таблиці налаштувати таким чином: в області стовпців розташувати поля «місце проживання область» та «місце розташування район», в області рядків – поля «група інфекцій» та «визначення інфекційного захворювання», в області підсумкових значень – поле «реєстраційний №» (змінити для нього підсумкову операцію з обчислення суми на обчислення кількості). Змінити текст в комірках-заголовках зведеної таблиці. В області фільтрації розташувати поле «тривале лікування».

Зразок. «Зведена таблиця 5.1». Розподіл хворих за групами інфекційних захворювань і визначенням випадку нозологічної форми та місцем проживання (область, район) за 31.08.2020–11.08.2021 р. представлено на *рис. 39*.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
4		Полтавська				Полтавська Підсумок	Сумська					Сумська Підсумок	Харківська				Харківська Підсумок	Загальний підсумок				
5	Інфекції	Кременчуцький	Лубенський	Миргородський	Полтавський		Котопольський	Охтирський	Роменський	Сумський	Шосткинський		Богодухівський	Валювський	Красноградський	Куп'янський	Лозівський	Харківський	Чугуївський			
6	Інфекції дихальних шляхів	3	2	3	3	11	2	2	3	3	1	11	1	1	1	1		6	3	13	35	
7	Дифтерія						1	1		2									1		1	5
8	Кашлюк		1	1	1	3		1				1						4		4	8	
9	Кір				1	1	1					1	1	1	1			2		5	7	
10	Менінгококова інфекція	2		1		3			1			1						1		1	5	
11	Паротит	1	1			2									1				1		2	4
12	Туберкульоз органів дихання			1	1	2			3		1	4										6
13	Інфекції зовнішніх покривів		1	1	1	3	2	1				3					1			3	9	
14	Короста		1	1	1	3	2	1				3		2			1			3	9	
15	Кишкові інфекції	2	3	2	4	11	1	3	2	4	1	11	1	2			1	7	3	14	36	
16	Бруцельоз		1		1	2			1	1		2							1		1	5
17	Вірусний гепатит А		1	1	1	3	1			1	1	3					1	1	1	3	9	
18	Сальмонельоз				1	1		2				2	1	1				1	1	1	4	7
19	Черевний тиф	1				1				1		1						2		2	4	
20	Шигельоз	1	1	1	1	4		1	1	1		3		1				3		4	11	
21	Кров'яні інфекції	1		1	1	3	1	2	1	2		6	1	3	4	1			2	11	20	
22	Кліщовий енцефаліт				1	1	1	1	1	1		3		1	2					3	7	
23	Рикетсіоз	1				1		1		1		2	1	1	1				1	4	7	
24	Туляремія			1		1	1					1		1	1	1			1	4	6	
25	Загальний підсумок	6	6	7	9	28	6	8	6	9	2	31	3	8	5	2	2	13	8	41	100	

Рис. 39. Вікно зведеної таблиці з угрупованням хворих за групами інфекційних захворювань і визначенням випадку нозологічної форми та місцем проживання (область, район)

Зразок. «Зведена таблиця 5.2». Розподіл хворих за групами інфекційних захворювань і визначенням випадку нозологічної форми та місцем проживання (область, район) за 31.08.2020–31.12.2020 р. представлено на *рис. 40* (тривале лікування).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Тривале лікування	так																
2		Область, район																
3	Кількість хворих																	
4		Полтавська		Полтавська Підсумок	Сумська			Сумська Підсумок	Харківська			Харківська Підсумок	Загальний підсумок					
5	Інфекції	Лубенський	Миргородський	Полтавський		Охтирський	Роменський	Сумський	Шосткинський		Валювський	Красноградський	Куп'янський	Харківський	Чугуївський			
6	Інфекції дихальних шляхів		1	1	2				1		1	1		2		3	6	
7	Кашлюк													2		2	2	
8	Кір			1	1							1				1	2	
9	Менінгококова інфекція		1		1												1	
10	Туберкульоз органів								1		1						1	
11	Інфекції зовнішніх покривів											1				1	1	
12	Короста											1				1	1	
13	Кишкові інфекції	1	1		2	1	1	2						2	1	3	9	
14	Бруцельоз	1	1		2			1							1	1	4	
15	Вірусний гепатит А							1			1						1	
16	Сальмонельоз						1				1						1	
17	Черевний тиф							1			1			1		1	2	
18	Шигельоз													1		1	1	
19	Кров'яні інфекції						1	1		2			1	1	1	3	5	
20	Кліщовий енцефаліт							1	1		2						2	
21	Рикетсіоз												1		1	2	2	
22	Туляремія													1		1	1	
23	Загальний підсумок	1	1	2	4	2	1	3	1	7	2	1	1	4	2	10	21	

Рис. 40. Вікно зведеної таблиці з угрупованням хворих за групами інфекційних захворювань і визначенням випадку нозологічної форми та областю і районом проживання та фільтрацією за ознакою тривалого лікування

3.1.3. Практичне заняття 3.

Визначення функції розподілу та основних числових статистичних характеристик показника

Тема: Дескриптивна статистика епідеміологічних даних. Визначення функції розподілу показника. Визначення основних числових статистичних характеристик.

Мета: удосконалення навичок з автоматизації розрахунків числових статистичних характеристик показників, визначення функції розподілу показника за даними вибірки.

Теоретичні відомості

Дескриптивна або описова статистика (англ. descriptive statistics) – розділ статистики, який займається обробкою емпіричних даних, їхньою систематизацією, наочним представленням у вигляді графіків та таблиць, а також їх кількісним описом через основні статистичні характеристики.

Статистичні характеристики показника залежать від типу даних. У статистиці розглядається два основних типи даних: атрибутивні (описові) – виражені текстом та кількісні – виражені числом. Кількісні дані поділяються на дискретні (які приймають скінчений перелік значень) і неперервні (які можуть приймати безліч значень). Відповідність між значенням показника та ймовірністю появи цього значення називається законом розподілу показника, який може бути заданий у вигляді таблиці, формули або графіка.

На наведених графіках (рис. 41) 50 значень дискретного показника та функції розподілу (гістограми частот) показника можна побачити різницю між цими поняттями:

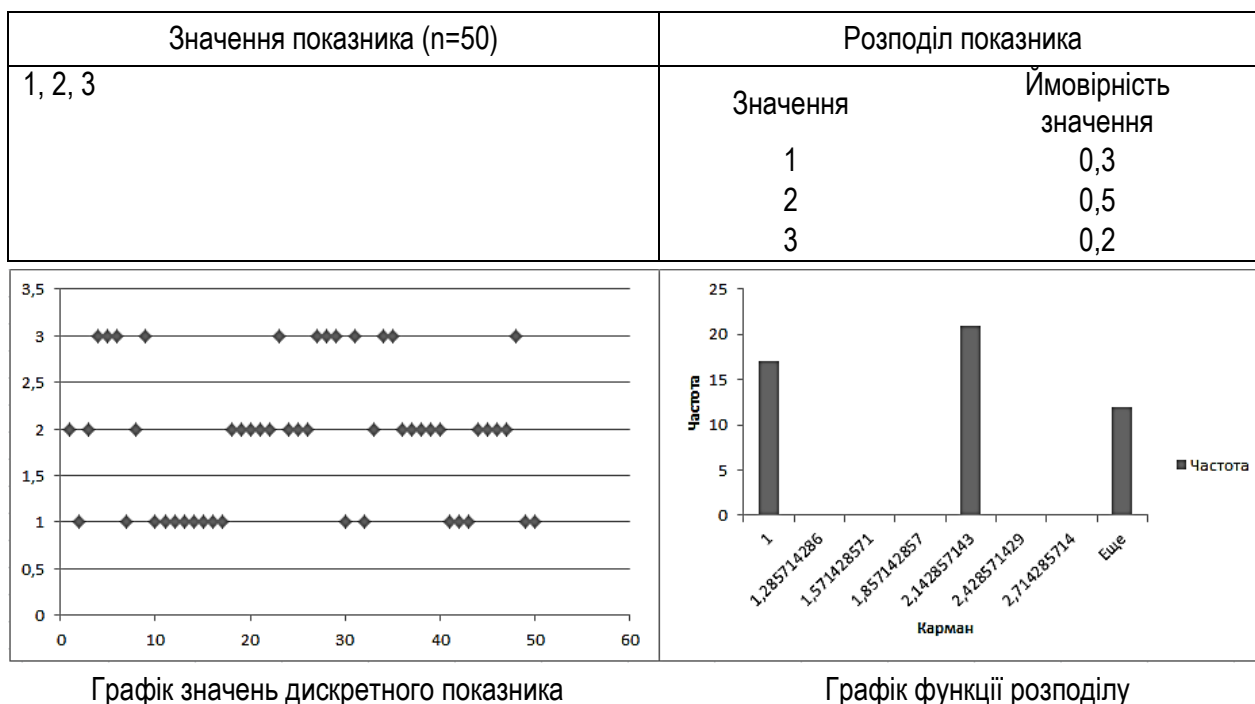
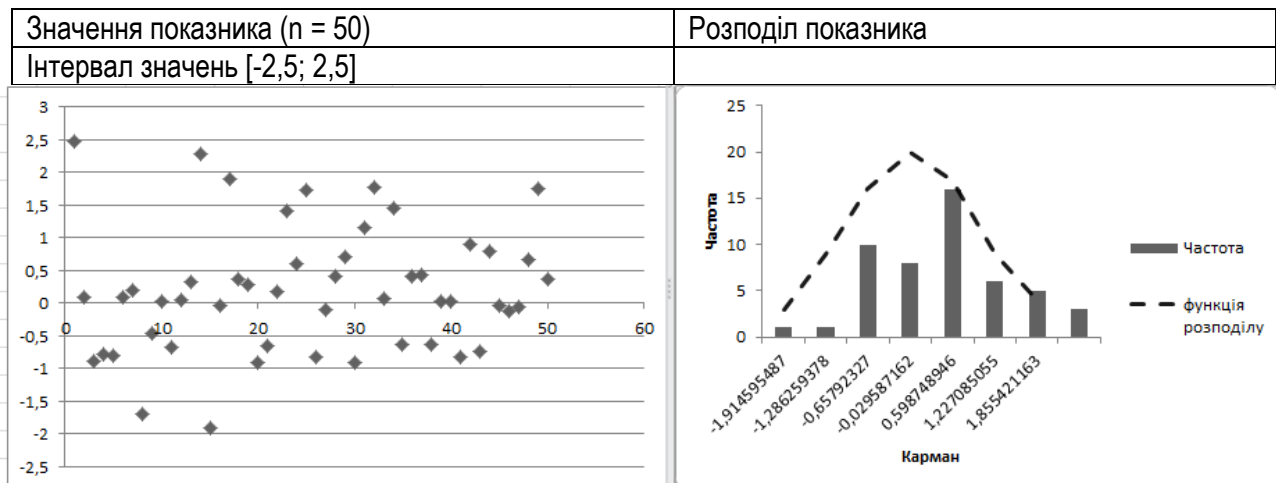


Рис. 41. Графіки значень та функції розподілу дискретного кількісного показника

На рис. 42 представлені графік 50 значень та функції розподілу неперервного показника.



Графік значень неперервного показника

Графік функції розподілу

Рис. 42. Графіки значень та функції розподілу неперервного кількісного показника

Для дискретного показника функція розподілу $f(x)$ = ймовірності того, що значення показника дорівнює x .

Для неперервного показника функція розподілу $f(x)$ = ймовірності того, що значення показника дорівнює або менше за x .

Окремим випадком неперервного розподілу є нормальний розподіл.

Закон розподілу показника визначає методи його подальшого дослідження: якщо показник має нормальний розподіл, використовуються методи параметричної статистики, в іншому випадку – непараметричні методи.

Висновки щодо функції розподілу можна зробити на підставі: 1) аналізу узагальнюючих числових статистичних характеристик ряду кількісних даних, що досліджуються; 2) графіка функції розподілу даних (гістограми частот) та діаграми розмаху (Box Plot); 3) графіка нормальної ймовірності (Normal Probability Plot або квантильний графік, або графік Q-Q); 4) перевірки статистичних критеріїв щодо нормальності даних.

Нехай x_1, x_2, \dots, x_n – значення кількісної ознаки, n – кількість значень.

Числові статистичні характеристики ряду кількісних даних можна поділити на декілька груп.

Характеристики міри положення (або центральної тенденції)

Найчастіше використовуються:

- Середня арифметична ряду даних (англ. Mean, M), обчислюється за формулою $M = \frac{1}{n} \sum_{i=1}^n x_i$.

- Медіана ряду даних (англ. Median, Me) – таке значення ознаки, для якого однаково ймовірно, що інші значення ряду даних виявляться меншими або більшими за нього. Щоб обчислити медіану, дані повинні бути впорядковані від найменшого до найбільшого. Якщо n – непарне число ($n=2k+1$), $Me = x_{k+1}$, тобто середнє число у впорядкованому ряду. Якщо n – парне число ($n=2k$), $Me = \frac{x_k + x_{k+1}}{2}$, тобто середнє арифметичне двох чисел, які знаходяться посередині впорядкованого ряду.

- Мода ряду даних (англ. Mode, Mo) – значення ознаки, яке найбільше повторюється у ряду даних.

• Середня геометрична ряду даних (англ. Geom.mean, G), обчислюється за формулою $G = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$.

• Квартилі (англ. Quartiles, Q_1, Q_2, Q_3). Щоб їх обчислити, дані повинні бути впорядковані від найменшого до найбільшого. Квартилі ділять впорядковані дані на чверті. Q_2 (50 %-й) = Ме. Q_1 (25 %-й) – Ме частини впорядкованого ряду від найменшого до Ме (Q_2). Q_3 (75 %-й) – Ме частини впорядкованого ряду від Ме (Q_2) до найбільшого.

Характеристики міри розсіювання (мінливості)

• Розмах (англ. Range, R) – різниця між найбільшим та найменшим значенням ряду, обчислюється за формулою $R = x_{max} - x_{min}$.

• Дисперсія (англ. Sample Variance, σ^2) – міра розсіювання даних навколо середнього значення, вимірює наскільки далеко розкидані дані від M , обчислюється за формулою $\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - M)^2$.

• Стандартне відхилення від середнього значення (англ. Standard Deviation, σ) – міра варіабельності ряду даних, середньоквадратичний розкид даних, дає уявлення, де знаходяться дані: вище або нижче середнього значення, використовується частіше, ніж дисперсія, обчислюється за формулою $\sigma = \sqrt{\sigma^2}$. У випадку нормального розподілу в інтервалі $M - \sigma < X < M + \sigma$ знаходиться 68 % усіх значень даних ряду, що досліджуються;

• Стандартна помилка середнього значення (англ. Standard Error of mean, m), обчислюється за формулою $m = \frac{\sigma}{\sqrt{n}}$.

• Коефіцієнт варіації (англ. Coefficient of variation, l) – відносна міра коливання варіаційного ряду, обчислюється за формулою $l = \frac{\sigma}{M} \cdot 100 \%$. Коефіцієнт варіації використовується для оцінки різноманітності даних у ряді, що досліджується. Якщо $l < 10 \%$, різноманітність даних вважається слабкою. Якщо $10 \% \leq l < 20 \%$, різноманітність даних вважається середньою. Якщо $20 \% \leq l \leq 33 \%$, різноманітність даних вважається значною. Якщо $l > 33 \%$, це свідчить про неоднорідність інформації та необхідність виключення з вибірки найбільших і найменших значень.

• Міжквартильний розмах (англ. Quartile range, QR), обчислюється за формулою $QR = Q_3 - Q_1$. У випадку закону розподілу відмінного від нормального мірою розсіювання даних є міжквартильний розмах: в інтервалі $Q_1 < X < Q_3$ знаходиться 50 % усіх значень даних, що досліджуються.

Характеристики міри скошеності (симетричності) функції розподілу

• Коефіцієнт асиметрії (англ. Skewness, As), дозволяє встановити симетричність розподілу даних навколо середнього значення, обчислюється за формулою:

$$As = \frac{1}{n\sigma^3} \sum_{i=1}^n (x_i - M)^3.$$

Характеристики міри гостровершинності функції розподілу

• Екссес (англ. Kurtosis, E), дозволяє встановити гостровершинність або плосковершинність функції розподілу, обчислюється за формулою:

$$E = \frac{1}{n\sigma^4} \sum_{i=1}^n (x_i - M)^4 - 3.$$

Якщо дані розподілені за нормальним законом, спостерігається наступне:

1. Майже усі (99,7 %) відхилення від середнього значення менше 3σ .
2. Дві третини (68,3 %) відхилень від середнього значення менше за σ .
3. Половина відхилень від середнього значення менше за $0,625\sigma$.

Крива (функція) нормального розподілу характеризується тим, що існує певне центральне максимальне значення досліджуваного параметра, і частота появлення інших значень тим менше, чим далі це значення від центрального.

Форма нормальної кривої має вид одновершинної симетричної кривої, вітки якої асимптотично наближаються до осі абсцис. Найбільша ордината кривої відповідає середньому значенню M . Графік має точки перетину при $x = M - \sigma$ та $x = M + \sigma$. При збільшенні σ крива щільності розподілу стає більш пологою. При зменшенні σ графік щільності розподілу більше стискається до осі Oy , стає гостровершинним (рис. 43).

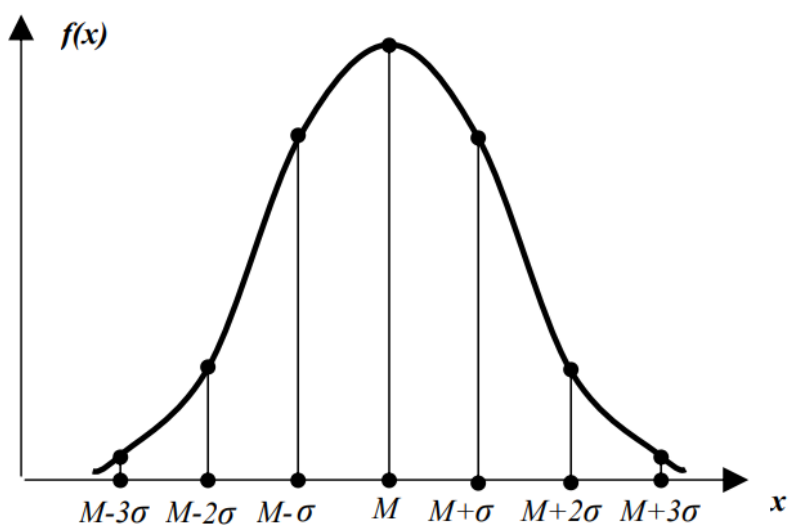


Рис. 43. Графік кривої нормального розподілу

Якщо показник має нормальний розподіл, то виконуються наступні співвідношення між числовими характеристиками:

- $M \approx Me \approx Mo$ (середнє значення, медіана та мода співпадають);
- $As \approx 0$ (коефіцієнт асиметрії близький до 0), взагалі $As \in [-2; 2]$;
- $E \approx 0$ (ексцес близький до 0), взагалі $E \in [-2; 2]$;
- $\sigma \approx 0,75(Q_3 - Q_1) \approx \frac{1}{6}R$ при $n > 30$.

M і σ називаються точковими оцінками параметрів нормального розподілу. Якщо $M=0$, $\sigma=1$, то розподіл називається стандартним нормальним розподілом.

Окрім точкової оцінки параметрів використовується інтервальна. Довірчий інтервал (англ. Confidence Level, Δ) – це діапазон значень, який із високою ймовірністю включає усі дані. Довірчий інтервал для середнього значення має вигляд $(M-\Delta, M+\Delta)$, де $\Delta = t_{\alpha, n} \frac{\sigma}{\sqrt{n}}$, n – кількість значень у вибірці, σ – стандартне відхилення, коефіцієнт t обирається з таблиці розподілу Стюдента і залежить від заданого дослідником рівня значущості α та кількості значень у вибірці, за якою обчислені M і σ . Довірча ймовірність пов'язана з рівнем значущості і дорівнює $1-\alpha$.

У потужних статистичних програмах (SPSS, Statistica, StatGraphics, EpiInfo, тощо) і програмі Microsoft Excel для перевірки нормальності розподілу ряду кількісних даних можна: 1) отримати описові статистичні характеристики та перевірити виконання співвідношень між ними; 2) побудувати гістограму частот та порівняти отриману гістограму з дзвоноподібним графіком нормального розподілу; 3) побудувати і дослідити діаграму розмаху; 4) побудувати за досліджуваними даними і проаналізувати графік нормальної ймовірності (Normal Probability Plot); 5) перевірити статистичні тести щодо нормальності. Пункти 1, 2, 3, 4 відносяться до так званого попереднього аналізу даних.

Задача-еталон

Дано: Інформація про випадки захворювань на гострі інфекції дихальних шляхів на території України у 2019 р. надана в таблиці Excel (джерело даних – звіт МОЗ про інфекційну захворюваність в Україні – форма 1) (табл. 4).

Таблиця 4

Стовпці (епідеміологічні показники) у таблиці Excel

№ з/п	Назва показника
1	Регіон України
2	Область
3	Населення
4	Інтенсивний показник на 100 000, все населення
5	Інтенсивний показник на 100 000, доросле населення
6	Інтенсивний показник на 100 000, діти 0–17 років
7	Інтенсивний показник на 100 000, діти 5–9 років
8	Інтенсивний показник на 100 000, діти 10–14 років
9	Інтенсивний показник на 100 000, діти 15–17 років

Дані для аналізу надаються у файлі Lesson 03D.xlsx.

Вправа 1. Обчислити числові статистичні характеристики показників, наданих у стовпцях 4–9.

- Область допустимих значень показника: x_{min} , x_{max} .
- Показники центральної тенденції: M ; Me ; Mo ; Q_1 , Q_3 .
- Ступінь варіації показника: σ^2 , σ , l , m , $R = x_{max} - x_{min}$, $QR = Q_3 - Q_1$.
- Зкошеність: As .
- Гостровершинність: E .
- Кількість значень (англ. Count): n .
- Довірчий інтервал для середнього значення Δ .

Для обчислення числових статистичних характеристик використати програму Microsoft Excel: статистичні функції або засіб «Аналіз даних → Описова статистика». Проаналізувати співвідношення між числовими статистичними характеристиками. Зробити висновки.

Алгоритм виконання:

1. Відкрити файл Lesson 03D.xlsx. Зберегти як Lesson 03R.xlsx.
2. Діапазон B31:J53 аркушу «Первинні дані» зарезервованій для обчислення таблиці основних статистичних характеристик (рис. 44).

	A	B	C	D
31		Мінімум	Мін	
32		1-й квартиль	Q1	
33		Медіана	Me	
34		3-й квартиль	Q3	
35		Максимум	Max	
36		Розмах	R	
37		Міжквартильний розмах	QR	
38		Середнє значення	M	
39		Відносна різниця між M та Me		
40		Мода	Mo	
41		Дисперсія	σ^2	
42		Стандартне відхилення	σ	
43			R/6	
44		Відносна різниця між σ та R/6		

	A	B	C	D
45			0,75QR	
46		Відносна різниця між σ та 0,75QR		
47		Коефіцієнт варіації, %	I	
48		Асиметрія	Az	
49		Ексіос	E	
50		Кількість значень	n	
51		Похибка середнього	m	
52		Відносна похибка середнього значення		
53		Довірчий інтервал для середнього значення	Δ	

Рис. 43. Шаблон таблиці основних статистичних числових характеристик

3. Порядок заповнення таблиці:

- у комірку D31 ввести формулу =MIN(D3:D27);
- у комірку D32 ввести формулу =QUARTILE.INC(D3:D27;1);
- у комірку D33 ввести формулу =MEDIAN(D3:D27);
- у комірку D34 ввести формулу =QUARTILE.INC(D3:D27;3);
- у комірку D35 ввести формулу =MAX(D3:D27);
- у комірку D36 ввести формулу =D35-D31;
- у комірку D37 ввести формулу =D34-D32;
- у комірку D38 ввести формулу =AVERAGE(D3:D27);
- у комірку D39 ввести формулу =ABS(D38-D33)/D38;
- у комірку D40 ввести формулу =MODE.SNGL(D3:D27);
- у комірку D41 ввести формулу =VAR.S(D3:D27);
- у комірку D42 ввести формулу =STDEV.S(D3:D27);
- у комірку D43 ввести формулу =D36/6;
- у комірку D44 ввести формулу =ABS(D42-D43)/D42;
- у комірку D45 ввести формулу =0,75*D37;
- у комірку D46 ввести формулу =ABS(D42-D45)/D42;
- у комірку D47 ввести формулу =(D42/D38)*100;
- у комірку D48 ввести формулу =SKEW(D3:D27);
- у комірку D49 ввести формулу =KURT(D3:D27);
- у комірку D50 ввести формулу =COUNT(D3:D27);
- у комірку D51 ввести формулу =D42/SQRT(D50);
- у комірку D52 ввести формулу =D51/D38;
- у комірку D53 ввести формулу =CONFIDENCE.T(0,05;D42;25).

4. Скопіювати формули з діапазону D31: D53 до діапазону E31:I31.

На рис. 44 представлені отримані результати.

	A	B	C	D	E	F	G	H	I
31		Мінімум	Min	4074,30	1188,59	16191,33	16942,67	16291,35	14244,35
32		1-й квартиль	Q1	13576,78	5157,36	45289,73	40207,40	40105,86	34596,29
33		Медіана	Me	16091,44	6646,65	54874,67	52057,25	47110,91	51430,56
34		3-й квартиль	Q3	17760,98	7449,86	67204,88	62653,59	53012,43	57191,61
35		Максимум	Max	26913,96	12985,76	85449,04	78892,28	69996,99	86352,72
36		Розмах	R	22839,66	11797,17	69257,72	61949,62	53705,64	72108,37
37		Міжквартильний розмах	QR	4184,20	2292,51	21915,15	22446,19	12906,57	22595,32
38		Середнє значення	M	15093,78	6309,89	53756,75	50622,72	45026,53	48092,29
39		Відносна різниця між M та Me		0,07	0,05	0,02	0,03	0,05	0,07
40		Мода	Mo	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
41		Дисперсія	σ^2	24534939,82	5653306,55	278333632,32	272755910,05	167289109,21	268162317,83
42		Стандартне відхилення	σ	4953,28	2377,67	16683,33	16515,32	12934,03	16375,66
43			R/6	3806,61	1966,20	11542,95	10324,94	8950,94	12018,06
44		Відносна різниця між σ та R/6		0,23	0,17	0,31	0,37	0,31	0,27
45			0,75QR	3138,15	1719,38	16436,37	16834,64	9679,93	16946,49
46		Відносна різниця між σ та 0,75QR		0,37	0,28	0,01	0,02	0,25	0,03
47		Коефіцієнт варіації, %	l	32,82	37,68	31,03	32,62	28,73	34,05
48		Асиметрія	As	-0,43	0,15	-0,44	-0,42	-0,61	-0,15
49		Екссес	E	1,31	2,19	-0,05	-0,52	0,18	0,27
50		Кількість значень	n	25	25	25	25	25	25
51		Похибка середнього значення	m	990,66	475,53	3336,67	3303,06	2586,81	3275,13
52		Відносна похибка середнього значення		0,07	0,08	0,06	0,07	0,06	0,07
53		Двоірочний інтервал для середнього значення	Δ	2044,61	981,45	6886,54	6817,19	5338,90	6759,54

Рис. 45. Результати обчислень числових статистичних характеристик

Перевірити правильність обчислень можна застосувавши засіб Microsoft Excel «Аналіз даних».

Засіб «Аналіз даних» є надбудовою і за замовчуванням не встановлюється. Якщо засіб встановлений, він розташований на вкладці «Дані» (рис. 46):

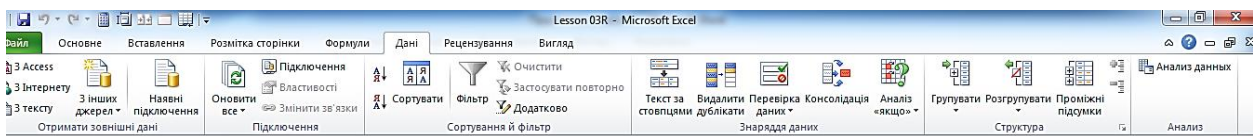


Рис. 46. Команди вкладки «Дані» меню табличного процесора Microsoft Excel

Для встановлення засобу «Аналіз даних» необхідно: 1) виконати команду «Файл → Параметри → Надбудови»; 2) перейти до керування надбудовами; 3) у вікні керування активізувати «Пакет аналізу» (рис. 47).

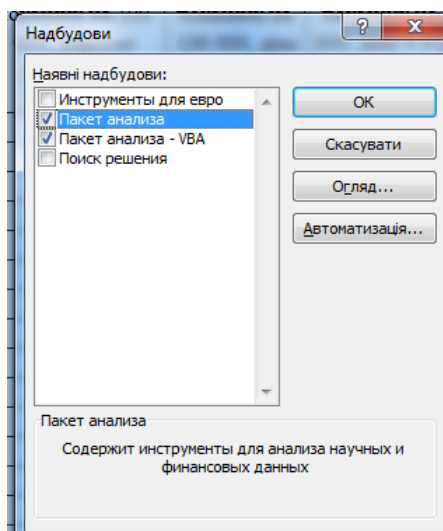


Рис. 47. Вікно керування надбудовами табличного процесора Microsoft Excel

Алгоритм виконання процедури «Аналіз даних → Описова статистика»:

1. Виконати команду «Дані → Аналіз даних → Описова статистика» (Data Analysis → Descriptive Statistics).
2. У вікні процедури встановити наступні параметри (рис. 48):

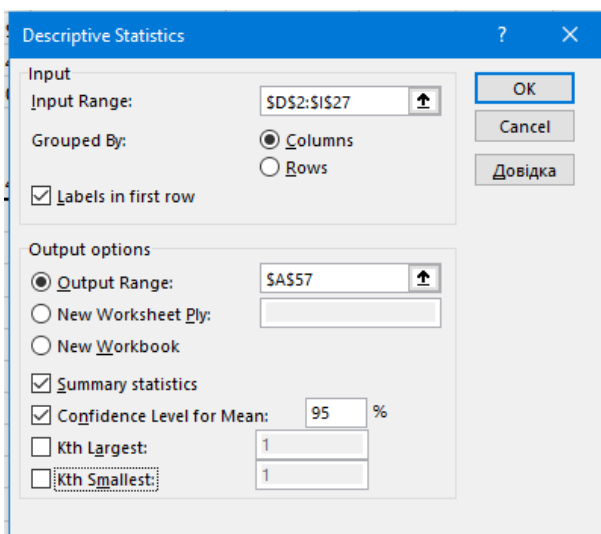


Рис. 48. Налаштування параметрів процедури Microsoft Excel «Аналіз даних → Описова статистика»

Вхідний і вихідний інтервал розташовані на аркуші «Первинні дані».

3. Після закриття вікна процедури на аркуші «Первинні дані», починаючи з комірки A57, з'явиться замовлена інформація.

4. Для зручності перегляду виконати такі процедури: 1) на вкладці «Основне» включити опцію «Вирівнювання → Перенесення тексту»; 2) для діапазону A57:L72 встановити формат «Число, 2 десяткових знаки».

Зразок результату представлений на рис. 49.

	A	B	C	D	E	F	G	H	I	J	K	L
57	Інтенсивний показник на 100 000, все населення		Інтенсивний показник на 100 000, доросле населення		Інтенсивний показник на 100 000, діти 0-17 років		Інтенсивний показник на 100 000, діти 5-9 років		Інтенсивний показник на 100 000, діти 10-14 років		Інтенсивний показник на 100 000, діти 15-17 років	
58												
59	Mean	15093,77805	Mean	6309,889994	Mean	53756,74595	Mean	50622,71592	Mean	45026,5303	Mean	48092,28737
60	Standard Error	990,6551331	Standard Error	475,5336601	Standard Error	3336,666794	Standard Error	3303,064698	Standard Error	2586,805823	Standard Error	3275,132473
61	Median	16091,43823	Median	6646,653846	Median	54874,6654	Median	52057,24849	Median	47110,91212	Median	51430,56293
62	Mode	#Н/Д	Mode	#Н/Д	Mode	#Н/Д	Mode	#Н/Д	Mode	#Н/Д	Mode	#Н/Д
63	Standard Deviation	4953,275665	Standard Deviation	2377,6683	Standard Deviation	16683,33397	Standard Deviation	16515,32349	Standard Deviation	12934,02912	Standard Deviation	16375,66236
64	Sample Variance	24534939,82	Sample Variance	5653306,546	Sample Variance	278333632,3	Sample Variance	272755910	Sample Variance	167289109,2	Sample Variance	268162317,8
65	Kurtosis	1,309800693	Kurtosis	2,192237814	Kurtosis	-0,054122186	Kurtosis	-0,515475296	Kurtosis	0,180872438	Kurtosis	0,268265917
66	Skewness	-0,433365372	Skewness	0,151818337	Skewness	-0,439388694	Skewness	-0,415890984	Skewness	-0,609390053	Skewness	-0,152793227
67	Range	22839,65878	Range	11797,1706	Range	69257,7157	Range	61949,61761	Range	53705,63736	Range	72108,36926
68	Minimum	4074,304867	Minimum	1188,590048	Minimum	16191,32784	Minimum	16942,66651	Minimum	16291,35122	Minimum	14244,35155
69	Maximum	26913,96364	Maximum	12985,76065	Maximum	85449,04353	Maximum	78892,28412	Maximum	69996,98858	Maximum	86352,72081
70	Sum	377344,4512	Sum	157747,2499	Sum	1343918,649	Sum	1265567,898	Sum	1125663,258	Sum	1202307,184
71	Count	25	Count	25	Count	25	Count	25	Count	25	Count	25
72	Confidence Level(95%)	2044,611704	Confidence Level(95%)	981,453237	Confidence Level(95%)	6886,541796	Confidence Level(95%)	6817,19048	Confidence Level(95%)	5338,904818	Confidence Level(95%)	6759,541199

Рис. 49. Результат виконання процедури «Аналіз даних → Описова статистика»

Висновки:

1. Мода для всіх показників не визначена, тобто немає повторюваних значень.
2. Відносна різниця між середнім значенням та медіаною незначна – 3,2 %.
3. Відносна різниця між стандартним відхиленням та 1/6 розмаху на рівні 0,17–0,37.
4. Відносна різниця між стандартним відхиленням та 0,75 міжквартильного розмаху на рівні 0,01–0,37.
5. Відносна похибка середнього значення досить мала.
6. Абсолютне значення асиметрії на рівні 0,15–0,61, менше за 2.

7. Абсолютне значення ексцесу на рівні 0,05–2,19, майже у всіх показників менше за 2, виняток – інтенсивний показник на 100 000, доросле населення.

8. Коефіцієнт варіації на рівні 0,28–0,32, що свідчить про значну неоднорідність даних.

9. Можна висувати гіпотези щодо нормальності показників.

Вправа 2. Для епідеміологічних показників у стовпцях 4–9 побудувати гістограми частот.

Варіант 1. Для побудови використати засіб Microsoft Excel «Аналіз даних → Гістограма» (процедура є в усіх версіях).

Алгоритм виконання процедури «Гістограма»:

1. Для кожного із стовпців 4–9 виконати команду «Дані → Аналіз даних → Гістограма» (Data Analysis → Histogram).

2. У вікні процедури послідовно для кожного стовпця встановити параметри «Первинних даних» (рис. 50).

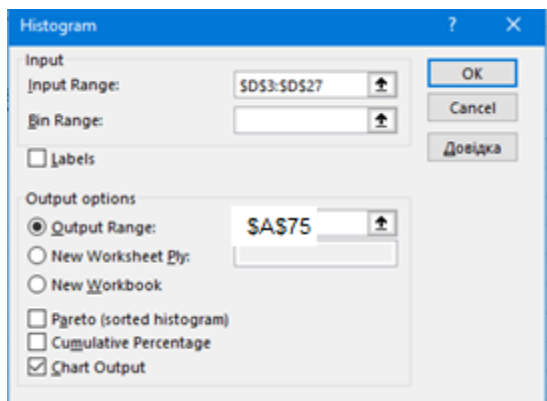


Рис. 50. Налаштування процедури Microsoft Excel «Аналіз даних → Гістограма»

3. Після закриття вікна процедури на аркуші «Первинні дані» з'явиться інформація для показника «Інтенсивний показник на 100 000, все населення».

Дані частот для показника, що побудовано за допомогою процедури Microsoft Excel «Аналіз даних → Гістограма» (рис. 51):

Карман	Частота	Тлумачення
4074,304867	1	< Кількість значень менших за 4074,304867 дорівнює 1
8642,236622	2	< Кількість значень від 4074,304867 до 8642,236622 дорівнює 2
13210,16838	2	...
17778,10013	14	...
22346,03189	5	...
Ше	1	< Кількість значень більших за 22346,03189 дорівнює 1

Рис. 51. Таблиця частот, побудована за допомогою процедури Гістограма

Таким чином для вхідних інтервалів D3:D27, E3:E27, F3:F27, G3:G27, H3:H27, I3:I27 можна отримати (рис. 52):

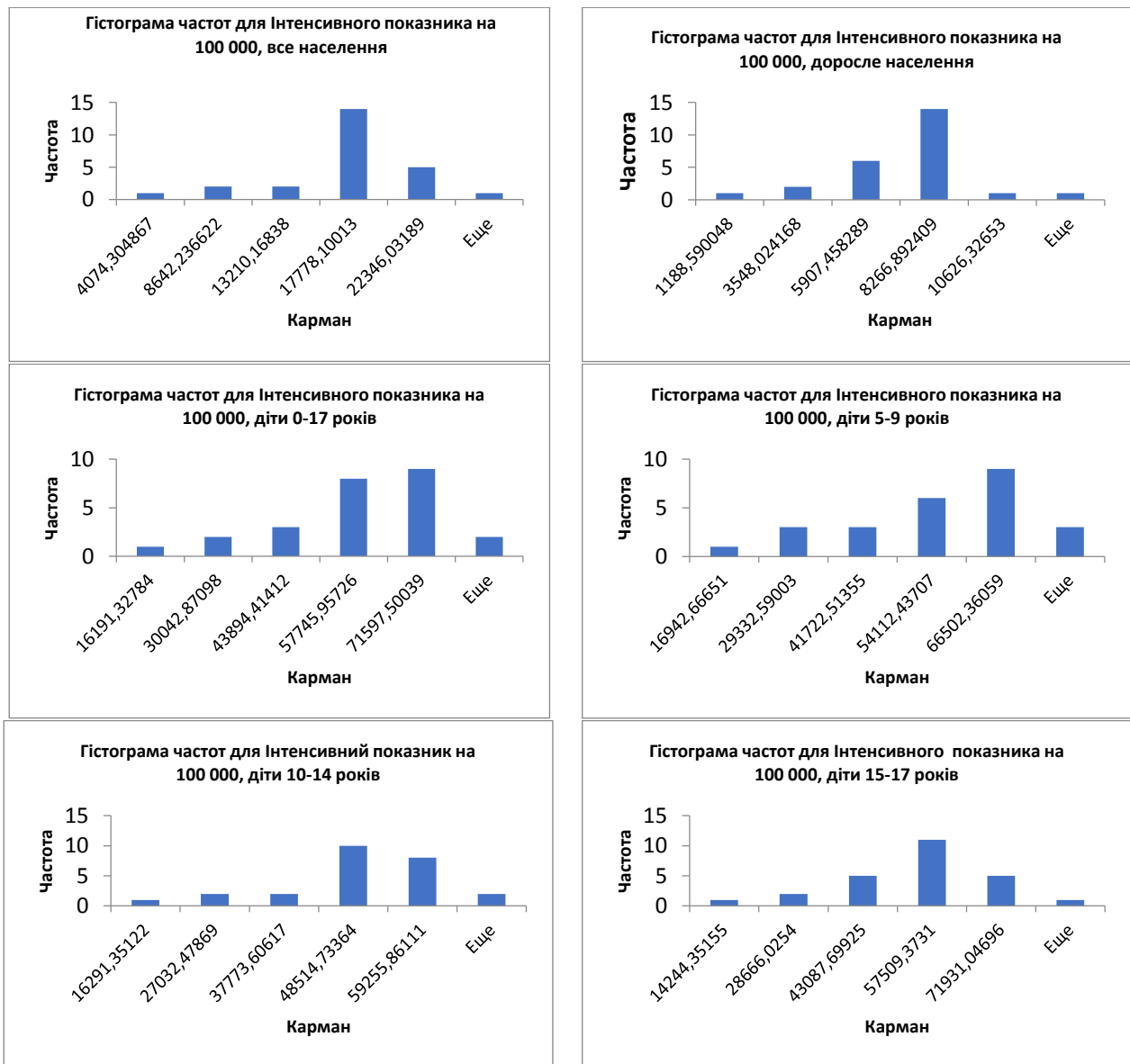


Рис. 52. Графіки гістограм частот для показників 4–9 (різних вікових груп), побудовані процедурою Microsoft Excel «Аналіз даних → Гістограма»

Висновок. Розподіл показників різних вікових груп близький до дзвоноподібного нормального.

Варіант 2. У версії Microsoft Excel 2016 і вище серед діаграм, що будуються за допомогою майстра, з'явився розділ «Статистичні діаграми», до якого входять «Гістограма» та «Діаграма розмаху» (рис. 53).

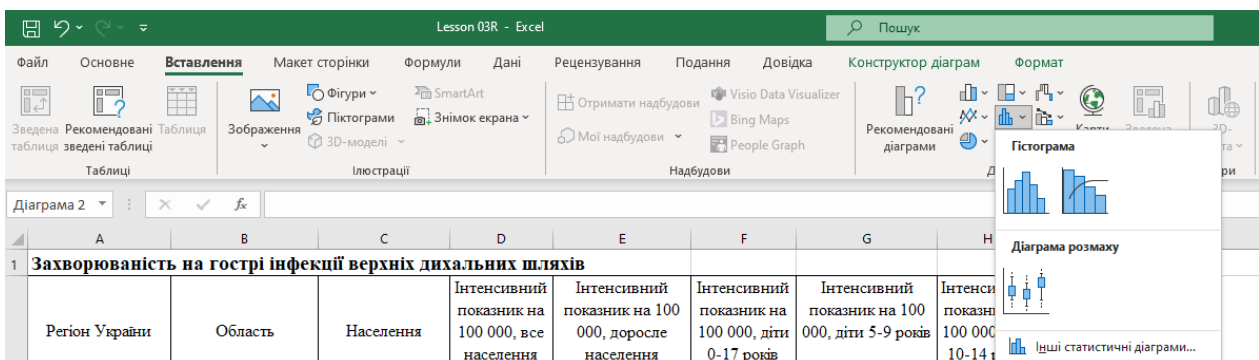


Рис. 53. Команди вкладки «Вставка» меню табличного процесора Microsoft Excel

На *рис. 54* представлена «Гістограма» для даних з діапазону D2:D27, побудована за допомогою цього засобу, яка виглядає таким чином:

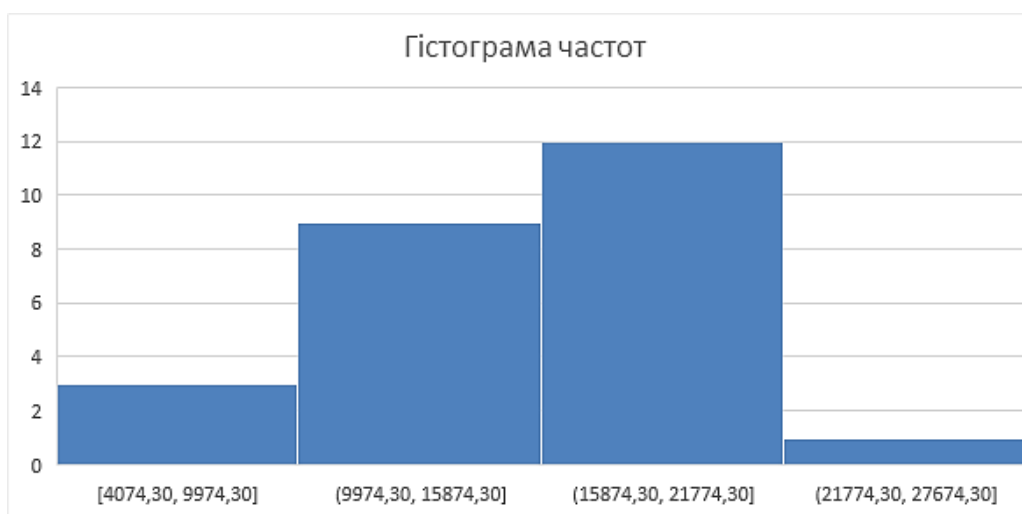


Рис. 54. Гістограма частот для показника 4 (захворюваність дорослого населення), яка побудована за допомогою команди «Вставлення → Статистичні діаграми → Гістограма»

Ця гістограма дещо відрізняється від тієї, що побудована за допомогою процедури «Аналіз даних → Гістограма», оскільки область значень поділена на меншу кількість інтервалів.

На *рис. 55* представлена діаграма розмаху для даних з діапазону D2:D27, яка виглядає таким чином:

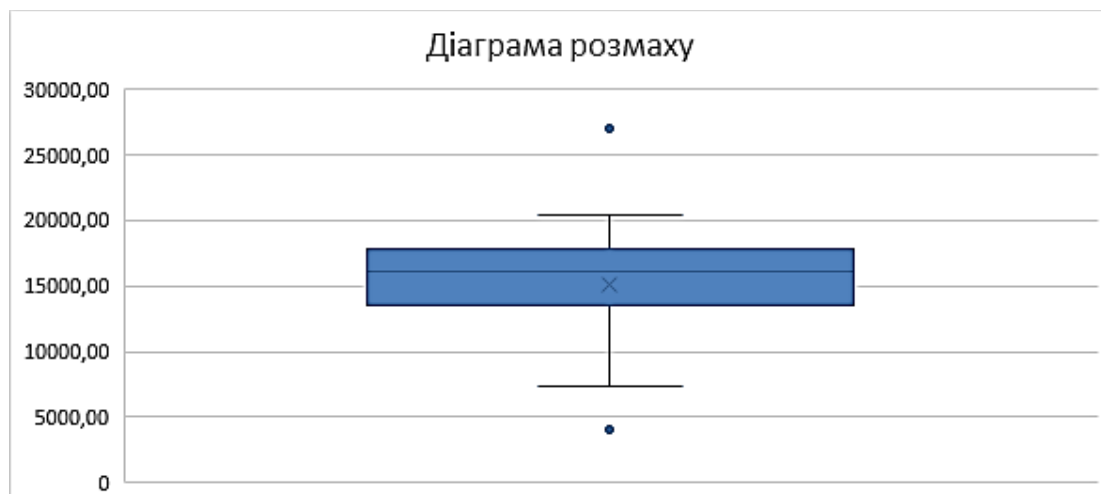


Рис. 55. Діаграма розмаху для показника 4 (захворюваність дорослого населення), яка побудована за допомогою команди «Вставлення → Статистичні діаграми → Діаграма розмаху»

Діаграма розмаху (Box Plot) – це графік, який дозволяє стисло зобразити основні параметри розподілу ймовірностей. Такий вид діаграми в зручній формі показує медіану, нижній та верхній квартилі, мінімальне та максимальне значення вибірки у діапазоні 1,5 міжквартильного розмаху, а також аутлаєри (викиди). Відстань між різними частинами ящика дозволяє визначити ступінь мінливості й асиметрію даних. Висота прямокутника відповідає ширині міжквартильного розмаху.

З діаграми видно, що серед значень показника є викиди – значення, які різко відрізняються від інших значень, але їх небагато.

Висновки. Аналіз гістограми частот і діаграми розмаху не спростовує нормальність розподілу показників, що досліджуються.

Вправа 3. Для даних у стовпцях 4–9 побудувати графіки нормальної ймовірності. Оскільки процедура побудови такого графіка потребує додаткових допоміжних дій, кожний графік слід будувати на окремому аркуші книги Lesson 03R.xlsx.

Алгоритм виконання:

1. Додати новий аркуш до книги Lesson 03R.xlsx, назвати його «Показник 4».
 2. Діапазон комірок D3:D27 з аркушу «Первинні дані» скопіювати до діапазону A1:A25 аркушу «Показник 4».
 3. Дані з A1:A25 округлити до цілих. Результат помістити в комірки B1:B25. Для цього в комірку B1 ввести формулу =ROUND(A1;0), скопіювати формулу з B1 до B2:B25.
 4. Скопіювати значення комірок B1:B25 до C1:C25 і відсортувати за зростанням.
 5. До комірок D1:D25 ввести числа 1, 2,...25.
 6. До комірки E1 ввести формулу =NORM.S.INV(D1/(25+1)) (обернене значення стандартного нормального розподілу для ймовірності D1/(25+1)).
 7. За допомогою команди «Вставлення → Діаграми → Точкова» побудувати графік нормальної ймовірності. Налаштування графіка: джерело даних – Ряд 1; значення ряду X =Аркуш1!\$C\$1:\$C\$25, значення ряду Y =Аркуш1!\$E\$1:\$E\$25
- Побудований графік має наступний вигляд (рис. 56):

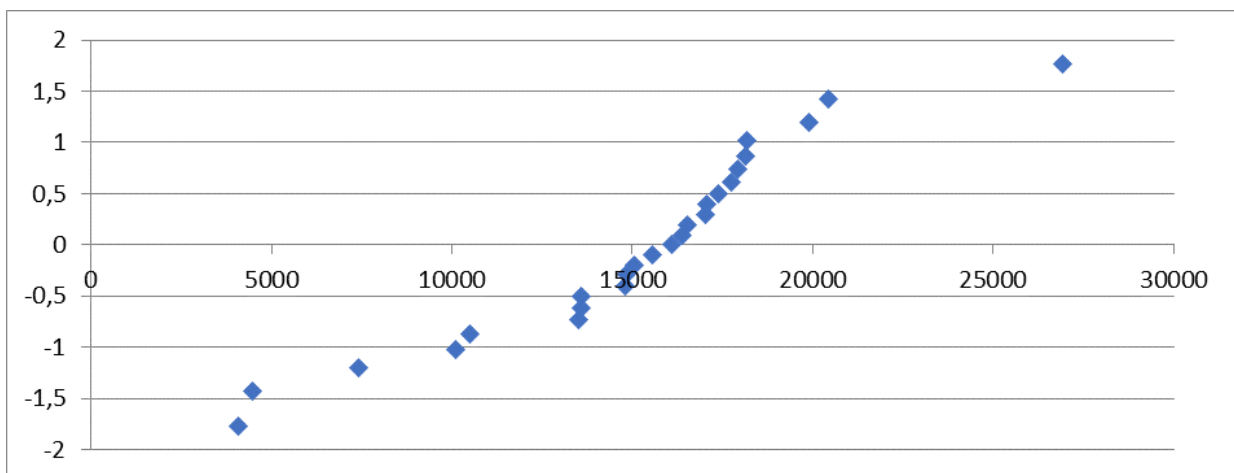


Рис. 56. Графік нормальної ймовірності (Normal Probability Plot) для показника 4

Якщо отримане зображення близьке до прямої лінії, можна стверджувати, що розподіл показника близький до нормального. Засіб дослідження нормальності розподілу показника за допомогою графіка нормальної ймовірності є у багатьох спеціалізованих статистичних програмах, таких як Statistica, SPSS та ін.

Висновок. Розподіл досліджуваного показника 4 близький до нормального.

Інколи в статистичних дослідженнях переходять від аналізу первинних кількісних даних до аналізу нормованих даних.

Нормовані дані – це дані, що отримані шляхом математичного перетворення первинних даних за формулою $z_i = \frac{x_i - M}{\sigma}$. Розподіл нормованих даних ідентичний розподілу первинних даних, проте $M=0$, а $\sigma=1$, що значно скорочує статистичні обчислення.

Серед кількісних даних окремих клас складають дихотомічні дані – ті, що приймають два значення 0 або 1. Для їх аналізу розроблені спеціальні методи. Принципова відмінність між дихотомічною та неперервною змінною полягає в тому, що перша представляє категорії, а друга вимірює.

У разі категоріальних даних використовують так звані атрибутивні розподіли. У цьому випадку основними числовими характеристиками є абсолютна (число повторювань) і відносна (P %) частота (доля) значення показника у первинних даних. Доля показника аналогічна середньому значенню за сукупністю. Інколи вказується M_0 – значення, яке найбільш повторюється.

Вправа 4. Обчислити числові статистичні характеристики атрибутивних даних зі стовпця 1 (регіон України).

Алгоритм виконання:

1. Створити зведену таблицю на новому аркуші за інформацією з джерела даних, що розташоване на аркуші «Первинні дані».

2. Макет зведеної таблиці налаштувати таким чином: в області рядків розташувати поле «регіон України», в області підсумкових значень – 2 рази поле «область», для 2-го поля «область» змінити налаштування параметрів значення (рис. 57).

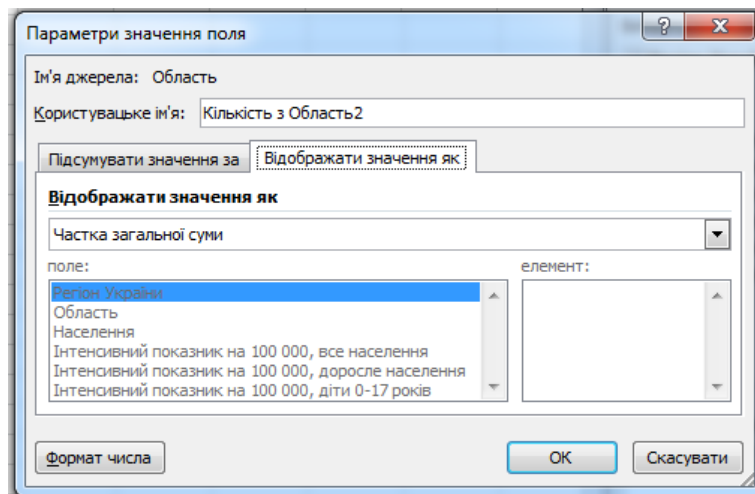


Рис. 57. Налаштування параметрів підсумкового поля зведеної таблиці

3. Змінити текст в комірках-заголовках зведеної таблиці.

На рис. 58 представлено зразок зведеної таблиці.

Регіон	кількість областей	частка, %
Західний	8	32,00%
Південний	3	12,00%
Північний	4	16,00%
Східний	5	20,00%
Центральний	5	20,00%
Загальний підсумок	25	100,00%

Рис. 58. Зведена таблиця з додатковим обчисленням відсотків за стовпцем

Висновки. Групування областей за регіонами нерівномірне: мінімальна кількість областей в регіоні – 3, максимальна – 8.

Вправа 5. Додати до таблиці дихотомічну змінну «рівень захворюваності», яку визначити за правилом:

Рівень захворюваності =

- {0 Інтенсивний показник на 100 000, все населення < середнього значення
- {1 Інтенсивний показник на 100 000, все населення \geq середнього значення

Обчислити числові характеристики дихотомічної змінної.

Алгоритм виконання:

1. Обчислити рівень захворюваності для кожної області. Для цього в комірку J3 ввести формулу =IF(D3<D\$38;0;1). (У комірці D\$38 знаходиться середнє значення інтенсивного показника на 100 000, все населення.) Скопіювати формулу на діапазон J4^J27.

2. Обчислити кількість областей з рівнем захворюваності =0. Для цього в комірку J28 ввести формулу =COUNTIF(J3:J27;0). У комірці K28 обчислити відносну частоту =J28/25 та встановити відсотковий формат числа в комірці K28.

3. Обчислити кількість областей з рівнем захворюваності =1. Для цього в комірку J29 ввести формулу =COUNTIF(J3:J27;1). У комірці K29 обчислити відносну частоту =J29/25 та встановити відсотковий формат числа в комірці K29.

4. Обчислити моду показника «Рівень захворюваності». Для цього у комірку J30 ввести формулу =MODE.SNGL(J3:J27).

Результат обчислень представлено у *табл. 5*.

Таблиця 5

Числові характеристики дихотомічної змінної

Рівень захворюваності	Кількість областей	Відносна частота
0	11	44,00 %
1	14	56,00 %

Висновок. Оскільки M_o показника «Рівень захворюваності» дорівнює 1, можна припустити, що в даних, які досліджуються, переважає підвищений рівень захворюваності.

3.1.4. Практичне заняття 4.

Перевірка гіпотез щодо нормального розподілу у вибірці.

Статистичні методи порівняння середніх значень

Тема: Інференційна статистика епідеміологічних даних. Статистичні гіпотези. Перевірка гіпотез щодо нормального розподілу у вибірці. Статистичні методи порівняння середніх значень.

Мета: удосконалення навичок з автоматизації розрахунків статистичних критеріїв за даними вибірки; перевірка гіпотез про нормальність розподілу та рівність середніх значень показників у двох або кількох нормальних вибірках.

Теоретичні відомості

Інференційна статистика (англ. inferencial statastics) – це галузь статистики, яка використовує дані вибірки для отримання висновків, прогнозів або узагальнень на дані генеральних сукупностей або популяцій. Одним із основних методів інференційної статистики є перевірка статистичних гіпотез.

Статистична гіпотеза – це деяке твердження щодо властивостей досліджуваних даних. Гіпотезу називають простою, якщо вона містить лише одне припущення, складною, якщо вона складається із скінченної чи нескінченної кількості простих гіпотез.

Перевірка гіпотез – формулювання гіпотез та використання вибірових даних для перевірки того, чи підтримуються або спростовуються ці гіпотези.

Статистичний тест, або статистичний критерій – суворе математичне правило, за яким приймається або відкидається статистична гіпотеза із заданим рівнем значущості. Статистичні критерії використовуються в інференційній статистиці для встановлення закону розподілу, порівняння груп, оцінки взаємозв'язків та здійснення прогнозів.

Статистичні критерії можна поділити на групи за призначенням:

- Критерії згоди. Використовуються для перевірки гіпотези про те, що досліджувана випадкова величина підпорядковується певному закону розподілу.
- Критерії на однорідність. Використовуються для перевірки гіпотези взаємної відповідності законів розподілу у двох або кількох вибірках.
- Критерії щодо рівності або відмінності центральної тенденції у двох або кількох вибірках.
- Критерії щодо рівності або відмінності дисперсії у двох або кількох вибірках.
- Критерії щодо значущості кореляційного зв'язка.
- Критерії щодо значущості коефіцієнтів регресії та ін.

Статистичні критерії поділяються на параметричні (які включають до розрахунків параметри ймовірнісного розподілу даних у вибірках – середні значення та дисперсії і потребують нормального закону розподілу даних у вибірках) й непараметричні (які засновані на оперуванні частотами або рангами).

Статистичні критерії можна класифікувати ще за такими ознаками:

- за типом даних (кількісні, дихотомічні, текстові);
- за обсягом вибірки (для великих або малих вибірок);
- за кількістю вибірок (дві або більше);
- за якістю вибірок (пов'язані або незалежні);
- за потужністю критерію.

Методика перевірки статистичних гіпотез

При перевірці статистичних гіпотез висувається нульова гіпотеза H_0 , зазвичай це твердження щодо відсутності ефекту (відмінності, зв'язку ..., тощо), та альтернативна гіпотеза H_1 – твердження щодо наявності ефекту (відмінності, зв'язку ..., тощо), яка приймається у випадку відхилення нульової гіпотези. Для перевірки статистичної гіпотези використовують статистичний критерій, що відповідає досліджуваній проблемі. Статистичний критерій – це неперервна випадкова величина, яка обчислюється за вибіркою і приймає різні дійсні значення.

Для перевірки гіпотези простір значень критерію поділяють на дві області, що не перетинаються: критичну та область прийняття. Критичною областю називають сукупність значень критерію, за яких нульову гіпотезу слід відхилити. Областю прийняття гіпотези (областю допустимих значень) називають сукупність значень критерію, за яких нульову гіпотезу приймають.

Перевірка гіпотези передбачає: розрахунок значення критерію, визначення меж критичної області і перевірку потрапляння критерію до критичної області. Вирізняють двобічні й однобічні (лівобічні, правобічні) критичні області. Їх використання залежить від вибору альтернативної гіпотези. Нехай K – критерій, обчислений за вибіркою. Двобічна критична область задається двома критичними точками $K_{кр}^1$ і $K_{кр}^2$, гіпотеза H_0 відхиляється, якщо $K < K_{кр}^1$ або $K > K_{кр}^2$. Лівобічна критична область задається однією критичною точкою $K_{кр}$, гіпотеза H_0 відхиляється, якщо $K < K_{кр}$. Правобічна критична область задається однією критичною точкою $K_{кр}$, гіпотеза H_0 відхиляється, якщо $K > K_{кр}$.

Оскільки дані вибірки випадкові, під час перевірки статистичної гіпотези можливі помилки. Є два роди помилок. Помилка першого роду полягає у тому, що відкидають нульову гіпотезу, тоді як насправді ця гіпотеза правильна. Ймовірність припуститися такої помилки називають «рівнем значущості» та позначають буквою α (альфа). Помилка другого роду полягає у тому, що приймають нульову гіпотезу, тоді як насправді ця гіпотеза хибна. Ймовірність зробити цю помилку позначають буквою β (бета). Значення $1-\beta$ називають «потужність критерію» – це ймовірність заперечення неправильної гіпотези.

На практиці зазвичай задають рівень значущості α , в епідеміології найчастіше вибирають значення $\alpha=0,05$. Від рівня значущості залежать значення меж критичної області. Найменше значення рівня значущості, для якого відхиляється основна нульова гіпотеза, називається «р-значенням (p-value), або досягнутим рівнем значущості».

Перевірка відповідності заданому закону розподілу проводиться за критеріями згоди. Умовно ці критерії можна поділити на два класи: критерії перевірки на відповідність нормальному закону розподілу та універсальні критерії, що перевіряють дані на відповідність будь-якому закону розподілу.

Існує понад 20 статистичних критеріїв щодо перевірки нормальності даних. Найбільш відомі: критерій Пірсона χ^2 (вимоги – обсяг вибірки > 30), критерій Колмогорова-Смирнова (вимоги – обсяг вибірки > 30), критерій Шапіро-Уїлка, (вимоги – обсяг вибірки > 3), критерій Ліллієфорса (вимоги – обсяг вибірки > 30). Ці тести передбачають значний обсяг обчислень і вбудовані в потужні статистичні пакети, такі як SPSS, Statistica, StatGraphics, EpiInfo та ін. У програмі Microsoft Excel вбудованих засобів для перевірки нормальності за цими критеріями немає, але існує багато посібників зі статистичних розрахунків за допомогою табличного процесора, де детально розглядаються такі приклади.

Найпростішими критеріями перевірки нормальності розподілу можна вважати такі: 1) простий тест «Девіда-Хартлі-Пірсона», або RS-тест, який полягає в обчисленні RS-критерію (нормалізованого розмаху) за формулою $RS = \frac{R}{\sigma}$ та порівнянні його з критичними значеннями з таблиці (додаток 1); 2) простий тест М. О. Плохінського щодо значень коефіцієнтів асиметрії та ексцесу нормального розподілу, який полягає в обчисленні помилок репрезентативності для

асиметрії та ексцесу за формулами: $m_{As} = \sqrt{\frac{6}{n+3}}$, $m_E = 2\sqrt{\frac{6}{n+3}}$, для нормального розподілу виконуються співвідношення $\frac{|As|}{m_{As}} \leq 3$ та $\frac{|E|}{m_E} \leq 3$;

3) простий тест Є. І. Пустильника щодо значень коефіцієнтів асиметрії та ексцесу нормального розподілу, який полягає в обчисленні критичних значень для асиметрії та ексцесу за формулами: $As_{\text{крит}} = \sqrt[3]{\frac{6(n-1)}{(n+1)(n+3)}}$, $E_{\text{крит}} = \sqrt[5]{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}$ та порівнянні їх з обчисленими за вибіркою; для нормального розподілу $As < As_{\text{крит}}$ та $E < E_{\text{крит}}$.

Задача-еталон

Дано: Інформація про випадки захворювань на гострі інфекції дихальних шляхів на території України у 2019 р. надана в таблиці Excel – файл Lesson 04D.xlsx, аркуш «Первинні дані».

Вправа 1. Перевірити нормальність розподілу кількісних даних у стовпцях 4–9 за RS-критерієм.

Алгоритм виконання:

1. Відкрити файл Lesson 04D.xlsx. Зберегти як Lesson 04R.xlsx.
 2. Для обчислення розмаху показника 4 в комірку D30 ввести формулу =MAX(D3:D27)-MIN(D3:D27).
 3. Для обчислення стандартного відхилення показника 4 в комірку D31 ввести формулу =STDEV.S(D3:D27).
 4. Для обчислення RS-критерію для показника 4 в комірку D32 ввести формулу =D30/D31.
 5. Скопіювати формули з комірок D30:D32 до комірок E30:I32.
 6. Скопіювати значення з комірок D32:I32 до комірок C39:C44 (для цього скористатися засобом «Спеціальне вставлення → Вставити значення → Операція транспонувати»).
 7. Для перевірки потрапляння обчисленого RS-критерію до критичної області в комірку F39 ввести формулу =AND(C39>D39;C39<E39). (Функція AND перевіряє одночасне виконання двох умов: $C39 > D39$ і $C39 < E39$. Результатом обчислення функції AND буде логічна константа «істина (true)» або «хиба (false)».)
 8. Скопіювати формулу з F39 до діапазону F40:F44
- Зразок результату представлений на *рис. 59*.

	A	B	C	D	E	F
37	RS - тест		$\alpha = 0,05$		$n = 25$	
38	№ стовпця	Показник	Rs	U_1	U_2	Rs - тест
39	4	Інтенсивний показник на 100 000, все населення	4,611021134	3,34	4,71	TRUE
40	5	Інтенсивний показник на 100 000, доросле населення	4,961655333	3,34	4,71	FALSE
41	6	Інтенсивний показник на 100 000, діти 0-17 років	4,151311472	3,34	4,71	TRUE
42	7	Інтенсивний показник на 100 000, діти 5-9 років	3,751038703	3,34	4,71	TRUE
43	8	Інтенсивний показник на 100 000, діти 10-14 років	4,152274351	3,34	4,71	TRUE
44	9	Інтенсивний показник на 100 000, діти 15-17 років	4,40338642	3,34	4,71	TRUE

Рис. 59. Результати перевірки RS-критерію для показників 4–9

Висновки. Підтверджена нормальність даних у всіх стовпцях, окрім показника захворюваності серед дорослого населення (5).

Вправа 2. Перевірити нормальність розподілу кількісних даних у стовпцях 4–9 за критерієм Плохінського.

Алгоритм виконання:

1. Для обчислення коефіцієнта асиметрії показника 4 в комірку D33 ввести формулу =SKEW(D3:D27).

2. Для обчислення ексцесу показника 4 в комірку D34 ввести формулу =KURT(D3:D27).

3. Скопіювати формули з комірок D33: D34 до комірок E33:I34.

4. Скопіювати значення з комірок D33: I34 до комірок G39:H44 (для цього скористатися засобом «Спеціальне вставлення → Вставити значення → Операція транспонувати»).

5. Згідно з кількістю даних у вибірці (значення комірки E37) обчислити критичне значення критерію Плохінського для асиметрії. Для цього до комірки C50 ввести формулу =SQRT(6/(E\$37+3)), а до комірки C51 формулу =3*C50.

6. Згідно з кількістю даних у вибірці (значення комірки E37) обчислити критичне значення критерію Плохінського для ексцесу. Для цього до комірки E50 ввести формулу =2*C50, а до комірки E51 формулу =3*E50.

Результати обчислень представлені у *табл. 6*.

Таблиця 6

Критичне значення критерію Плохінського для ексцесу та асиметрії

m_{As}	0,46291005	m_E	0,9258201
$3m_{As}$	1,38873015	$3m_E$	2,777460299

7. Для перевірки потрапляння асиметрії та ексцесу показника 4 до критичної області в комірку I39 ввести формулу =AND(G39<=C\$51;H39<=E\$51).

8. Скопіювати формулу з I39 до діапазону I40:I44.

Зразок результату представлено на *рис. 60*.

	A	B	G	H	I
38	№ стовпця	Показник	As	E	Тест Плохінського
39	4	Інтенсивний показник на 100 000, все населення	-0,433365372	1,309800693	TRUE
40	5	Інтенсивний показник на 100 000, доросле населення	0,151818337	2,192237814	TRUE
41	6	Інтенсивний показник на 100 000, діти 0-17 років	-0,439388694	-0,05412219	TRUE
42	7	Інтенсивний показник на 100 000, діти 5-9 років	-0,415890984	-0,5154753	TRUE
43	8	Інтенсивний показник на 100 000, діти 10-14 років	-0,609930053	0,180872438	TRUE
44	9	Інтенсивний показник на 100 000, діти 15-17 років	-0,152793227	0,268265917	TRUE

Рис. 60. Результати перевірки критерію Плохінського для показників 4–9

Висновки. Підтверджена нормальність даних у всіх стовпцях (4–9), зокрема показників захворюваності у всіх вікових групах, які досліджувалися, та загалом серед всього населення

Вправа 3. Перевірити нормальність розподілу кількісних даних у стовпцях 4–9 за критерієм Пустильника.

Алгоритм виконання:

1. Згідно з кількістю даних у вибірці (значення комірки E37) обчислити критичне значення критерію Пустильника для асиметрії. Для цього до комірки C57 ввести формулу $= (6 * (E37 - 1) / ((E37 + 1) * (E37 + 3)))^{(1/3)}$.

2. Згідно з кількістю даних у вибірці (значення комірки E37) обчислити критичне значення критерію Пустильника для ексцесу. Для цього до комірки E57 ввести формулу $= (24 * E37 * (E37 - 2) * (E37 - 3) / ((E37 + 1)^2 * (E37 + 3) * (E37 + 5)))^{(1/5)}$.

Результати обчислень:

As крит	0,582653515	E крит	13,04147975
---------	-------------	--------	-------------

Для перевірки потрапляння асиметрії та ексцесу показника 4 до критичної області в комірку J39 ввести формулу $= \text{AND}(G39 < \$C\$57; H39 < \$E\$57)$. Скопіювати формулу з J39 до діапазону J40:J44.

Зразок результату представлено на *рис. 61*.

	A	B	C	D	E	F	G	H	I	J
38	№ стовпця	Показник	Rs	U ₁	U ₂	Rs - тест	As	E	Тест Плохінського	Тест Пустильника
39	4	Інтенсивний показник на 100 000, все населення	4,611021134	3,34	4,71	TRUE	-0,433365372	1,309800693	TRUE	TRUE
40	5	Інтенсивний показник на 100 000, доросле населення	4,961655333	3,34	4,71	FALSE	0,151818337	2,192237814	TRUE	TRUE
41	6	Інтенсивний показник на 100 000, діти 0-17 років	4,151311472	3,34	4,71	TRUE	-0,439388694	-0,05412219	TRUE	TRUE
42	7	Інтенсивний показник на 100 000, діти 5-9 років	3,751038703	3,34	4,71	TRUE	-0,415890984	-0,5154753	TRUE	TRUE
43	8	Інтенсивний показник на 100 000, діти 10-14 років	4,152274351	3,34	4,71	TRUE	-0,609930053	0,180872438	TRUE	TRUE
44	9	Інтенсивний показник на 100 000, діти 15-17 років	4,40338642	3,34	4,71	TRUE	-0,152793227	0,268265917	TRUE	TRUE

Рис. 61. Результати перевірки критерію Пустильника для показників 4–9

Висновки. Підтверджена нормальність даних у всіх стовпцях (4–9), зокрема показників захворюваності у всіх вікових групах, які досліджувалися, та загалом серед всього населення.

Теоретичні відомості

t-критерій Стюдента – загальна назва статистичних критеріїв, що ґрунтуються на порівнянні даних з розподілом Стюдента. Найчастіше застосовується з метою перевірки рівності середніх значень у двох вибірках.

Нехай $x_{n1}^1, \dots, x_{n1}^1$ – значення кількісного показника в 1-й вибірці, $x_{n2}^2, \dots, x_{n2}^2$ – значення кількісного показника у 2-й вибірці.

Умови застосування t-критерію:

- достатня кількість значень у вибірках: $n1, n2 > 16$;
- значення у вибірках розподілені за нормальним законом.

На *рис. 62* представлена система модифікацій t-критерію Стюдента.



Рис. 62. Класифікація системи модифікацій t-критерію Стьюдента

Алгоритм використання t-критерію Стьюдента:

1. Перевірка даних в обох вибірках на відповідність закону нормального розподілу.

2. Визначення типу вибірок – незалежні або залежні (зв'язані):

- якщо вибірки незалежні, перевірка гіпотези про рівність дисперсій за критерієм F-Фішера;
- якщо вибірки залежні, перевірка наявності кореляційного зв'язку між вибірками.

3. Обчислення емпіричного t-критерію Стьюдента.

Обчислення (математичний вираз) критерію залежить від:

- вибірки – залежні чи незалежні;
- дисперсії – гомогенні чи гетерогенні;
- однаковий чи різний обсяг вибірок.

Результат обчислення позначається як $t_{\text{емпіричне}} (t_{\text{емп.}})$.

4. Вибір рівня значущості α , визначення числа ступенів свободи, знаходження критичного табличного значення $t_{\text{критичне}} (t_{\text{крит.}})$.

5. Формулювання висновку:

- якщо $|t_{\text{емп.}}| < t_{\text{крит.}}$ – приймаємо гіпотезу H_0 і вважаємо, що відмінності між вибірками статистично незначущі;
- якщо $|t_{\text{емп.}}| \geq t_{\text{крит.}}$ – приймаємо гіпотезу H_1 і вважаємо, що відмінності між вибірками статистично значущі.

Вправа 4. Порівняти середні значення епідеміологічних показників у стовпцях 5 та 6. Застосувати засіб Microsoft Excel «Аналіз даних».

Алгоритм виконання:

- Перевірка вибірок на нормальність була проведена у попередніх вправах.
- Вибірки незалежні. Кількість значень у вибірках однакова і дорівнює 25.
- Обчислити дисперсії даних у стовпцях 5 і 6 і з'ясувати, яка з них більша:

- a) у комірку E28 аркуша «Первинні дані» ввести формулу =VAR.S(E3:E27), результат 5653306,55;
- b) у комірку F28 аркуша «Первинні дані» ввести формулу =VAR.S(F3:F27), результат 278333632,32;
- c) у комірку G28 аркуша «Первинні дані» ввести формулу =MAX(E28:F28), результат 278333632,32.

4. Для перевірки гомогенності дисперсій виконати команду «Дані → Аналіз даних → Двовибірковий тест для дисперсії» (Data Analysis → F-Test Two-Sample For Variances).

5. У вікні процедури встановити наступні параметри (рис. 63):

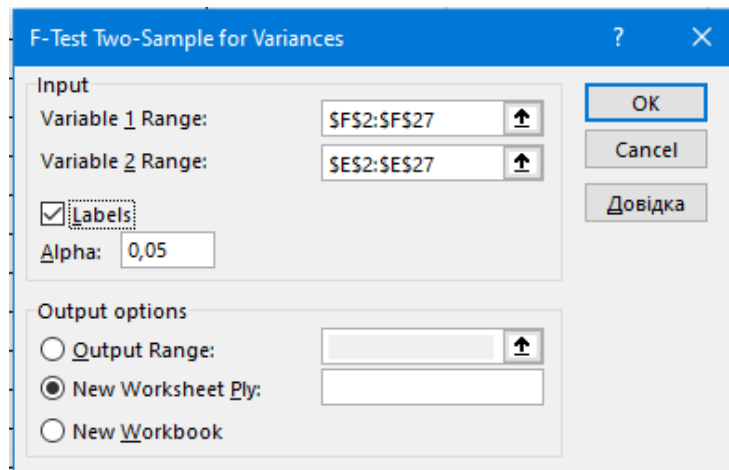


Рис. 63. Налаштування параметрів процедури Microsoft Excel «Аналіз даних → Двовибірковий тест для дисперсії»

Інтервал змінної 1 – це інтервал з більшою дисперсією.

Рівень значущості $\alpha = 0,05$

Результати виконання процедури виводяться на новий аркуш книги Lesson 04R.xlsx.

6. Після закриття вікна процедури на новому аркуші з'явиться замовлена інформація (рис. 64):

	A	B	C
1	F-Test Two-Sample for Variances		
2			
3		<i>Інтенсивний показник на 100 000, діти 0-17 років</i>	<i>Інтенсивний показник на 100 000, доросле населення</i>
4	Mean	53756,74595	6309,889994
5	Variance	278333632,3	5653306,546
6	Observations	25	25
7	df	24	24
8	F	49,23377673	
9	P(F<=f) one-tail	4,27072E-15	
10	F Critical one-tail	1,983759568	
11			

Рис. 64. Результат виконання процедури «Двовибірковий тест для дисперсії»

Значення $F_{\text{емп.}} = 49,23$ (комірка B8).

Значення $F_{\text{крит.}} = 1,98$ (комірка B10).

Оскільки $F_{\text{емп.}} > F_{\text{крит.}}$, приймаємо рішення про підтвердження гіпотези про відмінність дисперсій на заданому α -рівні.

7. Для порівняння середніх значень вибірок виконати команду «Дані → Аналіз даних → Двовибірковий t-тест з різними дисперсіями»(Data Analysis → t-Test Two-Sample Assuming Unequal Variances).

8. У вікні процедури встановити наступні параметри (рис. 65):

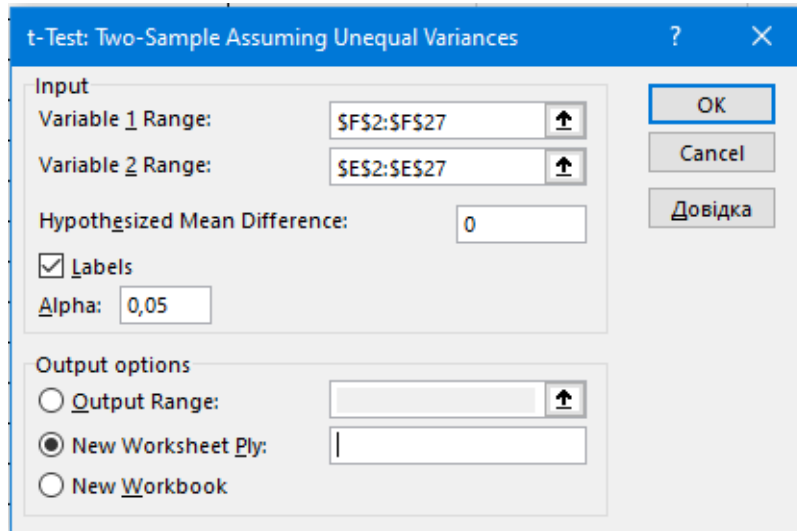


Рис. 65. Вікно налаштування процедури «Двовибірковий t-тест з різними дисперсіями»

Гіпотетична різниця між середніми значеннями – 0.

Рівень значущості $\alpha=0,05$.

Результати виконання процедури виводяться на новий аркуш книги Lesson 04R.xlsx.

9. Після закриття вікна процедури на новому аркуші з'явиться замовлена інформація (рис. 66):

	A	B	C
1	t-Test: Two-Sample Assuming Unequal Variances		
2			
3		Інтенсивний показник на 100 000, діти 0-17 років	Інтенсивний показник на 100 000, доросле населення
4	Mean	53756,74595	6309,889994
5	Variance	278333632,3	5653306,546
6	Observations	25	25
7	Hypothesized Mean Difference	0	
8	df	25	
9	t Stat	14,07758831	
10	P(T<=t) one-tail	1,09014E-13	
11	t Critical one-tail	1,708140761	
12	P(T<=t) two-tail	2,18029E-13	
13	t Critical two-tail	2,059538553	

Рис. 66. Результат виконання процедури «Двовибірковий t-тест з різними дисперсіями»

Значення $t_{\text{емп.}} = 14,08$ (комірка B9).

Значення $t_{\text{крит.}} = 2,06$ (комірка B13).

10. Оскільки $t_{\text{емп.}} > t_{\text{крит.}}$, приймаємо рішення про підтвердження гіпотези про відмінність середніх значень вибірок на заданому α -рівні.

Теоретичні відомості

У книзі Стентона Гланца «Медико-біологічна статистика», яка набула популярності на початку 2000-х років, наводиться найпростіший метод перевірки статистичної значущості різниці між середніми двох нормально розподілених вибірок. Якщо відома така інформація: M_1 – середнє значення показника 1-ї вибірки, m_1 – стандартна помилка середнього значення 1-ї вибірки, n_1 – обсяг 1-ї вибірки, M_2 – середнє значення показника 2-ї вибірки, m_2 – стандартна помилка середнього значення 2-ї вибірки, n_2 – обсяг 2-ї вибірки, то для встановлення статистичної значущості різниці між середніми треба:

- обчислити $t_{\text{емп.}} = \frac{\text{Різниця середніх}}{\text{Стандартна помилка різниці середніх}} = \frac{|M_1 - M_2|}{\sqrt{m_1^2 + m_2^2}}$;

- знайти табличне значення $t_{\text{крит.}}$ – квантиль розподілу Стьюдента з параметрами $\alpha = 0,05$, кількість ступенів свободи $df = n_1 + n_2 - 2$;

- порівняти $t_{\text{емп.}}$ і $t_{\text{крит.}}$.

Зробити висновки: якщо $t_{\text{емп.}} < t_{\text{крит.}}$, різниці немає, а як ні, різниця значуща.

Якщо для дослідження значущості різниці надається інформація про середні значення, що виражені в долях, використовується z-критерій, який полягає в наступному:

- обчислити $z_{\text{емп.}} = \frac{\text{Різниця долей}}{\text{Стандартна помилка різниці долей}} = \frac{|M_1 - M_2|}{\sqrt{m_1^2 + m_2^2}}$;

- знайти табличне значення $z_{\text{крит.}}$ – квантиль стандартного нормального розподілу з параметрами $\alpha = 0,05$;

- порівняти $z_{\text{емп.}}$ і $z_{\text{крит.}}$;

- зробити висновки: якщо $z_{\text{емп.}} < z_{\text{крит.}}$, різниці немає, а як ні, різниця значуща.

Вправа 5

Дано: Інформація про захворюваність на кишкові інфекції по двох областях надана на аркуші «Групові дані» файлу Lesson 04R.xlsx (рис. 67).

	А	В	С
1	Захворюваність на кишкові інфекції		
2		Область А	Область Б
3	Кількість районів	7	8
4	Середнє значення	2 643	3534
5	Помилка середнього значення	475,3	365,2
6			

Рис. 66. Дані для дослідження значущості різниці середніх

Завдання: перевірити значущість різниці середніх значень.

Алгоритм виконання:

1. Перейти на аркуш «Групові дані» файлу Lesson 04R.xlsx.
2. Обчислити $t_{\text{емп.}}$. Для цього в комірку B7 ввести формулу $=\text{ABS}(B4-C4)/\text{SQRT}(B5^2+C5^2)$. Результат 1,48648.
3. Знайти $t_{\text{крит.}}$. Для цього в комірку B8 ввести формулу $=\text{T.INV}(1-0,05/2;B3+C3-2)$. Результат 2,16036.
4. В комірці B9 обчислити логічний вираз $=B7>B8$. Результат FALSE (ХИБА).

Висновок. Оскільки $t_{\text{емп.}} < t_{\text{крит.}}$, можна зробити висновок, що різниці між середніми значеннями немає.

Вправа 6

Дано: Інформація про результати тестування лаборантів на носійство HBsAg по двом клініко-діагностичним лабораторіям надана на аркуші «Дані в долях» файлу Lesson 04R.xlsx (рис. 68).

	A	B	C
1	Тестування лікарів-лаборантів на носійство HBsAg		
2		Лабораторія 1	Лабораторія 2
3	Доля носіїв (%)	14,1	22,1
4	Помилка долі (%)	0,5	0,8
5			
6	z емп		
7	z крит		
8	Значимість різниці		

Рис. 68. Дані для дослідження значущості різниці доль

Завдання: перевірити значущість різниці середніх значень.

Алгоритм виконання.

1. Перейти на аркуш «Дані в долях» файлу Lesson 04R.xlsx.
2. Обчислити $z_{\text{емп.}}$. Для цього в комірку B6 ввести формулу $=\text{ABS}(B3-C3)/\text{SQRT}(B4^2+C4^2)$. Результат 8,48.
3. Знайти $z_{\text{крит.}}$. Для цього в комірку B7 ввести формулу $=\text{NORM.S.INV}(1-0,05/2)$. Результат 1,96.
4. В комірці B8 обчислити логічний вираз $=B6>B7$. Результат TRUE (ІСТИНА).

Висновок. Оскільки $z_{\text{емп.}} > z_{\text{крит.}}$, можна зробити висновок, що різниця між долями значуща.

Теоретичні відомості

Дисперсійний аналіз (ANOVA, від англ. Analysis of Variations) – це статистичний метод, який використовується для порівняння середніх значень декількох вибірок кількісних даних. Дисперсійний аналіз складається з кількох процедур, що призначені для перевірки гіпотез про зв'язок між певною ознакою та досліджуваними факторами, які не мають кількісного опису, а також для встановлення ступеня впливу факторів та їх взаємодії.

Факторами називають контрольовані чинники, що впливають на кінцевий результат. Рівнем фактора, або способом обробки, називають значення, що

характеризують конкретний прояв цього фактора. Ці значення зазвичай подають у номінальній або порядковій шкалі вимірювань. Значення вимірюваної ознаки називають «відгуком».

При однофакторному дисперсійному аналізі первинні дані подають у вигляді таблиць, у яких кількість стовпчиків дорівнює кількості рівнів фактора, а кількість значень у кожному стовпчику – кількості спостережень при відповідному рівні фактора. Завданням аналізу є перевірка нульової гіпотези про рівність середніх значень сукупностей, що розглядаються.

Основною метою дисперсійного аналізу є дослідження значущості відмінностей між середніми за допомогою порівняння (аналізу) дисперсій.

Для аналізу загальна дисперсія розглядається як сума міжгрупової та внутрішньогрупової дисперсії, зумовленої випадковими помилками. Розподіл загальної дисперсії на кілька джерел дозволяє порівняти дисперсію, викликану різницею між групами, з дисперсією, викликану внутрішньогруповою мінливістю. Якщо фактор є впливовим, дисперсія спостережень в окремій групі значно менша, ніж дисперсія всієї вибірки. Причина цього, очевидно, полягає в суттєвій різниці між середніми значеннями у групах.

Обмеження застосовності критерію:

- досліджувані змінні мають бути розподілені нормально;
- кількість значень у досліджуваних змінних: $3 \leq n_{1,2,3...} \leq 100$;
- гомогенність групових дисперсій.

Метод перевірки гомогенності дисперсій залежить від співвідношення кількості значень у вибірках, що розглядаються. Якщо кількість даних у вибірках однакова, застосовується критерій Кохрена, який полягає в обчисленні відношення максимальної вибіркової дисперсії до суми групових дисперсій та порівнянні отриманого відношення з табличним значенням (додаток 2), яке залежить від кількості факторів (вибірок) і кількості значень у вибірці.

Критерій Стюдента вважають окремим випадком дисперсійного аналізу при $n = 2$.

Вправа 7. Порівняти середні значення показників у стовпцях 7, 8 та 9. Використати засіб Microsoft Excel «Аналіз даних».

Алгоритм виконання:

1. Перевірка вибірок на нормальність була проведена у попередніх вправах.
2. Вибірки незалежні. Кількість значень у вибірках однакова і дорівнює 25.
3. Перевірка гомогенності дисперсій за допомогою критерію Кохрена:
 - a) у комірку G28 аркуша «Первинні дані» ввести формулу $=\text{VAR.S}(G3:G27)$, результат 272755910,05;
 - b) у комірку H28 аркуша «Первинні дані» ввести формулу $=\text{VAR.S}(H3:H27)$, результат 167289109,21;
 - c) у комірку I28 аркуша «Первинні дані» ввести формулу $=\text{VAR.S}(I3:I27)$, результат 268162317,83;
 - d) у комірку J28 ввести формулу $=\text{MAX}(G28:I28)$, результат 272755910,05;
 - e) у комірку K28 ввести формулу $=\text{SUM}(G28:I28)$, результат 708207337,09;

- f) у комірку L28 ввести формулу $=J28/K28$, результат 0,385, це критерій Кохрена емпіричний $G_{\text{емп.}}$;
- g) знайти критерій Кохрена критичний при $\alpha=0,05$, $f=25-1=24$ (кількість значень у вибірках), $k=3$ (кількість вибірок)(див. додаток 2), результат $G_{\text{крит.}} = 0,5153$;
- h) оскільки $G_{\text{емп.}} < G_{\text{крит.}}$, гіпотеза про гомогенність дисперсій підтвердилася, тож умови застосування параметричного дисперсійного аналізу Фішера виконані.

4. Для порівняння середніх значень вибірок виконати команду «Дані → Аналіз даних → Однофакторний дисперсійний аналіз» (Data Analysis → ANOVA Single Factor).

5. У вікні процедури встановити наступні параметри (рис. 69):

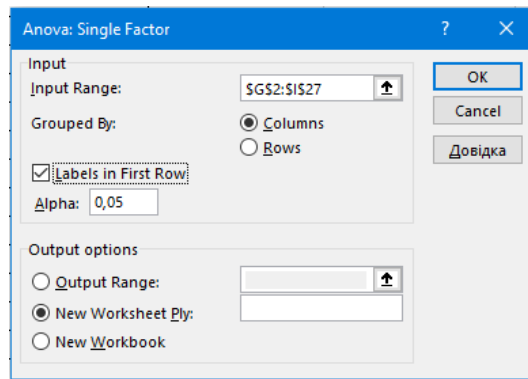


Рис. 69. Налаштування параметрів процедури Microsoft Excel «Аналіз даних → Однофакторний дисперсійний аналіз»

Рівень значущості $\alpha = 0,05$.

Результати виконання процедури виводяться на новий аркуш книги Lesson 04R.xlsx.

6. Після закриття вікна процедури на новому аркуші з'явиться замовлена інформація (рис. 70).

	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	Groups	Count	Sum	Average	Variance		
5	Інтенсивний показник на 100 000, діти 5-9 років	25	1265568	50622,72	2,73E+08		
6	Інтенсивний показник на 100 000, діти 10-14 років	25	1125663	45026,53	1,67E+08		
7	Інтенсивний показник на 100 000, діти 15-17 років	25	1202307	48092,29	2,68E+08		
8							
9							
10	ANOVA						
11	Source of Variation	SS	df	MS	F	P-value	F crit
12	Between Groups	3,93E+08	2	1,96E+08	0,831664	0,439463	3,123907
13	Within Groups	1,7E+10	72	2,36E+08			
14							
15	Total	1,74E+10	74				

Рис. 70. Результат виконання процедури «Однофакторний дисперсійний аналіз»

Значення $F_{\text{емп.}} = 0,832$ (комірка E12).

Значення $F_{\text{крит.}} = 3,124$ (комірка G12).

Оскільки $F_{\text{емп.}} < F_{\text{крит.}}$, приймаємо рішення про підтвердження гіпотези про рівність середніх значень вибірок на заданому α -рівні.

Вправа 8

Дано: Інформація про показники захворюваності на гепатит А по п'яти регіонах надана на аркуші «Декілька вибірок» файлу Lesson 04R.xlsx (рис. 71):

	A	B	C	D	E	F
	роки	показник захворюваності на гепатит А в регіоні X1 (вибірка X1)	показник захворюваності на гепатит А в регіоні X2 (вибірка X2)	показник захворюваності на гепатит А в регіоні X3 (вибірка X3)	показник захворюваності на гепатит А в регіоні X4 (вибірка X4)	показник захворюваності на гепатит А в регіоні X5 (вибірка X5)
3						
4	2017	0,29	1,24	0,21	1,23	0,75
5	2018	0,34	1,07	0,34	0,85	0,66
6	2019	0,36	0,97	0,43	0,76	0,59
7	2020	0,33	0,54	0,39	0,64	0,94
8	2021	0,41	0,86	0,54	1,11	0,77
9	2022	0,37	0,67	0,29	1,01	0,52

Рис. 71. Дані для дослідження різниці значень показника у п'яти вибірках

Завдання: визначити, чи відрізняються регіони рівнем захворюваності на гепатит А.

Метод рішення – дослідження рівності середніх значень показника захворюваності. Якщо середні значення показника захворюваності статистично рівні, то можна зробити висновок, що місце проживання не впливає на захворюваність на гепатит, в протилежному випадку впливає.

Оскільки досліджується 5 регіонів, для прийняття рішення використовується однофакторний дисперсійний аналіз.

Гіпотеза H_0 : Відмінність середніх значень несуттєва (впливу немає).

Гіпотеза H_1 : Відмінність середніх значень суттєва (вплив є).

Алгоритм виконання:

1. Перейти на аркуш «Декілька вибірок» файлу Lesson 04R.xlsx.
2. В комітках B10:F10 за допомогою функції Excel AVERAGE обчислити середнє значення показника захворюваності по кожному регіону.
3. В комітках B11:F11 за допомогою функції Excel MEDIAN обчислити медіану показника захворюваності по кожному регіону.
4. Отримані наступні результати (рис. 72):

	A	B	C	D	E	F
	роки	показник захворюваності на гепатит А в регіоні X1 (вибірка X1)	показник захворюваності на гепатит А в регіоні X2 (вибірка X2)	показник захворюваності на гепатит А в регіоні X3 (вибірка X3)	показник захворюваності на гепатит А в регіоні X4 (вибірка X4)	показник захворюваності на гепатит А в регіоні X5 (вибірка X5)
3						
10	Середнє значення	0,350	0,892	0,367	0,933	0,705
11	Медіана	0,350	0,892	0,367	0,933	0,705

Рис. 72. Результати обчислення середнього значення і медіани по кожному регіону

Оскільки для кожної вибірки середнє значення і медіана приблизно однакові, можна вважати, що закон розподілу у вибірках нормальний.

5. Перевірити тест на гомогенність дисперсій:

а) в комірках B12:F12 за допомогою функції Excel VAR.S обчислити дисперсію по кожному регіону; результат:

Дисперсія	0,002	0,067	0,013	0,050	0,022
-----------	-------	-------	-------	-------	-------

б) у комірці G12 за допомогою формули =MAX(B12:F12)/SUM(B12:F12) обчислити значення критерію Кохрена; результат $G_{емп.} = 0,435$;

в) визначити критичне значення критерію Кохрена за таблицею (додаток 2): $\alpha = 0,05$, $f = 6-1 = 5$ (кількість значень у вибірках), $k = 5$ (кількість вибірок); результат $G_{крит.} = 0,4783$;

г) оскільки $G_{емп.} < G_{крит.}$, гіпотеза про гомогенність дисперсій підтвердилася, тож умови застосування параметричного дисперсійного аналізу Фішера виконані.

6. Для порівняння середніх значень вибірок виконати команду «Дані → Аналіз даних → Однофакторний дисперсійний аналіз» (Data Analysis → ANOVA Single Factor).

Результати дисперсійного аналізу:

<i>F</i> емпіричне	<i>F</i> критичне	<i>P</i> -value
15,27123309	2,75871	1,91413E-06

7. Оскільки $F_{емп.} > F_{крит.}$, можна стверджувати, що різниця між середніми суттєва. Отже місце проживання впливає на захворюваність на гепатит А.

3.1.5. Практичне заняття 5.

Дослідження статистичних зв'язків

Тема: Дослідження статистичного зв'язку між даними. Вибір методу дослідження зв'язку залежно від типу первинних даних. Коефіцієнти кореляції та їх характеристики.

Мета: удосконалення навичок з автоматизації дослідження статистичного зв'язку між даними.

Теоретичні відомості

Окремий розділ інференційної статистики присвячений дослідженню статистичної залежності одного показника від іншого або від групи інших показників. Статистичною залежністю називається така залежність між показниками, коли зі зміною значення одного показника змінюється розподіл або параметри розподілу іншого.

Методи дослідження статистичних залежностей визначаються типом первинних даних і кількістю змінних (табл. 7).

Методи дослідження статистичних залежностей

Загальна кількість змінних	Шкала вимірювання		Закон розподілу	Метод
	факторна змінна	результативна змінна		
2	кількісна	кількісна	нормальний	Параметрична кореляція Пірсона
			відмінний від нормального	Непараметрична кореляція Спірмена
	хоча б одна порядкова			Непараметрична кореляція Спірмена або Кендала
3 і більше	порядкова	порядкова		Конкордація
2 і більше	найменувань	кількісна або порядкова	нормальний	Параметричний дисперсійний аналіз
			відмінний від нормального	Непараметричний дисперсійний аналіз
2	найменувань	порядкова		Рангово-бісеріальний коефіцієнт кореляції
2	найменувань	кількісна		Точково-бісеріальний коефіцієнт кореляції
2	найменувань, 2 × 2 значень			Таблиці спряженості. Коефіцієнт контингенції Пірсона ϕ . Коефіцієнт асоціації Юла Q
2	найменувань, 2 × k (> 2) значень			Коефіцієнт взаємної зв'язаності Чупрова K. Коефіцієнт взаємної зв'язаності Крамера V
2	найменувань, k1 × k2 значень, k1 > 2, k2 > 2			Коефіцієнт взаємної зв'язаності Пірсона C. Коефіцієнт взаємної зв'язаності Чупрова K. Коефіцієнт взаємної зв'язаності Крамера V

Кореляція – статистична залежність між кількісними змінними, за якої кожному значенню однієї змінної x відповідає певне очікуване значення іншої змінної y . Властивості статистичних зв'язків: 1) значущість; 2) направленість; 3) сила; 4) симетрія; 5) форма. Зв'язок називається «статистично значущим», якщо приймається гіпотеза про наявність зв'язку в генеральній сукупності.

Існують три типи кореляції: позитивна, негативна та нульова.

Позитивна (або пряма) кореляція відбувається, коли зі збільшенням значення однієї змінної збільшується значення іншої змінної.

Негативна (або зворотна) кореляція відбувається, коли зі збільшенням значення однієї змінної зменшується значення іншої змінної.

Нульова кореляція відбувається, коли немає залежності між двома змінними. Це означає, що змінні не взаємодіють між собою.

Більшість статистичних кореляційних критеріїв спрямовані на виявлення та інтерпретацію зв'язку лінійної форми. Найпростіший засіб виявлення форми зв'язку – аналіз діаграми розсіювання.

Діаграма розсіювання – графік, осі якого відповідають значенням двох змінних, а кожен досліджуваний є точкою. На підставі цієї діаграми можна зробити висновки про форму зв'язку (лінійний, нелінійний), напрямок (прямий, зворотний), наявність аутлаєрів (викидів).

На *рис. 73* представлені види діаграм розсіювання.

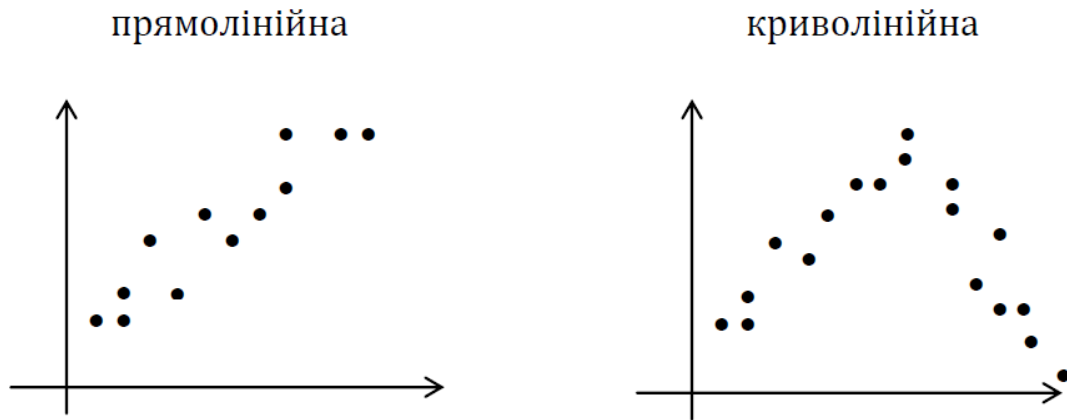


Рис. 73. Класифікація зв'язків за формою

Силу і напрям кореляції між двома кількісними змінними визначає коефіцієнт кореляції (r), який приймає значення в інтервалі від «-1» до «+1». Метод обчислення коефіцієнта кореляції залежить від розподілу змінних. Показником напрямку зв'язку є знак коефіцієнта кореляції (знак «+» відповідає прямій кореляції, знак «-» – зворотній). Показником сили зв'язку є абсолютна величина коефіцієнта кореляції. Сила зв'язку безпосередньо вказує, наскільки синхронно проявляється спільна мінливість досліджуваних змінних. Оцінити силу зв'язку можна за допомогою шкали Чедока або шкали Дж. Коена (табл. 8, 9).

Таблиця 8

Класифікація зв'язку за силою (шкала Чедока)

Якісна характеристика сили зв'язку	Кількісна міра сили зв'язку
дуже слабка	$0,00 < r \leq 0,10$
слабка	$0,10 < r \leq 0,30$
помірна	$0,30 < r \leq 0,50$
середня	$0,50 < r \leq 0,70$
сильна	$0,70 < r \leq 0,90$
дуже сильна	$0,90 < r \leq 1,00$

Розмір ефекту зв'язку між змінними можна визначити і з допомогою коефіцієнта детермінації, який дорівнює квадрату коефіцієнта кореляції $R = r^2$.

Таблиця 9

Оцінка зв'язку за класифікацією Дж. Коена

Якісна характеристика сили зв'язку	Кількісна міра сили зв'язку
несуттєвий	$0,00 \leq r^2 < 0,01$
малий	$0,01 \leq r^2 < 0,09$
середній	$0,09 \leq r^2 < 0,25$
великий	$0,25 \leq r^2 < 1,00$

Нехай змінна x має кількісні значення x_1, x_2, \dots, x_n , змінна y – y_1, y_2, \dots, y_n . x і y розподілені за нормальним законом, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. Коефіцієнт лінійної кореляції Пірсона визначається за формулою:
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

На величину коефіцієнта кореляції Пірсона не впливає те, в яких одиницях виміру представлені змінні. Будь-які лінійні перетворення змінних (помноження на константу, додавання константи) не змінюють значення коефіцієнта кореляції.

Значущість коефіцієнта кореляції Пірсона перевіряється наступним чином:

• обчислюється значення емпіричного критерію за формулою:

$$t_{\text{емп}} = \frac{|r|}{\sqrt{1-r^2}} \sqrt{n-1};$$

• для заданого рівня значущості α і числа ступенів свободи $=n-2$ знаходиться критичне значення критерію – квантиль розподілу Стьюдента (за таблицями або за допомогою функції Microsoft Excel).

Якщо $t_{\text{емп}} > t_{\text{крит}}$, можна стверджувати, що коефіцієнт кореляції значущий.

Якщо первинні дані не відповідають закону нормального розподілу або представлені в порядковій шкалі, то для дослідження зв'язку між ознаками використовуються коефіцієнти рангової кореляції r_s Спірмена або τ Кендалла. Коефіцієнт Спірмена більш точно враховує кількісну ступінь зв'язку між змінними. При розрахунках тих самих даних $r_s > \tau$. Для обчислення рангових коефіцієнтів переходять від значень змінних x_i до рангів цих значень rx_i , тобто порядкових номерів значень x_i у впорядкованій сукупності первинних даних. Коефіцієнт рангової кореляції Спірмена визначається за формулою $r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)}$, де $d_i = rx_i - ry_i$ – різниця рангів, n – кількість значень у вибірці. Однак коли для обчислення задіяна програма Microsoft Excel, достатньо перейти від значень до рангів і визначити для них коефіцієнт кореляції Пірсона. Значущість коефіцієнта кореляції Спірмена перевіряється таким же чином, як для коефіцієнта кореляції Пірсона.

Нехай змінна x має кількісні значення x_1, x_2, \dots, x_n , змінна y має кількісні значення y_1, y_2, \dots, y_n . Розподіл змінних відрізняється від нормального.

Алгоритм обчислення рангового коефіцієнта кореляції τ Кендалла:

1. Обчислити ранги для первинних значень показників rx_1, rx_2, \dots, rx_n та ry_1, ry_2, \dots, ry_n .

2. Впорядкувати таблицю рангів двох змінних за зростанням рангів змінної x rx_i .

3. Для кожного значення ry_i обчислити кількість узгоджених P і незгоджених Q пар. Пара (rx_i, ry_i) вважається узгодженою, якщо зміна за x збігається за направленістю зі зміною за y .

4. Коефіцієнт рангової кореляції τ Кендалла визначається за формулою $\tau = \frac{P-Q}{n(n-1)/2}$.

5. Значущість коефіцієнта кореляції τ Кендалла перевіряється таким же чином, як для коефіцієнта кореляції Пірсона.

Перевірку зв'язку можна здійснювати лише для пов'язаних вибірок. Це означає, що між елементами обох досліджуваних вибірок існує взаємно однозначна відповідність, а кількість елементів у вибірках є однаковою.

Задача-еталон

Дано: Інформація про випадки захворювань на гострі інфекції дихальних шляхів на території України у р. надана в таблиці Excel (джерело даних – звіт МОЗ про інфекційну захворюваність в Україні – форма 1).

У табл. 10 представлені показники, які внесені в таблицю Excel.

Таблиця 10

Дані стовпців (епідеміологічні показники) таблиці Excel

№ з/п	Назва показника	Примітка
1	Регіон України	
2	Область	
3	Населення	
4	Інтенсивний показник на 100 000, все населення	
5	Індекс урбанізації	Попередньо розрахований за формулою: Індекс урбанізації = кількість міського населення області*100 / / кількість всього населення області
6	Індекс групи ризику за віком	Попередньо розрахований за формулою: Індекс групи ризику за віком = (Кількість населення до 14 років + Кількість населення понад 65 років) * 100 / Кількість всього населення
7	Індекс жіночого населення	Попередньо розрахований за формулою: Індекс жіночого населення = Кількість жінок * 100 / Кількість всього населення
8	Щільність населення осіб/кв.км	Попередньо розрахований за формулою: Щільність населення = Кількість населення / Площу території в кв. км

Примітка: групи ризику за віком можуть бути визначені по різному на певних територіях

Дані для аналізу надаються у файлі Lesson 05D.xlsx, аркуш «Первинні дані».

Вправа 1. Дослідити наявність статистичного зв'язку між показниками 4, 5, 6, 7, 8.

Алгоритм виконання:

1. Відкрити файл Lesson 05D.xlsx. Зберегти як Lesson 05R.xlsx.

2. Для перевірки нормальності розподілу даних у стовпцях 4, 5, 6, 7, 8 на аркуші «Первинні дані» підготовлений шаблон. Скористайтеся ним і введіть значення відношення R/σ , A_s , E у відповідні комірки (C40:C44, G40;G44, H40:H44). Результат перевірки R_s критерію у комірках F40:F44. Результат перевірки критерію Плохінського у комірках K40:K44. Можна стверджувати, що розподіл первинних даних у стовпцях 4, 5, 6, 7 відповідає нормальному, а у стовпці 8 відрізняється.

3. Побудувати діаграми розсіювання для наступних пар показників: 4 і 5, 4 і 6, 4 і 7, 4 і 8 (рис. 74).



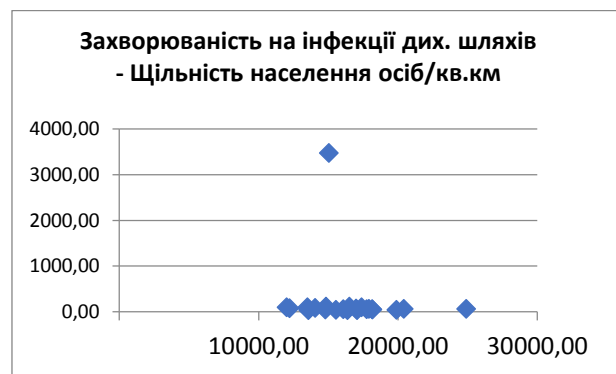
Зв'язок між інтенсивним показником захворюваності на інфекції дихальних шляхів та індексом урбанізації – нелінійний, прямий



Зв'язок між інтенсивним показником захворюваності на інфекції дихальних шляхів та індексом групи ризику за віком – близький до лінійного, прямий



Зв'язок між інтенсивним показником захворюваності на інфекції дихальних шляхів та індексом жіночого населення – нелінійний, прямий



Зв'язок між інтенсивним показником захворюваності на інфекції дихальних шляхів та щільністю населення – треба досліджувати без викидів



Діаграма розсіювання без даних м. Києва, де щільність населення перевищила щільність населення в областях у сотні разів. Зв'язок між інтенсивним показником захворюваності на інфекції дихальних шляхів та щільністю населення нелінійний

Рис. 74. Діаграми розсіювання між показником захворюваності на інфекції дихальних шляхів та демографічними показниками

Для ретельного дослідження дані треба розбивати на однорідні групи.

4. Існує 2 варіанти обчислення коефіцієнта кореляції Пірсона в програмі Microsoft Excel.

4.1. За допомогою вбудованої функції:

4.1.1. Виділіть комірку E26, зверніться до майстра функцій, оберіть функцію CORREL з розділу «Статистичні».

4.1.2. У вікні побудови функції задайте аргументи – діапазони з первинними даними інтенсивного показника на 100 000, все населення й індексу урбанізації (рис. 75).

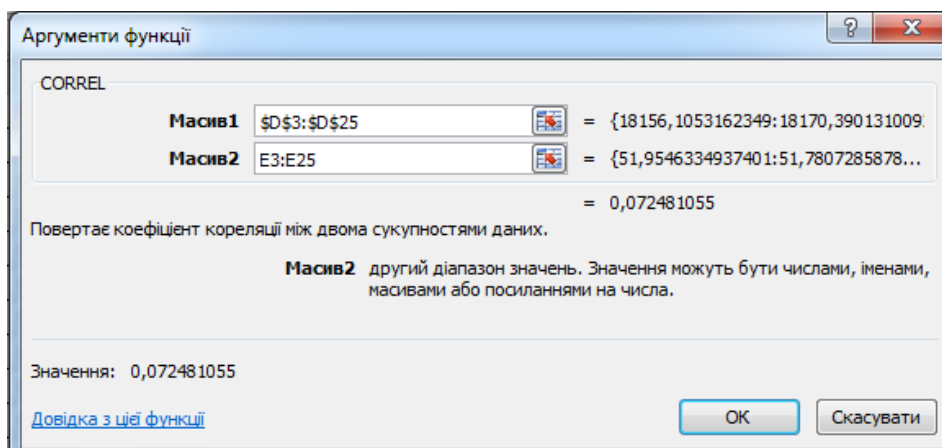


Рис. 75. Вікно побудови функції CORREL (коефіцієнт кореляції Пірсона)

4.1.3. Після закриття вікна побудови функції в комірці E26 буде розташований результат обчислення коефіцієнта кореляції за формулою =CORREL(\$D\$3:\$D\$25;E3:E25);

4.1.4. Скопіювати формулу з комірки E26 до комірок F26 і G26.

Отримані результати зазначені в табл. 11.

Таблиця 11

Результати розрахунку кількісної та якісної сили зв'язку

№ з/п	Індекси факторів ризику		Кількісна міра сили зв'язку	Якісна міра сили зв'язку
1	інтенсивний показник	індекс урбанізації	0,0725	дуже слабка
2	захворюваності на 100 000,	індекс групи ризику за віком	0,5693	середня
3	все населення	індекс жіночого населення	0,3260	помірна

4.2. За допомогою процедури «Аналіз даних → Кореляція»:

4.2.1. Виконати команду «Дані → Аналіз даних → Кореляція» (Data Analysis → Correlation);

4.2.2. У вікні налаштування процедури встановити наступні параметри (рис. 76):

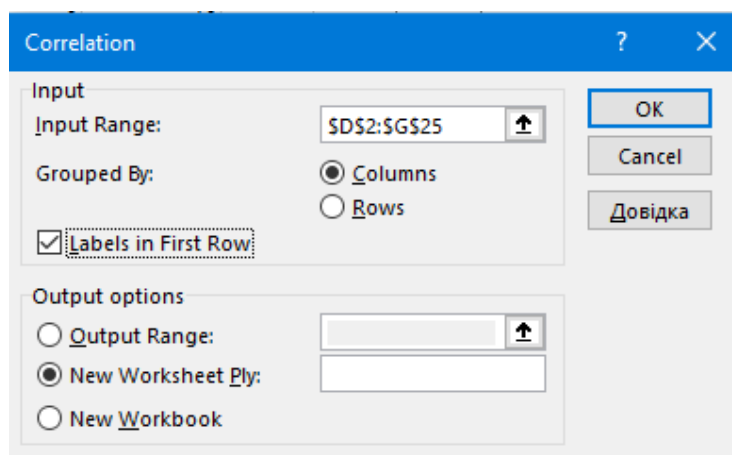


Рис. 76. Налаштування параметрів процедури Microsoft Excel «Аналіз даних → Кореляція»

Результати виконання процедури виводяться на новий аркуш книги Lesson 05R.xlsx.

4.2.3. Після закриття вікна процедури на новому аркуші з'явиться замовлена інформація (рис. 77).

	A	B	C	D	E
1		Інтенсивний показник на 100 000, все населення	Індекс урбанізації	Індекс групи ризику за віком (<14 і >65)	Індекс жіночого населення
2	Інтенсивний показник на 100 000, все населення	1			
3	Індекс урбанізації	0,072481055	1		
4	Індекс групи ризику за віком (<14 і >65)	0,569321729	0,096030208	1	
5	Індекс жіночого населення	0,325964504	0,665875362	0,395441201	1
6					

Рис. 77. Результат виконання процедури «Аналіз даних → Кореляція»

У таблиці (рис. 77) обчислені коефіцієнти кореляції Пірсона між усіма показниками.

5. Визначити значущість обчислених коефіцієнтів кореляції.

5.1. Для обчислення $t_{\text{емп}}$ до комірки E27 аркушу «Первинні дані» ввести формулу =SQRT(22)*ABS(E26)/SQRT(1-E26^2). Скопіювати її до комірок F27^G27.

5.2. Для визначення $t_{\text{крит}}$ до комірки E28 ввести формулу =T.INV.2T(0,05;21). Ця функція звертається до таблиці квантилів T розподілу для двобічної критичної області і видає табличне значення для $\alpha=0,05$ і $df=21$. Скопіювати формулу до комірок F28^G28.

5.3. Для перевірки значущості до комірки E29 ввести формулу =E27>E28. Скопіювати її до F29:G29.

Результати обчислень представлені на рис. 78.

	B	C	D	E	F	G
26			r	0,072481	0,569321729	0,325965
27			t емп	0,340863	3,248154101	1,617239
28			t крит	2,079614	2,079613845	2,079614
29			Значущість	FALSE	TRUE	FALSE

Рис. 78. Результати перевірки значимості коефіцієнтів кореляції

Висновки:

- Між інтенсивним показником захворюваності на гострі інфекції верхніх дихальних шляхів на 100 000 населення та індексом урбанізації зв'язок прямий, дуже слабкий та незначущий.

- Між інтенсивним показником захворюваності на гострі інфекції верхніх дихальних шляхів на 100 000 населення та індексом групи ризику за віком зв'язок прямий, середній та значущий.

• Між інтенсивним показником захворюваності на гострі інфекції верхніх дихальних шляхів на 100 000 населення та індексом жіночого населення зв'язок прямий, помірний та незначущий.

6. Обчислити ранговий коефіцієнт кореляції Спірмена для дослідження зв'язку між інтенсивним показником захворюваності на гострі інфекції верхніх дихальних шляхів на 100 000 населення та щільністю населення. Беручи до уваги діаграму розсіювання, кількість рядків, що досліджуються, дорівнює 22.

6.1. Для обчислення рангів значень «інтенсивного показника захворюваності» до комірки J3 ввести формулу =RANK(D3;\$D\$3:\$D\$24;1). Скопіювати її на діапазон J4:J24.

6.2. Для обчислення рангів значень «щільності населення» до комірки K3 ввести формулу =RANK(H3;\$H\$3:\$H\$24;1). Скопіювати її на діапазон K4:K24.

6.3. У комірку L3 ввести формулу =(J3-K3)^2. Скопіювати її на діапазон L4:L24.

6.4. У комірку L26 ввести формулу =SUM(L3:L24).

6.5. У комірку L27 ввести формулу =6*L26/(22*(22^2-1)).

6.6. У комірку L28 ввести формулу =1-L27. Обчислене в комірці L28 значення коефіцієнта кореляції Спірмена дорівнює (-0,20045).

6.7. Значення коефіцієнта кореляції Спірмена може бути отримане також за формулою =CORREL(J3:J24;K3:K24). Результат (-0,20045).

6.8. Для перевірки значущості скопіювати обчислене значення в комірку H26. У комірку H27 ввести формулу =SQRT(21)*ABS(H26)/SQRT(1-H26^2), а у комірку H28 =T.INV.2T(0,05;20). Скопіювати формулу з G29 до H29.

Результати обчислень представлені на *рис. 79*.

	A	B	C	D	E	F	G	H
26				r	0,072481	0,569321729	0,3259645	-0,200451722
27				t емп	0,340863	3,248154101	1,6172394	0,937615425
28				t крит	2,079614	2,079613845	2,0796138	2,085963447
29				Значущість	FALSE	TRUE	FALSE	FALSE

Рис. 79. Коефіцієнти кореляції та їх значущість

Висновки. Між інтенсивним показником захворюваності на гострі інфекції верхніх дихальних шляхів на 100 000 населення та щільністю населення зв'язок зворотний, слабкий та незначущий. Отже, проведений кореляційний аналіз показав, що на території дослідження за певний період часу перебіг епідемічного процесу гострих інфекцій верхніх дихальних шляхів залежить від величини групи ризику за віком (діти та особи похилого віку).

7. Обчислити ранговий коефіцієнт кореляції τ Кендалла для дослідження зв'язку між інтенсивним показником захворюваності на гострі інфекції верхніх дихальних шляхів на 100 000 населення та щільністю населення.

7.1. Значення рангів, що досліджуються, скопіювати з комірок J3:K24 до діапазону N3:O24. Скористатися засобом «Спеціальна вставка → Значення».

7.2. Для діапазону N3:O24 виконати команду «Дані → Сортувати».

7.3. У вікні налаштування процедури сортування встановити наступні параметри (*рис. 80*):

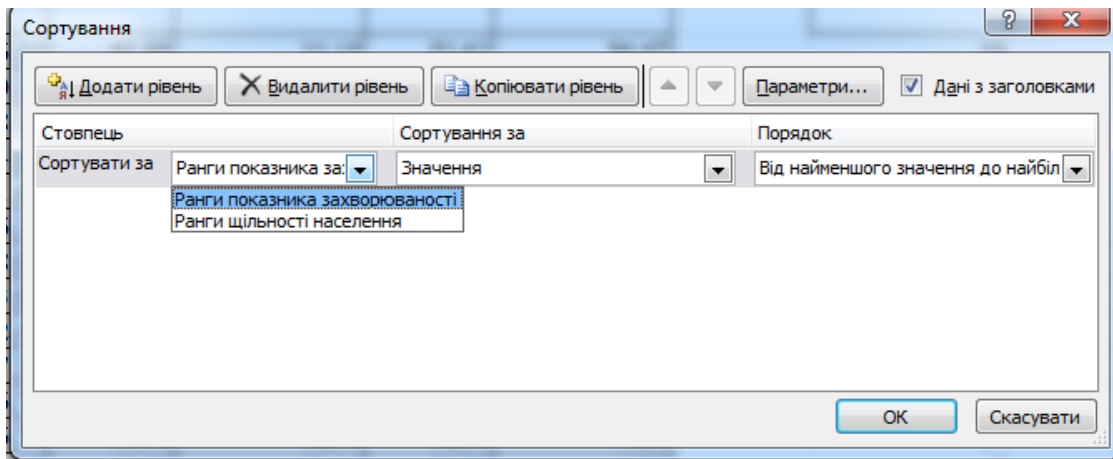


Рис. 79. Налаштування процедури сортування в Microsoft Excel

7.4. Вміст діапазону з рангами після виконання процедури сортування (рис. 81):

Ранги показника захворюваності	Ранги щільності населення
1	20
2	16
3	19
4	6
5	3
6	17
7	7
8	22
9	5
10	9
11	2
12	21
13	15
14	1
15	18
16	11
17	13
18	10
19	8
20	4
21	12
22	14

Рис. 81. Результат сортування таблиці рангів, яка створена для обчислення коефіцієнта кореляції τ Кендалла

7.5. Для підрахунку кількості узгоджених пар до комірки P3 ввести формулу =COUNTIF(O3:\$O\$24;">"&O3). Скопіювати її до діапазону P4:P24. У комірці P26 обчислити суму значень комірок P3:P24.

7.6. Для підрахунку кількості неузгоджених пар до комірки Q3 ввести формулу =COUNTIF(O3:\$O\$254;"<"&O3). Скопіювати її до діапазону Q4;Q24. У комірці Q26 обчислити суму значень комірок Q3:Q24.

7.7. У комірці Q27 обчислити різницю між кількістю узгоджених та кількістю неузгоджених пар за формулою =P26-Q26.

7.8. У комірці Q28 обчислити коефіцієнт кореляції τ Кендалла за формулою =Q27/(22*21/2). Результат дорівнює -0,14286, що за абсолютною величиною менше за значення коефіцієнта рангової кореляції Спірмена (-0,20045).

Висновки. Коефіцієнти рангової кореляції вказують на слабку зворотну незначущу кореляцію між інтенсивним показником захворюваності на гострі інфекції верхніх дихальних шляхів на 100 000 населення та щільністю населення. Отже, проведений кореляційний аналіз показав, що на території дослідження за певний період часу не виявлено впливу демографічних процесів (окрім груп ризику, що представлено в попередньому дослідженні) на перебіг епідемічного процесу гострих інфекцій верхніх дихальних шляхів.

Вправа 2

Дано: Інформація про захворюваність на кір в регіоні та охоплення плановими щепленнями від кору в цьому ж регіоні надана на аркуші «Кореляція Пірсона» файлу Lesson 05R.xlsx (рис. 82):

	А	В	С
	роки	захворюваність на кір на 100 тис. населення регіону	% щеплених від кору в регіоні
3			
4	2017	40,6	87,8
5	2018	90,8	86,8
6	2019	99,9	85,1
7	2020	97,1	87,1
8	2021	96,7	84
9	2022	125,2	86,7

Рис. 82. Дані для дослідження статистичного зв'язку епідеміологічних показників захворюваності на кір та відсотку щеплених

Завдання: використовуючи дані таблиці, визначити, чи є залежність між захворюваністю на кір та охопленням плановими профілактичними щепленнями.

Алгоритм виконання:

1. Перейти на аркуш «Кореляція Пірсона» файлу Lesson 05R.xlsx.
2. За даними комірок В4:В9 та С4:С9 побудувати діаграму розсіювання (рис. 83).

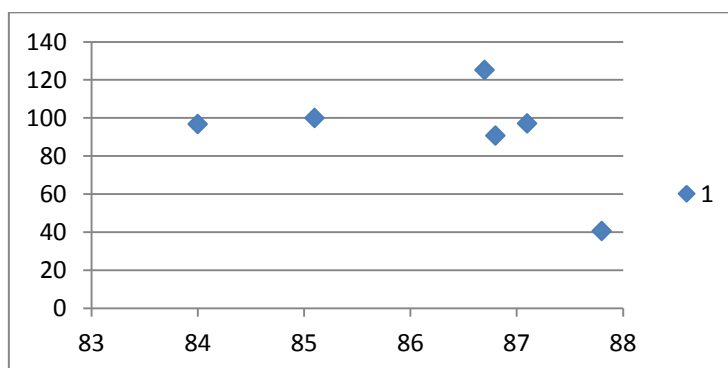


Рис.83. Діаграма розсіювання епідеміологічних показників: відсоток охоплення профілактичним щепленням (вісь X) та захворюваність на кір (вісь Y)

Із діаграми видно, що ймовірність лінійного зв'язку між показниками невелика.

3. Перевірити, чи є нормальним закон розподілу у наданих вибірках (рис. 84):
 - а) у комірках В10:С10 за допомогою функції AVERAGE обчислити середні значення показників;

- b) у комірках B11:C11 за допомогою функції MEDIAN обчислити медіани показників;
- c) за отриманими результатами зробити висновок, чи близький розподіл показників до нормального:

		захворюваність на кір на 100 тис. населення регіону	% щеплених від кору в регіоні
3			
10	Середнє значення	91,72	86,25
11	Медіана	96,90	86,75

Рис. 84. Результати обчислення числових статистичних характеристик: середнє значення і медіана

Оскільки середні значення і медіани приблизно рівні, можна припустити, що у наданих вибірках нормальний закон розподілу.

4. У комірці B13 за допомогою функції =CORREL(B4:B9;C4:C9) обчислити лінійний коефіцієнт кореляції Пірсона. Результат -0,41075.

5. Перевірити значущість коефіцієнта кореляції при рівні значущості $\alpha = 0,05$.

a) у комірці B14 для обчислення емпіричного значення критерію за формулою $t_{\text{емп}} = \frac{|r|}{\sqrt{1-r^2}} \sqrt{n-1}$ ввести функцію =ABS(B13*SQRT(6-2)/SQRT(1-B13^2)); результат 0,90101;

b) у комірці B15 для обчислення критичного значення критерію ввести функцію =T.INV.2T(0,05;6-2); результат 2,77644;

c) оскільки $t_{\text{емп}} < t_{\text{крит}}$, можна стверджувати, що коефіцієнт кореляції не є значущим, тобто лінійного зв'язку між досліджуваними показниками немає.

Вправа 3

Дано: Інформація про захворюваність на дизентерію дітей віком до 1 року в місті N та відсоток нестандартної молочної продукції в цьому ж місті надана на аркуші «Кореляція Спірмена» файлу Lesson 05R.xlsx (рис. 85):

	A	B	C	D	E
3	місяці року	захворюваність дітей на дизентерію (на 10000)	ранги	відсоток нестандартної молочної продукції	ранги
4	січень	78		10	
5	лютий	47		4	
6	березень	40		1,2	
7	квітень	51		3,3	
8	травень	60		9	
9	червень	58		13	
10	липень	53		4,8	
11	серпень	60		5,8	
12	вересень	655		5,3	
13	жовтень	74		15,3	
14	листопад	58		6,8	

Рис. 85. Дані для дослідження статистичного зв'язку епідеміологічних показників захворюваності на дизентерію та відсотку нестандартної молочної продукції

Завдання: використовуючи дані таблиці, визначити, чи є залежність між захворюваністю на дизентерію та відсотком нестандартної молочної продукції.

Алгоритм виконання:

1. Перейти на аркуш «Кореляція Спірмена» файлу Lesson 05R.xlsx.
2. За даними комірок B4:B14 та D4:D14 побудувати діаграму розсіювання (рис. 86).

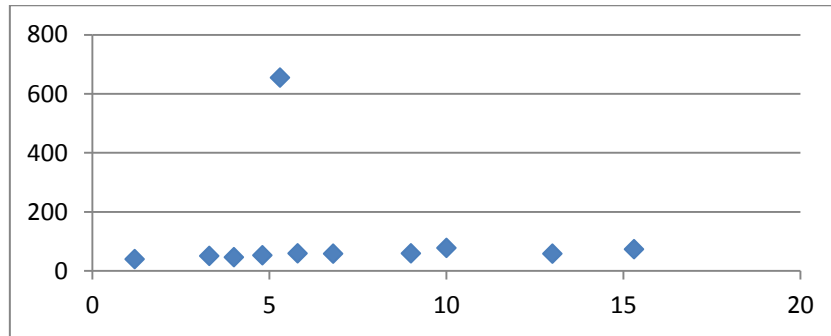


Рис. 86. Діаграма розсіювання показників: відсоток нестандартної молочної продукції (вісь X) та захворюваність на дизентерію (вісь Y)

З діаграми видно, що серед значень захворюваності є таке, що різко відрізняється від інших, тому є сенс перейти до дослідження рангів і для оцінки зв'язку використати коефіцієнт кореляції Спірмена.

3. У комірках C4:C14 обчислити ранги відповідних значень з діапазону B4:B14. Для цього в комірку C4 ввести формулу `=RANK.AVG(B4;B4:B14;1)`. Скопіювати формулу з C4 до діапазону C5:C14.

4. У комірках E4:E14 обчислити ранги відповідних значень з діапазону D4:D14. Для цього в комірку E4 ввести формулу `=RANK.AVG(D4;D4:D14;1)`. Скопіювати формулу з E4 до діапазону E5:E14.

Отримані такі результати (рис. 87):

	A	B	C	D	E
3	місяці року	захворюваність дітей на дизентерію (на 10000)	ранги	відсоток нестандартної молочної продукції	ранги
4	січень	78	10	10	9
5	лютий	47	2	4	3
6	березень	40	1	1,2	1
7	квітень	51	3	3,3	2
8	травень	60	7,5	9	8
9	червень	58	5,5	13	10
10	липень	53	4	4,8	4
11	серпень	60	7,5	5,8	6
12	вересень	655	11	5,3	5
13	жовтень	74	9	15,3	11
14	листопад	58	5,5	6,8	7

Рис. 87. Результати обчислення рангів значень показників захворюваності на дизентерію та відсотку нестандартної молочної продукції

5. За допомогою функції `CORREL` в комірці B16 обчислити коефіцієнт кореляції Спірмена. Результат 0,6895.

- б. Перевірити значущість коефіцієнта кореляції при рівні значущості $\alpha = 0,05$.
- а) у комірці B18 для обчислення емпіричного значення критерію за формулою $t_{\text{емп}} = \frac{|r|}{\sqrt{1-r^2}} \sqrt{n-1}$ вести функцію =ABS(B16*SQRT(11-2)/SQRT(1-B16^2)); результат 2,856;
- б) у комірці B19 для обчислення критичного значення критерію ввести функцію =T.INV.2T(0,05;11-2); результат 2,26216;
- с) оскільки $t_{\text{емп}} > t_{\text{крит}}$, можна стверджувати, що коефіцієнт кореляції значущий, тобто між досліджуваними показниками існує прямий зв'язок середньої сили ($r = 0,6895$).

Теоретичні відомості

Статистичні зв'язки між кількісними і порядковими змінними називаються кореляцією, а між номінальними показниками, які вимірюються за шкалою найменувань – асоціацією.

Шкала найменувань створюється шляхом привласнення «імен» об'єктам, які використовуються тільки щоб відрізнити один об'єкт від іншого і можуть бути позначені словами або числами. Найпростіша номінальна шкала – дихотомічна. У разі вимірювання за дихотомічною шкалою змінні можна кодувати двома символами або цифрами, наприклад бінарними значеннями 0 та 1, які зазвичай відповідають наявності або відсутності деякої ознаки.

Найбільш поширеними статистичними показниками, які використовуються для вимірювання зв'язку між бінарними (альтернативними) ознаками, є коефіцієнт контингенції Пірсона ϕ і коефіцієнт асоціації Юла Q . Значення цих коефіцієнтів може змінюватися в межах від «-1» до «+1».

Нехай змінна x має бінарні значення x_1, x_2, \dots, x_n , змінна y має бінарні значення y_1, y_2, \dots, y_n .

Алгоритм обчислення коефіцієнтів контингенції і асоціації:

1. На підставі первинних значень будується таблиця спряженості – це засіб подання спільного розподілу двох змінних, призначений для вивчення зв'язку між ними. Рядки таблиці відповідають значенням однієї змінної, стовпці – значенням іншої змінної. На перетині рядка та стовпця вказується частота спільної появи f_{ij} відповідних значень двох ознак x_i та y_j .

У таблицях спряженості можуть бути як абсолютні, так і відносні частоти (у частках чи відсотках). Ці таблиці бувають різного розміру. Найменші – чотириклітинні таблиці спряженості мають такий вигляд (*табл. 12*):

Таблиця 12

Таблиці спряженості для обчислення коефіцієнтів контингенції і асоціації

Вибірка	Відсутність ознаки (0)	Наявність ознаки (1)	Усього
Перша вибірка	a	b	$n1 = a + b$
Друга вибірка	c	d	$n2 = c + d$
Усього	$a + c$	$b + d$	$n = n1 + n2$

або

		Змінна x		Усього
		0	1	
Змінна y	0	a	b	$a + b$
	1	c	d	$c + d$
Усього		$a + c$	$b + d$	n

Для побудови таблиці спряженості можна використати засіб Microsoft Excel «Зведені таблиці» або відповідний засіб в потужних статистичних програмах. У таблиці спряженості a відповідає кількості записів первинних даних, у яких змінна x має значення 0, змінна y має значення 0, тобто кількості пар (0,0), b відповідає кількості пар (1,0), c відповідає кількості пар (0,1), d відповідає кількості пар (1,1).

2. Коефіцієнт контингенції Пірсона обчислюється за формулою:

$$\varphi = \frac{ad - bc}{\sqrt{(a + c)(b + d)(a + b)(c + d)}}$$

3. Коефіцієнт асоціації Юла обчислюється за формулою: $Q = \frac{ad-bc}{ad+bc}$

4. Значущість коефіцієнта контингенції перевіряється за критерієм χ^2 наступним чином:

4.1. $\chi^2_{\text{емп.}} = \varphi^2 \times n$, де n – кількість значень у первинних вибірках x і y .

4.2. $\chi^2_{\text{крит.}}$ залежить тільки від рівня значущості, оскільки число ступенів свободи $df = 1$: $\chi^2_{0,05} = 3,841$, $\chi^2_{0,01} = 6,635$, $\chi^2_{0,001} = 10,829$.

4.3. Якщо $\chi^2_{\text{емп.}} < \chi^2_{\text{крит.}}$, статистично значущої залежності між змінними немає.

4.4. Якщо $\chi^2_{\text{емп.}} \geq \chi^2_{\text{крит.}}$, залежність між змінними статистично значуща, тобто показники змінюються узгоджено.

4.5. Показником розміру ефекту виступає безпосередньо саме емпіричне значення коефіцієнта контингенції φ .

5. Значущість коефіцієнта асоціації перевіряється за критерієм χ^2 наступним чином:

5.1. $\chi^2_{\text{емп.}} = Q^2 \times n$.

5.2. $\chi^2_{\text{крит.}}$ обирається так само, як для коефіцієнта контингенції.

5.3. Висновки щодо значущості коефіцієнта асоціації робляться аналогічно п. 4.3 і 4.4.

Коефіцієнт контингенції можна обчислити в Microsoft Excel за формулою коефіцієнта лінійної кореляції Пірсона CORREL, якщо не переходити до аналізу таблиці спряженості. Коефіцієнт асоціації завжди більший коефіцієнта контингенції.

Для таблиці розміром 2×2 використовується ще один показник величини розміру ефекту – відношення шансів (odds ratio – OR): $OR = \frac{ad}{bc}$

Вправа 4. На підставі епідеміологічних показників 4, 5, 6, 7 таблиці з інформацією про випадки захворювань на гострі інфекції дихальних шляхів на території України у 2019 р. (аркуш «Первинні дані» файлу Lesson05R.xlsx) сформувані бінарні показники 9, 10, 11, 12, які будуть відображати перевищення середнього значення відповідного показника. Дослідити наявність статистичного зв'язку між показниками 9, 10, 11, 12.

Алгоритм виконання:

1. Створити у файлі Lesson 05R.xlsx новий аркуш з ім'ям «Якісні дані» і скопіювати до нього первинні дані (аркуш «Первинні дані», діапазон комірок A2:H25).

2. У комірках D26:H26 обчислити середні значення показників 4, 5, 6, 7. Статистичну функцію для обчислення середнього значення можна побудувати за допомогою майстра функцій Microsoft Excel (рис. 88).

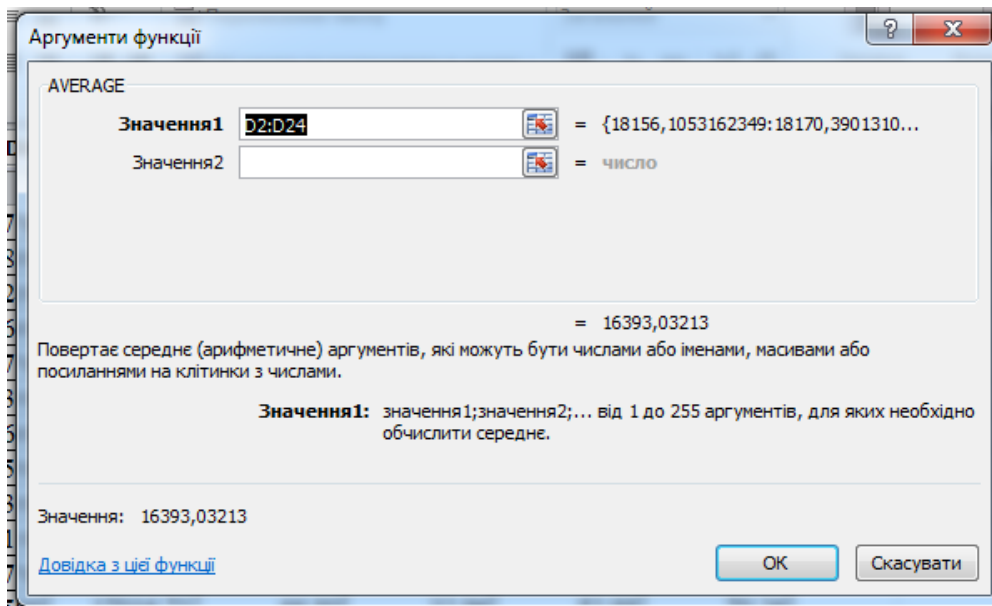


Рис. 87. Вікно побудови функції для обчислення середнього значення діапазону комірок

Формулу можна створити в комірці D26 і скопіювати до комірок E26:G26.

3. На базі стовпця «Інтенсивний показник захворюваності» створити стовпець «Бінарні дані рівня захворюваності» за наступним правилом: якщо показник захворюваності в області більший за середнє значення, рівень захворюваності =1, інакше 0. Для цього у комірці I2 побудувати логічну функцію ЯКЩО (англ. IF) (рис. 89).

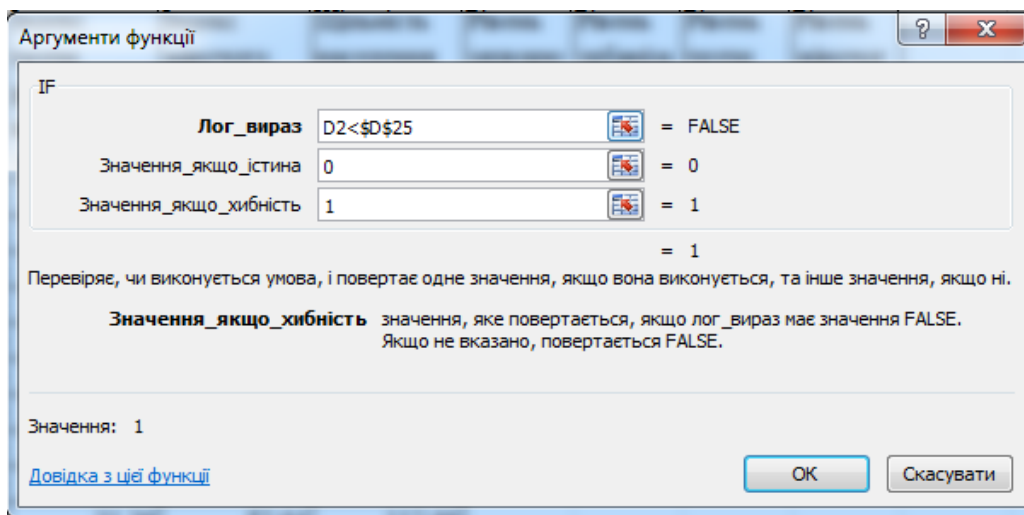


Рис. 89. Вікно побудови функції формування бінарної змінної «рівень захворюваності»

Функцію з комірки I2 скопіювати до діапазону I3:I24.

4. У комірці J2 аналогічно діям пункту 3 сформуванню функції для визначення рівня урбанізації області. Функцію скопіювати до діапазону J3:J24.

5. У комірці K2 аналогічно діям пункту 3 сформуванню функції для визначення рівня групи ризику за віком в області. Функцію скопіювати до діапазону K3:K24.

6. У комірці L2 аналогічно діям пункту 3 сформувати функцію для визначення рівня жіночого населення в області. Функцію скопіювати до діапазону L3:L24.

Таким чином буде отримана таблиця з бінарними даними для аналізу (рис. 90).

	A	B	C	D	E	F	G	H	I	J	K	L
	Регіон України	Область	Населення	Інтенсивний показник на 100 000, все населення	Індекс урбанізації	Індекс групи ризику за віком (<14 і >65)	Індекс жіночого населення	Щільність населення осіб/кв.км	Рівень захворюваності	Рівень урбанізації	Рівень групи ризику за віком	Рівень жіночого населення
1												
2	Центральний	Вінницька	1507700	18156,11	51,95	33,15	53,62	56,67	1	0	1	1
3	Західний	Волинська	1020800	18170,39	51,78	32,74	52,79	50,57	1	0	1	0
4	Східний	Дніпропетровська	3093200	17398,00	84,06	32,70	54,34	96,92	1	1	1	1
5	Центральний	Житомирська	1177600	19904,09	59,43	33,02	53,35	39,55	1	0	1	0
6	Західний	Закарпатська	1243700	12021,44	36,90	31,84	51,91	97,18	0	0	0	0
7	Східний	Запорізька	1636300	17936,32	77,30	32,68	54,34	60,25	1	1	1	1
8	Західний	Івано-Франківська	1350600	13518,92	44,20	31,24	52,72	97,06	0	0	0	0
9	Центральний	Київська	1795500	24913,96	61,34	32,83	54,20	63,61	1	0	1	1
10	Східний	Кіровоградська	902300	13592,47	63,50	33,36	53,83	36,49	0	1	1	1
11	Західний	Львівська	2476100	14847,34	60,71	31,36	52,64	112,66	0	0	0	0
12	Південний	Миколаївська	2950700	13576,78	68,53	32,41	53,61	44,36	0	1	0	1
13	Південний	Одеська	1090500	17016,71	66,80	32,96	52,88	70,26	1	1	1	0
14	Північний	Полтавська	2349700	14801,09	62,20	32,14	53,75	46,77	0	1	0	1
15	Західний	Рівненська	1350600	17760,98	47,04	33,01	52,48	56,90	1	0	1	0
16	Північний	Сумська	1140900	15550,06	69,54	32,00	54,08	43,37	0	1	0	1
17	Західний	Тернопільська	1034400	12227,25	45,82	31,28	53,15	73,68	0	0	0	0
18	Північний	Харківська	1021000	14074,30	81,11	31,01	53,60	82,23	0	1	0	1
19	Південний	Херсонська	2596300	16389,76	61,17	32,69	53,54	35,14	0	0	1	1
20	Західний	Хмельницька	1000400	20434,91	57,74	33,14	53,48	59,37	1	0	1	1
21	Центральний	Черкаська	1227500	16091,44	57,03	32,85	53,97	55,36	0	0	1	1
22	Західний	Чернівецька	1159200	16517,93	42,93	31,68	52,83	109,60	1	0	0	0
23	Північний	Чернігівська	889900	17079,27	65,54	33,54	54,53	29,84	1	1	1	1
24	Центральний	м.Київ	957700	15060,20	100,00	32,21	53,89	3469,60	0	1	0	1
25				16393,03	61,59	32,43	53,46					

Рис. 90. Таблиця кількісних і бінарних якісних показників, що досліджуються

7. У комірках J25:L25 обчислити коефіцієнти кореляції між рівнем захворюваності і рівнями демографічних показників (табл. 13).

Таблиця 13

Коефіцієнти кореляції між бінарними показниками

Показник 1	Показник 2	Функція для обчислення коефіцієнта кореляції	Значення r
Рівень захворюваності	Рівень урбанізації	=CORREL(\$I\$2:\$I\$24;J2:J24)	-0,13741
Рівень захворюваності	Рівень групи ризику за віком	=CORREL(\$I\$2:\$I\$24;K2:K24)	0,664141
Рівень захворюваності	Рівень жіночого населення	=CORREL(\$I\$2:\$I\$24;L2:L24)	-0,1240646

8. На аркуші «Якісні дані» побудувати зведену таблицю на базі даних у комірках A1:L24. Результат помістити починаючи з комірки A28.

8.1.Зведена таблиця для пари «рівень захворюваності – рівень урбанізації» формується наступним чином: в область рядків помістити поле «рівень захворюваності», в область стовпців – поле «рівень урбанізації», в область даних – поле «область». Підсумкова операція – кількість значень. Результат представлений на рис. 91.

27				
28	Кількість з Область	Рівень ...		
29	Рівень захворюваності		0 1	Загальний підсумок
30	0		6 6	12
31	1		7 4	11
32	Загальний підсумок		13 10	23
33				

Рис. 91. Таблиця спряженості показників «рівень захворюваності – рівень урбанізації»

8.2. Виділити діапазон комірок A28:D32 і за допомогою засобу «Спеціальне вставлення → Значення» скопіювати значення комірок до діапазону A34: D38.

8.3. Повернутися до конструктора зведеної таблиці (для цього достатньо клацнути по будь-якій комірці з діапазону A28:D32). Змінити поле стовпців зведеної таблиці на «рівень групи ризику». Результат представлений на *рис. 92*.

28	Кількість з Область	Рівень ...		
29	Рівень захворюваності	0	1	Загальний підсумок
30	0	9	3	12
31	1	1	10	11
32	Загальний підсумок	10	13	23

Рис. 92. Таблиця спряженості показників «рівень захворюваності – рівень групи ризику за віком»

Скопіювати значення комірок A28:D32 до діапазону A41: D45.

8.4. Повернутися до конструктора зведеної таблиці. Змінити поле стовпців зведеної таблиці на «рівень жіночого населення». Результат представлено на *рис. 93*.

28	Кількість з Область	Рівень ...		
29	Рівень захворюваності	0	1	Загальний підсумок
30	0	4	8	12
31	1	5	6	11
32	Загальний підсумок	9	14	23

Рис. 93. Таблиця спряженості показників «рівень захворюваності – рівень жіночого населення»

Скопіювати значення комірок A28:D32 до діапазону A48: D52.

9. На базі таблиці спряженості обчислити коефіцієнт контингенції Пірсона ϕ для пари «рівень захворюваності – рівень урбанізації». У комірку G35 ввести формулу $= (B36 * C37 - B37 * C36) / \text{SQRT}(D36 * B38 * C38 * D37)$.

10. У комірці H35 обчислити значення критерію $\chi^2_{\text{емп.}}$ за формулою $= G35^2 * D38$.

11. У комірку I35 ввести значення $\chi^2_{\text{крит.}} = 3,841$ для $\alpha = 0,05$.

12. У комірку J35 ввести формулу щодо формування висновку перевірки значущості коефіцієнта контингенції $= \text{IF}(H35 < I35; \text{«коefficientів контингенції і асоціації «залежність відсутня»}; \text{«залежність існує»})$.

13. На базі таблиці спряженості обчислити коефіцієнт асоціації Юла для пари «рівень захворюваності – рівень урбанізації». У комірку G36 ввести формулу $= (B36 * C37 - B37 * C36) / (B36 * C37 + B37 * C36)$.

14. У комірці H36 обчислити значення критерію $\chi^2_{\text{емп.}}$ за формулою $= G36^2 * D38$.

15. У комірку I36 ввести значення $\chi^2_{\text{крит.}} = 3,841$ для $\alpha = 0,05$.

16. У комірку J36 ввести формулу щодо формування висновку перевірки значущості коефіцієнта асоціації $= \text{IF}(H36 < I36; \text{«залежність відсутня»}; \text{«залежність існує»})$.

17. У комірці G37 обчислити відношення шансів OR для пари «рівень захворюваності – рівень урбанізації» за формулою $= B36 * C37 / (C36 * B37)$.

18. Формули з комірок G35:J37 скопіювати до діапазону G42:J44 і G49:J51. Результат обчислень представлений на *рис. 94*.

28	Кількість з Область	Рівень ...							
29	Рівень захворюваності	0	1	Загальний підсумок	Ф =	-0,13740858	0,434265734	3,841	залежність відсутня
30	0	4	8	12	Q=	-0,27272727	1,710743802	3,841	залежність відсутня
31	1	5	6	11	OR=	0,571428571			
32	Загальний підсумок	9	14	23					
33									
34	Кількість з Область	Рівень ...							
35	Рівень захворюваності	0	1	Загальний підсумок	Ф =	0,664141488	10,14493007	3,841	залежність є
36	0	6	6	12	Q=	0,935483871	20,12799168	3,841	залежність є
37	1	7	4	11	OR=	30			
38	Загальний підсумок	13	10	23					
39									
40									
41	Кількість з Область	Рівень ...							
42	Рівень захворюваності	0	1	Загальний підсумок	Ф =	-0,12406456	0,354016354	3,841	залежність відсутня
43	0	4	8	12	Q=	-0,25	1,4375	3,841	залежність відсутня
44	1	5	6	11	OR=	0,6			
45	Загальний підсумок	9	14	23					
46									
47									
48	Кількість з Область	Рівень ...							
49	Рівень захворюваності	0	1	Загальний підсумок	Ф =	-0,12406456	0,354016354	3,841	залежність відсутня
50	0	4	8	12	Q=	-0,25	1,4375	3,841	залежність відсутня
51	1	5	6	11	OR=	0,6			
52	Загальний підсумок	9	14	23					
53									

Рис. 94. Результати обчислень коефіцієнтів контингенції, асоціації та відношення шансів для пар бінарних змінних «рівень захворюваності – рівень урбанізації», «рівень захворюваності – рівень групи ризику», «рівень захворюваності – рівень жіночого населення»

Висновки:

1. Між рівнем захворюваності та рівнем урбанізації зв'язок зворотний, слабкий та незначущий, відношення шансів = 0,57.

2. Між рівнем захворюваності та рівнем групи ризику за віком зв'язок прямий, значущий, за коефіцієнтом контингенції – середній, за коефіцієнтом асоціації – дуже сильний, відношення шансів = 30.

3. Між рівнем захворюваності та рівнем жіночого населення зв'язок зворотний, слабкий та незначущий, відношення шансів = 0,6.

Отже, завдяки проведеним дослідженням було встановлено і статистично підтверджено залежність захворюваності на інфекції дихальних шляхів від величини групи ризику за віком (діти та особи похилого віку).

Вправа 5.

Дано: Інформація про захворюваність на дифтерію та наявність щеплення від цієї хвороби надана на аркуші «Асоціація» файлу Lesson 05R.xlsx у такому вигляді (*рис. 95*):

	A	B	C
3	групи спостереження	захворіло	не захворіло
4	щеплені	11	19890
5	не щеплені	18	3032

Рис. 95. Дані для дослідження статистичного зв'язку епідеміологічних показників захворюваності на дифтерію та наявності щеплення від цієї хвороби

Завдання. Використовуючи дані таблиці, визначити, чи є зв'язок між захворюваністю на дифтерію та наявністю щеплення. Визначити силу зв'язку. Обчислити індекс і коефіцієнт вакцинації.

Алгоритм виконання:

1. Перейти на аркуш «Асоціація» файлу Lesson 05R.xlsx.
2. За допомогою формул обчислити підсумки по рядках та стовпцях таблиці з даними. Результати представлені на *рис. 96*.

	A	B	C	D
3	групи спостереження	захворіло	не захворіло	Усього
4	щеплені	11	19890	19901
5	не щеплені	18	3032	3050
6	Усього	29	22922	22951

Рис. 96. Результати обчислення сум по рядках і стовпцях таблиці з даними про кількість захворювань та щеплень від дифтерії

3. Для перевірки гіпотези про наявність зв'язку використовується критерій ХІ-квадрат: гіпотеза Н₀: вибірки належать до однієї генеральної сукупності (немає значної різниці); гіпотеза Н₁: Н₀ неправильна, тобто різниця є, отже зв'язок існує:

- а) у комірці В8 обчислити критерій ХІ-квадрат емпіричний за формулою $\chi^2_{\text{емп.}} = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$; для цього ввести формулу =(D6*(B4*C5-C4*B5)^2)/(D4*D5*B6*C6); результат обчислення 59,95913;
- б) визначити критерій $\chi^2_{\text{крит.}}$; для цього в комірку В9 ввести формулу =CHISQ.INV(1-0,05;1); результат 3,841;
- в) оскільки $\chi^2_{\text{емп.}} > \chi^2_{\text{крит.}}$, вірна гіпотеза Н₁ і можна стверджувати, що зв'язок є.

4. Сила зв'язку визначається значенням коефіцієнта Юла $Q = \frac{ad-bc}{ad+bc}$. У комірку В10 ввести формулу =(B4*C5-C4*B5)/(B4*C5+C4*B5). Результат -0,8296. Таким чином, згідно з коефіцієнтом Юла зв'язок сильний, зворотний.

5. Для обчислення індексу і коефіцієнта вакцинації використовують значення захворюваності за групами щеплених і нещеплених:

- а) у комірці Е4 обчислити захворюваність за групою щеплених; для цього вводиться формула =B4/D4; результат 0,000553;
- б) у комірці Е5 обчислити захворюваність за групою нещеплених; для цього вводиться формула = B5/D5; результат 0,005902;
- в) індекс вакцинації обчислюється як відношення захворюваності групи нещеплених до захворюваності групи щеплених; у комірку Е7 вводиться формула =E5/E4; Результат 10,67714;

д) коефіцієнт вакцинації обчислюється за формулою $\frac{\text{захворюваність за групою нещеплених} - \text{захворюваність за групою щеплених}}{\text{захворюваність за групою нещеплених}}$; в комірку Е8 вводиться формула =(E5-E4)/E5; результат 0,9063.

3.1.6. Практичне заняття 6.

Числові характеристики та методи прогнозування часових рядів

Тема: Динамічні (часові) ряди та їхні числові характеристики. Методи статистичного прогнозування часових рядів.

Мета: удосконалення навичок з автоматизації дослідження часових рядів.

Теоретичні відомості

Динамічним або часовим рядом (англ. Time Series) називають сукупність спостережень досліджуваної величини, впорядковану в хронологічному порядку. Спостереження проводять через однакові періоди часу. Іншої інформації, крім спостережень, дослідник не має. На відміну від елементів випадкової вибірки, де індекси вводяться тільки для зручності ідентифікації, члени часового ряду не є статистично незалежними.

Нехай маємо $n+1$ спостережень деякої випадкової величини $y_0, y_1, y_2, \dots, y_n$. Для дослідження часового ряду використовують його числові характеристики, які представлені в *табл. 14*.

Таблиця 14

Числові характеристики часового ряду та формули їх розрахунку

Характеристики	Розрахункові формули	
	ланцюговий показник	базовий показник
1. Абсолютний приріст	$\Delta_i = y_i - y_{i-1}$	$\Delta_i = y_i - y_0$
2. Темп зростання	$K_i = \frac{y_i}{y_{i-1}}$	$K_i = \frac{y_i}{y_0}$
3. Темп приросту	$KP_i = \frac{y_i - y_{i-1}}{y_{i-1}}$	$KP_i = \frac{y_i - y_0}{y_0}$
4. Середня арифметична значень	$\bar{y} = \frac{1}{n} \sum_{i=0}^n y_i$	
5. Середній темп зростання	$T = \sqrt[n]{\prod_{i=1}^n T_i}$	
6. Стандартне відхилення	$\sigma = \sqrt{\frac{1}{n} \sum_{i=0}^n (y_i - \bar{y})^2}$	
7. Коефіцієнт варіації	$l = \frac{\sigma}{\bar{y}} \cdot 100\%$	

Першим кроком дослідження часових рядів епідеміологічних даних є виявлення моменту початку епідемії. Для кожного моменту часу обчислюються кумулятивна сума випадків захворювання та відсоток захворілих від загальної кількості населення. Початком епідемії будемо вважати момент, коли кількість хворих перевищила 1 % всього населення (це, як приклад, при певній нозологічній формі може бути інший показник).

Для визначення однорідності часового ряду обчислюється коефіцієнт варіації l . Якщо $l > 33\%$, це свідчить про значну неоднорідність інформації.

Для перевірки нормальності часового ряду та виявлення аномальних значень (викидів) застосовується метод Ірвіна, який ґрунтується на порівнянні сусідніх значень ряду та розрахунку характеристики $\lambda_t = \frac{|y_t - y_{t-1}|}{\sigma}$, $t = 1, 2, \dots, n$, σ – стандартне відхилення.

Розрахункові значення λ_i порівнюють із критичним значенням $\lambda_{\alpha, n}$ і якщо вони їх перевищують, то відповідні рівні y_i вважаються аномальними. Аномальність виявляється як несподіваний стрибок (або спад) із подальшим поступовим встановленням попереднього рівня. Критичні значення для рівня значущості $\alpha = 0,05$ (помилка 5 %) наведено в *табл. 15*.

Таблиця 15

Критичні значення критерію Ірвіна

n	2	3	10	20	30	50	100
$\lambda_{\alpha, n}$	2,8	2,3	1,6	1,3	1,2	1,1	1,0

У разі виявлення значної неоднорідності або викидів рекомендується провести згладжування ряду і повторно перевірити на наявність викидів. Метою згладжування є «гасіння» помилок.

Два найбільш поширених методи згладжування рядів – це метод ковзних (або плинних) середніх і метод експоненційного згладжування.

Метод ковзних (або плинних) середніх полягає в тому, що кілька значень часового ряду, які йдуть один за одним, замінюються їхнім середнім значенням.

Нехай маємо часовий ряд з $n+1$ спостережень деякої випадкової величини $y_0, y_1, y_2, \dots, y_n$. Задаємо деяке значення k ($k < n+1$). Для кожних k послідовних значень ряду можна розрахувати середнє значення: $x_1 = \frac{y_0 + y_1 + \dots + y_{k-1}}{k}$. Далі переходимо до розрахунку середнього для наступних значень ряду: $x_2 = \frac{y_1 + y_2 + \dots + y_k}{k}$ тощо. Кількість даних k називають порядком ковзної середньої. Стандартне значення $k=3$. Основна ідея такого згладжування – замінити фактичні дані часового ряду розрахунковими даними. Ряд стає коротшим, але більш згладженим. Ковзні середні можна використовувати не тільки для згладжування ряду, але й для прогнозування: k останніх даних і буде прогнозом для наступної дати.

Обчислення за допомогою цього методу достатньо точно відображають зміни основних показників попереднього періоду. Однак метод ковзних середніх має свої недоліки:

- 1) якщо ряд має значні викиди, то вони погано гасяться;
- 2) відбувається втрата $(k-1)$ значень ряду;
- 3) якщо порядок k є парним, то дати розрахункових значень не співпадають з датами реальних спостережень;
- 4) вплив кожного доданку в розрахунку середнього значення вважається однаковим, хоч зрозуміло, що останні значення ряду несуть більше інформації, ніж початкові.

Метод експоненційного згладжування також полягає в тому, що кілька послідовних значень часового ряду усереднюються ковзною середньою, але з вагою ($w_1x_t + w_2x_{t+1} + w_3x_{t+2}$), де $w_1 + w_2 + w_3 = 1$. Вагові коефіцієнти w_1, w_2, w_3 беруть у вигляді експоненційних функцій. Це дозволяє надавати більшої ваги останнім значенням часового ряду, ніж початковим.

Під час дослідження часового ряду використовується так званий метод декомпозиції, який полягає в наступному: будь-який часовий ряд можна представити як суму детермінованого та випадкового компонентів: $y_t = d_t + \varepsilon_t$. Детермінований компонент у свою чергу можна представити як суму (адитивна модель) або добуток (мультиплікативна модель) трьох компонентів – трендової (tr_t), циклічної (c_t) та сезонної (s_t): $d_t = tr_t + s_t + c_t$ або $d_t = tr_t \times s_t \times c_t$.

Тренд (від англ. Trend – тенденція) – загальна тенденція при різнонаправленому русі, визначена загальною спрямованістю змін показника часового ряду. Тренд є результатом дії довготривалих факторів.

В програмі Microsoft Excel для виділення трендової складової часового ряду та прогнозування можна використовувати такі засоби:

- команда «Додати лінію тренду» при роботі з графіком значень часового ряду;
- функції FORECAST (лінійний прогноз), GROWTH (обчислення прогнозованого експоненційного росту на основі наявних даних), LINEST (параметри лінійного наближення), LOGEST (параметри експоненційного наближення), TREND (повертає значення відповідно до лінійного тренду);
- процедура «Регресія» надбудови «Аналіз даних»;
- процедура «Ковзне середнє» надбудови «Аналіз даних»;
- процедура «Експоненційне згладжування» надбудови «Аналіз даних».

Сезонна компонента характеризує циклічні коливання показника, пов'язані з календарними циклами (тиждень, місяць, квартал, рік тощо). Сезонні коливання мають періодичний або близький до нього характер упродовж одного року. Для більшості інфекцій характерний сезонний підйом захворюваності, і найчастіше близькі значення показників повторюються рік у рік. Виражена сезонність враховується за будь-якого прогнозу терміном більше місяця. У найпростішому випадку дані за визначений календарний період (тиждень, місяць року) розглядають окремо. Прогнозне значення на визначений календарний період обчислюється як середнє значення всіх даних, відповідних періоду, які накопичені в епідеміологічній базі даних за декілька років. Такий спосіб прогнозування передбачає, що щорічна захворюваність носить незмінний характер та спостереження протягом кожного нового року – наступна реалізація одного й того самого випадкового процесу. Такі моделі можуть успішно застосовуватися для отримання грубих середньострокових та довгострокових проєкцій захворюваності.

Циклічна компонента характеризує циклічні коливання показника, не пов'язані з календарними циклами. Для створення аналітичних моделей прогнозування циклічних процесів захворюваності можна використовувати циклічну функцію виду, запропоновану Робертом Серфлінгом у 1963 р., $c(t) = a_0 + \sum_{i=1}^k (a_i \cos it + b_i \sin it)$, коефіцієнти моделі визначаються за значеннями первинного часового ряду: $a_0 = \frac{1}{n} \sum_{j=1}^n y_{tj}$, $a_i = \frac{2}{n} \sum_{j=1}^n y_{tj} \cos it$, $b_i = \frac{2}{n} \sum_{j=1}^n y_{tj} \sin it$.

Часовий ряд називається стаціонарним, якщо його середнє значення та дисперсія не залежать від часу. Стандартне середнє значення = 0, дисперсія = 1. У стаціонарного часового ряду є тільки випадкова складова.

Якщо за значенням досліджуваного показника побудувати графік, то за ним легко відстежити наявність основних компонентів (рис. 97).

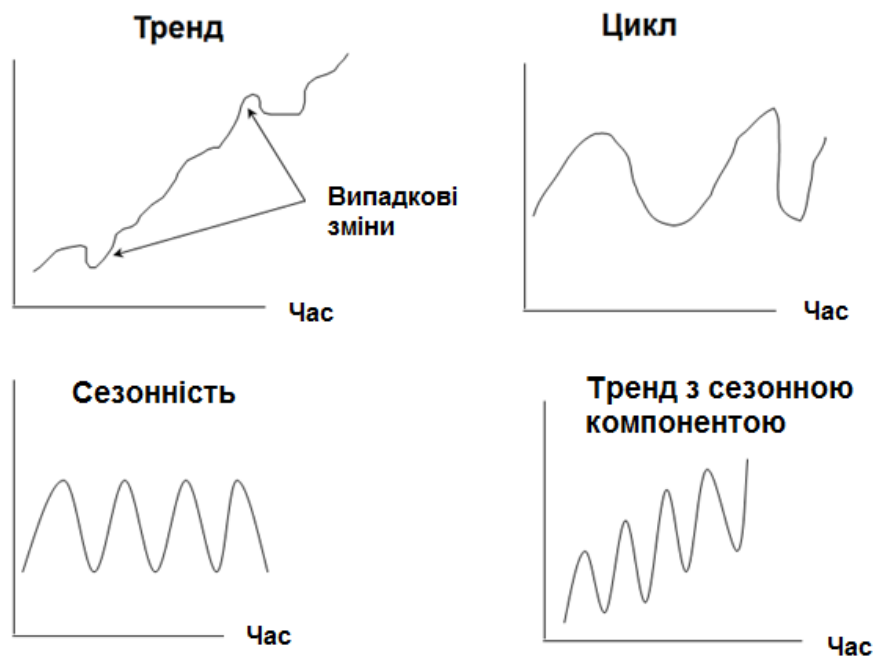


Рис. 97. Характеристика динамічного ряду захворюваності

Існує низка тестів перевірки наявності детермінованих компонентів часового ряду. Для перевірки наявності тренду часового ряду використовується тест на порівняння середніх рівнів ряду: ряд з n рівнів розбивається на 2 частини, в кожній з яких приблизно $n/2$ значень; за допомогою критерію Стьюдента перевіряється гіпотеза про суттєву відмінність середніх значень окремо взятих частин.

Навіть у разі негативного результату тесту наявності тренду в програмі Microsoft Excel можна побудувати графік значень часового ряду і спробувати підібрати форму тренду. Для цього треба виділити лінію з даними, викликати контекстне меню і виконати команду «Додати лінію тренду». Якість побудованого таким чином тренду визначається коефіцієнтом детермінації R^2 . Чим ближче R^2 до 1, тим якість краща.

Для перевірки наявності циклів у числовому ряді використовується критерій екстремальних точок. Нехай маємо $n+1$ спостережень деякої випадкової величини $y_0, y_1, y_2, \dots, y_n$. Для визначення екстремальної точки потрібні три послідовні спостереження. Екстремальною точкою («піком» або «ямою») називають значення, яке більше або менше і за попереднє, і за наступне. Індикатор екстремальної точки можна задати за допомогою функції :

$$x_i = \begin{cases} 1 & y_{i-1} < y_i \text{ і } y_{i+1} < y_i \text{ або } y_{i-1} > y_i \text{ і } y_{i+1} > y_i \text{ — умова (I)} \\ 0 & \text{якщо не виконується умова (I)} \end{cases}$$

$i = 1, \dots, n - 1$. Інтервал між екстремальними точками називають фазою. Рекомендується додатково переглянути перелік екстремальних точок і якщо у ряді виявилися дві рядом розташовані, то вважати їх за одну. Загальна кількість

екстремальних точок обчислюється за формулою $E = \sum_{i=1}^{n-1} x_i$. Для визначення статистики критерію окрім E розраховуються додаткові змінні $M(n) = \frac{2}{3}(n-1)$, $D(n) = \frac{16(n+1)-29}{90}$. Статистика критерію екстремальних точок $S = \frac{E-M(n)}{\sqrt{D(n)}}$.

Критичне значення береться з таблиці квантилів нормального розподілу при заданому α , нульовому середньому значенні та одиничній дисперсії $S_{кр} = u_{\alpha/2}$. Якщо $|S| \leq S_{кр}$ то часовий ряд випадковий, в протилежному випадку можна стверджувати, що ряд має цикли. Періодом коливань називається проміжок часу між двома послідовними максимальними відхиленнями від положення рівноваги. Найпростіший спосіб визначення приблизного періоду циклу – за допомогою графіка значень часового ряду. Більш точне значення обчислюється за допомогою спектрального аналізу Фур'є.

При дослідженні часового ряду потрібно визначити регулярні складові, задати їх рівняннями та одержати загальний вираз моделі ряду. Якщо під час формування аналітичного виразу моделі ряду виникають труднощі, часовий ряд можна задавати як систему рекурентних відношень, що відбивають залежність значень показника у деякий момент часу від значень того ж показника у попередні моменти. Така модель називається авторегресійною.

Авторегресійні моделі різних порядків – першого, другого, в загальному випадку k -ого можна описати рівняннями наступного вигляду:

$$Y_i = b_0 + b_1 Y_{i-1} + \varepsilon,$$

$$Y_i = b_0 + b_1 Y_{i-1} + b_2 Y_{i-2} + \varepsilon,$$

$$Y_i = b_0 + b_1 Y_{i-1} + b_2 Y_{i-2} + \dots + b_k Y_{i-k} + \varepsilon,$$

де Y_i – значення часового ряду в момент часу i ; $b_0, b_1, b_2, \dots, b_k$ – коефіцієнти авторегресії; ε – нескорельована випадкова величина, пов'язана з помилками спостереження і похибками моделі. Порядок авторегресії – це число попередніх значень у ряду, які використовуються для прогнозування значення в даний момент.

У 1960-х рр. науковці почали активно досліджувати сезонні та циклічні характеристики часових рядів. Д. Бокс та Г. Дженкінс систематизували результати цих робіт та запропонували комплексний підхід до моделювання і прогнозування часових рядів на підставі авторегресійних моделей та моделей ковзного середнього (ARMA/ARIMA – autoregressive integrated moving average), який використовують для опису і прогнозування процесів, що здійснюють випадкові коливання навколо певного середнього значення.

Необхідною умовою для побудови авторегресії є кореляційна залежність між послідовними рівнями часового ряду, яку називають автокореляцією рівнів ряду. Кількісно її можна визначити за допомогою лінійного коефіцієнта кореляції між рівнями цього ж ряду, зсунутими на декілька кроків у часі. Кількість періодів, за якими розраховується коефіцієнт кореляції, називається «лагом». Послідовність коефіцієнтів автокореляції рівнів першого, другого і більших порядків називають автокореляційною функцією часового ряду. Графік залежності її значень від величини лага (порядку коефіцієнта автокореляції) називається корелограмою. За допомогою аналізу автокореляційної функції та корелограми можна виявити структуру ряду. Якщо найвищим виявився коефіцієнт автокореляції першого

порядку, досліджуваний ряд містить лише тенденцію. Якщо найбільшим виявився коефіцієнт автокореляції порядку τ , ряд містить циклічні коливання з періодичністю τ моментів часу.

В останніх версіях табличного процесора Microsoft Excel (з 2016 р.) реалізована нова процедура прогнозування часового ряду за допомогою методу експоненційного згладжування – засіб «Аркуш прогнозу», розташований на вкладці «Дані». Побудова і дослідження авторегресійних моделей із застосуванням спектрального аналізу і автокореляційної функції реалізовані в програмі Statistica.

Задача-еталон

Дано: Кількість випадків захворювання COVID-19 у Харківській області щотижнево за 29.02.2020–21.01.2022 р.

Чисельність населення в Харківській області (осіб) – 2 658 461.

Дані знаходяться в файлі Lesson 06D.xlsx, аркуш «Первинні дані».

Таблиця 16

Структура даних для аналізу

№ з/п	Назва показника	Примітка
1	Номер тижня за порядком	
2	Початок тижня	Наданий у форматі дати ДД.ММ.РРРР
3	Кінець тижня	Наданий у форматі дати ДД.ММ.РРРР
4	Рік	Може бути обчислений за формулою: =Рік (Дата початку тижня)
5	Номер тижня в році	Може бути обчислений за формулою: =Номер тижня (Дата початку тижня)
6	Місяць	Може бути обчислений за формулою: =Місяць(Дата початку тижня)
7	Сезонність	Може бути обчислений за формулою: $\text{Сезон} = \begin{cases} \text{Зима} & \text{місяць} < 3 \\ \text{Весна} & 3 \leq \text{місяць} < 6 \\ \text{Літо} & 6 \leq \text{місяць} < 9 \\ \text{Осінь} & 9 \leq \text{місяць} < 12 \\ \text{Зима} & \text{місяць} > 11 \end{cases}$
8	Кількість випадків захворювання	Число
9	Інтенсивний показник захворюваності на 100 000 населення	Обчислюється за формулою: =Кількість випадків*100000/Чисельність населення

Вправа 1. Обчислити значення показників 4, 5, 6, 7, 9.

Алгоритм виконання:

1. Відкрити файл Lesson 06D.xlsx. Зберегти як Lesson 06R.xlsx.
2. Для обчислення показника «рік» в комірку таблиці D3 ввести формулу =YEAR(B3).
3. Для обчислення показника «номер тижня в році» в комірку таблиці E3 ввести формулу =WEEKNUM(B3).
4. Для обчислення показника «місяць» в комірку таблиці F3 ввести формулу =MONTH(B3).
5. Для обчислення показника «сезон» в комірку таблиці G3 ввести формулу =IF(F3<3;"Зима";IF(F3<6;"Весна";IF(F3<9;"Літо";IF(F3<12;"Осінь";"Зима")))).

6. Для обчислення показника «інтенсивний показник захворюваності на 100 000 населення» в комірку таблиці I3 ввести формулу $=H3*100000/SH\$1$.

7. Скопіювати формули з комірок D3: G3 до діапазону D4:G101.

8. Скопіювати формулу з комірки I3 до діапазону I4:I101.

9. Зберегти зміни у файлі Lesson 06R.xlsx.

Вправа 2. Виявити тиждень, коли кількість хворих перевищила 1 % населення області.

Алгоритм виконання:

1. Для обчислення кумулятивної суми випадків захворювання в комірку таблиці J3 ввести формулу $=H3$. В комірку J4 ввести формулу $=J3+H4$. Скопіювати формулу з комірки J4 до діапазону J5:J101.

2. Для обчислення відсотка хворих на початок тижня в комірку K3 ввести формулу $=J3/SH\$1$. Скопіювати її до діапазону K4:K101.

3. Виділити діапазон комірок K3:K101, встановити для комірок цього діапазону правило умовного форматування: якщо значення комірки перевищує 1 %, заливка комірки має бути рожевою, а колір тексту червоним. Таким чином можна встановити, що кількість хворих перевищить 1 % населення області на 28-му тижні від початку спостереження.

4. За допомогою команди «Дані → Фільтрація → Користувацький автофільтр» відібрати рядки таблиці первинних даних за умовою – значення показника (%) захворілих від усього населення більше за 1 % (рис. 98).

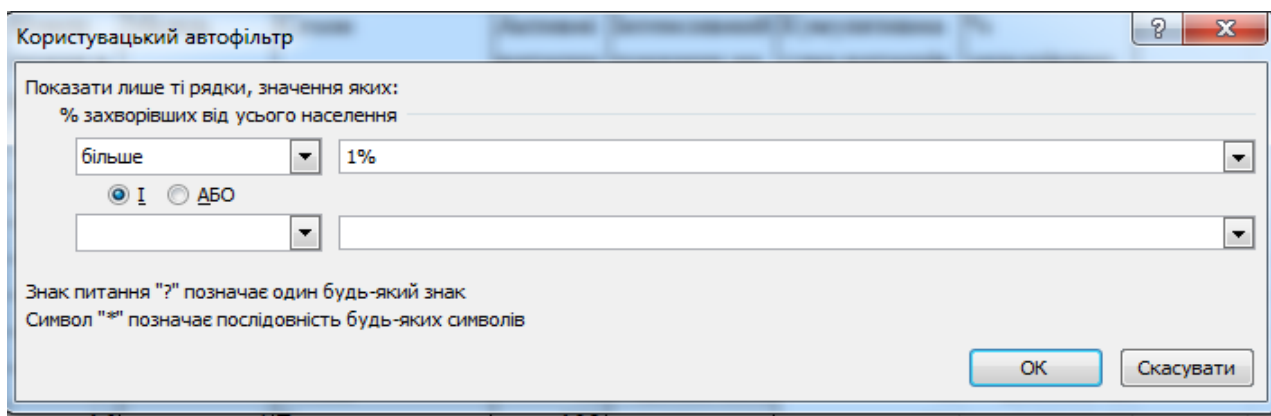


Рис. 98. Скріншот вікна налаштування параметрів користувацького автофільтра

5. Відібрану інформацію скопіювати до нового аркушу файлу Lesson 06R.xlsx. Аркуш перейменувати на «72 значення».

6. Зберегти зміни у файлі Lesson 06R.xlsx.

Вправа 3. Обчислити числові характеристики інтенсивного показника захворюваності. Дослідити наявність аномальних значень часового ряду. Провести згладжування часового ряду методом ковзної середньої.

Алгоритм виконання:

1. Зробити активним аркуш «72 значення».

2. Обчислити числові характеристики інтенсивного показника захворюваності (за діапазоном I3:I74):

2.1. У комірці I76 середнє значення – формула $=AVERAGE(I3:I74)$, результат 381,7167777.

- 2.2. У комірці I77 мінімум – формула =MIN(I3:I74), результат 20,01157813.
 2.3. У комірці I78 медіана – формула =MEDIAN(I3:I74), результат 304,0292861.
 2.4. У комірці I79 максимум – формула =MAX(I3:I74), результат 1280,703384.
 2.5. У комірці I80 стандартне відхилення – формула =STDEV.S(I3:I74), результат 310,5923454.
 2.6. У комірці I81 коефіцієнт варіації – формула =I80/I76, результат 0,813672239.

Висновки:

- Коефіцієнт варіації значно перевищує 0,33, що свідчить про значну неоднорідність даних.

- Мінімум значно менший за середнє значення і медіану.
- Максимум значно більший за середнє значення і медіану.

3. У діапазоні L4:L74 обчислити абсолютний приріст інтенсивного показника захворюваності. Для цього в комірку L4 ввести формулу =I4-I3. Скопіювати формулу до діапазону L5:L74.

4. У діапазоні M4:M74 обчислити значення критерію Ірвіна. Для цього в комірку M4 ввести формулу =ABS(L4)/\$I\$80. Скопіювати формулу до діапазону M5:M74.

5. У діапазоні N4:N74 обчислити індикатор аномального значення. Для цього у комірку N4 ввести формулу =IF(M4<1,05;" ";"V"). Скопіювати її до діапазону N5:N74. Результат: виявлене аномальне значення на 91-му тижні (20.11.2021–26.11.2021).

Оскільки виявлений високий коефіцієнт варіації первинних даних та аномальне значення, необхідно провести згладжування даних.

6. Виконати команду «Дані → Аналіз даних → Ковзне середнє» (Data Analysis → Moving Average). У вікні процедури встановити наступні параметри (рис. 99):

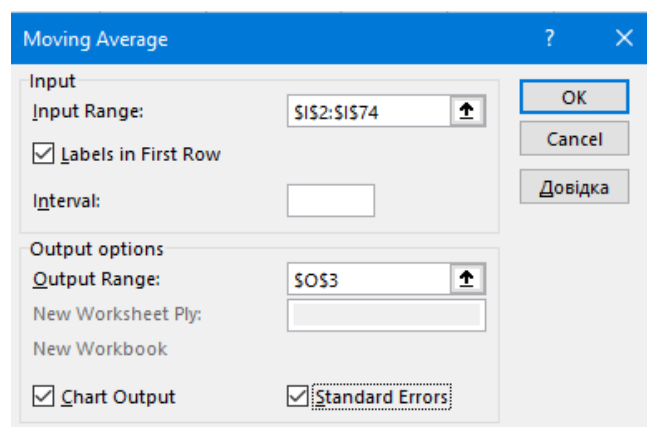


Рис. 99. Налаштування параметрів процедури Microsoft Excel «Аналіз даних → Ковзне середнє»

Вхідний і вихідний інтервал розташовані на аркуші «72 значення».

7. Після закриття вікна процедури на аркуші «72 значення» починаючи з комірки O3 з'явиться замовлена інформація:

7.1. У діапазоні O3:O74 – нові значення інтенсивного показника захворюваності після процедури згладжування порядку 3. Кількість значень зменшилася на 2. Значення розташовані в діапазоні O5:O74.

7.2. У діапазоні P3:P74 – стандартні похибки процедури згладжування.

7.3. Побудований графік первинних даних і прогнозованих за допомогою згладжування методом ковзного середнього (рис. 100).

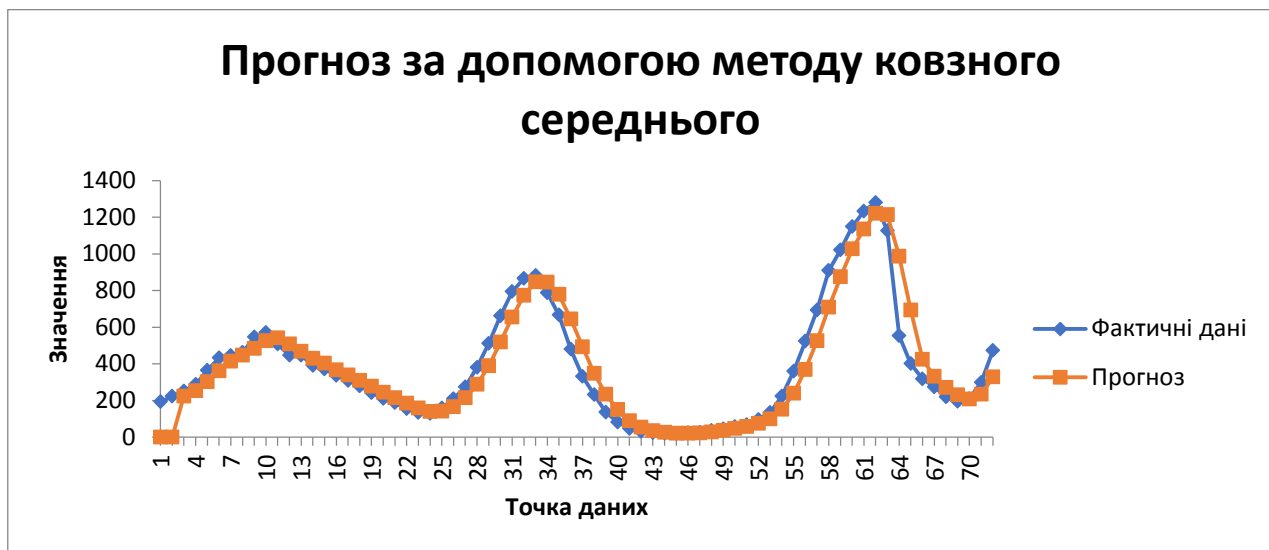


Рис. 100. Графік значень інтенсивного показника захворюваності після процедури згладжування

Після процедури згладжування необхідно повторити перевірку нових даних на однорідність та наявність аномальних значень.

8. Обчислити числові характеристики інтенсивного показника захворюваності (за діапазоном O5:O74):

8.1. У комірці O76 середнє значення – формула =AVERAGE(O5:O74), результат 383,7874 (було 381,7167777).

8.2. У комірці O77 мінімум – формула =MIN(O5:O74), результат 21,50367 (було 20,01157813).

8.3. У комірці O78 медіана – формула =MEDIAN(O5:O74), результат 317,9847 (було 304,0292861).

8.4. У комірці O79 максимум – формула =MAX(O5:O74), результат 1220,97 (було 1280,703384).

8.5. У комірці O80 стандартне відхилення – формула =STDEV.S(O5:O74), результат 302,834 (було 310,5923454).

8.6. У комірці O81 коефіцієнт варіації – формула =O80/O76, результат 0,789067 (було 0,813672239).

Висновки. Коефіцієнт варіації зменшився, але значно перевищує 0,33, що свідчить про неоднорідність даних і може вказувати на якість діагностики чи реєстрації захворювань.

9. Провести розрахунки для виявлення аномальних значень:

9.1. Абсолютний приріст – результат у комірках Q6:Q74.

9.2. Відношення критерію Ірвіна – результат у комірках R6:R74.

9.3. Індикатор аномальних точок – результат у комірках S6:S74.

Аномальних точок не виявлено. Далі будемо досліджувати ряд значень з діапазону O5:O74.

10. Обчислити числові характеристики часового ряду з діапазону O5:O74:

10.1. У діапазоні Т6:Т74 обчислити темп приросту. Для цього в комірку Т6 ввести формулу =Q6/O5. Скопіювати формулу до діапазону Т7:Т74.

10.2. У діапазоні U6:U74 обчислити темп приросту. Для цього в комірку U6 ввести формулу =O6/O5. Скопіювати формулу до діапазону U7:U74.

10.3. У комірці U76 обчислити середній темп зростання. Для цього в комірку U76 ввести формулу =GEOMEAN(U6:U74). Результат 1,00560102.

Висновок. Оскільки середній темп зростання часового ряду майже не відрізняється від 1, можна припустити стабілізацію епідемічної ситуації.

11. Зберегти зміни у файлі Lesson 06R.xlsx.

Вправа 4. Перевірити гіпотезу про наявність тренду часового ряду інтенсивного показника захворюваності за діапазоном O5:O74.

Алгоритм виконання:

1. Зробити активним аркуш «72 значення».

2. Для перевірки гіпотези поділити значення ряду на 2 частини по 35 значень:

2.1. У комірках A80:A114 записати числа від 1 до 35.

2.2. Скопіювати значення з комірок O5:O39 до діапазону B80:B114 (скористатися засобом «Спеціальна вставка → Значення»).

2.3. Скопіювати значення з комірок O40:O74 до діапазону C80:C114 (скористатися засобом «Спеціальна вставка → Значення»).

3. Обчислити числові характеристики частин числового ряду:

3.1. У комірці B116 – середнє значення 1-ї частини, формула =AVERAGE(B80:B114), значення 409,988371.

3.2. У комірці B117 – дисперсія 1-ї частини, формула =VAR.S(B80:B114), значення 40582,9098.

3.3. У комірці C116 – середнє значення 1-ї частини, формула =AVERAGE(C80:C114), значення 357,58641.

3.4. У комірці C117 – дисперсія 1-ї частини, формула =VAR.S(C80:C114), значення 144117,953.

4. Перевірити гіпотезу про відмінність дисперсій. Скористатися процедурою «Дані → Аналіз даних → Двовибірковий тест для дисперсії» (Data Analysis → F-Test Two-Sample for Variances) (рис. 101).

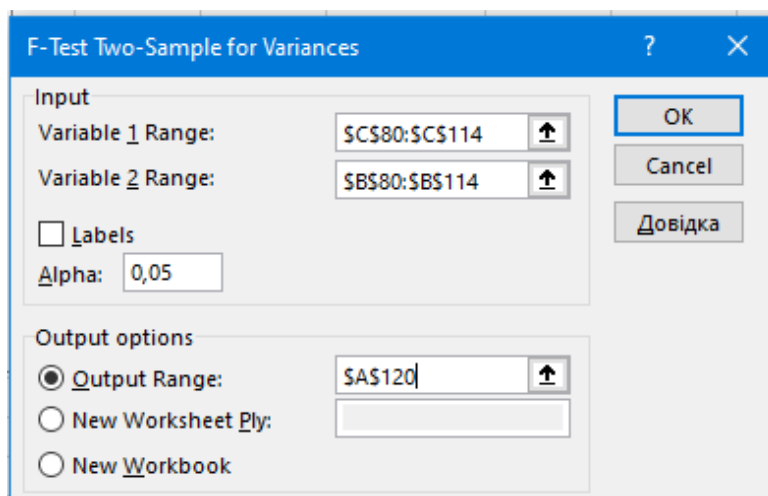


Рис. 101. Налаштування параметрів процедури Microsoft Excel «Аналіз даних → Двовибірковий тест для дисперсії»

Інтервал змінної 1 – це інтервал з більшою дисперсією.
Результати виконання процедури представлені в *табл. 17*.

Таблиця 17

Результати розрахунку двовибіркового F-тесту для дисперсії

Статистичні характеристики	Змінна 1	Змінна 2
Середнє	357,58641	409,988371
Дисперсія	144117,953	40582,9098
Спостереження	35	35
df	34	34
F	3,55119812	
P (F<=f) одностороннє	0,00018843	
F критичне одностороннє	1,77206648	

Оскільки $F > F$ критичне одностороннє, можна стверджувати, що дисперсії різні.

5. Перевірити гіпотезу про рівність середніх значень, скориставшись процедурою «Дані → Аналіз даних → Двовибірковий t-тест з різними дисперсіями» (Data Analysis → t-Test Two-Sample Assuming Unequal Variances) (*рис. 102*).

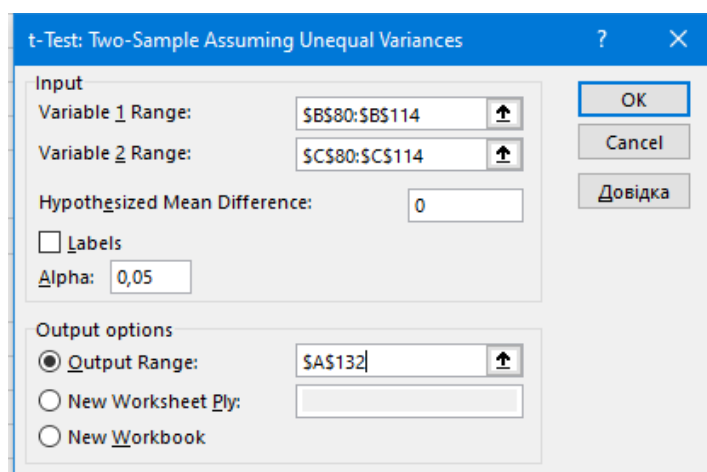


Рис. 102. Налаштування параметрів процедури Microsoft Excel «Аналіз даних → Двовибірковий t-тест з різними дисперсіями»

Результати виконання процедури представлені в *табл. 18*.

Таблиця 18

Результати розрахунку двовибіркового t-тесту з різними дисперсіями

Статистичні характеристики	Змінна 1	Змінна 2
Середнє	409,988371	357,58641
Дисперсія	40582,9098	144117,953
Спостереження	35	35
Гіпотетична різниця середніх	0	
Df	52	
t-статистика	0,72135178	
P(T<=t) одностороннє	0,23696206	
t критичне одностороннє	1,67468915	
P(T<=t) двостороннє	0,47392411	
t критичне двостороннє	2,00664681	

Оскільки $t < t$ критичне одностороннє, $t < t$ критичне двостороннє, можна стверджувати, що середні значення рівні, тож часовий ряд показує стабільність епідемічного процесу.

Висновок. Можна стверджувати, що запроваджені жорсткі режимно-обмежувальні заходи щодо COVID-19 в Україні з 2020 р. вплинули на перебіг епідемічного процесу і протягом майже двох років (з 29.02.2020 по 21.01.2022 р.) захворюваність була приблизно на однаковому рівні.

6. Зберегти зміни у файлі Lesson 06R.xlsx.

Вправа 5. Дослідити часовий ряд інтенсивного показника захворюваності (дані діапазону O5:O74) на наявність циклічної компоненти за критерієм екстремальних точок.

Алгоритм виконання:

1. Зробити активним аркуш «72 значення».

2. У діапазоні V6:V74 обчислити логічну змінну, яка відповідає результату перевірки умови екстремальності значення. Для цього в комірку V6 ввести формулу =OR(AND(O5<O6;O7<O6);AND(O5>O6;O7>O6)). Скопіювати формулу до діапазону V7: V74.

3. У діапазоні W6:W74 обчислити індикатор екстремальності значення. Для цього в комірку W6 ввести формулу =IF(V6;1;0). Скопіювати формулу до діапазону W7: W74.

4. У комірці W75 обчислити кількість екстремальних значень (7). Скорегувати значення, виключивши останню точку. Таким чином E = 6 (комірка W76).

5. Обчислити $M(n) = \frac{2}{3}(n - 1)$. n = 70, M(n) = 46 (комірка W77, формула =2*69/3).

6. Обчислити $D(n) = \frac{16(n+1)-29}{90}$, n = 70, D(n) = 12,3 (комірка W78, формула =(16*71-29)/90)

7. Обчислити $S = \frac{E-M(n)}{\sqrt{D(n)}}$, S = 11,405 (комірка W79, формула =ABS((W76-W77)/SQRT(W78))).

8. Обчислити $S_{\text{крит.}} = (-1,96)$ (комірка W80, формула =NORM.S.INV(0,05/2)).

9. Оскільки $S > S_{\text{крит.}}$, можна стверджувати, що часовий ряд має циклічну компоненту.

10. В Microsoft Excel (версія 2016 і вище) для визначення періоду циклічної компоненти можна скористатися функцією FORECAST.ETS.SEASONALITY (рис. 103).

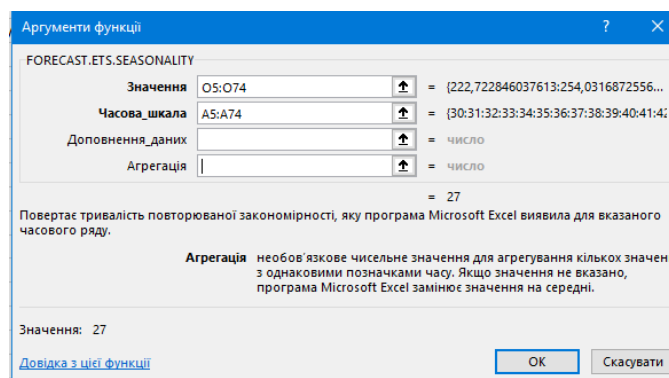


Рис. 103. Вікно функції FORECAST.ETS.SEASONALITY, яка повертає значення періоду циклічної компоненти

Період дорівнює 27 тижням, що відповідає приблизно половині року.

11. Зберегти зміни у файлі Lesson 06R.xlsx.

Вправа 6. За допомогою засобу Microsoft Excel (версія 2016 і вище) «Аркуш прогнозу» побудувати інтервальний прогноз інтенсивного показника захворюваності (дані діапазону O5:O74). Складові вкладки «Дані» представлені на *рис. 104*.

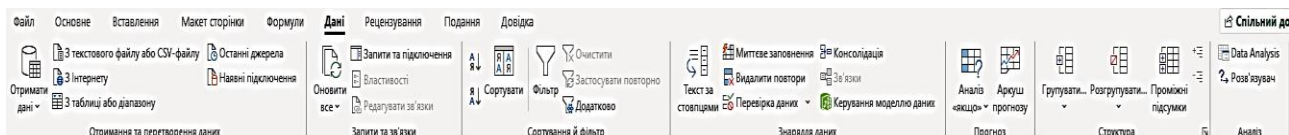


Рис. 104. Перелік засобів вкладки «Дані» Microsoft Excel (версія 2016 і вище)

Алгоритм виконання:

1. Зробити активним аркуш «72 значення».
2. В стовпцях Y, Z, AA сформувати область даних для запуску процедури «Аркуш прогнозу» (*рис. 105*).

Y	Z	AA
Номер тижня	Тиждень	Інтенсивний показник захворюваності
1	19.09.2020	222,722846
2	26.09.2020	254,0316873
3	03.10.2020	301,4275302
4	10.10.2020	361,6503433

Рис. 105. Зразок області даних для запуску процедури «Аркуш прогнозу»

Структура області:

- 2.1. У діапазоні Y3:Y72 числа від 1 до 70.
 - 2.2. Скопіювати значення дати початку тижня з діапазону B5:B74 до діапазону Z3:Z72.
 - 2.3. Скопіювати значення інтенсивного показника захворюваності з діапазону O5:O74 до діапазону AA3:AA72.
3. Виконати команду «Дані → Аркуш прогнозу». На першому етапі налаштовуються параметри прогнозування (*рис. 106*).

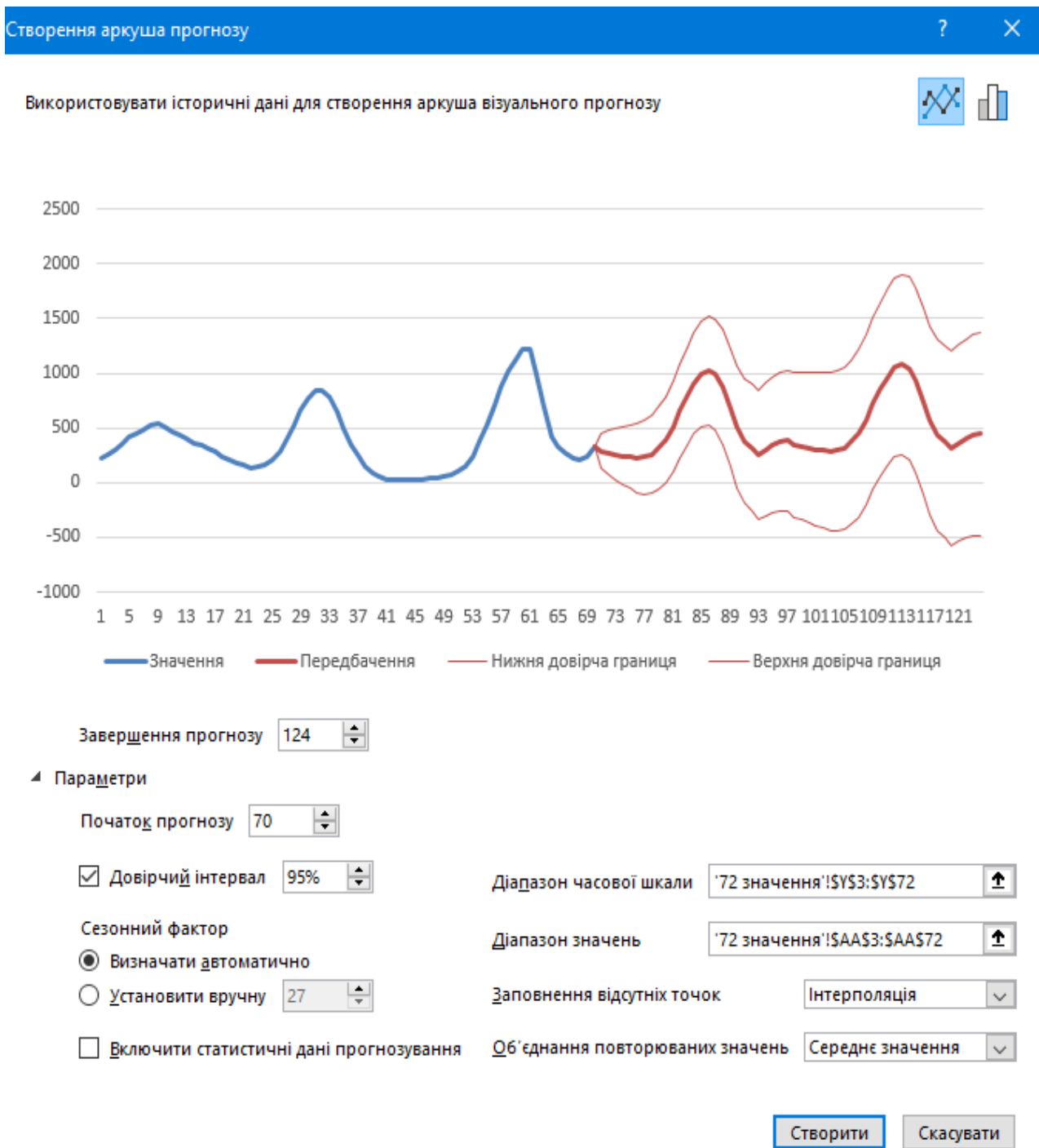


Рис. 106. Налаштування параметрів процедури Microsoft Excel «Аркуш прогнозу»

За замовчанням прогноз будується на 54 тижні, що дорівнює 2 періодам досліджуваного часового ряду. Точку завершення прогнозу можна скорегувати. У разі необхідності можна корегувати інші параметри процедури. Після натискання на клавішу «Створити» запускається процедура експоненційного згладжування, за допомогою якої у вигляді таблиці будується прогноз і його довірчі межі. Результат виводиться на новий аркуш книги. До 70-ї позиції в таблицю виводяться первинні дані з діапазону AA3:AA72 (рис. 107, 108).

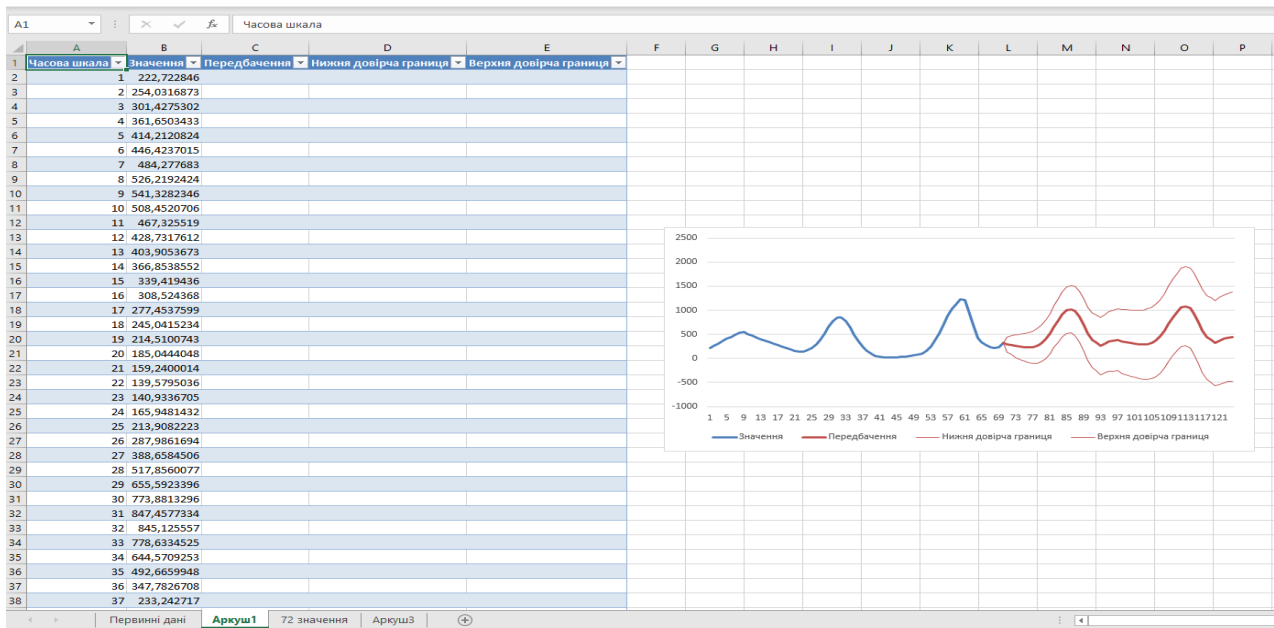


Рис. 107. Вміст «Аркушу прогнозу», таблиця з первинними даними (1-а частина)

Часова шкала	Значення	Передбачення	Нижня довірча границя	Верхня довірча границя	F
67	66	271,2220843			
68	67	230,1833028			
69	68	208,5793247			
70	69	234,6470383			
71	70	327,4450895	327,4450895	327,45	327,45
72	71		288,3566474	128,30	448,41
73	72		273,1095125	72,94	473,28
74	73		258,7634522	25,19	492,33
75	74		244,7547915	-18,08	507,59
76	75		233,8507439	-55,36	523,07
77	76		226,9688352	-86,48	540,42
78	77		234,451314	-101,54	570,44
79	78		259,0427177	-98,13	616,21
80	79		308,9287539	-68,29	686,14
81	80		389,969711	-6,33	786,26
82	81		515,9159511	101,37	930,46
83	82		658,2495235	226,18	1090,32
84	83		791,5678069	342,61	1240,52
85	84		908,1537708	442,89	1373,42
86	85		997,7557949	516,69	1478,82
87	86		1021,896319	525,49	1518,30
88	87		985,9511245	474,63	1497,27
89	88		869,4027444	343,55	1395,26
90	89		682,8000184	142,77	1222,83
91	90		511,0356617	-42,84	1064,92
92	91		377,1453643	-190,28	944,57
93	92		314,6930972	-266,00	895,38
94	93		254,3009787	-339,39	847,99
95	94		306,1836066	-300,26	912,63
96	95		346,8044544	-272,17	965,78
97	96		371,6772905	-259,60	1002,96
98	97		387,3962336	-255,98	1030,78
99	98		348,4105658	-315,60	1012,42
100	99		333,1634309	-342,42	1008,75
101	100		318,8173706	-368,18	1005,81
102	101		304,8087099	-393,44	1003,05
103	102		293,9046623	-415,44	1003,25
104	103		287,0227536	-433,28	1007,33

Рис. 108. Вміст «Аркушу прогнозу», таблиця з прогнозом і його довірчими межами (2-а частина)

За розрахованими значеннями побудований графік (рис. 109).

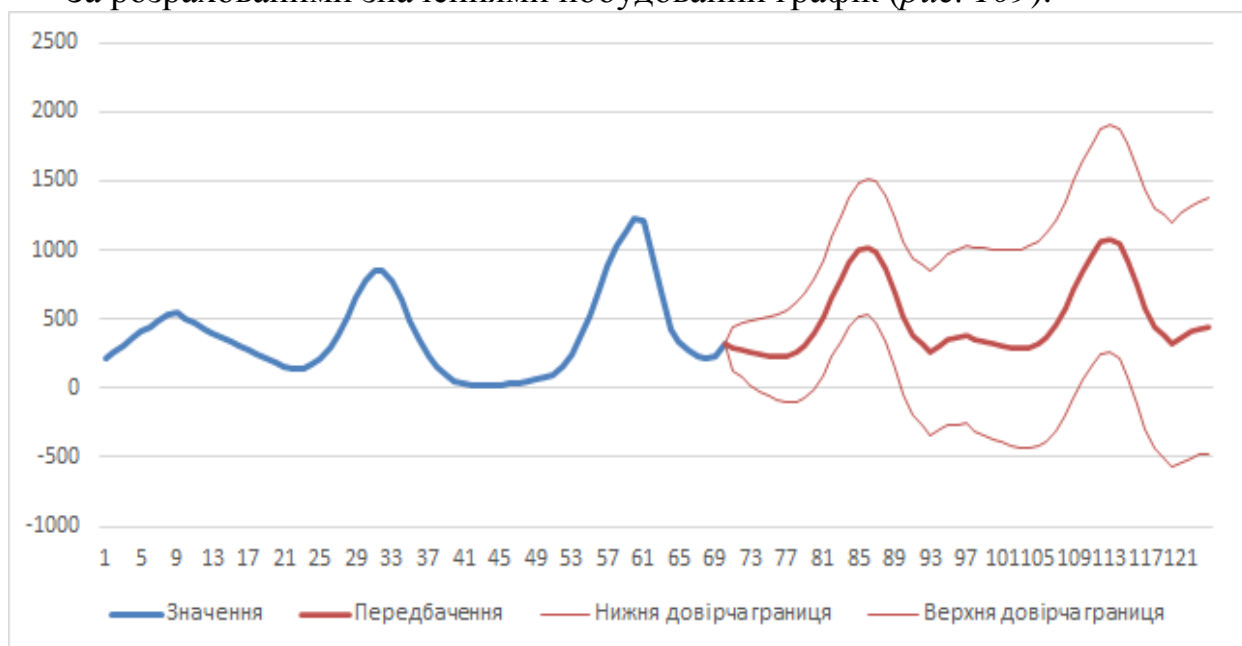


Рис. 109. Прогноз захворюваності на COVID-19, побудований за допомогою засобу «Аркуш прогнозу» програми Microsoft Excel

4. Перейменувати аркуш на «Аркуш прогнозу».

5. Зберегти зміни у файлі Lesson 06R.xlsx.

Висновки. Прогнозування за допомогою засобу «Аркуш прогнозу» програми Microsoft Excel підтвердило факт відсутності тренду досліджуваного часового ряду. Епідемічний процес розвивався циклічно з періодом коливань 27 тижнів. Процедура прогнозування визначила значення і довірчі межі для захворюваності на наступні після 15.01.2022 р. 54 тижні.

ТЕСТИ ДЛЯ ПІДСУМКОВОГО КОНТРОЛЮ РІВНЯ ЗНАНЬ

1. Епідеміологічна діагностика – це:
 - A. Сукупність науково обґрунтованих рекомендацій щодо профілактики епідемії.
 - B. Система спостереження за епідемічним процесом.
 - C. Система розпізнавання захворюваності й епідемічного стану населення.
 - D. Масові обстеження населення для отримання певних даних.
 - E. Системний аналіз у теорії управління протиепідемічною системою.
2. Вкажіть основні напрямки ретроспективного епідеміологічного аналізу.
 - A. Санітарно-епідеміологічна експертиза території.
 - B. Аналіз багаторічної динаміки захворюваності населення.
 - C. Аналіз річної (помісячної) динаміки захворюваності.
 - D. Аналіз захворюваності за факторами ризику.
 - E. Аналіз якості та ефективності профілактичних заходів.
 - F. Оперативне обстеження осередків з поодинокими випадками захворювання.
3. Соціальна значущість інфекційних хвороб оцінюється за:
 - A. Тенденцією захворюваності.
 - B. Витратами на стаціонарне лікування захворілих.
 - C. Показниками смертності та інвалідизації.
 - D. Розповсюдженням інфекційної захворюваності.
 - E. Витратами на протиепідемічні заходи.
4. До непрямих економічних збитків від інфекційних хвороб відносять:
 - A. Затрати на лікування захворілих.
 - B. Виплати з тимчасової непрацездатності у зв'язку з інфекційною хворобою.
 - C. Вартість недоотриманої суспільством продукції в результаті втрати працездатності від інфекційної хвороби.
 - D. Витрати на проведення протиепідемічних заходів в осередках інфекційних хвороб.
 - E. Виплати по інвалідності у зв'язку з інфекційною хворобою.
5. Який з видів графічного зображення можна використовувати для відображення динаміки захворюваності:
 - A. Стовпчикова діаграма.
 - B. Картограма.
 - C. Лінійна діаграма.
 - D. Бульбашкова діаграма.
 - E. Секторна діаграма.
6. За допомогою картограми можна наочно зобразити:
 - A. Динаміку захворюваності.
 - B. Захворюваність, яка має циклічну закономірність.
 - C. Щомісячні коливання захворюваності.
 - D. Структуру захворюваності.
 - E. Рівень захворюваності по територіям.
7. Напрямок профілактики хвороби можна визначити на етапі:
 - A. Збору і первинної обробки епідеміологічної інформації.
 - B. Дескриптивного епідеміологічного дослідження.
 - C. Аналітичного епідеміологічного дослідження.
 - D. Епідеміологічного експерименту.
 - E. Математичного моделювання епідемічного процесу.

8. Для аналізу рівня та частоти захворюваності використовують такі показники:
- Екстенсивні епідеміологічні показники.*
 - Середні арифметичні величини.*
 - Абсолютні числа кількості захворювань.*
 - Інтенсивні епідеміологічні показники.*
9. При плануванні протиепідемічних заходів за ознакою спрямованості на епідемічний процес виділяють наступні групи:
- Група заходів, спрямованих на джерело збудників інфекції.*
 - Група заходів, спрямованих на шляхи і фактори передачі інфекції.*
 - Група заходів, спрямованих на сприйнятливість людей до інфекції.*
 - Група загальних заходів.*
10. Одна з основних ознак, за якою проводять вивчення розподілу захворюваності по території:
- Клімат території.*
 - Ландшафт території.*
 - Адміністративний поділ території.*
 - Рівень економічного розвитку території.*
 - Рівень медичного обслуговування території.*
11. Прогнозування проявів епідемічного процесу на основі використання математичних формул, виведених лише з цифрових матеріалів, що характеризують фактичні прояви епідемічного процесу, це:
- Логічне моделювання.*
 - Формально-математичне моделювання.*
 - Кількісне моделювання.*
 - Інтуїтивне моделювання.*
 - Фізичне моделювання.*
12. Епідемічна тенденція – це:
- Закономірне підвищення захворюваності в певні місяці року.*
 - Регулярні коливання рівня захворюваності в багаторічній динаміці.*
 - Короткочасне підвищення захворюваності на обмеженій території.*
 - Регулярні коливання рівня захворюваності в річній динаміці.*
 - Основна спрямованість змін інтенсивності епідемічного процесу за ряд років.*
13. Вкажіть види статистичних досліджень, що залежать від часу реєстрації даних:
- | | |
|----------------------|---------------------------|
| <i>А. Поточне.</i> | <i>Д. Ретроспективне.</i> |
| <i>В. Суцільне.</i> | <i>Е. Оперативне.</i> |
| <i>С. Вибіркове.</i> | <i>Ф. Одночасне.</i> |
14. Вкажіть види статистичних досліджень, що залежать від ступеня охоплення об'єкта дослідження:
- | | |
|----------------------|-------------------------|
| <i>А. Вибіркове.</i> | <i>Д. Суцільне.</i> |
| <i>В. Одночасне.</i> | <i>Е. Оперативне.</i> |
| <i>С. Поточне.</i> | <i>Ф. Проспективне.</i> |
15. Статистична методологія включає:
- Загальні поняття та категорії статистики.*
 - Методи збору та систематизації даних.*
 - Набір статистичних характеристик.*

- D. Методи порівняння центральних тенденцій.*
- E. Методи виявлення статистичних зв'язків.*
- F. Методи прогнозування.*
- G. Математичне моделювання.*
- H. Методи керування.*

16. Формами статистичного спостереження є:

- A. Статистична звітність.*
- B. Спеціальні статистичні дослідження.*
- C. Скринінг.*
- D. Витяг даних з державних реєстрів.*
- E. Спілкування в соціальних мережах.*

17. Яким вимогам повинен відповідати первинний матеріал для розрахунку статистичних характеристик епідеміологічного показника?

- A. Достатня амплітуда значень.*
- B. Відсутність коливань значень.*
- C. Наявність групових ознак.*
- D. Достатня кількість спостережень.*
- E. Якісно однорідна сукупність.*

18. Що з переліченого є недоліком спостереження?

- A. Невелика кількість даних.*
- B. Особистісні особливості спостерігача.*
- C. Єдина форма для збору інформації.*
- D. Велика кількість показників.*
- E. Наявність атрибутивних даних.*

19. Статистика вивчає ознаки, які:

- A. Мають кількісний вираз (кількісні).*
- B. Не мають кількісного виразу (атрибутивні).*
- C. Кількісні і атрибутивні.*

20. Середня величина кількісної ознаки – це:

- A. Значення ознаки, що знаходиться в середині ряду розподілу.*
- B. Значення ознаки, що трапляється найчастіше.*
- C. Узагальнена характеристика ознаки в даному ряді значень.*

21. Для вимірювання центральної тенденції атрибутивної ознаки застосовують:

- A. Медіану.*
- C. Квартилі.*
- B. Моду.*
- D. Ексцес.*

22. Для вимірювання варіації значення кількісної ознаки застосовують такі статистичні характеристики:

- A. Мода.*
- D. Дисперсія.*
- B. Медіана.*
- E. Ексцес.*
- C. Стандартне відхилення.*
- F. Розмах.*

23. Статистична гіпотеза – це:

- A. Будь-яке припущення, що використовується у статистичному дослідженні.*
- B. Припущення, яке можна перевірити за допомогою розрахунків з використанням наявної статистичної інформації.*
- C. Наукове припущення, що висувається для пояснення будь-якого явища і вимагає емпіричної перевірки.*

- 24.** Статистичний критерій – це:
- A. Відмітна ознака, що приймається за норму.
 - B. Набір правил, які служать для перевірки статистичної гіпотези.
 - C. Середнє значення ознаки.
 - D. Значення ознаки, що трапляється найчастіше.
- 25.** Критична область значень – це:
- A. Максимальні значення ознаки.
 - B. Мінімальні значення ознаки.
 - C. Область значень ознаки, для якої відхиляється статистична гіпотеза.
- 26.** Середнє значення ознаки у двох сукупностях однакове. Чи може бути різною варіація ознаки в цих сукупностях?
- A. Так.
 - B. Ні.
- 27.** Варіація ознаки у двох сукупностях однакова. Чи може бути різною середнє значення ознаки в цих сукупностях?
- A. Так.
 - B. Ні.
- 28.** За допомогою якого статистичного метода можна підтвердити достовірність різниці кількісного показника у трьох групах?
- A. Кореляційного аналізу
 - B. Критерію Стьюдента.
 - C. Регресійного аналізу.
 - D. Дисперсійного аналізу.
- 29.** Які методи для оцінки достовірності відмінностей необхідно використовувати у випадку, коли первинні дані не мають нормального розподілу?
- A. Описові методи
 - B. Параметричні методи
 - C. Непараметричні методи
 - D. Багатовимірні методи
 - E. Рутинні методи
- 30.** Яка статистична характеристика визначає силу зв'язку між ознаками?
- A. Коефіцієнт варіації.
 - B. Коефіцієнт регресії.
 - C. Коефіцієнт детермінації.
 - D. Коефіцієнт кореляції.
- 31.** Факторна ознака – це:
- A. Ознака, що змінюється під впливом інших ознак.
 - B. Ознака, що впливає на зміну інших ознак.
- 32.** У випадку, коли коефіцієнт кореляції дорівнює 1, зв'язок між ознаками:
- A. Відсутній.
 - B. Зворотний.
 - C. Помірний.
 - D. Функціональний.
- 33.** При якому зв'язку напрям зміни результативної ознаки збігається із напрямом зміни ознаки-фактора?
- A. Прямому.
 - B. Зворотному
 - C. Криволінійному.
 - D. Атрибутивному.

- 34.** Які статистичні показники використовуються тільки для аналізу часових рядів?
- A. Абсолютний приріст. D. Середній темп зростання.*
B. Перший кuartиль. E. Амплітуда.
C. Коефіцієнт варіації. F. Період.
- 35.** Динамічний ряд, в якому його рівні коливаються в значних межах, підлягає вирівнюванню. Який з наведених методів є методом вирівнювання динамічного ряду?
- A. Метод «Хі-квадрат».*
B. Метод найменших квадратів.
C. Метод ковзної середньої.
D. Метод стандартизації.
- 36.** Чи може бути відсутньою циклічна компонента динамічного ряду?
- A. Так. B. Ні.*
- 37.** Час фактичного існування осередку інфекційного захворювання:
- A. Час від моменту захворювання до моменту ізоляції хворого і проведення заключної дезінфекції.*
B. Час від моменту захворювання до моменту ізоляції хворого плюс максимальний інкубаційний період.
C. Час від моменту захворювання до початку епідобстеження осередку.
D. Час від моменту забору матеріалу від хворого для лабораторного дослідження до виписки із стаціонару.
- 38.** Для багаторічної динаміки захворюваності характерним є такий прояв:
- A. Сезонність.*
B. Щорічна захворюваність.
C. Циклічність.
D. Спорадична, розрізнена захворюваність.
E. Епідемічний спалах.
- 39.** Епідемічна тенденція – це:
- A. Закономірний підйом захворюваності в певні місяці року.*
B. Регулярні коливання рівня захворюваності в багаторічній динаміці.
C. Короточасний підйом захворюваності на обмеженій території.
D. Регулярні коливання рівня захворюваності в річній динаміці.
E. Основна спрямованість змін інтенсивності епідемічного процесу за ряд років.
- 40.** Метою аналітичного прийому епідеміологічного методу є:
- A. Визначення пріоритетних проблем профілактики.*
B. Формулювання первинних гіпотез щодо чинників ризику.
C. Оцінка гіпотез методом випадок-контроль або когортним дослідженням.
D. Перевірка гіпотез в контрольованому експерименті.
E. Перевірка гіпотез в природньому експерименті.
- 41.** Який із розділів ретроспективного епідеміологічного аналізу дозволяє оцінити повноту і своєчасність проведення заходів щодо осіб, які спілкувалися з джерелом інфекції?
- A. Аналіз річної динаміки захворюваності.*
B. Аналіз багаторічної динаміки захворюваності.
C. Аналіз якості та ефективності протиепідемічних заходів.
D. Аналіз захворюваності за соціально-віковими групами населення.
E. Аналіз якості та ефективності профілактичних заходів.

42. Який з розділів ретроспективного епідеміологічного аналізу дозволяє визначити сезонність захворювання?
- A. *Аналіз якості й ефективності профілактичних заходів.*
 - B. *Аналіз якості й ефективності протиепідемічних заходів.*
 - C. *Аналіз захворюваності за соціально-віковими групами населення.*
 - D. *Аналіз річної динаміки захворюваності.*
 - E. *Аналіз багаторічної динаміки захворюваності.*
43. Який з розділів ретроспективного епідеміологічного аналізу дозволяє встановити циклічність (періодичність) епідемічного процесу?
- A. *Аналіз річної динаміки захворюваності.*
 - B. *Аналіз багаторічної динаміки захворюваності.*
 - C. *Аналіз захворюваності за соціально-віковими групами населення.*
 - D. *Аналіз якості й ефективності профілактичних заходів.*
 - E. *Аналіз якості й ефективності протиепідемічних заходів.*
44. Який з розділів ретроспективного епідеміологічного аналізу дозволяє оцінити своєчасність госпіталізації джерел інфекції?
- A. *Аналіз річної динаміки захворюваності.*
 - B. *Аналіз багаторічної динаміки захворюваності.*
 - C. *Аналіз якості й ефективності протиепідемічних заходів.*
 - D. *Аналіз захворюваності за соціально-віковими групами населення.*
 - E. *Аналіз якості й ефективності профілактичних заходів.*
45. Який з розділів ретроспективного епідеміологічного аналізу дозволяє оцінити повноту та своєчасність проведення флюорографічного обстеження населення?
- A. *Аналіз захворюваності за соціально-віковими групами населення.*
 - B. *Аналіз багаторічної динаміки захворюваності.*
 - C. *Аналіз річної динаміки захворюваності.*
 - D. *Аналіз якості та ефективності профілактичних заходів.*
 - E. *Аналіз якості та ефективності протиепідемічних заходів.*
46. Який з розділів ретроспективного епідеміологічного аналізу дозволяє виявити колективи «ризик» відносно поширення інфекційного захворювання?
- A. *Аналіз річної динаміки захворюваності.*
 - B. *Аналіз багаторічної динаміки захворюваності.*
 - C. *Аналіз захворюваності за соціально-віковими групами населення.*
 - D. *Аналіз якості та ефективності профілактичних заходів.*
 - E. *Аналіз якості й ефективності протиепідемічних заходів.*
47. Який з розділів ретроспективного епідеміологічного аналізу дозволяє встановити тривалість міжепідемічних періодів епідемічного процесу?
- A. *Аналіз річної динаміки захворюваності.*
 - B. *Аналіз багаторічної динаміки захворюваності.*
 - C. *Аналіз захворюваності за соціально-віковими групами населення.*
 - D. *Аналіз якості й ефективності профілактичних заходів.*
 - E. *Аналіз якості й ефективності протиепідемічних заходів.*

48. Який з розділів ретроспективного епідеміологічного аналізу дозволяє виявити вікові групи «ризик» у загальній структурі інфекційної захворюваності?

- A. Аналіз річної динаміки захворюваності.
- B. Аналіз багаторічної динаміки захворюваності.
- C. Аналіз захворюваності за соціально-віковими групами населення.
- D. Аналіз якості й ефективності профілактичних заходів.
- E. Аналіз якості й ефективності протиепідемічних заходів.

49. Який з розділів ретроспективного епідеміологічного аналізу дозволяє оцінити якість осередкової дезінфекції?

- A. Аналіз захворюваності за соціально-віковими групами населення.
- B. Аналіз багаторічної динаміки захворюваності.
- C. Аналіз річної динаміки захворюваності.
- D. Аналіз якості й ефективності протиепідемічних заходів.
- E. Аналіз якості й ефективності профілактичних заходів.

50. Який з розділів ретроспективного епідеміологічного аналізу дозволяє оцінити індекс і коефіцієнт сезонності?

- A. Аналіз захворюваності за соціально-віковими групами населення.
- B. Аналіз річної динаміки захворюваності.
- C. Аналіз багаторічної динаміки захворюваності.
- D. Аналіз якості й ефективності профілактичних заходів.
- E. Аналіз якості й ефективності протиепідемічних заходів.

Правильні відповіді:

1	2	3	4	5	6	7	8	9	10	11	12	13
C	A, B, C, D, E	C	C	C	E	C	D	A, B, C, D	C	B	E	A, D, E, F
14	15	16	17	18	19	20	21	22	23	24	25	26
A, D	A, B, C, D, E, F, G	A, B, C, D	D	A, B	C	C	B	C, D, F	B	B	C	A
27	28	29	30	31	32	33	34	35	36	37	38	39
A	D	C	D	B	D	A	A, D	C	A	A	C	E
40	41	42	43	44	45	46	47	48	49	50		
C	C	D	B	C	D	C	B	C	D	B		

ДОДАТКИ

Додаток 1

Критичні межі $U_1(\alpha)$ і $U_2(\alpha)$ критерію нормальності розподілу Девіда-Хартлі-Пірсона (RS-критерію)

n	$\alpha = 0,01$		$\alpha = 0,05$		$\alpha = 0,10$	
	U_1	U_2	U_1	U_2	U_1	U_2
3	1,737	2,000	1,758	1,999	1,782	1,997
4	1,870	2,445	1,980	2,429	2,040	2,409
5	2,020	2,803	2,150	2,753	2,220	2,712
6	2,150	3,095	2,200	3,012	2,370	2,949
7	2,260	3,338	2,400	3,222	2,490	3,143
8	2,350	3,543	2,500	3,399	2,590	3,308
9	2,440	3,720	2,590	3,552	2,680	3,449
10	2,510	3,875	2,670	3,685	2,760	3,570
11	2,580	4,012	2,740	3,800	2,840	3,680
12	2,640	4,134	2,800	3,910	2,900	3,780
13	2,700	4,244	2,860	4,000	2,960	3,870
14	2,750	4,340	2,920	4,090	3,020	3,950
15	2,800	4,440	2,970	4,170	3,070	4,020
16	2,840	4,520	3,010	4,240	3,120	4,090
17	2,880	4,600	3,060	4,310	3,170	4,150
18	2,920	4,670	3,100	4,370	3,210	4,210
19	2,960	4,740	3,140	4,430	3,250	4,270
20	2,990	4,800	3,180	4,490	3,290	4,320
25	3,150	5,060	3,340	4,710	3,450	4,530
30	3,270	5,260	3,470	4,890	3,590	4,700
35	3,380	5,420	3,580	5,040	3,700	4,840
40	3,470	5,560	3,670	5,160	3,790	4,960
45	3,550	5,670	3,750	5,260	3,880	5,060
50	3,620	5,770	3,830	5,350	3,950	5,140
55	3,690	5,860	3,900	5,430	4,020	5,220
60	3,750	5,940	3,960	5,510	4,080	5,290
65	3,800	6,010	4,010	5,570	4,140	5,350
70	3,850	6,070	4,060	5,630	4,190	5,410
75	3,900	6,130	4,130	5,680	4,240	5,460
80	3,940	6,180	4,150	5,730	4,280	5,510
85	3,990	6,230	4,200	5,780	4,330	5,560
90	4,020	6,270	4,240	5,820	4,360	5,600
95	4,060	6,320	4,270	5,860	4,400	5,640
100	4,100	6,360	4,310	5,900	4,440	5,680

Критерій Кохрена для перевірки гомогенності дисперсій

Критерій Кохрена G

Рівень значущості $\alpha = 0,01$								
$\begin{matrix} f \\ k \end{matrix}$	1	3	6	10	16	36	144	∞
2	0,9999	0,9794	0,9172	0,8539	0,7949	0,7067	0,6062	0,5000
3	0,9933	0,8831	0,7606	0,6743	0,6059	0,5153	0,4230	0,3333
4	0,9676	0,7814	0,6410	0,5536	0,4884	0,4057	0,3251	0,2500
5	0,9279	0,6957	0,5531	0,4697	0,4094	0,3351	0,2644	0,2000
6	0,8828	0,6258	0,4866	0,4084	0,3529	0,2858	0,2229	0,1667
7	0,8376	0,5685	0,4347	0,3616	0,3105	0,2494	0,1929	0,1429
8	0,7945	0,5209	0,3932	0,3248	0,2779	0,2214	0,1700	0,1250
9	0,7544	0,4810	0,3592	0,2950	0,2514	0,1992	0,1521	0,1111
10	0,7175	0,4469	0,3308	0,2704	0,2297	0,1811	0,1376	0,1000
15	0,5747	0,3317	0,2386	0,1918	0,1612	0,1251	0,0934	0,0667
20	0,4799	0,2654	0,1877	0,1501	0,1248	0,0960	0,0709	0,0500
30	0,3632	0,1913	0,1327	0,1054	0,0867	0,0658	0,0480	0,0333
40	0,2940	0,1508	0,1033	0,0816	0,0668	0,0503	0,0363	0,0250
60	0,2151	0,1069	0,0722	0,0567	0,0461	0,0344	0,0245	0,0167
120	0,1225	0,0585	0,0387	0,0302	0,0242	0,0178	0,0125	0,0083
∞	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Рівень значущості $\alpha = 0,05$								
$\begin{matrix} f \\ k \end{matrix}$	1	3	6	10	16	36	144	∞
2	0,9985	0,9392	0,8534	0,7880	0,7341	0,6602	0,5813	0,5000
3	0,9669	0,7977	0,6771	0,6025	0,5466	0,4748	0,4031	0,3333
4	0,9065	0,6841	0,5598	0,4884	0,4366	0,3720	0,3093	0,2500
5	0,8412	0,5981	0,4783	0,4118	0,3645	0,3066	0,2513	0,2000
6	0,7808	0,5321	0,4184	0,3568	0,3135	0,2612	0,2129	0,1667
7	0,7271	0,4800	0,3726	0,3154	0,2756	0,2278	0,1833	0,1429
8	0,6798	0,4377	0,3362	0,2829	0,2462	0,2022	0,1616	0,1250
9	0,6385	0,4027	0,3067	0,2568	0,2226	0,1820	0,1446	0,1111
10	0,6020	0,3733	0,2823	0,2353	0,2032	0,1655	0,1308	0,1000
15	0,4709	0,2758	0,2034	0,1671	0,1429	0,1144	0,0889	0,0667
20	0,3894	0,2205	0,1602	0,1303	0,1108	0,0879	0,0675	0,0500
30	0,2929	0,1593	0,1137	0,0921	0,0771	0,0604	0,0457	0,0333
40	0,2370	0,1259	0,0887	0,0713	0,0595	0,0462	0,0347	0,0250
60	0,1737	0,0895	0,0623	0,0497	0,0411	0,0316	0,0234	0,0167
120	0,0998	0,0495	0,0337	0,0260	0,0218	0,0165	0,0120	0,0083
∞	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

**Вміст аркушу «Первинні дані» файлу Lesson01D.xlsx
Коронавірус COVID-19 в Україні 2020–2022**

Регіон	Область	Населення	Інфіковано (абс. показник)	Смертельні випадки (абс. показник)	Госпіталізовано (абс. показник)
Україна	Україна	41130700	5014929	105505	955558
Західний	Волинська	1020800	142206	2394	24278
Західний	Закарпатська	1243700	116515	2724	33783
Західний	Івано-Франківська	1350600	185573	3476	52122
Західний	Львівська	2476100	307547	6569	59020
Західний	Рівненська	1140900	175663	2530	29026
Західний	Тернопільська	1021000	149090	2391	28342
Західний	Хмельницька	1227500	211466	3703	41918
Західний	Чернівецька	889900	155406	3290	31953
Південний	Миколаївська	1090500	156458	3530	29386
Південний	Одеська	2349700	331929	5917	55053
Південний	Херсонська	1000400	110186	2881	20283
Північний	Полтавська	1350600	199299	4181	25371
Північний	Сумська	1034400	186176	3208	28672
Північний	Харківська	2596300	307426	6858	66775
Північний	Чернігівська	957700	127853	2751	26894
Східний	Дніпропетровська	3093200	319171	9454	66709
Східний	Донецька	4056400	222794	5607	39843
Східний	Запорізька	1636300	227984	5866	36356
Східний	Кіровоградська	902300	47284	1741	10768
Східний	Луганська	2101700	80676	2392	13048
Центральний	Вінницька	1507700	153965	3546	42247
Центральний	Житомирська	1177600	198261	3538	38998
Центральний	Київська	1795500	243025	5019	39837
Центральний	м. Київ	2950700	480302	8966	87772
Центральний	Черкаська	1159200	178674	2973	27104

Вміст аркушу «Первинні дані» файлу Lesson02D.xlsx

Дані про реєстрацію випадків інфекційних захворювань по областях

Реєстраційний №	Дата реєстрації	Стать	Дата народження	Місце проживання, область	Місце проживання, район	Місце проживання, населений пункт	Професія за класифікацією	Професійна група ризику
1	2	3	4	5	6	7	8	9
1	19.03.2021	ж	16.05.1963	Харківська	Харківський	Харків	Працівники ділової та фінансової сфер	так
2	01.04.2021	ж	19.06.1974	Сумська	Сумський	Суми	Освітні професії	так
3	08.04.2021	ж	09.11.1977	Полтавська	Полтавський	Полтава	Промисловість	ні
4	26.03.2021	ж	29.10.1962	Сумська	Конотопський	Конотоп	Робітники у сфері транспорту	так
5	09.04.2021	ч	20.11.1974	Харківська	Богодухівський	Валки	Співробітники у сфері архітектури та інженерії	ні
6	09.04.2021	ч	14.08.1981	Полтавська	Миргородський	Миргород	Працівники ділової та фінансової сфер	так
7	19.02.2021	ж	29.07.1952	Харківська	Чугуївський	Чугуїв	Не працює	ні
8	05.04.2021	ж	13.06.1980	Полтавська	Кременчуцький	Кременчук	Працівники ділової та фінансової сфер	так
9	18.04.2021	ж	27.07.1960	Харківська	Богодухівський	Золочів	Співробітники у сфері архітектури та інженерії	ні
10	29.04.2021	ж	26.08.1976	Полтавська	Полтавський	Полтава	Співробітники у сфері архітектури та інженерії	ні
11	19.04.2021	ж	23.05.1952	Сумська	Конотопський	Путивль	Робітники з офісно-адміністративної підтримки	ні
12	19.03.2021	ж	16.05.1963	Харківська	Куп'янський	Куп'янськ	Робітники з офісно-адміністративної підтримки	ні
13	27.04.2021	ч	24.04.1986	Харківська	Лозівський	Лозова	Не працює	ні
14	18.04.2021	ж	19.01.1960	Сумська	Сумський	Білопілля	Працівники ділової та фінансової сфер	так
15	14.04.2021	ч	02.02.1992	Сумська	Охтирський	Охтирка	Не працює	ні
16	01.05.2021	ж	18.07.1965	Полтавська	Лубенський	Пирятин	Робітники у сфері транспорту	так
17	13.04.2021	ж	18.04.1962	Полтавська	Кременчуцький	Глобине	Освітні професії	так
18	16.04.2021	ч	26.06.1962	Полтавська	Лубенський	Лубни	Співробітники у сфері архітектури та інженерії	ні
19	12.03.2021	ж	11.07.1986	Харківська	Харківський	Безлюдівка	Співробітники з вирішення юридичних питань	ні
20	21.03.2021	ж	25.11.1994	Харківська	Богодухівський	Коломак	Освітні професії	так
21	09.04.2021	ч	05.06.1949	Сумська	Сумський	Краснопілля	Пенсіонер	ні
22	15.04.2021	ж	21.11.1993	Сумська	Охтирський	Глухів	Співробітники поліції та охоронних служб	так
23	16.04.2021	ж	08.03.1976	Харківська	Харківський	Циркуни	Медичні працівники та допоміжний персонал	так
24	12.04.2021	ж	19.09.1979	Сумська	Конотопський	Конотоп	Робітники у сфері громадських та соціальних послуг	так

1	2	3	4	5	6	7	8	9
25	30.04.2021	ж	28.10.1954	Сумська	Охтирський	Охтирка	Освітні професії	так
26	14.04.2021	ж	24.01.1981	Харківська	Ізюмський	Борова	Працівники ділової та фінансової сфер	так
27	19.04.2021	ч	11.01.1978	Харківська	Харківський	Дергачі	Працівники ділової та фінансової сфер	так
28	20.04.2021	ж	21.11.1964	Полтавська	Полтавський	Диканька	Промисловість	ні
29	05.05.2021	ж	06.11.1959	Полтавська	Лубенський	Лубни	Не працює	ні
30	13.04.2021	ч	01.05.1995	Харківська	Лозівський	Лозова	Працівники ділової та фінансової сфер	так
31	23.03.2021	ж	06.10.1979	Харківська	Харківський	Мерефа	Освітні професії	так
32	29.03.2021	ж	21.03.2000	Сумська	Конотопський	Конотоп	Співробітники поліції та охоронних служб	так
33	23.04.2021	ж	10.12.1961	Сумська	Сумський	Суми	Промисловість	ні
34	08.04.2021	ж	23.08.1954	Сумська	Роменський	Ромни	Не працює	ні
35	26.04.2021	ч	25.08.1965	Сумська	Сумський	Суми	Робітники з офісно-адміністративної підтримки	ні
36	11.04.2021	ч	10.02.1968	Харківська	Харківський	Солоницівка	Освітні професії	так
37	26.12.2020	ж	16.09.1977	Харківська	Ізюмський	Борова	Медичні працівники та допоміжний персонал	так
38	08.05.2021	ж	03.05.1978	Харківська	Ізюмський	Ізюм	Не працює	ні
39	12.03.2021	ж	22.02.1987	Полтавська	Полтавський	Полтава	Медичні працівники та допоміжний персонал	так
40	29.05.2021	ж	28.07.1971	Полтавська	Полтавський	Полтава	Не працює	ні
41	11.08.2021	ч	11.08.1951	Сумська	Конотопський	Путівль	Пенсіонер	ні
42	22.03.2021	ж	24.06.1966	Харківська	Чугуївський	Вовчанськ	Освітні професії	так
43	26.04.2021	ч	01.01.1965	Харківська	Красноградський	Кегичівка	Працівники ділової та фінансової сфер	так
44	28.09.2020	ж	01.01.1992	Харківська	Харківський	Дергачі	Освітні професії	так
45	11.05.2021	ч	01.01.1999	Полтавська	Лубенський	Пирятин	Працівники ділової та фінансової сфер	так
46	24.05.2021	ж	01.01.1990	Сумська	Роменський	Ромни	Працівники ділової та фінансової сфер	так
47	27.10.2020	ж	01.01.1964	Сумська	Роменський	Ромни	Робітники сфери продаж	так
48	13.05.2021	ч	01.01.1952	Сумська	Охтирський	Охтирка	Співробітники у сфері архітектури та інженерії	ні
49	11.05.2021	ж	01.01.1972	Харківська	Чугуївський	Зміїв	Медичні працівники та допоміжний персонал	так
50	04.05.2021	ж	01.01.1954	Харківська	Харківський	Мерефа	Співробітники у сфері архітектури та інженерії	ні
51	02.05.2021	ж	01.01.1971	Харківська	Ізюмський	Ізюм	Робітники в сфері «Догляд за собою та обслуговування»	так
52	22.04.2021	ж	09.06.1951	Полтавська	Миргородський	Миргород	Пенсіонер	ні
53	14.04.2021	ж	01.01.1999	Полтавська	Миргородський	Велика Багачка	Освітні професії	так
54	11.02.2021	ж	01.01.1972	Харківська	Чугуївський	Малинівка	Медичні працівники та допоміжний персонал	так
55	07.05.2021	ж	01.01.1970	Харківська	Харківський	Люботин	Співробітники у сфері архітектури та інженерії	ні
56	14.03.2021	ч	01.01.1981	Харківська	Харківський	Циркуни	Працівники ділової та фінансової сфер	так

1	2	3	4	5	6	7	8	9
57	11.05.2021	ж	01.01.1988	Полтавська	Кременчуцький	Кременчук	Робітники сфери продаж	так
58	26.04.2021	ж	01.01.1965	Сумська	Шосткінський	Шостка	Працівники ділової та фінансової сфер	так
59	11.05.2021	ж	01.01.1945	Сумська	Шосткінський	Шостка	Освітні професії	так
60	13.03.2021	ж	01.01.1968	Сумська	Роменський	Яснопільщина	Робітники з офісно-адміністративної підтримки	ні
61	05.04.2021	ж	08.03.1951	Полтавська	Кременчуцький	Козельщина	Медичні працівники та допоміжний персонал	так
62	25.03.2021	ж	26.10.1964	Полтавська	Полтавський	Диканька	Робітники сфери продаж	так
63	01.04.2021	ж	16.03.1966	Сумська	Охтирський	Глухів	Освітні професії	так
64	23.03.2021	ж	18.10.1991	Харківська	Харківський	Безлюдівка	Працівники ділової та фінансової сфер	так
65	22.03.2021	ж	05.05.1972	Сумська	Сумський	Краснопілля	Промисловість	ні
66	23.03.2021	ж	26.07.1976	Харківська	Чугуївський	Вовчанськ	Не працює	ні
67	23.03.2021	ч	25.07.1978	Харківська	Красноградський	Зачепилівка	Освітні професії	так
68	25.05.2021	ч	18.08.1988	Харківська	Ізюмський	Балаклія	Промисловість	ні
69	14.05.2021	ж	25.05.1945	Сумська	Охтирський	Глухів	Пенсіонер	ні
70	05.04.2021	ж	16.03.1974	Сумська	Конотопський	Конотоп	Працівники у сфері комп'ютерно-математичних, фізичних та соціальних наук	ні
71	14.04.2021	ж	01.01.1941	Полтавська	Лубенський	Пирятин	Пенсіонер	ні
72	08.05.2021	ч	22.02.2004	Полтавська	Миргородський	Миргород	Освітні професії	так
73	31.08.2020	ж	07.06.1943	Полтавська	Полтавський	Зінки	Медичні працівники та допоміжний персонал	так
74	26.04.2021	ж	13.06.1959	Харківська	Куп'янський	Куп'янськ	Робітники сфери продаж	так
75	04.05.2021	ч	17.10.1977	Харківська	Харківський	Люботин	Промисловість	ні
76	08.06.2021	ж	06.04.1972	Харківська	Чугуївський	Зміїв	Не працює	ні
77	07.06.2021	ж	23.07.1979	Полтавська	Полтавський	Опішня	Медичні працівники та допоміжний персонал	так
78	01.06.2021	ч	08.02.1981	Полтавська	Миргородський	Миргород	Не працює	ні
79	29.04.2021	ж	26.06.1953	Сумська	Сумський	Краснопілля	Пенсіонер	ні
80	17.04.2021	ж	05.07.1963	Сумська	Сумський	Білопілля	Медичні працівники та допоміжний персонал	так
81	04.06.2021	ж	15.01.1984	Сумська	Охтирський	Охтирка	Медичні працівники та допоміжний персонал	так
82	25.05.2021	ж	14.02.1946	Сумська	Роменський	Недригайлів	Пенсіонер	ні
83	22.04.2021	ж	08.02.1951	Сумська	Роменський	Ромни	Пенсіонер	ні
84	30.05.2021	ж	07.07.1951	Харківська	Красноградський	Красноград	Пенсіонер	ні
85	20.04.2021	ж	09.04.1971	Харківська	Ізюмський	Балаклія	Робітники у сфері громадських та соціальних послуг	так
86	05.04.2021	ж	08.04.1968	Харківська	Чугуївський	Малинівка	Співробітники з вирішення юридичних питань	ні
87	13.04.2021	ж	11.12.1967	Полтавська	Лубенський	Лубни	Промисловість	ні
88	05.04.2021	ж	14.06.1978	Полтавська	Кременчуцький	Кременчук	Медичні працівники та допоміжний персонал	так

1	2	3	4	5	6	7	8	9
89	01.04.2021	ж	29.10.1962	Полтавська	Миргородський	Шишаки	Працівники ділової та фінансової сфер	так
90	13.04.2021	ж	10.03.1960	Полтавська	Кременчуцький	Кременчук	Працівники ділової та фінансової сфер	так
91	18.04.2021	ч	15.06.1953	Харківська	Ізюмський	Борова	Робітники сфери «Монтаж, обслуговування та ремонт»	ні
92	01.04.2021	ж	19.06.1974	Харківська	Харківський	Липці	Освітні професії	так
93	30.03.2021	ж	07.12.1992	Харківська	Чугуївський	Зміїв	Освітні професії	так
94	11.04.2021	ж	09.11.1977	Харківська	Ізюмський	Балаклія	Промисловість	ні
95	14.04.2021	ж	08.06.1975	Сумська	Сумський	Суми	Співробітники поліції та охоронних служб	так
96	16.04.2021	ж	10.03.1959	Сумська	Охтирський	Охтирка	Працівники ділової та фінансової сфер	так
97	12.03.2021	ж	24.09.1956	Харківська	Красноградський	Зачепилівка	Робітники сфери продаж	так
98	16.03.2021	ч	01.01.1963	Харківська	Красноградський	Красноград	Робітники в сфері транспорту	так
99	04.05.2021	ж	23.07.1955	Полтавська	Полтавський	Опішня	Робітники з офісно-адміністративної підтримки	ні
100	28.05.2021	ж	10.11.1952	Полтавська	Миргородський	Шишаки	Пенсіонер	ні

Регістраційний №	Визначення інфекційного захворювання	Група інфекцій	Дата визначення інфекційного захворювання	Місце лікування	Завершення лікування, дата	Результат лікування
1	2	3	4	5	6	7
1	Черевний тиф	Кишкові інфекції	19.03.2021	Амб	05.04.2021	одужання
2	Дифтерія	Інфекції дихальних шляхів	01.04.2021	Амб	14.04.2021	одужання
3	Кашлюк	Інфекції дихальних шляхів	08.04.2021	Амб	25.04.2021	одужання
4	Вірусний гепатит А	Кишкові інфекції	26.03.2021	Амб	14.04.2021	одужання
5	Сальмонельоз	Кишкові інфекції	09.04.2021	Амб	24.04.2021	одужання
6	Короста	Інфекції зовнішніх покривів	09.04.2021	Амб	27.04.2021	одужання
7	Вірусний гепатит А	Кишкові інфекції	17.02.2021	Амб	15.03.2021	одужання
8	Черевний тиф	Кишкові інфекції	05.04.2021	Амб	16.04.2021	одужання
9	Рикетсіоз	Кров'яні інфекції	18.04.2021	Амб	14.05.2021	одужання
10	Короста	Інфекції зовнішніх покривів	29.04.2021	Амб	12.05.2021	одужання
11	Дифтерія	Інфекції дихальних шляхів	19.04.2021	Амб	14.05.2021	одужання
12	Паротит	Інфекції дихальних шляхів	19.03.2021	Амб	05.04.2021	одужання
13	Вірусний гепатит А	Кишкові інфекції	27.04.2021	Амб	15.05.2021	одужання
14	Черевний тиф	Кишкові інфекції	18.04.2021	Амб	25.05.2021	одужання
15	Короста	Інфекції зовнішніх покривів	14.04.2021	Амб	30.04.2021	одужання
16	Кашлюк	Інфекції дихальних шляхів	01.05.2021	Амб	14.05.2021	одужання
17	Паротит	Інфекції дихальних шляхів	13.04.2021	Амб	01.05.2021	одужання

1	2	3	4	5	6	7
18	Короста	Інфекції зовнішніх покривів	16.04.2021	Амб	01.05.2021	одужання
19	Шигельоз	Кишкові інфекції	12.03.2021	Амб	30.03.2021	одужання
20	Кір	Інфекції дихальних шляхів	21.03.2021	Амб	01.04.2021	одужання
21	Вірусний гепатит А	Кишкові інфекції	20.03.2021	Амб	20.04.2021	одужання
22	Кашлюк	Інфекції дихальних шляхів	15.04.2021	Амб	11.05.2021	одужання
23	Черевний тиф	Кишкові інфекції	16.04.2021	Амб	20.05.2021	помер
24	Кір	Інфекції дихальних шляхів	12.04.2021	Амб	30.04.2021	одужання
25	Дифтерія	Інфекції дихальних шляхів	30.04.2021	Амб	22.05.2021	одужання
26	Короста	Інфекції зовнішніх покривів	14.04.2021	Амб	14.05.2021	одужання
27	Шигельоз	Кишкові інфекції	18.04.2021	Амб	30.04.2021	одужання
28	Сальмонельоз	Кишкові інфекції	20.04.2021	Амб	30.04.2021	одужання
29	Вірусний гепатит А	Кишкові інфекції	03.05.2021	Амб	16.05.2021	одужання
30	Короста	Інфекції зовнішніх покривів	13.04.2021	Амб	27.04.2021	одужання
31	Кір	Інфекції дихальних шляхів	23.03.2021	Амб	19.04.2021	одужання
32	Короста	Інфекції зовнішніх покривів	29.03.2021	Амб	16.04.2021	одужання
33	Шигельоз	Кишкові інфекції	23.04.2021	Амб	30.04.2021	одужання
34	Туберкульоз органів дихання	Інфекції дихальних шляхів	08.04.2021	Амб	25.04.2021	одужання
35	Дифтерія	Інфекції дихальних шляхів	25.04.2021	Амб	05.05.2021	одужання
36	Кашлюк	Інфекції дихальних шляхів	11.04.2021	Амб	15.05.2021	одужання
37	Короста	Інфекції зовнішніх покривів	26.12.2020	Амб	05.01.2021	одужання
38	Кір	Інфекції дихальних шляхів	01.05.2021	Амб	12.06.2021	одужання
39	Туберкульоз органів дихання	Інфекції дихальних шляхів	12.03.2021	Амб	05.04.2021	одужання
40	Вірусний гепатит А	Кишкові інфекції	22.05.2021	Амб	12.06.2021	одужання
41	Короста	Інфекції зовнішніх покривів	31.07.2021	Стац	17.08.2021	одужання
42	Сальмонельоз	Кишкові інфекції	22.03.2021	Амб	07.04.2021	одужання
43	Кір	Інфекції дихальних шляхів	22.04.2021	Амб	15.05.2021	одужання
44	Кашлюк	Інфекції дихальних шляхів	24.09.2020	Амб	29.10.2020	одужання
45	Паротит	Інфекції дихальних шляхів	10.05.2021	Амб	25.05.2021	одужання
46	Шигельоз	Кишкові інфекції	21.05.2021	Амб	07.06.2021	одужання
47	Туберкульоз органів дихання	Інфекції дихальних шляхів	27.10.2020	Амб	14.11.2020	одужання
48	Сальмонельоз	Кишкові інфекції	10.05.2021	Амб	09.06.2021	одужання
49	Паротит	Інфекції дихальних шляхів	10.05.2021	Амб	30.05.2021	одужання
50	Вірусний гепатит А	Кишкові інфекції	01.05.2021	Амб	28.05.2021	одужання
51	Рикетсіоз	Кров'яні інфекції	02.05.2021	Амб	31.05.2021	одужання

1	2	3	4	5	6	7
52	Вірусний гепатит А	Кишкові інфекції	22.04.2021	Амб	11.05.2021	одужання
53	Туберкульоз органів дихання	Інфекції дихальних шляхів	11.04.2021	Амб	29.04.2021	одужання
54	Дифтерія	Інфекції дихальних шляхів	10.02.2021	Амб	02.03.2021	одужання
55	Кір	Інфекції дихальних шляхів	06.05.2021	Амб	31.05.2021	одужання
56	Кашлюк	Інфекції дихальних шляхів	11.03.2021	Амб	27.03.2021	одужання
57	Менінгококова інфекція	Інфекції дихальних шляхів	10.05.2021	Амб	28.05.2021	одужання
58	Туберкульоз органів дихання	Інфекції дихальних шляхів	24.04.2021	Амб	25.05.2021	одужання
59	Вірусний гепатит А	Кишкові інфекції	09.05.2021	Амб	29.05.2021	одужання
60	Туберкульоз органів дихання	Інфекції дихальних шляхів	13.05.2021	Амб	01.06.2021	одужання
61	Шигельоз	Кишкові інфекції	05.04.2021	Амб	17.04.2021	одужання
62	Кір	Інфекції дихальних шляхів	23.03.2021	Амб	22.04.2021	одужання
63	Рикетсіоз	Кров'яні інфекції	29.03.2021	Амб	20.04.2021	одужання
64	Сальмонельоз	Кишкові інфекції	20.03.2021	Амб	15.04.2021	одужання
65	Кліщовий енцефаліт	Кров'яні інфекції	15.03.2021	Амб	15.04.2021	одужання
66	Менінгококова інфекція	Інфекції дихальних шляхів	20.03.2021	Амб	10.04.2021	одужання
67	Кліщовий енцефаліт	Кров'яні інфекції	19.03.2021	Амб	12.04.2021	одужання
68	Шигельоз	Кишкові інфекції	22.05.2021	Амб	05.06.2021	одужання
69	Сальмонельоз	Кишкові інфекції	14.05.2021	Амб	04.06.2021	одужання
70	Туляремія	Кров'яні інфекції	02.04.2021	Амб	26.04.2021	одужання
71	Бруцельоз	Кишкові інфекції	11.04.2021	Амб	24.05.2021	одужання
72	Кашлюк	Інфекції дихальних шляхів	08.05.2021	Амб	22.05.2021	одужання
73	Шигельоз	Кишкові інфекції	31.08.2020	Амб	26.09.2020	одужання
74	Туляремія	Кров'яні інфекції	19.04.2021	Амб	25.05.2021	одужання
75	Шигельоз	Кишкові інфекції	01.05.2021	Амб	05.06.2021	одужання
76	Бруцельоз	Кишкові інфекції	29.05.2021	Амб	01.07.2021	одужання
77	Кліщовий енцефаліт	Кров'яні інфекції	04.06.2021	Амб	25.06.2021	одужання
78	Туляремія	Кров'яні інфекції	31.05.2021	Амб	21.06.2021	одужання
79	Менінгококова інфекція	Інфекції дихальних шляхів	26.04.2021	Амб	12.05.2021	одужання
80	Рикетсіоз	Кров'яні інфекції	17.04.2021	Амб	02.05.2021	одужання
81	Шигельоз	Кишкові інфекції	03.06.2021	Амб	14.06.2021	одужання
82	Кліщовий енцефаліт	Кров'яні інфекції	23.05.2021	Амб	10.06.2021	одужання
83	Бруцельоз	Кишкові інфекції	15.04.2021	Амб	22.05.2021	одужання
84	Кліщовий енцефаліт	Кров'яні інфекції	30.05.2021	Амб	14.06.2021	одужання
85	Сальмонельоз	Кишкові інфекції	20.04.2021	Амб	12.05.2021	одужання

1	2	3	4	5	6	7
86	Туляремія	Кров'яні інфекції	03.04.2021	Амб	19.04.2021	одужання
87	Шигельоз	Кишкові інфекції	12.04.2021	Амб	30.04.2021	одужання
88	Рикетсіоз	Кров'яні інфекції	31.03.2021	Амб	24.04.2021	одужання
89	Шигельоз	Кишкові інфекції	26.03.2021	Амб	14.04.2021	одужання
90	Менінгококова інфекція	Інфекції дихальних шляхів	12.04.2021	Амб	29.04.2021	одужання
91	Туляремія	Кров'яні інфекції	12.04.2021	Амб	30.04.2021	одужання
92	Кашлюк	Інфекції дихальних шляхів	01.04.2021	Амб	14.04.2021	одужання
93	Рикетсіоз	Кров'яні інфекції	29.03.2021	Амб	30.04.2021	помер
94	Кліщовий енцефаліт	Кров'яні інфекції	08.04.2021	Амб	26.04.2021	помер
95	Бруцельоз	Кишкові інфекції	13.04.2021	Амб	27.04.2021	одужання
96	Кліщовий енцефаліт	Кров'яні інфекції	14.04.2021	Амб	14.05.2021	одужання
97	Туляремія	Кров'яні інфекції	10.03.2021	Амб	02.04.2021	одужання
98	Рикетсіоз	Кров'яні інфекції	15.03.2021	Амб	23.04.2021	помер
99	Бруцельоз	Кишкові інфекції	03.05.2021	Амб	11.06.2021	одужання
100	Менінгококова інфекція	Інфекції дихальних шляхів	28.05.2021	Амб	28.06.2021	одужання

Додаток 5

**Вміст аркушу «Первинні дані» файлу Lesson03D.xlsx
Захворюваність на гострі інфекції верхніх дихальних шляхів**

Регіон України	Область	Населення	Інтенсивний показник на 100 тис., все населення	Інтенсивний показник на 100 тис., доросле населення	Інтенсивний показник на 100 тис., діти 0-17 років	Інтенсивний показник на 100 тис., діти 5-9 років	Інтенсивний показник на 100 тис., діти 10-14 років	Інтенсивний показник на 100 тис., діти 15-17 років
Центральний	Вінницька	1507700	18156,11	8602,76	60495,78	55172,65	52025,20	64164,22
Західний	Волинська	1020800	18170,39	8195,44	51349,16	40207,40	42387,26	57615,84
Східний	Дніпропетровська	3093200	17398,00	7342,51	63370,44	62546,87	51720,23	51848,16
Східний	Донецька	4056400	7408,61	2800,94	34672,80	29268,32	31252,71	27751,08
Центральний	Житомирська	1177600	19904,09	7853,14	69512,89	62653,59	55295,90	64636,16
Західний	Закарпатська	1243700	10521,44	4947,98	28883,73	25560,08	24636,86	29161,80
Східний	Запорізька	1636300	17936,32	7102,91	70431,10	72128,28	54659,12	53413,56
Західний	Івано-Франківська	1350600	13518,92	5458,26	45289,73	47287,47	45234,71	34596,29
Центральний	Київська	1795500	26913,96	12985,76	85449,04	78892,28	69996,99	86352,72
Східний	Кіровоградська	902300	13592,47	5157,36	52487,00	52057,25	43879,08	42229,17
Східний	Луганська	2101700	4496,11	1188,59	25869,54	21553,78	20465,28	19251,72
Західний	Львівська	2476100	14847,34	8156,96	42772,70	35577,04	37912,36	47272,96
Південний	Миколаївська	2950700	13576,78	5559,17	49546,89	51100,39	44333,58	43468,91
Південний	Одеська	1090500	17016,71	6167,55	61630,06	63287,39	47110,91	50933,60
Північний	Полтавська	2349700	14801,09	6576,40	55950,22	47915,08	51693,73	50426,90
Західний	Рівненська	1350600	17760,98	7092,77	51517,61	44526,23	47799,30	56548,87
Північний	Сумська	1140900	15550,06	5919,34	67204,88	65274,97	57289,69	55450,71
Західний	Тернопільська	1034400	10127,25	4724,77	33398,77	33463,23	27340,11	29102,84
Північний	Харківська	1021000	4074,30	1746,28	16191,33	16942,67	16291,35	14244,35
Південний	Херсонська	2596300	16389,76	4468,82	67564,54	60238,49	56835,81	65958,88
Західний	Хмельницька	1000400	20434,91	7449,86	77061,40	73141,72	60895,25	60843,36
Центральний	Черкаська	1227500	16091,44	7493,27	59418,49	56150,24	46155,41	51430,56
Західний	Чернівецька	1159200	16517,93	6674,13	54874,67	51322,37	47334,15	57191,61
Північний	Чернігівська	889900	17079,27	7435,63	67311,36	65010,69	53012,43	54526,94
Центральний	м. Київ	957700	15060,20	6646,65	51664,53	54289,43	40105,86	33885,99

Вміст аркушу «Первинні дані» файлу Lesson04D.xlsx

Захворюваність на гострі інфекції верхніх дихальних шляхів

Дані такі ж, як у додатку 5.

Вміст аркушу «Первинні дані» файлу Lesson05D.xlsx

Захворюваність на гострі інфекції верхніх дихальних шляхів

Регіон України	Область	Населення	Інтенсивний показник на 100 000, все населення	Індекс урбанізації	Індекс групи ризику за віком (< 14 i > 65)	Індекс жіночого населення	Щільність населення осіб/кв.км
Центральний	Вінницька	1507700	18156,11	51,95	33,15	53,62	56,67
Західний	Волинська	1020800	18170,39	51,78	32,74	52,79	50,57
Східний	Дніпропетровська	3093200	17398,00	84,06	32,70	54,34	96,92
Центральний	Житомирська	1177600	19904,09	59,43	33,02	53,35	39,55
Західний	Закарпатська	1243700	12021,44	36,90	31,84	51,91	97,18
Східний	Запорізька	1636300	17936,32	77,30	32,68	54,34	60,25
Західний	Івано-Франківська	1350600	13518,92	44,20	31,24	52,72	97,06
Центральний	Київська	1795500	24913,96	61,34	32,83	54,20	63,61
Східний	Кіровоградська	902300	13592,47	63,50	33,36	53,83	36,49
Західний	Львівська	2476100	14847,34	60,71	31,36	52,64	112,66
Південний	Миколаївська	2950700	13576,78	68,53	32,41	53,61	44,36
Південний	Одеська	1090500	17016,71	66,80	32,96	52,88	70,26
Північний	Полтавська	2349700	14801,09	62,20	32,14	53,75	46,77
Західний	Рівненська	1350600	17760,98	47,04	33,01	52,48	56,90
Північний	Сумська	1140900	15550,06	69,54	32,00	54,08	43,37
Західний	Тернопільська	1034400	12227,25	45,82	31,28	53,15	73,68
Північний	Харківська	1021000	14074,30	81,11	31,01	53,60	82,23
Південний	Херсонська	2596300	16389,76	61,17	32,69	53,54	35,14
Західний	Хмельницька	1000400	20434,91	57,74	33,14	53,48	59,37
Центральний	Черкаська	1227500	16091,44	57,03	32,85	53,97	55,36
Західний	Чернівецька	1159200	16517,93	42,93	31,68	52,83	109,60
Північний	Чернігівська	889900	17079,27	65,54	33,54	54,53	29,84
Центральний	м. Київ	957700	15060,20	100,00	32,21	53,89	3469,60

Вміст аркушу «Первинні дані» файлу Lesson06D.xlsx

Епідемія COVID-19 Україна, Харківська область, Населення(осіб) – 2658461

Номер тижня за порядком	Початок тижня	Кінець тижня	Активні випадки, кількість
1	29.02.2020	06.03.2020	0
2	07.03.2020	13.03.2020	0
3	14.03.2020	20.03.2020	0
4	21.03.2020	27.03.2020	1
5	28.03.2020	03.04.2020	4
6	04.04.2020	10.04.2020	8
7	11.04.2020	17.04.2020	62
8	18.04.2020	24.04.2020	109
9	25.04.2020	01.05.2020	268
10	02.05.2020	08.05.2020	409
11	09.05.2020	15.05.2020	499
12	16.05.2020	22.05.2020	391
13	23.05.2020	29.05.2020	332
14	30.05.2020	05.06.2020	388
15	06.06.2020	12.06.2020	479
16	13.06.2020	19.06.2020	631
17	20.06.2020	26.06.2020	685
18	27.06.2020	03.07.2020	703
19	04.07.2020	10.07.2020	71
20	11.07.2020	17.07.2020	782
21	18.07.2020	24.07.2020	836
22	25.07.2020	31.07.2020	1005
23	01.08.2020	07.08.2020	1495
24	08.08.2020	14.08.2020	2481
25	15.08.2020	21.08.2020	3224
26	22.08.2020	28.08.2020	3607
27	29.08.2020	04.09.2020	4414
28	05.09.2020	11.09.2020	5140
29	12.09.2020	18.09.2020	5933
30	19.09.2020	25.09.2020	6690
31	26.09.2020	02.10.2020	7637
32	03.10.2020	09.10.2020	9713
33	10.10.2020	16.10.2020	11493
34	17.10.2020	23.10.2020	11829
35	24.10.2020	30.10.2020	12282
36	31.10.2020	06.11.2020	14512
37	07.11.2020	13.11.2020	15174
38	14.11.2020	20.11.2020	13487
39	21.11.2020	27.11.2020	11890
40	28.11.2020	04.12.2020	11894
41	05.12.2020	11.12.2020	10409
42	12.12.2020	18.12.2020	9910
43	19.12.2020	25.12.2020	8939
44	26.12.2020	01.01.2021	8221
45	02.01.2021	08.01.2021	7446
46	09.01.2021	15.01.2021	6461
47	16.01.2021	22.01.2021	5636
48	23.01.2021	29.01.2021	5011
49	30.01.2021	05.02.2021	4111

Номер тижня за порядком	Початок тижня	Кінець тижня	Активні випадки, кількість
50	06.02.2021	12.02.2021	3578
51	13.02.2021	19.02.2021	3443
52	20.02.2021	26.02.2021	4219
53	27.02.2021	05.03.2021	5573
54	06.03.2021	12.03.2021	7268
55	13.03.2021	19.03.2021	10127
56	20.03.2021	26.03.2021	13602
57	27.03.2021	02.04.2021	17572
58	03.04.2021	09.04.2021	21112
59	10.04.2021	16.04.2021	23036
60	17.04.2021	23.04.2021	23440
61	24.04.2021	30.04.2021	20926
62	01.05.2021	07.05.2021	17733
63	08.05.2021	14.05.2021	12748
64	15.05.2021	21.05.2021	8811
65	22.05.2021	28.05.2021	6178
66	29.05.2021	04.06.2021	3613
67	05.06.2021	11.06.2021	2190
68	12.06.2021	18.06.2021	1261
69	19.06.2021	25.06.2021	833
70	26.06.2021	02.07.2021	656
71	03.07.2021	09.07.2021	551
72	10.07.2021	16.07.2021	532
73	17.07.2021	23.07.2021	632
74	24.07.2021	30.07.2021	710
75	31.07.2021	06.08.2021	953
76	07.08.2021	13.08.2021	1228
77	14.08.2021	20.08.2021	1541
78	21.08.2021	27.08.2021	1830
79	28.08.2021	03.09.2021	2561
80	04.09.2021	10.09.2021	3584
81	11.09.2021	17.09.2021	5938
82	18.09.2021	24.09.2021	9542
83	25.09.2021	01.10.2021	13922
84	02.10.2021	08.10.2021	18421
85	09.10.2021	15.10.2021	24181
86	16.10.2021	22.10.2021	27162
87	23.10.2021	29.10.2021	30542
88	30.10.2021	05.11.2021	32788
89	06.11.2021	12.11.2021	34047
90	13.11.2021	19.11.2021	29921
91	20.11.2021	26.11.2021	14728
92	27.11.2021	03.12.2021	10660
93	04.12.2021	10.12.2021	8457
94	11.12.2021	17.12.2021	7309
95	18.12.2021	24.12.2021	5865
96	25.12.2021	31.12.2021	5184
97	01.01.2022	07.01.2022	5586
98	08.01.2022	14.01.2022	7944
99	15.01.2022	21.01.2022	12585

ЛІТЕРАТУРА

1. Акіменко В.В. Прикладні задачі інтелектуального аналізу даних (DATA MINING). Київ : КНУ ім. Тараса Шевченка, 2018. 152 с.
2. Андрійчук О.М. Збірник ситуаційних задач та тестових завдань по епідеміології : навч. посіб. Київ : ДП «Видавничий дім «Персонал», 2020. 118 с.
3. Білобородько О.І., Ємел'яненко Т.Г. Аналіз динамічних рядів [Текст]: навч. посіб. Дніпро : РВВ ДНУ, 2014. 80 с.
4. Боснюк В.Ф. Математичні методи в психології : курс лекцій. Мультимедійне навчальне видання. Харків: НУЦЗУ, 2020. 141 с.
5. Виноград Н.О. Загальна епідеміологія : навч. посіб. для мед. ун-тів, інст., акад. 4-те вид., випр. Київ, 2021. 152 с.
6. Голованова І.А., Белікова І.В., Ляхова Н.О. Основи медичної статистики : навч. посіб. для аспірантів та клінічних ординаторів. Полтава, 2017. 113 с.
7. Джигирей І.М., Складанний Д.М. Спеціальні розділи математики. Статистичний аналіз даних у середовищі STATISTICA : навч. посіб. для студ. спеціальності 151 «Автоматизація та комп'ютерно-інтегровані технології», спеціалізації «Комп'ютерно-інтегровані технології сталих хімічних виробництв». Київ : КПІ ім. Ігоря Сікорського, 2019. 74 с. URL: <https://ela.kpi.ua/server/api/core/bitstreams/c75ec1ec-3398-44ba-85b6-33c00772a4bc/content>
8. Епідеміологія : підручник / А.М. Андрейчин та ін. ; за ред. І.П. Колеснікової. Вінниця : Нова книга, 2012. 576 с.
9. Корнієнко І.О., Воронова О.Ю. Статистичні методи у психології : курс лекцій з дисципліни для студентів денної та заочної форми навчання спеціальності 053 Психологія ОС «Магістр». Мукачево : МДУ, 2019. 44 с.
10. Костецька В.В., Кізілова Н.М. Математичне моделювання динаміки пандемії COVID-19. *Вісник Харківського національного університету імені В.Н. Каразіна, серія «Математичне моделювання. Інформаційні технології. Автоматизовані системи управління»*. 2020. Вип. 48. С. 65–71.
11. Мармоза А.Т. Теорія статистики : підруч. Київ : «Центр учбової літератури», 2013. 592 с.
12. Методичні рекомендації до практичних занять за предметом «Біостатистика» / розроб. Г.О. Слабкий, І.С. Миронюк, В.Й. Білак-Лук'яничук та ін. Ужгород, 2020. 155 с.
13. Миронюк І.С., Брич В.В. Методичні рекомендації до практичних занять з дисципліни «Громадське здоров'я з основами епідеміології». Ужгород, 2018. 58 с.
14. Нессонова Т.Д., Подаваленко А.П., Білера Н.В. Дослідження зв'язку перебігу пандемії COVID-19 з економічними показниками областей України: матеріали міжнар. міждисциплінарної наук.-практ. Інтернет-конф. «Медичні та біологічні науки: міждисциплінарний аспект» (Харків, 17 жовт., 2024). Харків : СГ НТМ «Новий курс», 2024. С. 221–225.
15. Основне про аналіз даних: посібник для початківців. URL: <https://www.microsoft.com/uk-ua/microsoft-365/excel/data-analysis>
16. Панченко Л.Ф. Математичні та статистичні методи аналізу соціологічної інформації: Практикум. Київ : КПІ імені І. Сікорського, 2018. 289 с.
17. Подаваленко А.П., Малиш Н.Г. Епідеміологічна діагностика – методична основа епідеміологічного нагляду. *Превентивна медицина. Теорія і практика*. 2023. № 2. С. 34–39.
18. Подаваленко А.П., Нессонова Т.Д., Білера Н.В. Вивчення зв'язку перебігу пандемії COVID-19 з демографічними процесами в областях України : матеріали наук.-практ. конф. з між-нар. участю, присвяч. щорічним читанням пам'яті академіка Л.В. Громашевського «Інфекційні хвороби сучасності: етіологія, епідеміологія, діагностика, лікування, профілактика, біобезпека» (Київ, 15 жовт. 2024). Київ : МОЗ, НАМН, ДУ «ІЕІХ ім. Л.В. Громашевського». С. 113–115.
19. Про затвердження критеріїв, за якими визначаються випадки інфекційних та паразитарних захворювань, які підлягають реєстрації : наказ МОЗ України № 905 від 28.12. 2015 р.

20. Про затвердження Порядку ведення обліку, звітності та епідеміологічного нагляду (спостереження) за інфекційними хворобами та Переліку інфекційних хвороб, що підлягають реєстрації: наказ МОЗ України № 1726 від 30.07.2020 р. URL: <https://zakon.rada.gov.ua/laws/show/z1333-20#Text>

21. Про захист населення від інфекційних хвороб : Закон України від 06.04.2000 № 1645-III. *Відомості Верховної Ради України (ВВР)*. 2000. № 29. Ст. 228. URL: <https://zakon.rada.gov.ua/laws/show/1645-14#Text>

22. Про систему громадського здоров'я : Закон України від 06.09.2022 № 2573-IX. *Відомості Верховної Ради України (ВВР)*. 2023. № 26. Ст. 93.

23. Томашевський О.В., Сніжної Г.В., Антонченко Н.А. Методичні вказівки до лабораторних та практичних робіт з дисципліни «Комп'ютерні технології статистичної обробки даних» для студентів спеціальностей 175 «Інформаційно-вимірювальні технології», 176 «Мікрота наносистемна техніка» денної та заочної форм навчання. Запоріжжя : НУ «Запорізька політехніка», 2023. 44 с.

24. Чемич М.Д., Малиш Н.Г., Чемич О.М., Ільїна Н.І. Епідеміологія у схемах : навч. посіб. Вінниця: Нова книга, 2020. 256 с.

25. Яровий А.Т., Страхов Є.М. Аналіз часових рядів : навч.-метод. посіб. Одеса: Освіта України, 2019. 109 с.

26. Alexei Korzh, Marine Georgiyants, Alla Podavalenko, Tetyana Nessonova, Valentyna Maslova, Vladislav Bereznyakov, Irina Asoyan. Conditions for the expansion of COVID-19 in Ukraine. *Bangladesh Journal of Medical Science*. 2023. Vol. 22, № 4. P. 778–784. DOI: 10.3329/bjms.v22i4.67108

27. Alla P. Podavalenko, Tatjana D. Nessonova, Victorija I. Zadorozhna, Larisa M. Hrytsenko. Epidemiological analysis of pertussis morbidity in Ukraine. *Wiadomosci Lekarskie*. 2021. Vol. 74(7). P. 1628–1633. DOI: 10.36740/WLek202107114

28. Höhle M., Paul M. Count data regression charts for the monitoring of surveillance time series. *Comput Stat Data Anal*. 2008. Vol. 52. P. 4357–4368.

29. Serfling R.E. Methods for Current Statistical Analysis of Excess Pneumonia-influenza Deaths. *Public Health Reports*. 1963. Vol. 78, № 6. P. 494–506.

30. Siettos C.I., Russo L. Mathematical modeling of infectious disease dynamics. *Virulence*. 2013. Vol. 4. № 4. P. 295–306.

ЕЛЕКТРОННІ РЕСУРСИ

1. Офіційний сайт Міністерства Охорони Здоров'я України. URL: <http://www.moz.gov.ua/>

2. Рамка цифрової компетентності працівника охорони здоров'я України: МОЗ України. URL: <https://moz.gov.ua/uploads/ckeditor/База%20знань%20EHEALTH/Рамка%20цифрової/Рамка%20цифрової%20компетентності%20працівника%20охорони%20здоров'я%20України.%20Версія%201.0.pdf>

3. Сайт центру громадського здоров'я МОЗ України. URL: <https://phc.org.ua/>

4. Сайт міжнародного центру профілактики та контролю захворювань. URL: <https://www.cdc.gov/>

5. Сайт European Centre for Disease Prevention and Control. URL: <https://www.ecdc.europa.eu/en>

Навчальне видання

Подаваленко Алла Павлівна
Нессонова Тетяна Дмитрівна
Білера Наталія Владиславівна

ЕПІДЕМІОЛОГІЧНА ДІАГНОСТИКА. СТАТИСТИЧНІ МЕТОДИ ДОСЛІДЖЕННЯ

**Навчальний посібник
для слухачів курсів тематичного удосконалення,
слухачів циклів спеціалізації та лікарів-інтернів**

Відповідальний за випуск А. П. Подаваленко



Редактор М. В. Тарасенко
Комп'ютерна верстка О. Ю. Лавриненко

Формат А4. Ум. друк. арк. 17,0. Зам. № 25-66.

**Редакційно-видавничий відділ
ХНМУ, пр. Науки, 4, м. Харків, 61022
izdatknmurio@gmail.com, vid.redact@knu.edu.ua**

Свідоцтво про внесення суб'єкта видавничої справи до Державного реєстру видавництв, виготівників і розповсюджувачів видавничої продукції серії ДК № 3242 від 18.07.2008 р.