МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені І. І. МЕЧНИКОВА

Д. В. МЕЛКОНЯН, В. В. ЯВОРСЬКА

СТАТИСТИКА В ТУРИЗМІ

НАВЧАЛЬНИЙ ПОСІБНИК

ОДЕСА
ОНУ
2021

УДК 624.131.4
М47

Рецензенти:

О. О. Любіцева, доктор географічних наук, професор, завідувач кафедри країнознавства та туризму Київського національного університету імені Тараса Шевченка;

О. А. Жерновникова, доктор педагогічних наук, професор, завідувач кафедри математики Харківського національного педагогічного університету імені Г. С. Сковороди.

Рекомендовано вченою радою
ОНУ імені І. І. Мечникова.
Протокол № 13 від 29.06.2021 р.

Мелконян Д. В.

М47 Статистика в туризмі : навч. посіб. / Д. В. Мелконян, В. В. Яворська. – Одеса : Одес. нац. ун-т ім. І. І. Мечникова, 2021. – 195 с.

ISBN 978-617-689-431-5

В навчальному посібнику наведено основні поняття, категорії і методи статистичної науки, принципи застосування методів статистичного дослідження для розв'язання конкретних задач в сфері туризму. Розглядається застосування табличного процесора MS Excel для статистичного аналізу даних сфери туризму. Кожен розділ супроводжується прикладами, питаннями для самоконтролю та завданнями для виконання під час аудиторних занять, а також для самостійної роботи студентів.

Для здобувачів вищої освіти ступеня бакалавра спеціальності 242 «Туризм».

УДК 624.131.4

ISBN 978-617-689-431-5

ЗМІСТ

ВСТУП

Відповідно до Стандарту вищої освіти за спеціальністю 242 «Туризм» для першого (бакалаврського) рівня вищої освіти, процес навчання здобувачів ступеня бакалавра спрямований, з-поміж іншого, на формування такої фахової компетентності, як здатність здійснювати моніторинг, інтерпретувати, аналізувати та систематизувати туристичну інформацію, уміння презентувати туристичний інформаційний матеріал. Одним з видів інформації, яким користується фахівець-туризмознавець, є статистична інформація, яку він повинен вміти узагальнювати, аналізувати, кількісно оцінювати і робити висновки, важливі з погляду ефективності організації і розвитку туристичної діяльності.

Формування у здобувачів вищої освіти ступеня бакалавра теоретичних знань і практичних навичок у галузі статистики для подальшого використання в професійній діяльності забезпечує навчальна дисципліна «Статистика в туризмі». В результаті навчання дисципліни студент повинен знати джерела статистичної інформації; статистичні методи кількісного і якісного дослідження процесів і явищ в туризмі; способи практичного розв’язання статистичних задач зі сфери туризму з використанням комп’ютерної техніки. Навчальна дисципліна прищеплює вміння: давати статистичну оцінку характеристик процесів та явищ в галузі туризму; проводити дисперсійний, кореляційний і регресійний аналіз статистичних даних у сфері туризму; визначати достовірність оцінок кількісних характеристик процесів та явищ в туризмі; логічно і аргументовано формулювати висновки за результатами статистичного дослідження; застосовувати комп’ютерну техніку для обробки статистичних даних, які характеризують процеси та явища в галузі туризму.

Навчальний посібник укладений у відповідності до Навчальної та Робочої програм з дисципліни «Статистика в туризмі» і складається з семи розділів.

У першому розділі розглядаються основні поняття і категорії

статистики; стадії статистичного дослідження: статистичне спостереження; зведення і групування результатів спостереження; аналіз отриманих узагальнюючих показників.

Другий розділ присвячений таким поняттям математичної статистики, як генеральна сукупність і вибірка; особливі форми групування одиниць за тією чи іншою ознакою – статистичні ряди розподілу та їх обробка. Наведено графічне зображення рядів розподілу і статистичних даних у сфері туризму.

В третьому розділі розглядаються основні статистичні характеристики варіаційних рядів: середні величини, показники варіації і умови їх застосування.

В четвертому розділі наводяться найбільш використовувані для вивчення соціально-економічних явищ закони розподілу випадкових величин і основні параметри цих законів.

П’ятий розділ присвячено розгляду кореляційного і регресійного аналізу, тобто способів встановлення залежності між ознаками і виду цієї залежності.

В шостому розділі висвітлюються питання перевірки гіпотез про вид закону розподілу, значущість коефіцієнта кореляції і рівняння регресії.

Сьомий розділ присвячений статистичному аналізу даних у сфері туризму за допомогою пакета програм MS Excel.

Питання для самоконтролю і завдання для самостійного виконання по кожному розділу спрямовані на закріплення теоретичного матеріалу і отримання навичок самостійного дослідження процесів і явищ у сфері туризму.

У навчальному посібнику розглянуті далеко не всі використовувані сьогодні на практиці статистичні методи. Автори навели лише ті методи, які є необхідними при обробці даних зі сфери туризму і аналізі отриманих результатів.

Навчальний посібник розрахований на 40 годин аудиторних занять та 35 годин самостійної роботи.

Автори навчального посібника висловлюють щиру подяку

рецензентам – доктору географічних наук, професору Любіцевій О. О. (Київський національний університет імені Тараса Шевченка) та доктору педагогічних наук, професору Жерновниковій О. А. (Харківського національного педагогічного університету імені Г. С. Сковороди) за конструктивні зауваження та рекомендації.

РОЗДІЛ 1
ОСНОВНІ КАТЕГОРІЇ СТАТИСТИКИ. СТАДІЇ СТАТИСТИЧНОГО ДОСЛІДЖЕННЯ

1.1. Основні поняття і категорії

Статистика – наука, яка вивчає кількісний аспект якісно визначених масових соціально-економічних, природних явищ і процесів, їхню структуру та розподіл у часі і просторі, виявляє діючі кількісні залежності, тенденції та закономірності.

Статистика вивчає свій предмет за допомогою певних *категорій*, тобто понять, що відображають найзагальніші та суттєві властивості, ознаки, зв'язки і відношення предметів та явищ об'єктивного світу.

Найважливішою категорією статистики є *статистична сукупність* – велика кількість відносно однорідних, але індивідуально розрізнених об'єктів, що характеризуються набором ознак і об'єднуються для спільного (групового) вивчення. Кожен елемент даної множини називається *одиницею* статистичної сукупності і характеризується спільними властивостями, які іменуються в статистиці *ознаками* і є важливою категорією статистичної науки. Для системи туроператорів і турагентств одиницею сукупності може бути окремий туроператор або турагентство. У деяких випадках для однієї й тієї самої сукупності можна виділити різні групи одиниць. Наприклад, при вивченні статево-вікової структури туристів одиницею є окремий турист, а при вивченні їхніх витрат у туристичній поїздці одиницею може бути середня витрата туриста.

Ознака – це об'єктивна характеристика одиниці статистичної сукупності, характерна риса або властивість, яка може бути визначена або виміряна. Саме значення різних ознак спостерігаються і реєструються на першій стадії статистичного дослідження – стадії статистичного спостереження.

Ознаками, що характеризують індустрію туризму, є виручка від реалізації туристської продукції, прибуток, витрати, чисельність

співробітників, кількість туристів, яких вони обслужили та ін.

Ознаками туриста є вік, стать, місце проживання, професія, витрати під час турпоїздки, вибір країни для туризму та ін.

Загальна кількість одиниць, що утворюють статистичну сукупність, називається *об'ємом сукупності*.

Об'єм сукупності слід відрізняти від *об'єму ознаки*, тобто сумарного значення ознаки за всіма одиницями досліджуваної сукупності. Наприклад, кількість туроператорів – це об'єм сукупності, а загальна кількість всіх турпакетів, реалізованих туроператорами, – це об'єм ознаки.

Найважливішою характеристикою статистичної сукупності є її *однорідність*. Однорідною є сукупність, одиниці якої близькі між собою за значеннями ознак, істотних для даного дослідження або ж належать до одного й того ж типу. Багато методів і прийомів статистичного дослідження можуть бути застосовані тільки до однорідних сукупностей. Важливо зазначити, що однорідність не означає повної відповідності всіх одиниць сукупності, а лише припускає наявність для всіх одиниць сукупності основної властивості, якості, типовості. Одна й та сама сукупність одиниць може бути однорідною за однією ознакою і неоднорідною – за іншою. Однорідність одиниць статистичної сукупності формується під впливом різних факторів, які створюють те спільне, що об'єднує одиниці сукупності. Однак ці ж чинники формують те, що відрізняє одну одиницю сукупності від іншої. Інакше кажучи, ці фактори створюють однорідність в цілому і неоднорідність в точці. У статистичній сукупності ці відмінності (тобто неоднорідності) частіше мають кількісну природу.

Кількісна зміна значення статистичної ознаки при переході від одного її елемента до іншого називається *варіацією*. Варіація – це одне з найважливіших властивостей статистичної сукупності, яка виникає під впливом різних факторів, як зовнішніх, так і внутрішніх. Статистика кількісно оцінює вплив кожного фактора на варіацію конкретної ознаки. Соціально-економічні явища, в тому число

туризм, як правило, мають велику варіацію. Наприклад, варіація туроператорів і турагентств країни за кількістю обслуговуваних ними туристів складається під впливом великої кількості факторів: економічних, політичних, соціальних, екологічних та ін.

Важливою категорією статистики є *статистична закономірність*, яка притаманна багатьом явищам і яка чітко виявляється лише при досить великій кількості спостережень.

Закономірністю в цілому прийнято називати повторюваність, послідовність і порядок змін в явищах. Оскільки статистична закономірність виявляється в результаті вивчення масових статистичних даних, це обумовлює її взаємозв'язок з Законом великих чисел, тобто статистичні закономірності є наслідком дії Закону великих чисел. Закон великих чисел у простому формулюванні гласить, що при великій кількості спостережень випадкові відхилення узагальнюючого показника у спостережуваних одиниць взаємно погашаються, в результаті чого чітко проявляються найбільш суттєві сторони досліджуваного явища або процесу. Інакше кажучи, поодинокі явища більшою мірою схильні до дії випадкових і несуттєвих факторів, ніж маса явищ в цілому.

Таким чином, закономірність, виявлену на основі масового спостереження, називають *статистичною закономірністю*. Вона проявляється не в кожному окремому, індивідуальному явищі, а в масі однорідних явищ, при узагальненні даних статистичної сукупності, тобто в середньому.

Статистичне дослідження, незалежно від його масштабів і цілей, завершується розрахунком і аналізом різних за видом і формою вираження статистичних показників. *Статистичний показник* – це узагальнена кількісна характеристика деякої властивості досліджуваної сукупності в цілому. Статистичний показник характеризує всю сукупність в цілому і може розраховуватися як сума абсолютних значень ознаки (наприклад, кількість турпакетів, реалізованих усіма туроператорами і турагентствами конкретної країни за певний рік або кількість співробітників туристичної компанії); як середнє значення

ознаки за сукупністю (наприклад, середня заробітна плата працівників великих туроператорів); як відносна величина (різні індекси; ступінь мінливості тривалості перебування туристів по регіонах країни та ін.).

Необхідність статистичного вивчення туризму обумовлена потребою в отриманні об'єктивної та достовірної інформації про стан і розвиток туристичної галузі та оцінки її внеску в загальну величину валового регіонального продукту. Крім цього, статистичне вивчення туризму необхідне для оцінки туристичних потоків і навантаження на міську інфраструктуру, для задоволення туристського попиту і відповідності споживчих очікувань пропозиції на ринку туристичних послуг.

Туризм є об'єктом статистичного дослідження, а *статистичне дослідження в туризмі* являє собою науково організований процес збирання даних про досліджувані явища (процеси, об'єкти) в туризмі і отримання статистичної інформації, важливої з погляду туризму.

Предметом вивчення статистики в туризмі є кількісна характеристика розвитку туризму і туристських послуг, його результатів і чинників, а також оцінка внеску туризму в економіку країни.

Статистичний аналіз пов'язаний з обробкою великої кількості даних, оскільки надійність висновків підвищується при збільшенні кількості оброблюваних даних. Туризм є інформаційно насиченою діяльністю. Однією з характерних особливостей туристської діяльності є велика кількість і різноманітність інформаційних потоків, які супроводжуються високою швидкістю обмінних операцій. Збір, обробка і передача інформації в індустрії туризму є надзвичайно важливою для її щоденного функціонування.

Інформація про успішність діяльності будь-якого суб'єкта економіки, в тому числі туристського сектора, міститься в статистичних даних, що характеризують його стан і розвиток. У зв'язку з цим необхідно проводити *статистичні дослідження*, використовувати статистичні методи обробки даних, щоб отримати надійні висновки для прийняття ефективних економіко-управлінських рішень. В нинішній

час для обробки та аналізу інформації в галузі туризму використовують різні наукові методи, в тому числі *методи математичної статистики*.

1.2. Статистичне спостереження

Статистичне дослідження складається з трьох стадій: 1) *статистичне спостереження*; 2) *зведення і групування* результатів спостереження; 3) *аналіз* отриманих узагальнюючих показників. Всі три стадії пов'язані між собою, і на кожній з них використовуються спеціальні методи, обумовлені змістом виконуваної роботи.

Процес статистичного дослідження починається з етапу збору первинного статистичного матеріалу, перевірки його повноти і достовірності.

Спостереження пов'язані з реєстрацією природних і соціально-економічних процесів, явищ, які піддаються вивченню. Статистичний аналіз в переважній більшості випадків оперує кількісними даними.

Джерелами статистичної інформації в туризмі є спостереження і звітності різних організацій. При цьому дані систематично подаються у вигляді квартальної, піврічної і річної звітності: 1. Статистична звітність державного і регіонального спостереження з різних галузей економіки: готелі та ресторани, транспорт і зв'язок, оренда і надання послуг, охорона здоров'я і надання соціальних послуг, надання інших комунальних, соціальних і персональних послуг. 2. Звітність міністерств і відомств (Міністерства розвитку економіки, торгівлі та сільського господарства України, Міністерства культури України, Міністерства інфраструктури України, Державної прикордонної служби України та інших відомств). 3. Вибіркові обстеження, що проводяться Державною службою статистики України і територіальними органами статистики, а також іншими установами та організаціями.

Для *статистичного спостереження* за діяльністю туристичних фірм використовується система показників статистики туризму, серед яких найпростішими за змістом є обсяг туристичних потоків; тривалість поїздок туристів; об'єм туристських витрат; діяльність

підприємств туризму і гостинності (наприклад, число туристських і готельних підприємств, номерний фонд готелів та їх завантаження, коефіцієнт використання готельного фонду) та ін.

Якщо при зборі первинних даних (про туристські потоки, доходи та витрати, діяльність туристських підприємств і т. п.) допущена помилка, або матеріал виявився неякісним, то це може вплинути на ефективність використання інших методів, на правильність і достовірність висновків.

Для розкриття закономірностей розвитку будь-якого соціально-економічного процесу або явища, в тому числі туризму, використання тільки методів статистичного спостереження є недостатнім. Зібраний в процесі статистичного спостереження матеріал являє собою розрізнені первинні відомості про окремі одиниці досліджуваного явища. У такому вигляді матеріал ще не характеризує явище або процес у цілому: не дає уявлення ані про величину (чисельність) явища, ані про його склад, ані про розмір характерних ознак, ані про суть зв’язків цього явища з іншими і т. д. Інакше кажучи, інформацію потрібно не тільки зібрати, але й систематизувати, а потім і проаналізувати. Для цього статистична наука використовує групу методів, пов’язаних зі зведенням і групуванням даних. Ці методи формують другий етап статистичного дослідження в туризмі – етап систематизації та класифікації зібраної інформації.

1.3. Зведення і групування статистичних даних

Зведення і групування статистичних даних передбачає створення деяких сховищ, в яких дані систематизовані і класифіковані, тобто організовані в певні структури. На цьому етапі відбувається упорядкування даних про одиниці сукупності, тобто об’єкти, що характеризують туристську сферу (про відвідувачів, підприємства, послуги тощо). При цьому структура баз даних і способи роботи з ними багато в чому визначаються цілями статистичного аналізу і ступенем апріорних знань про властивості досліджуваних об’єктів, явищ і процесів.

Якщо при статистичному спостереженні отримують відомості, що описують кожну одиницю, то дані зведення і групування характеризують всю статистичну сукупність або окремі її частини. У методологічному плані на цій стадії відбувається перехід від характеристик одиничного факту до характеристик їх сукупності.

Зведення – це наукова обробка первинних даних з метою отримання узагальнених характеристик досліджуваного явища за низкою суттєвих для нього ознак. За глибиною і точністю обробки матеріалу розрізняють зведення *просте* і *складне*. *Просте зведення* – це операція з підрахунку загальних підсумків за сукупністю одиниць спостереження. *Складне зведення* – це комплекс операцій, що охоплюють групування отриманих при спостереженні матеріалів, складання системи показників для характеристики типових груп і підгруп досліджуваної сукупності явищ, підрахунок кількості одиниць і підсумків в групах та підгрупах і оформлення результатів цієї роботи у вигляді статистичних таблиць.

В цілому зведення є найпростішою систематизацією одиничних фактів, операцією з підрахунку загальних підсумків. Наприклад, дані про кількість туристів, яких обслужили туроператори за певний рік, можна подати спочатку за окремим туроператором, потім звести їх на рівні конкретного регіону і лише потім підрахувати загальний підсумок за всіма туроператорам. Проте, такий спосіб підходить швидше для оперативного обліку або поточного моніторингу ситуації. Він не дозволяє провести серйозний аналіз, виявити закономірності в структурі або динаміці явищ, тому статистика активно використовує метод *групування*.

Суть *групування* полягає не просто в підведенні підсумку, а в розбивці досліджуваної сукупності на групи залежно від значень варіювальної ознаки. Туристів, яких обслужили туроператори, можна не тільки перерахувати окремо (в термінах статистики – застосувати просте зведення), але й розділити на групи за такими ознаками, як вік, стать, освіта, мета поїздки і т. д. Туристські організації можна групувати за прибутком і рентабельністю, формами власності;

туристичні послуги – за якістю, ціною, споживачами; країни – за розміром туристського валового внутрішнього продукту, рівнем безпеки і т. п. Отже, статистичні групування дозволяють виокремити з маси вихідного статистичного матеріалу однорідні групи одиниць, що мають спільну схожість в якісному і кількісному відношеннях.

Основне призначення групування полягає в тому, що даний метод забезпечує систематизацію інформації, отриманої в результаті спостереження; узагальнення та подання даних в компактному і наочному вигляді. Крім того, групування створює основу для подальшого аналізу і прогнозування.

Існують різні види статистичних групувань. Залежно від кількості групувальних ознак групування ділять на *прості* і *багатовимірні*. Наприклад, всіх туристів, яких обслужило певне турагентство, можна розділити за статтю і підрахувати кількість туристів-чоловіків і кількість туристів-жінок. В цьому разі отримаємо *просте групування*. Проте, сформовані групи можна поділити на підгрупи вже за іншою ознакою, наприклад, за освітою: скільки чоловіків і скільки жінок мають вищу, а скільки – середню освіту. У цьому разі буде утворено складне (комбіноване) групування. Якщо продовжити таке дроблення ознак і зробити їх більше чотирьох, то таке групування буде називатися *багатовимірним групуванням*. При цьому потрібно зважати на те, що групування з великою кількістю підгруп стає надмірно навантаженим інформацією, і тому його складно читати.

За допомогою методу групувань вирішуються такі основні завдання: виокремлення типів явищ; вивчення структури явища і структурних зрушень, що відбуваються в ньому; виявлення зв’язку і залежності між явищами.

Статистичні групування за завданнями, що вирішуються за їхньою допомогою, поділяються на: *типологічні, структурні* й *аналітичні*.

Типологічне групування – це поділ різнорідної сукупності на окремі якісно однорідні групи (класи, типи, види). Як приклад можна

розглянути групування організацій (за роками і за кількістю), які здійснюють свою діяльність в індустрії туризму та гостинності, розбивши їх на різні типи: готелі та аналогічні засоби розміщення, санаторно-курортні організації та організації відпочинку, установи культурно-дозвіллєвого типу, дитячі оздоровчі установи і т. п.

Структурне групування – це поділ якісно однорідної сукупності на групи за певними ознаками, що характеризують її склад і структуру.

За допомогою структурних угруповань можна вивчати структуру капіталу і фондів туристських підприємств, структуру туристського внутрішнього валового продукту, статевовіковий склад споживачів туристичних послуг, структуру витрат при здійсненні туристських поїздок і т. д.

Прикладом структурного групування можуть служити поїздки, які здійснили громадяни України за кордон і іноземні громадяни – в Україну, з різними цілями і в різні роки. За такими даними можна виявити структурне зрушення в перевагах, тобто, з якою метою (службова поїздка, туризм та ін.) більше або менше поїздок здійснюють українські або іноземні громадяни.

Аналітичне групування – це групування, що виявляє взаємозв'язок ознак, які характеризують одиниці однієї й тієї самої сукупності. Дані групуються за однією ознакою (незалежна характеристика), поряд з якою фіксуються значення іншої ознаки (залежна характеристика). На основі такого групування проводиться аналіз поведінки (зростання або зниження) залежної характеристики при зміні незалежної. Можна вивчати, наприклад, як активи впливають на прибуток, доходи – на витрати, продуктивність – на собівартість; як впливає курс гривні відносно якоїсь іноземної валюти на кількість українських туристів, що виїжджають за кордон, і т. д.

Результати статистичного групування подаються у вигляді статистичних таблиць, які є найбільш раціональною, систематизованою і наочною формою репрезентації даних про одиниці сукупності, тобто об'єкти, що характеризують туристську

сферу (відвідувачів, підприємства, послуги тощо). Такі таблиці наведені, наприклад, на сайті Державної служби статистики України (http://www.ukrstat.gov.ua), в розділі «Туризм». У них відображена туристична діяльність України в різні роки: кількість суб'єктів туристичної діяльності за регіонами; кількість туристів, обслугованих туроператорами та турагентами, за регіонами; кількість і вартість реалізованих туроператорами та турагентами туристичних пакетів; розподіл виїзних і внутрішніх туристів, обслугованих туроператорами та тур агентами, за метою поїздки та ін.

1.4. Статистичний аналіз і узагальнення статистичних даних

Аналіз статистичних даних – це завершальна стадія статистичного дослідження. Його суть полягає в здійсненні науково обґрунтованої інтерпретації результатів статистичного дослідження, у виявленні структурних і динамічних особливостей досліджуваного об'єкта (явища, процесу), закономірностей його поведінки і суттєвих взаємозв'язків, а також його місця в системі відношень з іншими суміжними фактами і процесами.

Етапами статистичного аналізу є формулювання мети аналізу; критична оцінка даних; порівняльна оцінка і забезпечення порівнянності даних; формування узагальнюючих показників; фіксація і обґрунтування істотних властивостей, особливостей, подібностей і відмінностей, зв'язків і закономірностей досліджуваних явищ і процесів; формулювання висновків про стан і розвиток досліджуваного явища або процесу.

Таким чином, статистичний аналіз в широкому сенсі слова – це статистичне дослідження, яке охоплює застосування всіх специфічних (спеціальних) методів статистики, а саме: статистичне спостереження, статистичне зведення і групування, виведення узагальнюючих статистичних показників до самого статистичного аналізу.

Статистика в туризмі використовує весь арсенал сучасних

математико-статистичних методів для встановлення логічного зв’язку показників – абсолютних, відносних, середніх величин та індексів, які можуть дати уявлення про зміну в часі різних факторів, що впливають на діяльність туристських організацій. Найчастіше використовуються такі статистичні (математико-статистичні) методи: метод перевірки гіпотез, кореляційний, регресійний аналізи, методи багатовимірної статистики, статистичне моделювання та ін.

Реалізація методів статистичного аналізу в сучасних умовах передбачає використання комп’ютерних технологій. В нинішній час існує досить велика кількість програмної підтримки методів статистичних досліджень, наприклад: Microsoft Excel, Statistica, MatLab, MathCAD та ін.

Питання для самоконтролю

1.1. Назвіть основні поняття і категорії статистики.
1.2. Що таке статистична сукупність?
1.3. Що таке ознака?
1.4. Чим відрізняється об’єм сукупності від об’єму ознаки?
1.5. Наведіть приклади ознак, які характеризують індустрію туризму.
1.6. Що являє собою однорідна статистична сукупність?
1.7. В чому виражається варіація статистичної ознаки?
1.8. Що таке статистична закономірність?
1.9 Що таке статистичний показник? Наведіть приклади показників статистики зі сфери туризму.
1.10. Чим обумовлена необхідність статистичного вивчення туризму?
1.11. Що є предметом вивчення статистики туризму?
1.12. Що являє собою статистичне дослідження в туризмі?
1.13. Перелічить основні етапи статистичного дослідження.
1.14. У чому полягають сутність і особливість статистичного спостереження в туризмі?
1.15. Що є джерелами статистичної інформації в туризмі?
1.16. Які помилки можуть виникати при статистичному спостереженні, і чи можуть ці помилки вплинути на правильність і достовірність висновків?

1.17. В чому полягає суть і значення зведення і групування статистичних даних в туризмі?

1.18. Які існують види статистичних групувань?

1.19. Які завдання вирішує статистика за допомогою методу групувань?

1.20. На які типи поділяють статистичні групування за кількістю групувальних ознак і за завданнями?

1.21. Що являє собою типологічне групування?

1.22. Що являє собою структурне групування?

1.23. Як називається групування, яке дозволяє виявити зв'язки між явищами, що вивчаються ?

1.24. В чому полягає суть аналізу статистичних даних?

1.25. Який розділ сучасної математики найбільш широко використовується у статистиці туризму?

1.26. Які математико-статистичні методи найбільш часто використовуються в туризмі?

1.27. Перелічить сучасні статистичні пакети прикладних програм.

Завдання для самостійного виконання

Завдання 1.1. Подайте в таблиці 1 *типологічне групування* на основі даних, що характеризують діяльність українських підприємств в індустрії туризму і гостинності, розбивши їх на різні типи: готелі та аналогічні засоби розміщення, санаторно-курортні організації і організації відпочинку, дитячі оздоровчі установи та ін. Для заповнення таблиці використовуйте дані Державної служби статистики України.

Таблиця 1

Групування українських підприємств індустрії туризму і гостинності

Тип підприємства	Рік				
	2015	2016	2017	2018	2019

Завдання 1.2. Подайте в таблиці 2 *структурне групування* на основі даних, що характеризують поїздки, які здійснили громадяни України за кордон і іноземні громадяни – в Україну з різними цілями і в різні роки, наприклад, службова, туризм, приватна поїздка та ін. Джерело даних – Державна служба статистики України.

Таблиця 2

Поїздки громадян України за кордон та іноземних громадян в Україну

Мета поїздки	Кількість поїздок українських громадян за кордон, % від загальної кількості		Кількість поїздок іноземних громадян в Україну, % від загальної кількості	
	2018	2019	2018	2019

Завдання 1.3. Наведіть в таблиці 3 *аналітичне групування* на основі даних, що характеризують залежність ціни за номер від кількості місць і рівня «зірковості» готелів, які бронює туроператор для майбутнього розміщення своїх туристів.

Таблиця 3

Умови проживання в готелях

Тип готелю (рівень «зірковості»)	Середня ціна за номер, євро		
	Одномісний номер (SNGL)	Двомісний номер (DBL)	Трьохмісний номер (TRPL)

РОЗДІЛ 2
СТАТИСТИЧНІ СУКУПНОСТІ. ОСНОВНІ ПОНЯТТЯ І ВИЗНАЧЕННЯ

2.1. Генеральна сукупність і вибірка

Найважливішим поняттям математичної статистики є *генеральна сукупність* – множина, що складається з однорідних елементів, які підлягають дослідженню щодо певної ознаки. Генеральну сукупність також визначають як повний набір всіх можливих значень, які може приймати випадкова величина. У зв'язку з тим, що елементами генеральної сукупності можуть бути як люди, предмети, міста і т. д., так і числа, і оскільки в кінцевому підсумку обробці підлягають значення, які приймає випадкова величина, то в подальшому викладі поняття генеральної сукупності і сукупності всіх значень випадкової величини розрізнятися не будуть.

Відібрана тим чи іншим способом частина одиниць генеральної сукупності називається *вибірковою сукупністю* або *вибіркою*. Вибірку вивчають для того, щоб зробити висновок про всю генеральну сукупність.

При статистичному аналізі передбачається, що вибіркова сукупність відповідає вимогам однорідності, незалежності і випадковості.

О*б'ємом* генеральної (N) або вибіркової (n) сукупності називають загальну суму членів цих сукупностей. Наприклад, якщо з 500 туристських компаній відібрано для дослідження 50, то об'єм генеральної сукупності $N = 500$, а об'єм вибірки $n = 50$.

Встановлено, що чим більшим є об'єм вибірки, тим краще вибірка відбиває генеральну сукупність. Звичайно, оптимальний об'єм вибірки є пропорційним ступеню мінливості ознаки. Якщо ознака сильно змінюється, то для отримання надійних результатів кількість вимірювань повинна бути достатньо великою. Якщо ознака змінюється слабо, то надійний результат може бути отриманий і при

малому об’ємі вибірки. У більшості випадків досить точні результати можна отримати при об’ємі вибірки не менше 30-35.

Добір елементів до вибірки повинен задовольняти такому обов’язковому правилу: кожна одиниця генеральної сукупності має однакову можливість бути відібраною. Така вимога виключає суб’єктивізм, упередженість в дослідженнях. Наприклад, вивчення діяльності лише великих туристських компаній може створити хибне уявлення про розвиток туризму в досліджуваній країні.

Таким чином, у вибірці повинні бути представлені всі можливі значення досліджуваної величини і, приблизно, в тих же пропорціях, з тими ж відносними частотами, що і в генеральній сукупності.

Вибірка повинна бути *репрезентативною,* тобто такою, по якій можна з упевненістю визначити досліджувані характеристики генеральної сукупності. Іншими словами, репрезентативність вибірки полягає в тому, що вона достовірно і повною мірою відбиває всі характеристики генеральної сукупності, частиною якої вона є.

Для отримання репрезентативної вибірки необхідно чітко визначити, що розуміється під генеральною сукупністю. Її кількісний і якісний склад залежить від об’єктів і цілей дослідження, що проводиться. Наприклад, якщо ми хочемо отримати дані про кількість туристів, які в’їжджали в усі регіони України в певному році, то туристи даного конкретного регіону – це вибірка з більш широкої генеральної сукупності – туристів всіх регіонів України. При цьому, не обов’язково, що ця вибірка виявиться репрезентативною. У зв’язку з цим слід намагатися зробити вибірку так, щоб вона найкращим чином представляла всю генеральну сукупність.

Отже, вирішення завдань на ринку туризму з використанням методів математичної статистики починається із складання вибірок, які повинні бути показовими відносно генеральної сукупності.

2.2. Статистичні ряди розподілу та їх обробка

Вибірка є множиною, важко доступною для огляду. Для

подальшого вивчення вибірку піддають перегрупуванню. Будучи основним засобом статистичного аналізу при розрахунку абсолютних і середніх величин, групування дає можливість, наприклад, отримати узагальнюючі характеристики досліджуваних процесів туристського обслуговування з різних напрямків.

Особливою формою групування є так звані *статистичні ряди розподілу*, які утворюються при групуванні одиниць за тією чи іншою ознакою.

Ряди розподілу – це впорядкований розподіл одиниць досліджуваної сукупності на групи за будь-якою варіювальною ознакою. Наприклад, розподіл туристів за цілями поїздки, віком, тривалістю проживання в готелях та ін.

Залежно від ознаки, покладеної в основу утворення ряду розподілу, тобто залежно від того, які ознаки вивчаються, розрізняють *атрибутивні і варіаційні ряди розподілу*.

Атрибутивні ряди – це ряди розподілу, побудовані за якісними ознаками, тобто ознаками, які не мають числового вираження. Вони відбивають стан одиниці сукупності (наприклад, стать туриста, освіта, форма власності туристської компанії і т. д.).

Варіаційні ряди – це ряди розподілу, побудовані за кількісною ознакою, тобто ознакою, що має числове вираження (наприклад, кількість турагентств або туристів, дохід турагентства, вік туриста і т. д.). Крім цього, варіаційним рядом або рядом розподілу називають подвійний ряд чисел, що показує, яким чином числові значення ознаки пов’язані з їхньою повторюваністю в даній статистичній сукупності. Наприклад, у 2017 р. кількість українських туристів, що виїжджали в деякі зарубіжні країни з метою організованого туризму (http://www.ukrstat.gov.ua) представлена нами таким чином (табл. 2.1):

Розташуємо ці дані (дані стовпців 3 та 6, табл. 2.1) в ряд з урахуванням повторюваності варіант в цій сукупності. Як видно, деякі цифри зустрічаються в даному ряду по кілька разів, наприклад, число 4 – 4 рази, 10 – 2 рази, 1 – 8 разів і т. д. Отже, з огляду на кількість повторень, даний ряд можна поділити на кластери і подати в

більш компактній формі (табл. 2.2). Тоді стовпчики 2 і 3 в табл. 2.2 формують варіаційний ряд. З таблиці видно, що, наприклад, по 10 % від загальної кількості туристів виїжджали до Білорусі і Болгарії (рядок 2); по 2 % виїжджали до Ізраїлю і Румунії (рядок 5) і т. д.

Таблиця 2.1

Кількість українських туристів (відсотки від загального турпотоку за 2017 р.), що виїжджали за кордон з метою організованого туризму

№ з/п	Країна	Кількість туристів, %	№ з/п	Країна	Кількість туристів, %
1	**2**	**3**	**4**	**5**	**6**
1.	Австрія	4	11.	Литва	1
2.	Білорусь	10	12.	Німеччина	4
3.	Болгарія	10	13.	Об’єднані Арабські Емірати	7
4.	Вірменія	1	14.	Польща	4
5.	Греція	4	15.	Російська Федерація	1
6.	Грузія	1	16.	Румунія	5
7.	Естонія	1	17.	Туніс	1
8.	Єгипет	12	18.	Туреччина	24
9.	Ізраїль	5	19.	Чехія	3
10.	Латвія	1	20.	Чорногорія	1

Таблиця 2.2

Варіаційний ряд

Кластери	Варіанти, x_i (%)	Число варіант, f_i (частота)
1	**2**	**3**
1. (Австрія, Греція, Німеччина, Польща)	4	4
2. (Білорусь, Болгарія)	10	2
3. (Вірменія, Грузія, Естонія, Латвія, Литва, Російська Федерація, Туніс, Чорногорія)	1	8
4. (Єгипет)	12	1
5. (Ізраїль, Румунія)	5	2
6. (Об'єднані Арабські Емірати)	7	1
7. (Туреччина)	24	1
8. (Чехія)	3	1

Варіаційний ряд дозволяє привести в певний порядок численні вимірювання кожної ознаки комплексу; він наочно показує коливання ознак та їхню різноманітність.

Варіація – це коливання значень величин в ряду розподілу (або мінливість ознак в окремих одиницях сукупності).

Варіаційний ряд складається з двох елементів: *варіант* і *частот*.

Варіанти – окремі значення ознаки, які вона приймає в варіаційному ряді. Їх позначають як $x_1, x_2, x_3, ..., x_n$. Інакше кажучи, $x_1, x_2, x_3, ..., x_n$ – це результати незалежних спостережень величини x_i. Ці результати, записані в порядку їх отримання, утворюють *статистичний ряд*.

Частота f (або *вага*) варіант визначається як кількість окремих варіант або кожної групи варіаційного ряду. Іншими словами, це числа, що показують, скільки разів зустрічаються окремі варіанти в ряду розподілу. Сума всіх частот варіаційного ряду являє собою *загальну кількість спостережень n* (*об'єм вибіркової сукупності* або *вибірки*):

$$f_1 + f_2 + f_3 + ... f_k = \sum_{i=1}^{k} f_i = n. \qquad (2.1)$$

Назва *вага* виражає той факт, що різні значення окремих варіант x_i мають неоднакову *важливість* при розрахунку тієї чи іншої характеристики ряду.

Частоти (ваги) виражають не тільки абсолютними, але й відносними числами – в частках одиниці або у відсотках (наприклад, в табл. 2.1) від загальної кількості варіант, що складають дану сукупність. У таких випадках їх називають *відносними частотами* або *частостями*. Загальна сума відносних частот дорівнює 1 або 100 % (якщо частоти виражені у відсотках від загальної кількості спостережень *n*) і називається *накопиченою частістю*:

$$\frac{f_1}{n}+\frac{f_2}{n}+\frac{f_3}{n}+...+\frac{f_k}{n}=\frac{\sum_{i=1}^{k} f_i}{n}=1 \quad \text{або} \quad \frac{\sum_{i=1}^{k} f_i}{n}\cdot 100=100\ \%. \qquad (2.2)$$

Заміна частот частостями не є обов'язковою, проте виявляється корисною і навіть необхідною тоді, коли доводиться зіставляти один з одним варіаційні ряди, які сильно відрізняються за їхніми об'ємами.

Залежно від характеру варіації ознаки розрізняють на *дискретні* та *інтервальні варіаційні ряди*.

Дискретний варіаційний ряд – це ряд розподілу, в якому групи складені за ознакою, що змінюється безперервно, тобто через певну кількість одиниць. Наприклад, кількість туристів, обслугованих туроператорами по різних регіонах України за певний рік, або дані про кількість туристів, обслугованих турагентами за різними цілями поїздки по Київський області за 2019 р., наведені у таблиці 2.3. Дані запозичені з сайту Державної служби статистики України і оброблені авторами навчального посібника.

Таблиця 2.3

Розподіл туристів, обслугованих турагентами за різними цілями поїздки по Київський області (2019 р.)

Мета поїздки	Кількість туристів
службова, ділова, навчання	1588
дозвілля, відпочинок	55339
лікування	285
спортивний туризм	42

Якщо число f_i (див. рівняння 2.1 або 2.2) є завеликим або близьким до n, то доцільно скласти *інтервальний варіаційний ряд*.

Інтервальний варіаційний ряд розподілу – це ряд розподілу, в якому групувальна ознака, що становить основу групування, може приймати в певному інтервалі будь-які значення, тобто в інтервальному варіаційному ряду варіанти коливаються в певних межах (наприклад, склад групи туристів за віком або ціни в готельних номерах різного рівня комфортності).

Після того, як визначено основу групування, слід вирішити питання про кількість груп, на які треба розбити досліджувану сукупність.

В таблиці 2.4 наведено приклад інтервального варіаційного ряду, який репрезентує вік і кількість туристів, що в'їхали у Львів 10.05.2010 р. (за даними Львівської міської ради, https://www.city-adm.lviv.ua).

Таблиця 2.4

Вік та кількість туристів (%), що в'їхали у Львів (10.05.2010 р.)

Вік туриста, років	Кількість туристів (% від загальної кількості туристів)
18-25	28
26-35	30
36-45	19
46-55	15
56-65	8

При визначенні меж інтервалів статистичних групувань часто виходять з того, що зміна кількісної ознаки призводить до появи нової якості. В цьому разі межа інтервалу встановлюється там, де відбувається перехід від однієї якості до іншої.

В цілому, кількість груп залежить від завдань дослідження і виду показника, покладеного в основу групування, об'єму сукупності, ступеня варіації ознаки. Наприклад, готельне господарство країни може бути представлено різними формами власності: муніципальними підприємствами, приватними, в змішаній власності без іноземного капіталу, відомчими і в державній власності з іноземною участю тощо.

Для побудови інтервального варіаційного ряду необхідно визначити довжину інтервалу та їхню кількість. Якщо варіація ознаки проявляється у порівняно вузьких межах, і розподіл носить

рівномірний характер, то будують групування з рівними інтервалами. *Довжина рівного інтервалу* h визначається за такою формулою:

$$h = \frac{R}{k} = \frac{x_{\max} - x_{\min}}{k}, \qquad (2.3)$$

де R – розмах варіації; k – кількість інтервалів; x_{max}, x_{min} – максимальне та мінімальне значення ознаки в вибіркової сукупності.

Кількість інтервалів k, їхні довжини можуть варіювати залежно від розв'язуваних задач. Для орієнтовної оцінки кількості інтервалів зазвичай використовують формулу Стерджеса:

$$k \approx 1 + 3{,}3221\ lgn, \qquad (2.4)$$

де k – кількість інтервалів (або груп); n – кількість одиниць сукупності (або об'єм вибірки). Згідно з цією формулою, вибір кількості інтервалів залежить від об'єму вибірки.

Кількість інтервалів k також можна приблизно намітити, користуючись такою схемою:

Таблиця 2.5

Кількість інтервалів залежно від об'єму вибірки

Об'єм вибірки, n	Кількість інтервалів, k
25-40	5-6
40-60	6-8
60-100	7-10
100-200	8-12
> 200	10-15

Якщо розмах варіації ознаки в сукупності великий, і значення ознаки варіюють нерівномірно, то використовують групування з нерівними інтервалами. Нерівні інтервали можуть бути прогресивно зростаючими і прогресивно спадними в арифметичній або геометричній прогресії. Величина інтервалу, що змінюється в арифметичній прогресії, визначається за формулою: $h_{i+1} = h_i + a$, де a – константа і може приймати додатне і від'ємне значення. Величина інтервалу, що змінюється в геометричній прогресії, визначається за

формулою: $h_{i+1} = h_i \times b$, де b – константа, яка може бути більше або менше одиниці.

Застосування нерівних інтервалів зумовлено тим, що в перших групах невелика різниця в показниках має велике значення, а в останніх групах ця різниця не є суттєвою. Наприклад, при побудові групування туристських компаній за показником чисельності працівників, який варіює від 20 до 2000 людей, доцільно розглядати нерівні інтервали, тому що враховуються як малі, так і великі компанії: 20-50, 50-110, 110-200, 200-320, 320-470, 470-650, 650-860, 860-1100, 1100-1370, 1370-1670, 1670-2000, тобто величина кожного наступного інтервалу більше попереднього і збільшується в арифметичній прогресії.

Іноді для певних цілей може бути прийнята й інша послідовність розташування результатів спостережень, наприклад, в порядку зростання (або зменшення) числових значень членів вибірки. Такий ряд називають *ранжованим рядом*. Така систематизація вихідних даних дозволяє, з-поміж іншого, легко побачити діапазон зміни ознаки, тобто розмах варіації.

Розглянемо приклад групування статистичних даних.

***Приклад* 2.1.** Необхідно зробити групування суб'єктів туристичної діяльності за об'ємом витрат на вироблення туристського продукту (табл. 2.6, складено авторами).

Таблиця 2.6

Характеристика суб'єктів туристичної діяльності за об'ємом витрат на вироблення туристського продукту

№	Об'єм	№	Об'єм
1	60	9	35
2	40	10	24
3	20	11	45
4	15	12	40
5	12	13	21
6	25	14	50
7	75	15	30
8	100	16	90

Відповідно до завдання групувальною ознакою є об'єм витрат суб'єкта туристичної діяльності на вироблення туристського продукту.

Кількість груп, згідно з формулою Стерджеса (2.4), – $k \approx 1 + 3{,}3221 \cdot lgn = 1 + 3{,}322 \times \lg(16) = 5$. Отже, сукупність необхідно розбити на 5 груп. Мінімальне значення становить 12 млн. грн., максимальне – 100 млн. грн.

Величину рівного інтервалу групування розраховують за формулою (2.3):

$$h = \frac{R}{k} = \frac{x_{\max} - x_{\min}}{k} \quad \frac{100 - 12}{5} = 17{,}6 \text{ млн. грн.}$$

Таким чином, величина інтервалу становить 17,6 млн. грн. Визначимо межі груп (табл. 2.7).

Таблиця 2.7

Розрахунок меж груп і частот

№ групи	Нижня межа	Верхня межа
1.	12	29,6 (12+17,6)
2.	29,6	47,2 (29,6+17,6)
3.	47,2	64,8 (47,2+17,6)
4.	64,8	82,4 (64,8+17,6)
5.	82,4	100 (82,4+17,6)

Результати розрахунків наведено в табл. 2.8. Згідно з даними табл. 2.6 і 2.7, варіаційний ряд розподілу суб'єктів туристичної діяльності за об'ємом витрат на вироблення туристського продукту має такий вигляд:

Таблиця 2.8

Групування суб'єктів туристичної діяльності за об'ємом витрат на вироблення туристського продукту

№ групи	Групи суб'єктів туристичної діяльності за об'ємом витрат	Кількість суб'єктів	Середній об'єм витрат, млн. грн.
1	12 - 29,6	6	20,8

2	29,6 - 47,2	5	38,4
3	47,2 - 64,8	2	56,0
4	64,8 - 82,4	1	73,6
5	82,4 - 100	2	91,2

Таким чином, в результаті проведення зведення і групування статистичної інформації ми розбили сукупність, що складається з 16 суб'єктів туристичної діяльності, на 5 груп. Як групувальну ознаку вибрали показник об'єму витрат суб'єкта. По кожній групі суб'єктів розрахували середній об'єм витрат суб'єкта туристичної діяльності

2.3. Графічне зображення рядів розподілу і статистичних даних у сфері туризму

Графічне представлення статистичних даних є способом наочного зображення результатів статистичного зведення та обробки масового матеріалу. В цілому це логічне продовження статистичних таблиць в процесі обробки статистичних даних. Правильно побудований графік є виразним і доступним, сприяє аналізу явищ, їх узагальненню та виявленню притаманних їм закономірностей.

Ряди розподілу зручніше аналізувати за допомогою їх графічного зображення, оскільки це дозволяє судити про форму розподілу і про характер зміни частот варіаційного ряду. Попередньо дані *групуються*, тобто розміщуються в варіаційний ряд, або розбиваються на певні інтервали або класи.

Ряди розподілу найчастіше зображуються у вигляді *полігону, гістограми, кумуляти* (кумулятивних кривих) та інших графіків, що характеризують особливості їх розподілу.

При зображенні *дискретних* варіаційних рядів використовують *полігон*, при зображенні *інтервальних* варіаційних рядів – *гістограму*. Полігони частот відрізняються від гістограм тим, що точки, що відповідають значенням частот або частостей, з'єднуються відрізками

прямих ліній.

За допомогою *кумуляти* зображується ряд накопичених частот.

Для побудови полігону частот або відносних частот необхідно на осі абсцис відкласти впорядковані за зростанням або спаданням (тобто ранжовані) значення варіант x_i, а на осі ординат – відповідні їм значення частот f_i або значення відносних частот p_i.

Наприклад, на рис. 2.1 наведено *полігон частот*, що характеризує кількість реалізованих туристичним агентством турпакетів різної ціни.

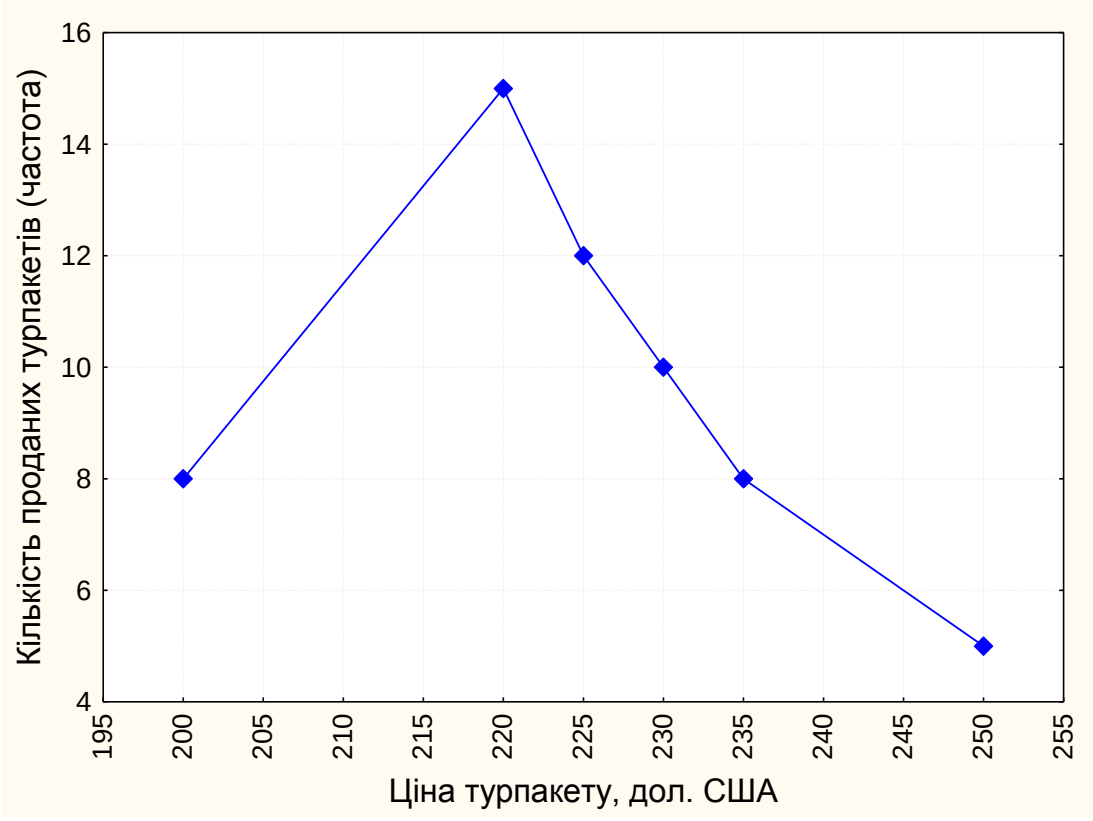


Рис. 2.1. Полігон розподілу турпакетів за ціною їх реалізації

Гістограма (стовпчасті і смугові діаграми) має вигляд ступінчастої фігури з прямокутників. При побудові гістограми на осі абсцис відкладаються величини інтервалів Δx, а частоти f_i зображуються прямокутниками на осі ординат, побудованими на відповідних інтервалах. Висота стовпчиків повинна бути пропорційною частотам. При перевірці сума всіх частот дорівнює числу всіх визначень, сума відносних частот дорівнює одиниці (або

100 %). Якщо величини інтервалів $\Delta x \to 0$, то при $n \to \infty$ гістограма перетворюється в криву розподілу.

На рис 2.2 наведена гістограма розподілу жителів Європи різних вікових груп (у відсотках до підсумку), які здійснили туристські поїздки в 2018 р. Графік побудований за опрацьованими нами даними з сайту Statistics/Eurostat (https://ec.europa.eu/eurostat/web/tourism/).

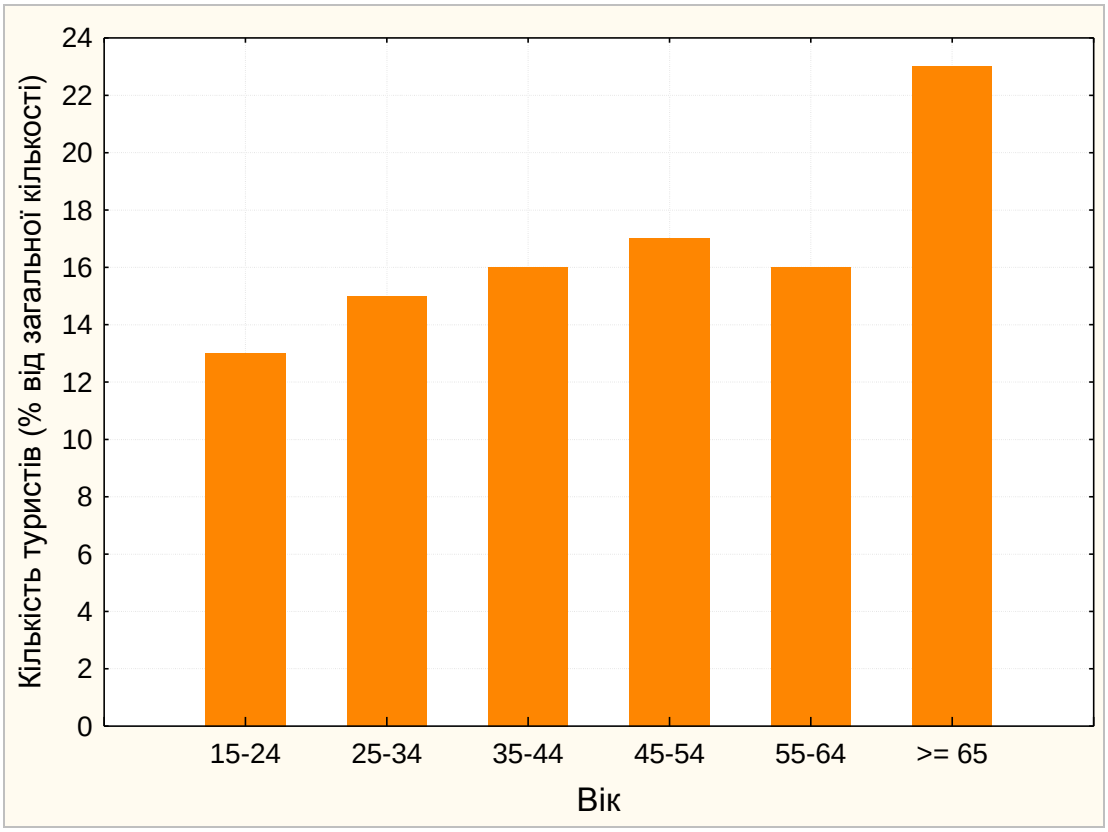


Рис. 2.2. Розподіл туристів (громадян країн-членів ЄС) за різними віковими групами (у відсотках до підсумку) за 2018 р.

Кумулята є ламаною, яка з'єднує точки з координатами (x_i, f_{x_i}) (де f_{x_i} – накопичені частоти) для дискретного ряду, або точки з координатами a_i, f_{a_i} – для інтервального ряду.

Кумулятивна крива будується для накопичених частот або накопичених відносних частот. Накопичені частоти знаходять послідовним підсумовуванням частот або кумуляцією (від лат. cumulatio – збільшення, скупчення) в напрямку від першого елемента

до кінця варіаційного ряду. Накопичені частоти показують, скільки одиниць сукупності мають значення ознаки, які є не більшими за значення, що розглядається.

Для побудови кумуляти на осі абсцис наносять значення варіант (або точки дискретного ряду, або межі інтервалів), а на осі ординат – відповідні кумулятивні частоти.

Побудову кумуляти розглянемо на прикладі оброблених нами даних про ділові поїздки за рік, які здійснили громадяни Великої Британії (за 2019 рік, за даними UK National Travel Survey, табл. 2.9; https://www.gov.uk/government/collections/national-travel-survey-statistics).

Таблиця 2.9

Кумулятивний інтервальний варіаційний ряд

Вік туриста	17-20	21-29	30-39	40-49	50-59	60-69	70+
Частота ділових поїздок за 2019 р., f_i	9	32	46	52	59	22	3
Накопичена частота, $\sum f_i$ (або кумулята частот)	9	41	87	139	198	220	223

Кумулята (або крива накопичених частот) наведена на рис. 2.3, з якого видно (ліва вісь ординати), що більшу частину ділових поїздок в 2019 р. здійснили громадяни у віці 17-59 років. Це 90 % від усієї кількості ділових поїздок (див. вісь для відносних частот).

Маючи в розпорядженні криву накопичених частот, легко підрахувати кількість спостережень, що лежать в будь-якому інтервалі досліджуваного діапазону. Вона знаходиться як різниця ординат кривої у відповідних точках. Наприклад, кількість спостережень інтервалу (21-29) – (60-69) дорівнює 220 - 41 = 179, рис. 2.3.

Крім цього, за допомогою кумулятивної кривої можна легко визначити відносну кількість спостережень, що не перевищують заданої кількості або переважають цю кількість. Для цього кумуляту необхідно представити в шкалі відносних частот (див. праву вісь рис.

2.3). Проводячи горизонтальну лінію, відповідну деякій відносній частоті, наприклад, 62 %, можна бачити, що це число забезпечується не більше ніж 139 спостереженнями. Особливо важливою є точка на кривій, відносно якої 50 % частот спостережень лежать зліва і 50 % – справа. Величина абсциси цієї точки називається *медіаною*.

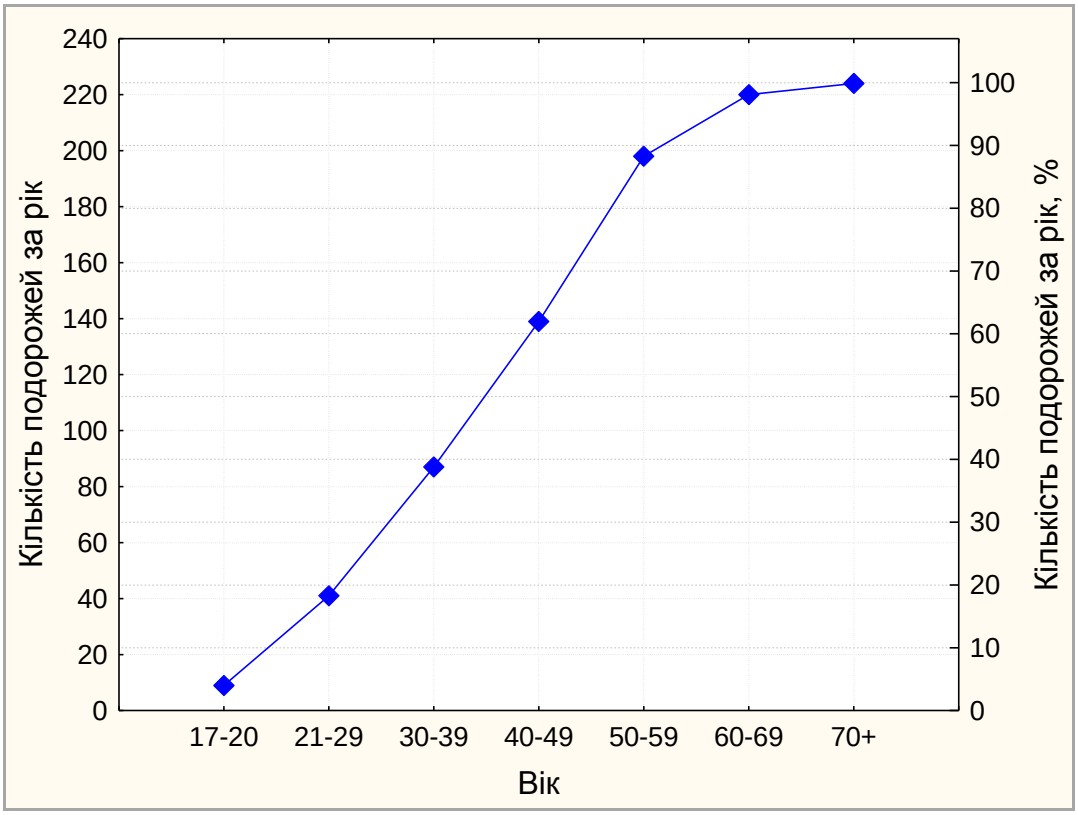


Рис. 2.3. Кумулята ділових поїздок громадян Великої Британії за один рік (2019 р.)

Інший приклад кумуляти відносних частот побудований нами і наведений на рис. 2.4. Рисунок відображає кількість туристських поїздок (на осі ординат – кумулятивна відносна частота поїздок, %) і тривалість цих поїздок в тижнях за наборами даних Foursquare (https://doi.org/10.1007/s40558-020-00170-6). Графік надає важливу інформацію з погляду туризму: наприклад, в цілому 90 % всіх поїздок коротше 30 днів, а 60 % – коротше 2-х тижнів.

Таким чином, гістограма, полігон частот і кумулята дозволяють реально побачити в загальних рисах закономірності розподілу досліджуваної генеральної сукупності.

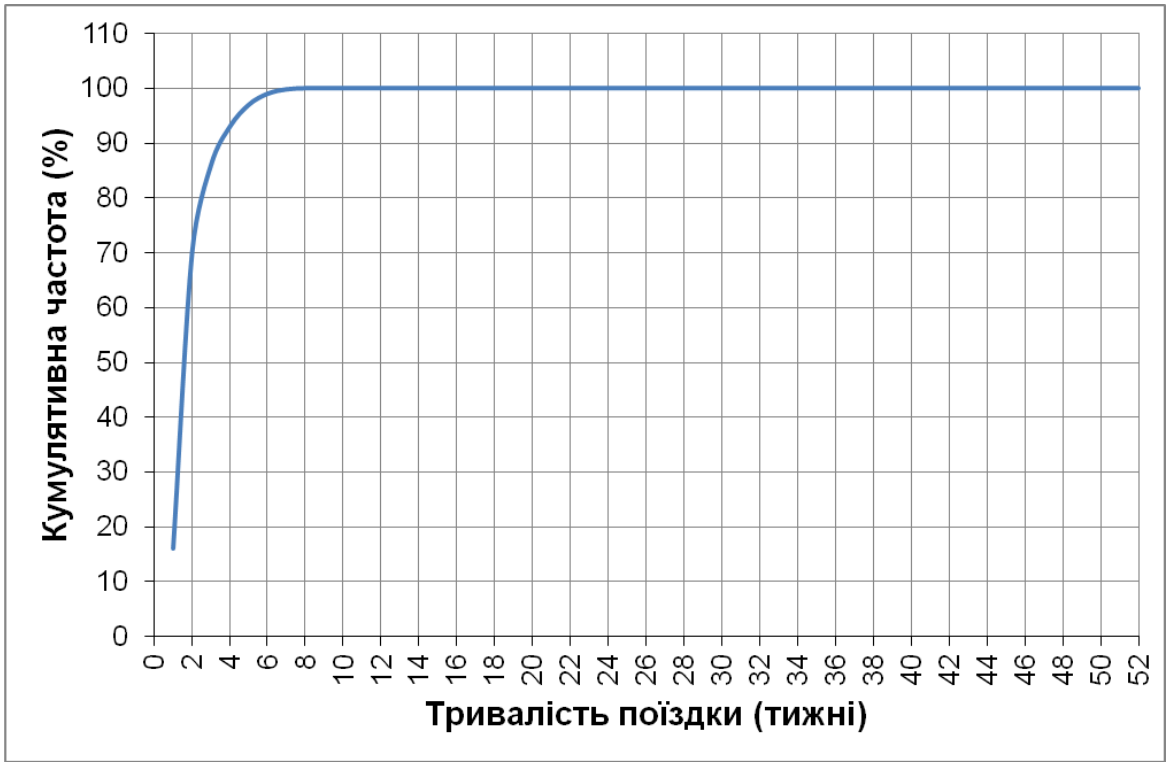


Рис. 2.4. Кумулята розподілу тривалості туристських поїздок (%) за наборами даних Foursquare

На первинному етапі аналізу графічні форми подання статистичних даних є важливим робочим методом статистики. У процесі узагальнення і аналізу статистичної інформації важливим є вибір форми графічних зображень. Правильний вибір графіка забезпечує логічне продовження статистичних таблиць в процесі узагальнення і аналізу статистичної інформації.

Нижче наведено приклади застосування інших графічних методів відображення статистичної інформації в сфері туризму. Наприклад, *лінійні діаграми* широко використовуються для характеристики змін явищ у часі. На рис. 2.5 графік відбиває динаміку кількості українських туристів, які виїжджали за кордон з 2006 по 2017 рр. (організований туризм). Дані запозичені з сайту Державної служби статистики України і оброблені авторами навчального посібника. З рисунку видно, наприклад, що, в цілому, за аналізований період кількість туристів зменшується. Крім цього, максимальна кількість туристів, які виїжджали за кордон за

аналізований період, спостерігається в 2007 р, а мінімальна – в 2017 р.

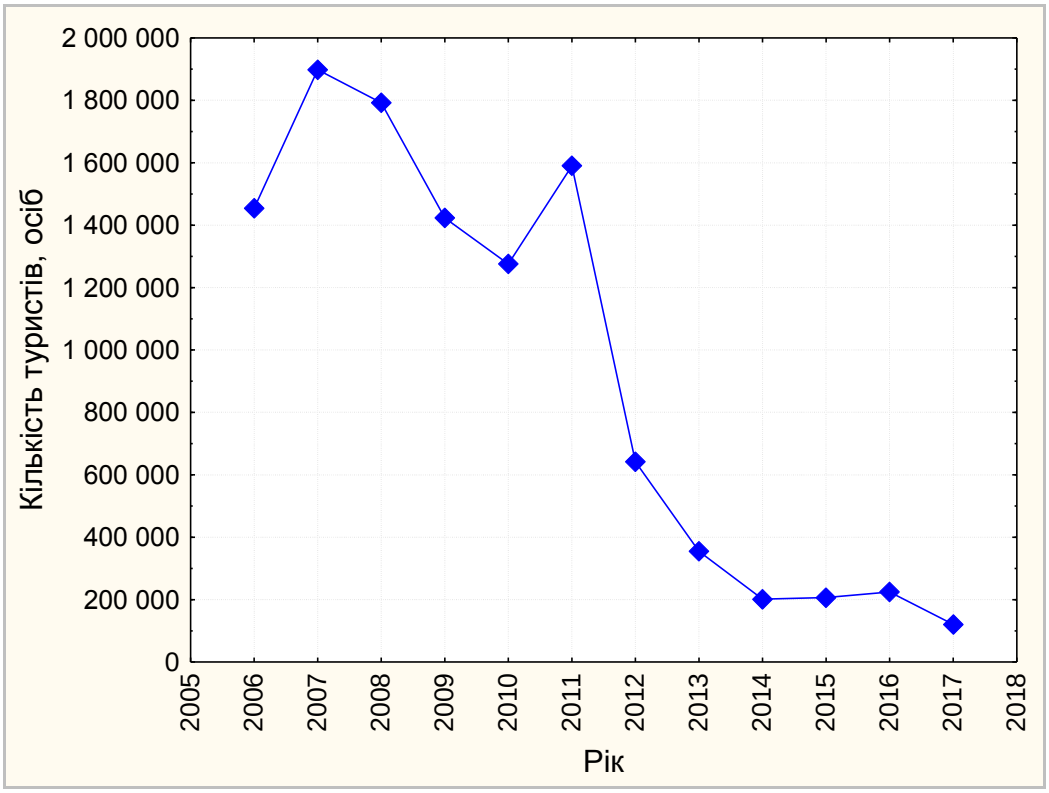


Рис. 2.5. Динаміка кількості українських туристів, які виїжджали за кордон в період 2006-2017 рр. (організований туризм)

Інший приклад відображення інформації в сфері туризму – у вигляді *об'ємної гістограми* – наведено на рис. 2.6. Тривимірний графік відбиває мінливість кількості туристів в часі і в просторі. Побудований нами графік надає інформацію про кількість українських туристів, яких обслужили туроператори в деяких областях України в період з 2006 по 2017 рр. З графіка видно, що з усіх розглянутих областей туроператори найбільше обслуговували туристів в Дніпропетровській, Запорізькій, Львівській, Одеській та Харківській областях. При цьому, в цих самих областях аналізовані показники вище у 2017 р. в порівнянні з 2015 і 2016 рр. Крім цього, видно, що за аналізований період кількість туристів зростає.

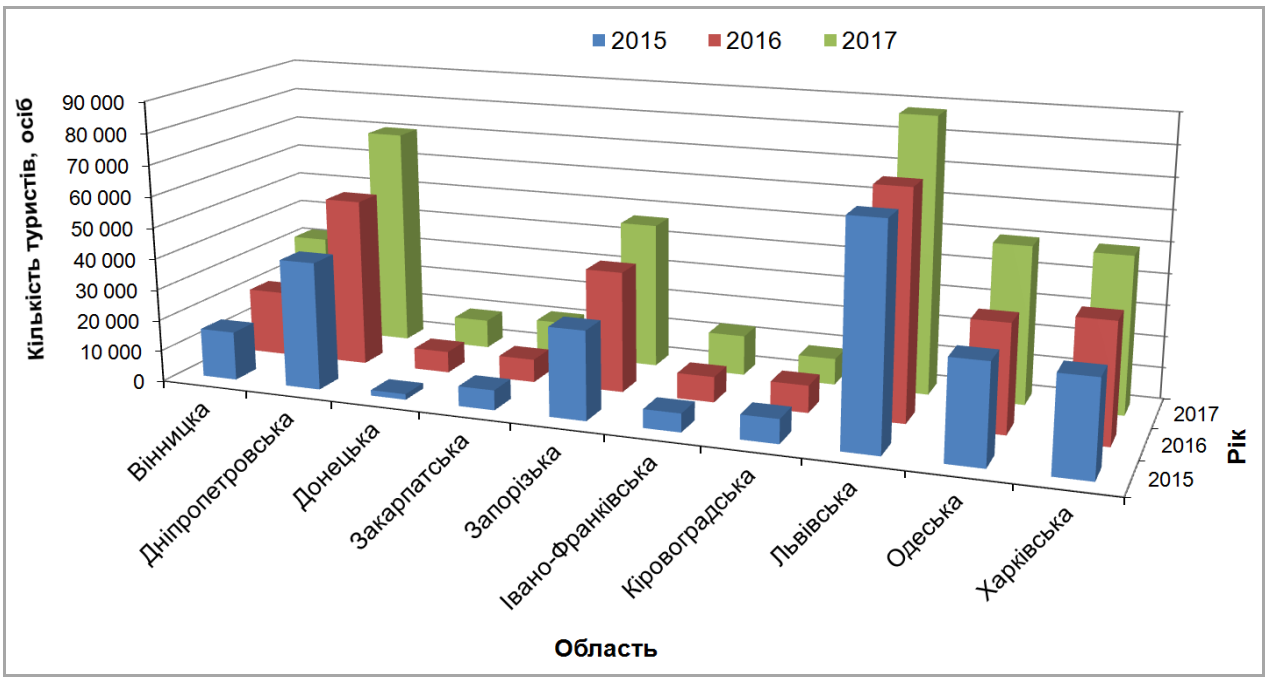


Рис. 2.6. Кількість туристів, обслугованих туроператорами за регіонами за період 2015-2017 рр.

Кругові (секторні) *діаграми* зручно використовувати, коли необхідно показати частку кожної величини в загальному об'ємі. Тут величина кожного значення зображується у вигляді сектора (круга), площа якого відповідає внеску цього значення в суму значень.

Приклад витрат суб'єктів туристичної діяльності на послуги сторонніх організацій, що використовуються при виробленні туристського продукту, поданий у вигляді об'ємної кругової діаграми (рис. 2.7). Графік побудований авторами за даними Державної служби статистики України). З діаграми видно, що в 2018 р. найбільші витрати суб'єктів туристичної діяльності становлять такі послуги: розміщення і проживання, транспортне обслуговування, медичне обслуговування.

В туризмі широко використовують також *картограми* (або *картодіаграми*) – зображення розташування і інтенсивності досліджуваного явища на основі контурної географічної карти за допомогою графічних символів, таких як забарвлення, штрихування, точки та ін. Прикладами графіків такого роду можуть бути

картограми-показники окремих видів туристських послуг в регіонах України.

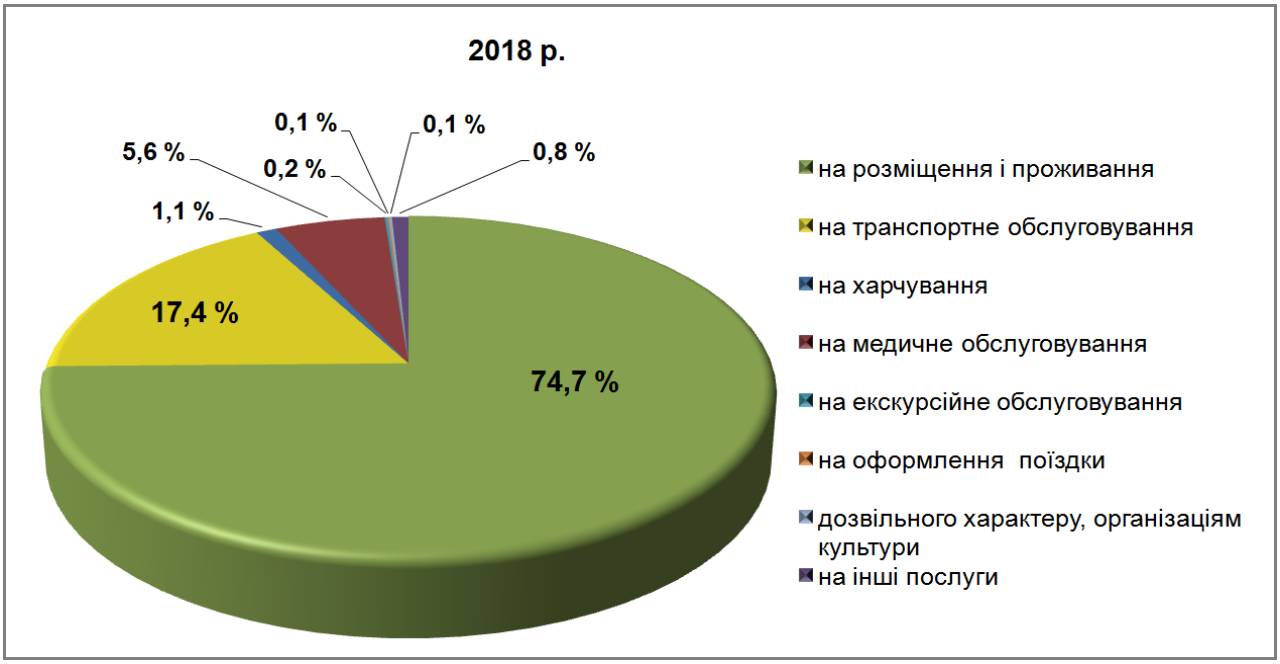


Рис. 2.7. Витрати суб'єктів туристичної діяльності на послуги сторонніх організацій, що використовуються при виробленні туристського продукту

Наприклад, на рис. 2.8 наведені показники туристичної діяльності по регіонах України станом на 2019 р. Карта складена авторами в результаті обробки даних Державної служби статистики України. Карта відображає вартість реалізованих туристичних пакетів, а також кількість туристів, обслугованих туроператорами та турагентами у 2019 році за регіонами. З карти видно, що в 2019 р. кількість обслугованих туристів та доходи від реалізованих турпакетів найбільші у м. Києві і Київській області, Львівській, Одеській та Дніпропетровській областях. Ці показники туристичних послуг значні також в Харківській, Запорізькій і Івано-Франківській і Донецькій областях.

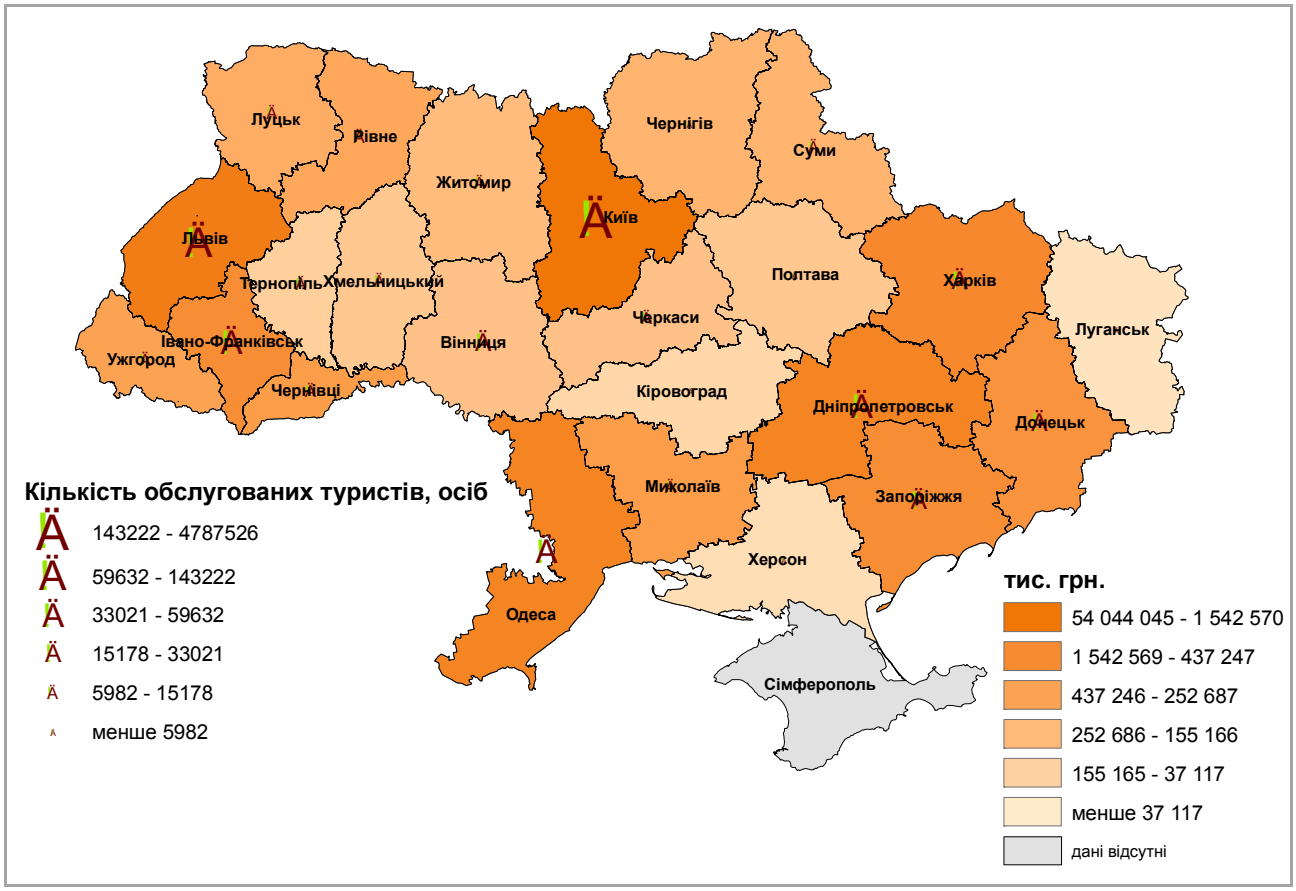


Рис. 2.8. Кількісна характеристика діяльності туроператорів та турагентств України за регіонами у 2019 році

Туристський попит піддається істотним сезонним коливанням, тому становить інтерес аналіз, наприклад, динаміки здійснення поїздок туристів по місяцях. Вплив сезонності на ринок туристичних послуг можна вивчити за допомогою аналізу часових рядів, в тому числі їх графічного зображення – *діаграм часових рядів*. Коли дані спостережень являють собою послідовність, що залежить від часу, їх називають *часовим рядом*.

Як приклад ми навели на рис. 2.9 дані про щомісячну кількість в'їздів туристів в Австрію за період 2010-2019 рр. Аналізовані дані, запозичені з сайту Statistics/Eurostat (https://ec.europa.eu/eurostat/web/tourism/), оброблені і графічно представлені авторами навчального посібника. Важливо зазначити, що за кількість прибулих туристів була прийнята кількість фактичних проживань в готелях та інших установах розміщення туристів.

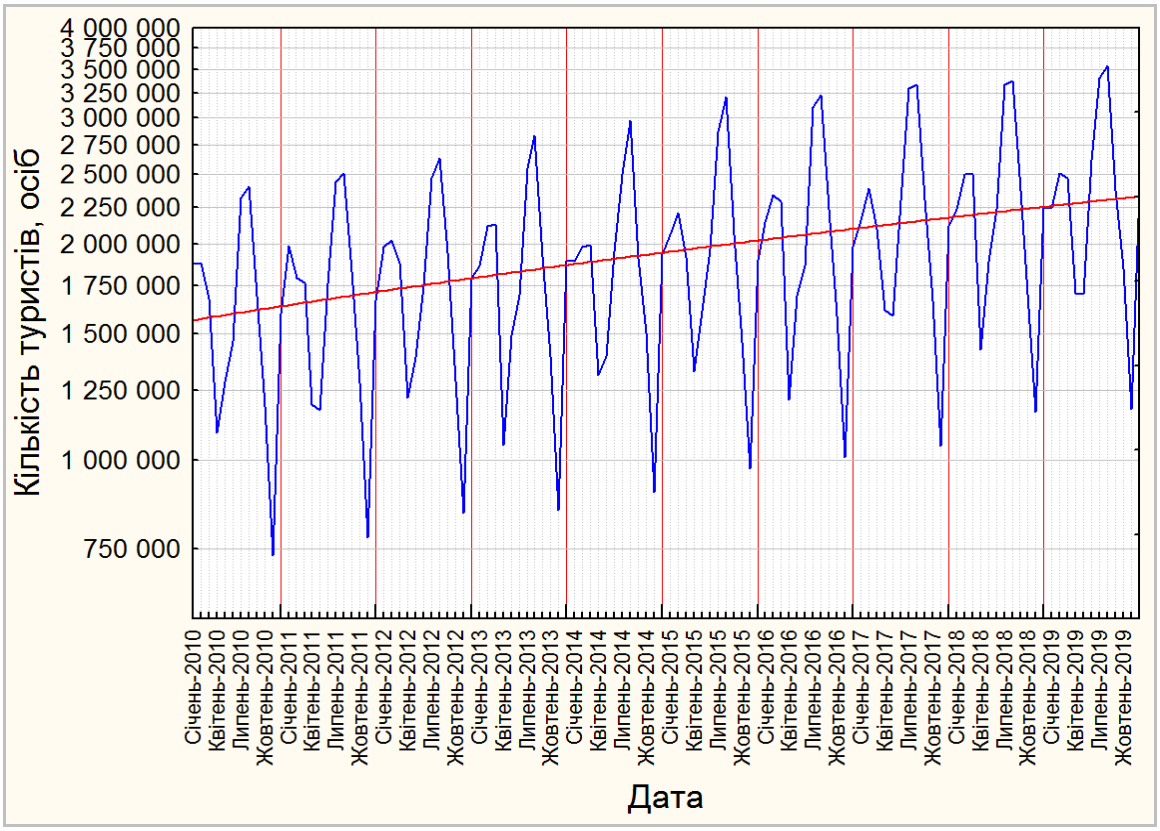


**Рис. 2.9. Сезонна динаміка туристів,
які прибули до Австрії за період 2010-2019 рр.**

На рисунку чітко видно сезонні коливання. Найбільша кількість туристів приїжджала в серпні і вересні, найменша – у листопаді. Крім цього, сезонна динаміка туристів має явно виражений висхідний тренд за період 2010-2019 рр.: кількість прибуттів туристів постійно зростала. З іншого боку, розраховані значення сезонних діапазонів досить стабільні. Це означає, що сезонність з плином часу змінилася неістотно.

Зазначимо, що існує велика кількість інших графічних методів подання статистичних рядів і, в цілому, статистичної інформації. Ми розглянули лише ті, які найчастіше використовуються в сфері туризму.

Питання для самоконтролю

2.1. Що називається статистичною сукупністю?

2.2. Що називається генеральною сукупністю? вибіркою?

2.3. Який об’єм вибірки вважається оптимальним?
2.4. З якою метою вдаються до формування вибірки?
2.5. Що означає репрезентативність вибірки?
2.6. Що являють собою статистичні ряди розподілу? Наведіть приклади зі сфери туризму.
2.7. Які ряди розподілу називають атрибутивними? варіаційними?
2.8. З яких елементів складається варіаційний ряд?
2.9. Що таке варіанта, варіація?
2.10. Що називається частотою, а що – відносною частотою?
2.11. Що являє собою об’єм сукупності?
2.12. Чому дорівнює сума частот варіаційного ряду?
2.13. Які ряди називають дискретними? інтервальними? Наведіть приклади зі сфери туризму.
2.14. Як визначається довжина рівного інтервалу?
2.15. Що дозволяє визначити формула Стерджеса?
2.16. У вигляді яких графіків зображуються ряди розподілу?
2.17. Схарактеризуйте техніку побудови гістограми, полігону і кумуляти.
2.18. Що таке ранжований статистичний ряд?
2.19. В яких випадках використовується групування з нерівними інтервалами?
2.20. Як називається ряд, в якому величина кількісної ознаки приймає тільки цілі значення?
2.21. Який ряд зображує кумуляту?
2.22. Як називаються графічні зображення дискретних та інтервальних варіаційних рядів?

Завдання для самостійного виконання

Завдання 2.1. Заповніть таблицю 1, яка характеризує витрати суб’єктів туристичної діяльності України на послуги сторонніх організацій *за ознакою транспортне обслуговування* (залізничний транспорт, повітряний транспорт та ін.) за період 2015-2019 рр. Для заповнення таблиці використовуйте дані Державної служби статистики України.

Таблиця 1

Витрати суб'єктів туристичної діяльності на транспортне обслуговування, тис. грн.

Тип транспорту	Рік				
	2015	2016	2017	2018	2019

Завдання 2.2. Складіть варіаційний ряд (табл. 2) за даними загальної характеристики колективних засобів розміщення туристів у 2015 році по України (наприклад: готелі, санаторії, пансіонати з лікуванням та ін.). Оцініть внесок (у відсотках) кожного засобу розміщення в загальну кількість колективних засобів розміщення. Джерело даних – Державна служба статистики України.

Таблиця 2

Загальна характеристика колективних засобів розміщення туристів у 2015 р.

Засоби розміщення	Варіанти, x_i (%)	Частота, f_i

Завдання 2.3. Спираючись на статистичні дані, визначте частку різних регіонів України в міжнародних туристських прибуттях в 2009, 2014 і 2019 рр., а також відсоток її зміни в 2019 році порівняно з 2009 і 2014 рр. Джерело даних – Державна служба статистики України.

Завдання 2.4. Складіть інтервальний варіаційний ряд (табл. 3), що подає вікові групи і кількість туристів, які в'їхали в країну *K* за певний рік (країну і рік вибрати самостійно).

Таблиця 3

Вік і кількість туристів (%)

Вік туриста, років	Кількість туристів (% від загальної кількості туристів)

Завдання 2.5. Побудуйте полігон частот на основі виконаного завдання 2.2 (табл. 2).

Завдання 2.6. Побудуйте гістограму розподілу туристів за різними віковими групами на основі виконаного завдання 2.4 (табл. 3).

Завдання 2.7. Складіть кумулятивний інтервальний варіаційний ряд на основі виконаного завдання 2.4 (табл. 3) і побудуйте кумулятивну криву розподілу вікових груп туристів. Встановіть вік туристів, які складають 80 % від загальної кількості туристів.

Завдання 2.8. За даними Всесвітньої туристської організації (ЮНВТО) побудуйте криву розподілу доходів від міжнародного туризму за період 2008-2020 рр. (табл. 4). Встановіть величини падіння доходів в певних роках в порівнянні з попередніми роками, а також здійсніть оцінку цих величин у відсотках відносно доходів попередніх років.

Таблиця 4

Рік	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Долари США, млрд.													

Завдання 2.9. Побудуйте об’ємну гістограму, яка характеризує мінливість (в часі і в просторі) кількості іноземних туристів, яких обслужили туроператори України з 2015 по 2019 рр. по всіх регіонах. Для виконання завдання використовуйте дані Державної служби статистики України.

Завдання 2.10. Побудуйте кругову діаграму, яка характеризує витрати суб’єктів туристичної діяльності на послуги сторонніх організацій, що використовуються при виробленні туристського продукту (наприклад, витрати на: розміщення і проживання, транспортне обслуговування, екскурсійне обслуговування та ін.). Для виконання завдання використовуйте дані про суб’єктів туристичної діяльності України за 2019 р. з сайту Державної служби статистики України.

Завдання 2.11. Побудуйте карту, яка покаже кількість суб’єктів туристичної діяльності України і загальні доходи від надання туристичних послуг за 2018 р. Джерело даних – Державна служба статистики України.

Завдання 2.12. Спираючись на дані Statistics/Eurostat (https://ec.europa.eu/eurostat/web/tourism/) побудуйте графік сезонної динаміки туристів, які прибули в певну країну Європи (на вибір) за період 2010-2020 рр.

РОЗДІЛ 3
ОСНОВНІ СТАТИСТИЧНІ ХАРАКТЕРИСТИКИ ВАРІАЦІЙНИХ РЯДІВ

Варіаційні ряди та їхнє графічне зображення дають наочне уявлення про варіювання ознак, проте, вони недостатні для повного опису вибірки. У зв'язку з цим використовують логічно і теоретично обґрунтовані числові показники, які називають *статистичними характеристиками*. До них зазвичай зараховують так звані *середні величини* (характеристики середнього положення або центральної тенденції ряду); *показники варіації* (характеристики розсіювання, розкиду даних щодо середніх величин); *показники форми розподілу* (характеристики асиметрії і ексцесу).

3.1. Середні величини

Середні величини є одними з найпоширеніших узагальнюючих статистичних показників. *Середня величина* – це узагальнена кількісна характеристика статистичної сукупності за будь-якою варіюючою ознакою, представлена в одній величині.

Показник у вигляді середньої величини виражає типові риси і дає узагальнюючу характеристику однотипних явищ за однією з варіюючих ознак; відображає рівень цієї ознаки відносно всіх одиниць досліджуваної сукупності. Він дозволяє порівнювати рівні однієї і тієї самої ознаки в різних сукупностях і знаходити причини їх розбіжностей.

Сутність середньої величини полягає в тому, що в ній взаємно погашаються відхилення значень ознаки окремих одиниць сукупності, зумовлені дією випадкових факторів, і враховуються зміни, викликані дією основних чинників. Це дозволяє середній величині відбивати типовий рівень ознаки і абстрагуватися від індивідуальних особливостей, властивих окремим одиницям.

На відміну від індивідуальних числових характеристик середні величини мають велику сталість, здатні характеризувати цілу групу

однорідних одиниць одним (середнім) числом, тобто відбивати те спільне, що притаманне всім одиницям досліджуваної сукупності. Вони є *характеристиками положення* – біля них групуються спостережувані (виміряні) значення досліджуваної величини.

Середні величини поділяються на два класи: *степеневі* (середня арифметична, середня гармонійна, середня геометрична, середня квадратична і т. д.) і *структурні* (мода, медіана, квартилі, децилі та ін.).

Нижче розглядаються ті середні характеристики, які найчастіше використовуються в дослідженнях і можуть бути застосовані у сфері туризму.

3.1.1. Степеневі середні величини

Степеневі середні обчислюють за допомогою загальної формули:

$$\overline{x} = \sqrt[k]{\frac{\sum x_i^k}{n}} = \left(\frac{\sum x_i^k}{n} \right)^{1/k}, \qquad (3.1)$$

де $\overline{x}$ – степенева середня k-ого порядку; x_i – варіанта; n – кількість спостережень (або кількість варіант); k – величина, за якою визначають вигляд (форму) середньої: при $k = 1$, згідно з рівнянням (3.1), виходить середня арифметична, при $k = 2$ – середня квадратична, при $k = 3$ – середня кубічна, при $k = -1$ утворюється середня гармонійна і т. д.

Степеневі середні залежно від подання вихідних даних обчислюються в двох формах: *простій* та *зваженій*.

Якщо в ряду розподілу кожна одиниця зустрічається один раз або однакову кількість разів, тобто немає необхідності групувати дані, тоді як середню обчислюють *середню арифметичну просту* (*незважену*).

Середню арифметичну просту визначають як суму всіх варіант

ряду, поділену на їх загальну кількість, тобто просту середню арифметичну ($\bar{x}$) значень $x_1, x_2, x_3,..., x_n$ ряду розподілу визначають як:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad (3.2)$$

де n – загальна кількість варіант.

Якщо в ряду розподілу окремі значення повторюються неоднакову кількість разів, то як середня арифметична визначається *середня арифметична зважена*, тобто розрахунок середньої здійснюється за згрупованими даними, які можуть бути дискретними або інтервальними. Інакше кажучи, середню величину називають зваженою, оскільки для згрупованих даних кожну варіанту враховують (тобто «зважують») за її частотою. Наприклад, якщо x_1 зустрічається в ряду розподілу f_1 разів, x_2 – f_2 разів, …, x_k – f_k разів (причому $f_1 + f_2 + ... + f_k = n$), тобто, якщо значення $x_1, x_2, x_3,..., x_k$ мають частоти $f_1, f_2, f_3,..., f_k$, при цьому $f_1 \neq f_2 \neq f_3 \neq ... \neq f_k$, тоді сума значень спостережень для першого інтервалу визначається як $f_1 \cdot x_1$, другого інтервалу – $f_2 \cdot x_2$ і т. д. Для цих згрупованих даних середню арифметичну визначають за формулою (3.3) і називають ***середньою арифметичною зваженою***:

$$\bar{x}_{зв.} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + ... + f_kx_k}{f_1 + f_2 + f_3 + ... + f_k} = \frac{\sum_{i=1}^{k}(f_i \cdot x_i)}{\sum_{i=1}^{k} f_i}, \qquad (3.3)$$

де k – кількість груп.

***Приклад* 3.1.** За даними про об’єм продажів туристських пакетів (табл. 3.1) необхідно визначити середню ціну їх реалізації.

Таблиця 3.1

Розподіл туристських пакетів різної ціни (євро)

Порядковий номер, i	1	2	3	4	5
Ціна турпакету, x_i	220	225	230	235	250
Кількість проданих турпакетів, од., f_i	10	10	10	10	10

Оскільки частоти мають однакові значення (10 од.), правомірно застосувати формулу для середньої арифметичної простої:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{n} = \frac{220 + 225 + 230 + 235 + 250}{5} = 232 \text{ євро.}$$

***Приклад* 3.2.** Розглянемо ті ж самі дані, які наведені в табл. 3.1, проте з іншою кількістю проданих турпакетів, тобто з різними частотами варіант x_1, x_2, x_3, x_4, x_5 (табл. 3.2).

Таблиця 3.2

Розподіл туристських пакетів різної ціни в різній кількості

Порядковий номер	1	2	3	4	5
Ціна турпакету, євро, x_i	220	225	230	235	250
Кількість проданих турпакетів, од., f_i	15	12	10	8	5

Оскільки в цьому разі значення ознаки, що усереднюється, тобто ціни турпакетів, зустрічаються неоднакову кількість разів (частота їх різна), це означає, що будь-яка варіанта цієї ознаки неоднаково впливає на середню величину на відміну від випадку в прикладі 3.1. Для зрівноваження зазначених впливів використовують середню арифметичну зважену величину (рівняння 3.3):

$$\bar{x}_{зв.} = \frac{\sum(f_i \cdot x_i)}{\sum f_i} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + f_4x_4 + f_5x_5}{f_1 + f_2 + f_3 + f_4 + f_5} =$$

$$= \frac{15 \cdot 220 + 12 \cdot 225 + 10 \cdot 230 + 8 \cdot 235 + 5 \cdot 250}{15 + 12 + 10 + 8 + 5} = 228{,}60 \text{ євро.}$$

З прикладів 3.1 і 3.2 видно, що різниця між середньоарифметичним і середньозваженим значеннями є доволі помітною: 232 євро та 228,60 євро, відповідно.

Таким чином, якщо кожна варіанта в ряду розподілу зустрічається один раз або однакову кількість разів, то як середня визначається проста середня арифметична. Якщо окреме значення ознаки повторюється неоднакову кількість разів, визначається середня арифметична зважена. Розрахунок середніх показників за формулою простої середньої арифметичної замість середньої арифметичної зваженої може призвести до серйозних помилок.

Очевидно, що якщо частоти (ваги) значень однакові, то середня арифметична зважена дорівнює простій середній арифметичній. У цьому можна переконатися шляхом заміни різних частот $f_1, f_2, f_3, ..., f_k$ на одну й ту саму частоту f. Нехай значення $x_1, x_2, x_3, ..., x_k$ мають однакові частоти $f_1 = f_2 = f_3 ... = f_k = f$, тобто $f_1 = f_2 = f_3 ... = f_k = f$, тоді з рівняння (3.3) маємо:

$$\bar{x}_{зв.} = \frac{f_1 x_1 + f_2 x_2 + f_3 x_3 + ... + f_k x_k}{f_1 + f_2 + f_3 + ... + f_k} = \frac{f \cdot \sum x_i}{f \cdot k} = \frac{\sum x_i}{k}. \qquad (3.4)$$

Середня квадратична. Середню квадратичну використовують, наприклад, при обчисленні середньої величини сторін декількох квадратів, середніх діаметрів і т. п. Зокрема, середня квадратична найбільш широко застосовується при розрахунку показників варіації, при вивченні взаємозв'язку явищ.

Середня квадратична проста (*незважена*) n додатних чи від'ємних величин $x_1, x_2, x_3, ..., x_n$ визначається за формулою:

$$\bar{x}_{кв.} = \sqrt{\frac{\sum x_i^2}{n}} = \sqrt{\frac{x_1^2 + x_2^2 + ... + x_n^2}{n}}. \qquad (3.5)$$

Для *згрупованих* даних обчислюють *середню квадратичну зважену*:

$$\bar{x}_{кв.зв.} = \sqrt{\frac{\sum_{i=1}^{k} f_i x_i^2}{\sum_{i=1}^{k} f_i}} = \sqrt{\frac{f_1 x_1^2 + f_2 x_2^2 + ... + f_k x_k^2}{f_1 + f_2 + ... + f_k}}\ . \qquad (3.6)$$

Середня арифметична має низку *важливих властивостей*, які більш повно розкривають її сутність і в ряді випадків дозволяють спрощувати обчислення. Розглянемо деякі з них.

1. Якщо кожну варіанту x_i ряду зменшити або збільшити на яке-небудь постійне число A, то середня арифметична нового ряду, відповідно, зменшується або збільшується на ту ж саму величину:

$$\frac{\sum (x_i \pm A) f_i}{\sum f_i} = \frac{\sum x_i f_i}{\sum f_i} \pm \frac{\sum A f_i}{\sum f_i} = \bar{x} \pm A.$$

2. Якщо кожну варіанту x_i поділити або помножити на якесь одне й те саме число A, то середня також, відповідно, зменшиться або збільшиться в A разів:

$$\frac{\sum \frac{x_i}{A} f_i}{\sum f_i} = \frac{\frac{1}{A}\sum x_i f_i}{\sum f_i} = \frac{\bar{x}}{A}. \qquad \frac{\sum (x_i \cdot A) f_i}{\sum f_i} = \frac{A \sum x_i f_i}{\sum f_i} = A \cdot \bar{x}.$$

3. Якщо всі ваги f_i зменшити або збільшити в A разів, то середня арифметична від цього не зміниться:

$$\frac{\sum x_i \frac{f_i}{A}}{\sum \frac{f_i}{A}} = \frac{\frac{1}{A}\sum x_i f_i}{\frac{1}{A}\sum f_i} = \bar{x}\ .$$

Виходячи з даної властивості, можна зробити висновок, що якщо всі ваги рівні між собою, то розрахунки за середньою арифметичною простою і середньою арифметичною зваженою приведуть до одного й того самого результату.

4. Сума добутків відхилень варіант від їх середньої арифметичної на відповідні їм частоти дорівнює нулю:

$$\sum \left[f_i \left(x_i - \bar{x} \right) \right] = \sum f_i x_i - \bar{x} \sum f_i = \bar{x} n - \bar{x} n = 0.$$

5. Сума квадратів відхилень варіант від їхньої середньої менше суми квадратів відхилень тих же самих варіант від будь-якої іншої величини A, що не дорівнює $\bar{x}$ $(A \neq \bar{x})$:

$$\sum (x_i - \bar{x})^2 < \sum \left(x_i - A\right)^2 .$$

$$\sum (x_i - A)^2 f_i = \sum (x_i - \bar{x} + \bar{x} - A)^2 f_i = \sum [(x_i - \bar{x}) + (\bar{x} - A)]^2 f_i =$$

$$= \sum [(x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - A) + (\bar{x} - A)^2] f_i = \sum (x_i - \bar{x})^2 f_i +$$

$$+ 2(\bar{x} - A)\sum (x_i - \bar{x}) f_i + (\bar{x} - A)^2 f_i = \sum (x_i - \bar{x})^2 f_i + 2(\bar{x} - A) \cdot 0 + \sum (\bar{x} - A)^2 f_i .$$

Отже, сума квадратів відхилень індивідуальних значень ознаки від довільної величини A більше суми квадратів їх відхилень від своєї середньої на величину $\sum (\bar{x} - A)^2 f_i$.

3.1.2. Структурні середні величини

Середня арифметична є важливою характеристикою ряду, проте вона не позбавлена недоліків, оскільки є дуже чутливою до збільшення або зменшення кількості спостережень за рахунок варіант, які за своєю величиною різко відрізняються від переважної більшості варіант. У зв'язку з цим на величину середньої арифметичної можуть значно впливати крайні елементи ранжованого варіаційного ряду, які є найменш характерними для нього, тобто є сильно відхиленими за своєю величиною від більшості варіант. Можна припустити, що ці елементи, що знаходяться на початку і в кінці зростального (або спадного) ряду, мають меншу точність. Однак немає підстав для того, щоб відкинути будь-які з них, оскільки немає явних ознак їх помилковості. З іншого боку, наявність таких значень, які відхиляються, в ряду розподілу, створює певні проблеми при аналізі. У зв'язку з цим в таких випадках як середню характеристику доцільно використовувати *структурні середні*, які представлені, в основному, *медіаною* та *модою*. Їх називають структурними (або *позиційними*) середніми, оскільки їх величини залежать від будови

ряду розподілу.

У сфері туризму із структурних середніх *медіана* та *мода* є найбільш використовуваними.

***Медіаною* (*Me*)** називається таке середнє значення, яке ділить упорядкований варіаційний ряд на дві рівні за кількістю членів (варіант) частини, причому в одній з них всі значення менше медіани, а в другій – більше.

Для визначення медіани дискретного варіаційного ряду необхідно ранжувати дані (тобто впорядкувати за зростанням або за спаданням).

При непарній кількості варіант ряду, тобто при $n = 2m+1$ (де n – кількість варіант, m – порядковий номер варіанти), медіаною буде значення варіанти, розташованої посередині ряду, тобто $Me = x_{m+1}$. Наприклад, якщо величина x_i приймає значення $x_1, x_2, x_3, ..., x_{15}$, то $n = 15$, тоді $m = (n-1)/2 = (15-1)/2 = 7$, а медіаною буде варіанта з порядковим номером 8, тобто $Me = x_{m+1} = x_{7+1} = x_8$. Наведемо приклад: в ряду: 2,30, 2,80, 2,81, 2,82, 2,83 медіаною буде число: $Me = x_3 = 2{,}81$.

При парній кількості членів ряду, тобто при $n = 2m$, за медіану приймається середнє арифметичне двох значень x_m та x_{m+1}, які знаходяться в середині ряду: $Me = \frac{x_m + x_{m+1}}{2}$. Наприклад, якщо x_i приймає значення $x_1, x_2, x_3, ..., x_{18}$, то $n = 18$, тоді $m = n/2 = 18/2 = 9$, а медіаною буде величина середньої арифметичної елементів ряду з порядковими номерами 9 та 10: $Me = \frac{x_m + x_{m+1}}{2} = \frac{x_9 + x_{10}}{2}$. Наведемо приклад: в такому ряду: 2,30, 2,80, 2,82, 2,84, 2,85, 2,86 медіаною буде число: $Me = \frac{x_3 + x_4}{2} = \frac{2{,}82 + 2{,}84}{2} = 2{,}83$.

Таким чином, якщо в даному ранжованому ряду величини x_i, досить віддалені від медіани, піддаються змінам, медіана не

зміниться, тоді як середня арифметична зміниться. При цих умовах значення середньої арифметичної виявляється менш надійним, ніж значення медіани, тому в таких випадках як середню краще використовувати медіану, а не середню арифметичну. Пояснимо це на конкретному прикладі.

***Приклад* 3.3.** Необхідно визначити середньооблікову кількість штатних працівників певного готелю за рік, якщо відома їх кількість за певний період, табл. 3.3.

Таблиця 3.3

Середньооблікова кількість штатних працівників готелю за рік

Рік	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Кількість працівників	210	1286	1316	1312	1316	1318	1319	1318	1320	1321	1320

Розташуємо дані про кількість працівників в порядку зростання (табл. 3.4).

Таблиця 3.4

Ранжований ряд розподілу за кількістю штатних працівників готелю

Кількість працівників	210	1286	1312	1316	1316	1318	1318	1319	1320	1320	1321
Рік	2008	2009	2011	2010	2012	2013	2015	2014	2016	2018	2017

Зауважимо, що число, яке знаходиться на початку зростаючого ряду, значно відрізняється від інших. При цих умовах значення середньої арифметичної $\bar{x} = 1214$ осіб, виявляється менш надійним, ніж значення медіани, яке становить: $Me = Me_6 = 1318$ осіб.

Для *інтервальних варіаційних рядів* розподілу *медіана* визначається як:

$$Me = x_0 + h \frac{\frac{\sum f_i}{2} - S_{m-1}}{f_m}, \qquad (3.7)$$

де x_0 – нижня межа (нижнє значення) медіанного інтервалу, тобто нижня

межа інтервалу, в який потрапляє медіана;

h – довжина медіанного інтервалу;

f_m – частота медіанного інтервалу;

$\sum f_i$ – сума частот у всіх інтервалах;

S_{m-1} – сума частот, накопичених до медіанного інтервалу.

***Модою* (*Мо*)** називається варіанта, яка з найбільшою частотою зустрічається в ряду розподілу. Якщо таких варіант більше однієї, розподіл називається бімодальним, а інакше він одномодальний. Деякі розподіли можуть не мати моди, тобто, якщо жодна з варіант не повторюється, мода відсутня. Наприклад, послідовність даних 1, 3, 6, 8, 12, 16 моди немає, а послідовність 3, 6, 8, 8, 8, 8, 16 є одномодальною, і мода дорівнює 8. Послідовність 2, 6, 6, 6, 6, 7, 8, 8, 8, 8, 18 є бімодальною, і мода дорівнює 6 і 8.

Для інтервальних варіаційних рядів розподілу *мода* розраховується за формулою:

$$Mo = x_0 + h \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})}, \qquad (3.8)$$

де x_0 – нижня межа модального інтервалу;

h – величина модального інтервалу;

f_m – частота модального інтервалу;

f_{m-1} – частота інтервалу, що передує модальному інтервалу;

f_{m+1} – частота інтервалу, що є наступним після модального інтервалу.

Алгоритм розрахунку медіани і моди для інтервального ряду розглянемо на прикладі кількості ділових поїздок жінок Великої Британії за рік.

***Приклад* 3.4.** За даними щодо кількості ділових поїздок за рік (за 2019 р.), які здійснили жінки Великої Британії різного віку (табл. 3.5), необхідно визначити медіанний і модальний вік туристів.

Таблиця 3.5

Інтервальний варіаційний ряд*

Вікові групи жінок, роки, x_i	Частота (кількість ділових поїздок), f_i	Сума накопичених частот, S
до 20	6	6
20-29	35	41
30-39	38	79
40-49	51	**130**
50-59	**52**	182
60-69	17	199
70 і більше	2	201
Разом	**201**	

* *Джерело: складено авторами за даними* UK National Travel Survey

В даному прикладі медіанний інтервал знаходиться в межах вікової групи 50-59 років, оскільки на цей інтервал припадає найбільша частота (52). Далі підставляємо в формулу (3.7) необхідні числові дані і обчислюємо медіанний вік туристів:

$$Me = x_0 + h\frac{\frac{\sum f_i}{2} - S_{m-1}}{f_{\text{m}}} = 50 + 9\frac{\frac{201}{2} - 130}{52} = 45 \text{ років}.$$

Розрахуємо модальний вік туристів за формулою (3.8):

$$M_0 = x_0 + h\frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} = 50 + 9\frac{52 - 51}{(52 - 51) + (52 - 17)} = 50 \text{ років}.$$

Модальний вік туристів становить 50 років.

Розглянуті вище узагальнюючі показники центру розподілу не розкривають характер послідовної зміни частот. У зв'язку з цим в аналізі закономірностей розподілу крім моди і медіани використовуються також інші структурні характеристики варіаційного ряду, так звані *квантилі*, що відсікають в межах ряду певну частину його членів. До них зараховують *квартилі*, *децилі і перцентилі*.

Якщо за накопиченою частотою медіана ділить діапазон даних на дві половини, то квартилі ділять його на 4 рівні частини, децилі – на 10, перцентилі – на 100 частин.

***Квартилі* (*Q*)** – це значення варіант, які ділять упорядкований ряд за об'ємом на чотири рівні частини. Отже, в ряду розподілу виділяють три квартиля (Q_1. Q_2, Q_3). З них 25 % одиниць сукупності за своєю величиною будуть менше Q_1; 25 % одиниць будуть знаходитись між Q_1 та Q_2; 25 % – між Q_2 та Q_3, а решта 25 % перевищать Q_3. Медіана є водночас другим квартилем.

Розрахунок квартилів базується на кумулятивних частотах (частостях), і визначаються перший і третій квартилі за формулами:

Перший квартиль:

$$Q_1 = x_{Q_1} + i\frac{0{,}25\sum f_i - S_{Q_1-1}}{f_{Q_1}}. \qquad (3.9)$$

Третій квартиль:

$$Q_3 = x_{Q_3} + i\frac{0{,}75\sum f_i - S_{Q_3-1}}{f_{Q_3}}, \qquad (3.10)$$

де x_{Q_1} – нижня межа інтервалу, який містить нижній квартиль (інтервал визначається за накопиченою частотою, що першою перевищує 25 %);

x_{Q_3} – нижня межа інтервалу, який містить верхній квартиль (інтервал визначається за накопиченою частотою, що першою перевищує 75 %);

i – величина інтервалу;

S_{Q_1-1} – накопичена частота інтервалу, який передує інтервалу, що містить нижній квартиль;

S_{Q_3-1} – накопичена частота інтервалу, який передує інтервалу, що містить верхній квартиль;

f_{Q_1} – частота інтервалу, який містить нижній квартиль;

f_{Q_3} – частота інтервалу, який містить верхній квартиль.

***Приклад* 3.5.** В таблиці 3.6 наведені витрати суб'єктів туристичної діяльності України (за 2019 р.) на послуги сторонніх організацій, що використовуються при виробленні туристського продукту. Необхідно визначити значення першого і третього квартилів за таблицею 3.6, тобто необхідно визначити витрати (25 % і 75 %) на вироблення туристського продукту. Таблиця складена авторами за даними Державної служби статистики України.

Таблиця 3.6

Витрати суб'єктів туристичної діяльності на послуги сторонніх організацій, що використовуються при виробленні туристського продукту (2019 р.)

Витрати суб'єктів туристичної діяльності на послуги (млн. грн.)	Кількість послуг до підсумку (частота, *f*)	Накопичена частота (*S*)	Накопичений відсоток послуг до підсумку
0,015 - 0,020	1	1	12,5
0,020 - 0,077	2	3	37,5
0,077 - 2,156	3	6	75,0
2,156 - 6,776	1	7	87,5
6,776 - 24,332	1	8	100
Разом	8		312,5

Нагадаємо, що інтервал, який містить нижній квартиль Q_1, визначається за накопиченою частотою, яка першою перевищує 25 %. За даними табл. 3.6, перші 25 % послуг входять до інтервалу 0,020-0,077 млн. грн., тоді значення першого квартиля Q_1, відповідно до рівняння (3.9), становить:

$$Q_1 = x_{Q_1} + i\frac{0{,}25\sum f_i - S_{Q_1-1}}{f_{Q_1}} = 0{,}020 + 0{,}057\frac{0{,}25\cdot 8 - 1}{2} = 0{,}05 \text{ млн. грн.}$$

Значення третього квартиля Q_3 знаходиться в інтервалі – 2,156-

6,776 млн. грн. (інтервал визначається за накопиченою частотою, яка першою перевищує 75 %). Тоді, згідно з формулою (3.10), маємо:

$$Q_3 = x_{Q_3} + i\frac{0{,}75\sum f_i - S_{Q_3-1}}{f_{Q_3}} = 2{,}156 + 4{,}62 \cdot \frac{0{,}75 \cdot 8 - 6}{1} = 2{,}16 \text{ млн. грн.}$$

Отже, в ряду розподілу за даними про витрати суб'єктів туристичної діяльності на послуги сторонніх організацій перший квартиль становить 50 тис. грн., а третій – 2,16 млн. грн., тобто на 25 % послуг витрати суб'єктів туристської діяльності (це, в основному, витрати: на оформлення поїздки; культурно-освітнього, культурно-дозвільного характеру) становлять не більше 50 тис. грн., а на 75 % послуг (це витрати на медичне обслуговування; на екскурсійне обслуговування; на харчування; на інші послуги) витрати не перевищують 2,16 млн. грн. Третій квартиль можна також інтерпретувати як мінімальні витрати на останні 25 % послуг (це послуги на транспортне обслуговування та на розміщення і проживання).

Одним із способів графічного подання середньої величини, в тому числі структурних середніх, а саме квартилів, є діаграма – так званий «ящик з вусами» (box and whisker plot, боксплот), – схема якої побудована нами і наведена на рис. 3.1.

Боксплоти можуть розташовуватися як вертикально, так і горизонтально. Прямі лінії, що йдуть від ящика, називаються «вусами» і використовуються для відображення дисперсії (тобто ступеня розкиду) за межами верхнього і нижнього квартилів. Викиди відображаються у вигляді окремих точок, що знаходяться на одній лінії з вусами.

Діаграми допомагають порівнювати групи даних і отримувати таку інформацію: значення середнього показника, в тому числі медіани; симетричні дані – якщо середня величина лежить в межах довірчого інтервалу медіани, то це непрямо вказує на те, що розподіл досліджуваної величини є симетричним; чи зміщені дані і, якщо так, то в якому напрямку – по розташуванню квартилів і медіани можна зробити висновок про асиметрію в розподілі; чи існують викиди і якими є їхні

значення – «вуса» діаграми обмежують вибірку зверху і знизу (горизонтальні рисочки на кінці «вусів» – максимальне і мінімальне значення), і, якщо значення виходять за ці «вуса», то вони вважаються викидами, тобто аномальними значеннями, не характерними для ряду в цілому, але, тим не менш, вони є присутніми в ньому.

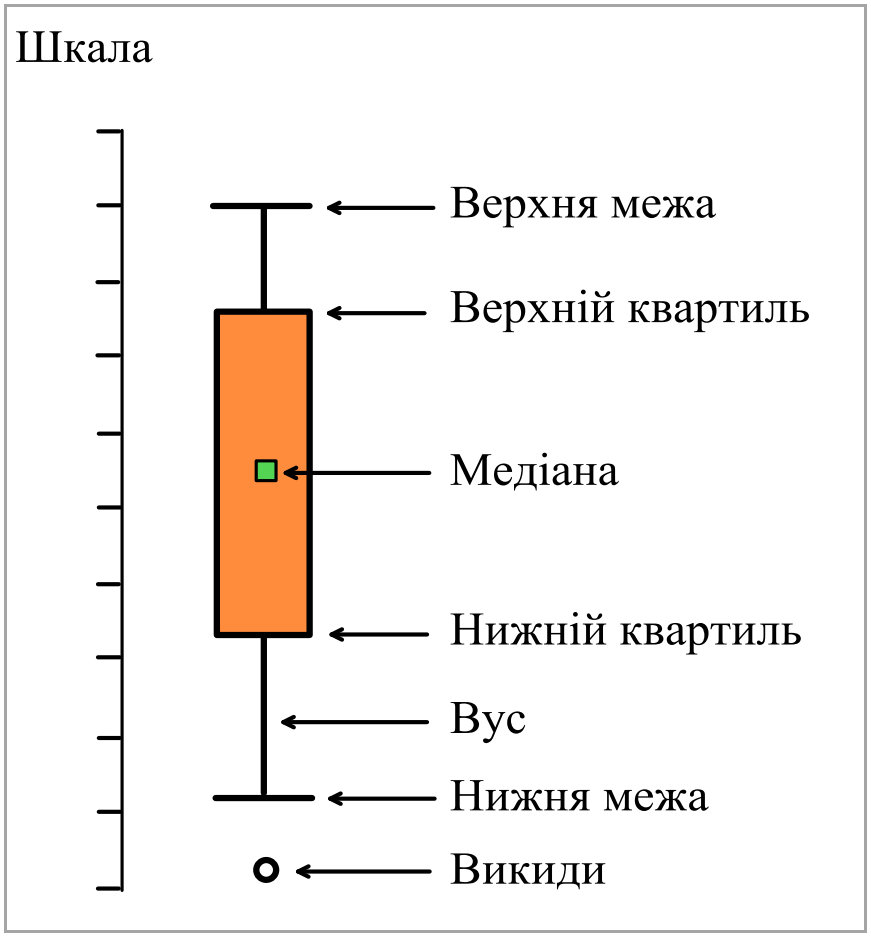


Рис. 3.1. Діаграма «ящик з вусами» (боксплот, box&whisker plot)

За боксплотами викиди – це точки, що виходять за межі 1,5 інтерквартильного розмаху. Відстань між верхнім і нижнім квартилями (значеннями, що відділяють 1/4 і 3/4 частини ряду) називається інтерквартильним розмахом. Усередині цього діапазону лежить 50 % спостережень.

За допомогою діаграми «ящик з вусами» можна також візуально порівняти розподіли і квартилі декількох вибірок. Наприклад, на рис. 3.2 наведені боксплоти, що характеризують кількість туристичних пакетів, реалізованих туроператорами та турагентами у 2018 та 2019 рр. за регіонами України. Боксплоти побудовані авторами

навчального посібника в результаті обробки та узагальнення даних сайту Державної служби статистики України.

З графіка видно, що в 2019 році порівняно з 2018 роком, в цілому, збільшилися медіана, нижній і верхній квартилі, а також мінімальна та максимальна кількість турпакетів, реалізованих туроператорами і турагентами. Тут значення першого квартиля показує максимальну середню кількість турпакетів, які реалізували туроператори і турагенти в перших 25 % регіонів і мінімальна їх кількість – в перших 75 % регіонів. Третій квартиль характеризує максимальну кількість турпакетів, реалізованих туроператорами і турагентами перших 75 % регіонів і мінімальну кількість турпакетів, реалізованих останніми 25 % регіонів даної сукупності.

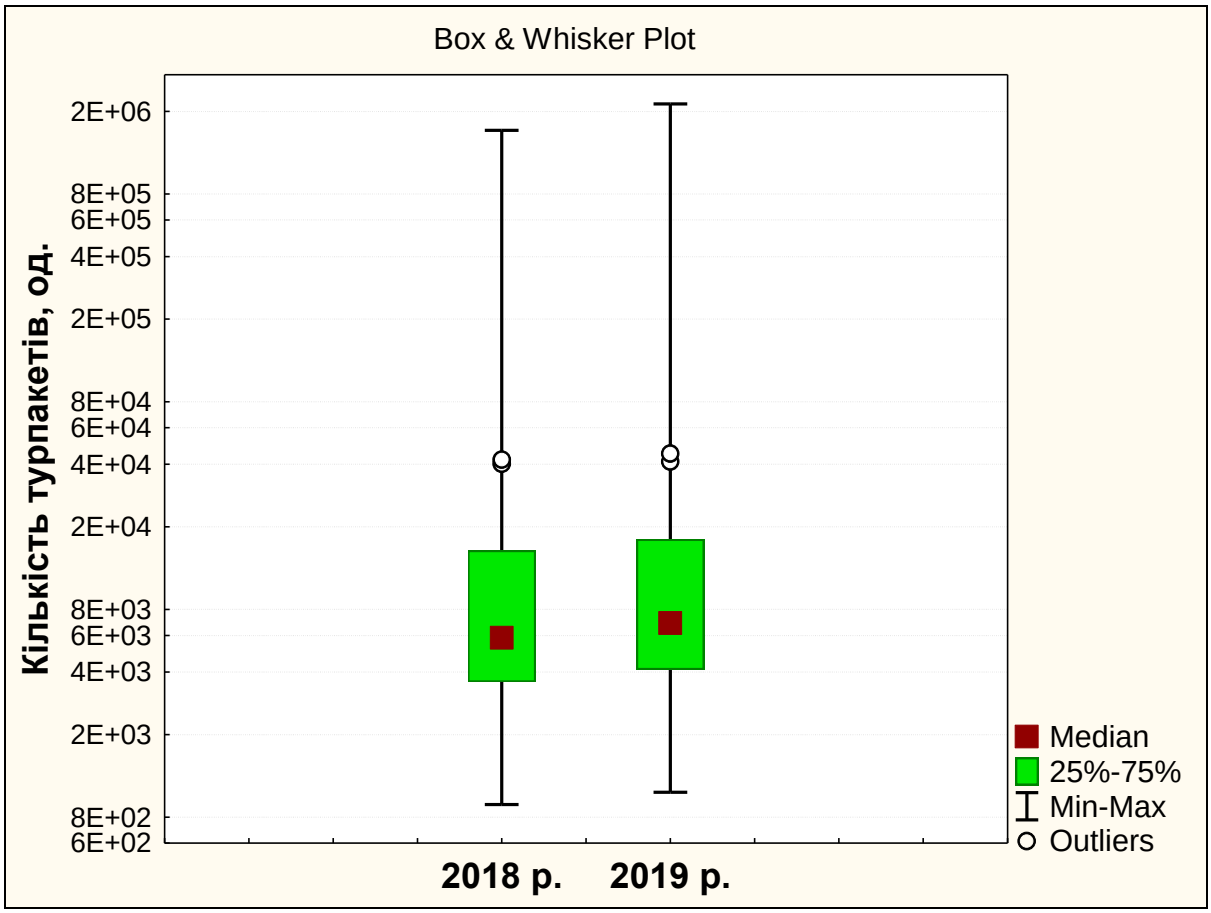


Рис. 3.2. Розподіл кількості турпакетів, реалізованих туроператорами та турагентами у 2018 та 2019 рр. за регіонами України

***Децилі* (*D*)** – це значення варіант, які ділять упорядкований ряд за об'ємом на десять рівних частин. У ряду розподілу виділяють

дев’ять децилів, оскільки медіана є одночасно п’ятим децилем. Розрахунок децилів оснований на кумулятивних частотах (частостях). Визначаються децилі за формулами:

$$D_1 = x_{D_1} + i\frac{0{,}1\sum f_i - S_{D_1-1}}{f_{D_1}}. \qquad (3.11)$$

$$D_2 = x_{D_2} + i\frac{0{,}2\sum f_i - S_{D_2-1}}{f_{D_2}} \qquad (3.12)$$

і т. д.

Дев’ятий дециль D_9 буде розраховуватись в інтервалі, куди входять накопиченим підсумком перші 90 % членів ряду:

$$D_9 = x_{D_9} + i\frac{0{,}9\sum f_i - S_{D_9-1}}{f_{D_9}}. \qquad (3.13)$$

В рівняннях (3.11)-(3.13) D_i – децилі (i – номер дециля); x_{D_i} – нижня межа інтервалу, що містить дециль; f_i – частота інтервалу, що містить дециль; S_{D_i-1} – накопичена частота інтервалу, який передує інтервалу, що містить дециль.

Таким чином, перший дециль D_1 буде розраховуватись в інтервалі, куди входять перші 10 % членів ряду; дев’ятий дециль D_9 буде розраховуватись в інтервалі, куди входять накопичені підсумком перші 90 % членів ряду. При цьому перший дециль D_1 буде показувати максимальне значення ознаки у перших 10 % членів ряду і мінімальне значення ознаки – у 90 %, які залишилися. Дев’ятий дециль D_9 буде показувати мінімальне значення ознаки в останніх 10 % одиниць ряду і максимальне значення ознаки – у перших 90 % членів ряду розподілу.

У практиці також використовують *перцентилі* P_3, P_{10}, P_{25}, P_{50}, P_{75}, P_{90} і P_{97}. Причому P_{25} та P_{75} відповідають першому і третьому квартилям, між якими знаходиться 50 % всіх членів ряду, а P_{50}

відповідає другому квартилю і дорівнює медіані, тобто $P_{50} = Me$.

3.2. Показники варіації

Середні величини дають узагальнюючу характеристику ознак статистичної сукупності. Проте, вони не показують, як розташовуються біля неї окремі значення (варіанти) ознаки, зосереджені вони поблизу середньої або значно відхиляються від неї. Інакше кажучи, для вимірювання варіації ознаки в сукупності потрібні міри, які оцінюють *мінливість, неоднорідність, розкид значень в групі даних*, що у відомому сенсі є невизначеністю.

Варіація, тобто зміни значень ознаки в часі або в просторі, є результатом впливу на кожну з одиниць сукупності великої кількості різних факторів, які по-різному поєднуються в кожному окремому разі, в результаті чого кожна з одиниць сукупності має своє власне, індивідуальне значення ознаки, що вивчається.

Середні величини не є універсальними характеристиками ознак статистичної сукупності (або варіаційних рядів). Дві статистичні сукупності можуть мати однакові середні значення ознаки, але в одній з них значення варіант можуть незначно відрізнятися від середньої, а в іншій – ці відмінності можуть бути великими, тобто в одному випадку варіація ознаки мала, а в іншому – велика. У зв'язку з цим, крім середніх величин, для характеристики ознак статистичної сукупності (або для оцінки ступеня варіювання ознак) використовують і *показники варіації*.

Вивчення характеру і ступеня варіації ознак є найважливішим питанням будь-якого статистичного дослідження, яке дозволяє оцінити не тільки коливання значень досліджуваної ознаки, але й її взаємозв'язок з іншими ознаками, схожість і відмінність сукупностей, випробувати гіпотези та ін.

Для глибокого аналізу досліджуваного соціально-економічного процесу або явища, в тому числі туризму, необхідно враховувати не тільки середні рівні досліджуваних показників, але й їхні варіації. В

туризмі найбільшою мірою варіації піддаються, наприклад, об'єми попиту та пропозиції.

Статистичні характеристики варіації поділяють на *абсолютні* і *відносні*. Найчастіше з ***абсолютних показників*** варіації використовуються *розмах варіації, середнє лінійне відхилення, дисперсія, стандартне відхилення (середньоквадратичне відхилення),* а з ***відносних показників*** *– коефіцієнт варіації, коефіцієнт осциляції, відносне лінійне відхилення, відносний показник квартильної варіації.*

3.2.1. Абсолютні показники варіації

Найпростішим показником варіації є ***розмах варіації*** (R), який являє собою різницю між крайніми членами варіаційного ряду спостережень: $R = x_{\max} - x_{\min}$. Чим сильніше варіює ряд, тим більше розмах варіації, і навпаки, чим слабкішою є варіація ряду, тим меншим є розмах варіації.

Розмах вимірює на числовій шкалі відстань, в межах якої змінюється варіанта і, в основному, використовується для вибірок невеликого об'єму. Він дає дуже поверхове уявлення про мінливість досліджуваного явища або процесу, не враховує кожне окреме значення. Він не є стабільним і змінюється від вибірки до вибірки.

Недоліком даного показника є те, що він оцінює лише межі варіації ознаки і не відбиває її коливання всередині цих меж, що можна показати на такому прикладі:

x_{1i}	2	10	16	21	28	32	37	42	48	54	$\bar{x}_1 = 29$
x_{2i}	2	10	29	29	33	32	32	34	35	54	$\bar{x}_2 = 29$

За *кількістю варіант* (n = 10), *лімітами* (значення максимальної та мінімальної варіант сукупності: $x_{\max} = 54$, $x_{\min} = 2$), *розмахом варіації* $R = x_{\max} - x_{\min}$ = 54-2 = 52 ці ряди не відрізняються один від

одного. Їх середні також є рівними між собою ($\bar{x}_1 = 29$, $\bar{x}_2 = 29$). Вони відрізняються одна від одної за ступенем варіювання, тобто ряди варіюють слабо або сильно відносно один одного. Однак ця особливість ніяк не відбивається на лімітах і розмаху варіації через те, що розмах варіації є грубою характеристикою мінливості спостережень (або ознаки), оскільки використовує всього два елементи ряду – $x_{\max}$ та $x_{\min}$.

Різний характер варіювання цих рядів можемо розглянути на прикладі простого показника варіації – ***середнього лінійного відхилення***, який будується на основі відхилень варіант від їх середньої, тобто $|x_i - \bar{x}| = d$. Сума таких відхилень, взята без урахування знаків і віднесена до кількості спостережень *n*, називається *середнім лінійним відхиленням*: $\bar{d} = \frac{\sum |x_i - \bar{x}|}{n}$. Тоді, якщо взяти суми відхилень варіант від їх середньої ($\bar{x} = 29$) для першого x_{1i} і другого x_{2i} рядів, отримаємо:

d_1	27	19	13	8	1	3	8	13	19	25	$\sum d_1 = 136$
d_2	27	19	0	0	4	3	3	5	6	25	$\sum d_2 = 92$

Звідси $\bar{d}_1$ = 136/10 = 13,6 і $\bar{d}_2$ = 92/10 = 9,2. Таким чином, перший ряд варіює сильніше, ніж другий. З наведених обчислень видно, що при однакових лімітах і розмаху варіації середнє лінійне відхилення виявилося неоднаковим для цих рядів: на величині цих показників позначився різний характер варіювання рядів.

Розрізняють *просте середнє лінійне відхилення* для *незгрупованих* даних:

$$\bar{d} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}, \qquad (3.14)$$

(де *n* – кількість членів ряду або кількість спостережень; x_i – *i*-е

значення варіанти) та *зважене середнє лінійне відхилення* для *згрупованих* даних або, інакше кажучи, для інтервальних варіаційних рядів:

$$\bar{d} = \frac{\sum_{i=1}^{k} \left| x_i - \bar{x} \right| \cdot f_i}{\sum_{i=1}^{k} f_i}, \qquad (3.15)$$

де f_i – частота i-ої варіанти, k – кількість часткових інтервалів (або кількість груп індивідуальних значень ознаки статистичної сукупності).

Перевагою середнього лінійного (абсолютного) відхилення є його розмірність, оскільки воно виражається в тих самих одиницях, що і значення досліджуваної величини.

Істотний недолік середнього абсолютного відхилення полягає в тому, що при його розрахунку не враховується знак різниці ($x_i - \bar{x}$), а, отже, внесок малих і великих відхилень x_i від $\bar{x}$ враховується однаково, що дещо знижує цінність даного параметра як міри мінливості.

Дисперсія та її властивості. Незважаючи на явну перевагу середнього лінійного відхилення перед розмахом варіації, цей показник не отримав широкого використання. Найпридатнішим є показник, який позбавлений недоліків розмаху варіації і побудований не на відхиленнях варіант від їхніх середніх, а на квадратах цих відхилень. Цей показник називають *дисперсією* (від лат. *dispersio* – розсіювання).

Дисперсія – це ступінь відхилення (або міра розкиду) варіант ознаки щодо її середнього значення. Чим більшою є дисперсія, тим сильнішим є розсіювання, тобто тим далі відхиляються значення варіант від середнього.

Дисперсію розраховують як середню арифметичну квадратів відхилень варіант від середньої величини.

Для *незгрупованих даних* обчислюється проста дисперсія за

формулою:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{1}{n}\cdot[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2]; \qquad (3.16)$$

для *згрупованих даних* обчислюється зважена дисперсія:

$$\sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot f_i}{\sum_{i=1}^{k} f_i}. \qquad (3.17)$$

В практичних обчисленнях для дисперсії часто зручнішою є формула, що представляє різницю між середньою арифметичною квадратів варіант і квадратом середньої арифметичної:

$$\sigma^2 = \overline{x^2} - (\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2, \qquad (3.18)$$

яка для згрупованих даних має такий вигляд:

$$\sigma^2 = \overline{x^2} - (\bar{x})^2 = \frac{\sum_{i=1}^{k} x_i^2 \cdot f_i}{\sum_{i=1}^{k} f_i} - \left(\frac{\sum_{i=1}^{k} x_i \cdot f_i}{\sum_{i=1}^{k} f_i}\right)^2. \qquad (3.19)$$

Цінність дисперсії полягає в тому, що будучи мірою варіювання числових значень ознаки навколо їх середньої величини, вона вимірює і внутрішню мінливість значень ознаки, яка залежить від різниць між спостереженнями.

Перевагою дисперсії перед іншими показниками варіації є те, що вона розкладається на складові компоненти, дозволяючи тим самим оцінювати вплив різних чинників на величину досліджуваної ознаки.

Дисперсія має низку важливих властивостей, з-поміж яких зазначимо такі:

1. Дисперсія постійної величини дорівнює нулю:

$$\sigma_A^2 = 0.$$

2. Дисперсія не змінюється, якщо кожну варіанту сукупності зменшити (або збільшити) на одне й те саме постійне число A:

$$\sigma^2 = \frac{1}{n}\sum[(x_i - A) - (\bar{x} - A)]^2 = \frac{1}{n}\sum(x_i - \bar{x})^2;$$

$$\sigma^2 = \frac{1}{n}\sum[(x_i + A) - (\bar{x} + A)]^2 = \frac{1}{n}\sum(x_i - \bar{x})^2.$$

3. Якщо кожну варіанту сукупності поділити (або помножити) на одне й те саме постійне число A, то дисперсія зменшиться (або збільшиться) в A^2 разів:

$$\sigma^2 = \frac{1}{n}\sum\left(\frac{x_i}{A} - \frac{\bar{x}}{A}\right)^2 = \frac{1}{A^2}\,\frac{\sum(x_i - \bar{x})^2}{n};$$

$$\sigma^2 = \frac{1}{n}\sum\left(x_i \cdot A - \bar{x} \cdot A\right)^2 = A^2\,\frac{\sum(x_i - \bar{x})^2}{n}.$$

4. Дисперсія від середньої завжди менше дисперсій, обчислених від будь-яких інших величин, тобто вона має властивість мінімальності:

$$\sigma^2 < \frac{\sum(x_i - A)^2}{n}, \quad \text{якщо} \quad A \neq \bar{x}.$$

Важливо зазначити, що дисперсія при малих значеннях об'єму (приблизно $n \leq 30$) і при великих значеннях ($n > 30$) обчислюється за різними формулами:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad \textit{при малому об'ємі вибірки } (n \leq 30); \qquad (3.20)$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \quad \textit{при великому об'ємі вибірки } (n > 30). \qquad (3.21)$$

Заміну дільника n на $n\text{-}1$ здійснюють для того, щоб усунути систематичну помилку для вибірок з малим об'ємом або «зміщення» щодо дисперсії всієї генеральної сукупності. У зв'язку з цим, показники намагаються оцінювати таким чином, щоб їх оцінки були незміщеними. Для вирішення проблеми зміщення вибіркової

дисперсії в її розрахунок вносять корегування – множать на $\frac{n}{n-1}$. Інакше кажучи, при розрахунку в знаменник ставлять не *n*, а *n*-1, як в рівнянні (3.20). Неважко помітити, що з ростом *n* (тобто об'єму вибірки) $\frac{n}{n-1}$ наближається до 1, тобто різниця між значеннями вибіркової і генеральної дисперсій зменшується.

Недоліком показника дисперсії є те, що його одиниця виміру не збігається з одиницею. виміру аналізованої ознаки. Цей недолік усувається шляхом розрахунку і аналізу *стандартного відхилення*. Ця величина в ряді випадків виявляється більш зручною характеристикою варіювання, ніж дисперсія, через те, що виражається в тих самих одиницях, що і середня величина.

Стандартне відхилення – це показник, який являє собою корінь квадратний з дисперсії:

для *незгрупованих даних* $$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}; \qquad (3.22)$$

для *згрупованих даних* $$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2 \cdot f_i}{\sum f_i}}. \qquad (3.23)$$

Таким чином, абсолютні показники варіації – середнє лінійне відхилення, дисперсія і стандартне відхилення – показують, наскільки в середньому відхиляються конкретні варіанти ознаки від її середнього значення.

3.2.2. Відносні показники варіації

Часто при дослідженнях виникає необхідність порівняння мінливості рядів, утворених з характеристик, які істотно розрізняються за величиною середніх. А оскільки показники варіації характеризують відхилення варіант від середнього, то величини лінійного відхилення різних варіаційних рядів можна порівнювати лише тоді, коли ці ряди характеризуються приблизно однаковими середніми. У практиці не завжди буває так, тому абсолютні

показники варіації аналізованих рядів виявляються непорівнянними.

Нехай, наприклад, є два ряди, які представляють кількість суб'єктів туристичної діяльності з 2000 по 2019 рр. в Харківській та Одеській областях. Необхідно визначити, в якій області сильніше варіює аналізована ознака, тобто кількість суб'єктів туристичної діяльності за цей період. Імовірно, на основі порівняння стандартних відхилень на це питання відповісти не можна, оскільки середні величини ознаки в цих двох рядах з великою ймовірністю будуть відрізнятися. Крім того, через те, що стандартне відхилення – це відхилення від середнього, некоректним буде порівнювати величини цих відхилень щодо різних середніх. Коректні результати в цьому разі можна буде отримати, якщо середні двох рядів матимуть однакові значення.

Зіставлення мінливості подібних рядів здійснюється за допомогою *відносних показників варіації*. До них належать:

1. *Коефіцієнт осциляції*, який відбиває відносне коливання крайніх значень ознаки навколо середньої:

$$K_R = \frac{R}{\bar{x}} \cdot 100\,\%, \qquad (3.24)$$

де R – розмах варіації.

2. *Лінійний коефіцієнт варіації* (або *відносне лінійне відхилення*), який характеризує частку усередненого значення абсолютних відхилень від середньої величини:

$$K_L = \frac{\bar{d}}{\bar{x}} \cdot 100\,\%, \qquad (3.25)$$

де $\bar{d}$ – середнє лінійне відхилення.

3. *Відносний показник квартильної варіації*:

$$K_{d_Q} = \frac{d_Q}{2Me} \cdot 100\,\%, \qquad (3.26)$$

де d_Q – абсолютний показник квартильної варіації.

4. ***Коефіцієнт варіації*** (C_V) – найбільш поширений показник, який являє собою відсоткове відношення стандартного відхилення до

середнього значення:

$$C_{\text{v}} = \frac{\sigma}{\bar{x}} \cdot 100\,\%. \qquad (3.27)$$

Якщо центр розподілу представлений медіаною (*Me*), то використовують *квартильний коефіцієнт варіації*:

$$C_Q = \frac{Q_3 - Q_1}{2Me}. \qquad (3.28)$$

Коефіцієнт варіації є величиною безрозмірною, що зручно для порівняльних оцінок варіації однієї й тієї самої ознаки в різних сукупностях, в яких ознака має різні середні значення і виражена в різних одиницях виміру. Виражається коефіцієнт варіації в частках одиниці або у відсотках

Виходячи з величини коефіцієнта варіації, можна судити про ступінь варіації ознаки, і, отже, про однорідність сукупності за цією ознакою. Чим більшою є величина коефіцієнта варіації, тим більшим є розкид значень ознаки навколо середнього, тим більшою є неоднорідність сукупності.

Значення коефіцієнта варіації, що перевищують 30 %, є характерними для якісно неоднорідної сукупності.

За величиною коефіцієнта варіації розрізняють такі ступені однорідності сукупності (табл. 3.7):

Таблиця 3.7

Градації коефіцієнта варіації

Коефіцієнт варіації, C_{v} (%)	Ступінь однорідності сукупності
< 30	однорідна
30-60	середня
> 60	неоднорідна

Розглянемо *приклад розрахунку показників варіації*.

***Приклад* 3.6.** Є дані про середню тривалість перебування у колективних засобах розміщення (діб) по регіонах України (2019 р.),

табл. 3.8. Таблиця складена авторами навчального посібника за даними державного статистичного спостереження «Колективні засоби розміщення».

Необхідно визначити: середню арифметичну, середнє лінійне відхилення, дисперсію, стандартне відхилення, коефіцієнт варіації тривалості перебування туристів у колективних засоби розміщення.

Таблиця 3.8

Розрахунок показників варіації

Номер варіанти, *i*	Значення (діб), x_i	Частота, f_i	$x_i \cdot f_i$	$\lvert x_i - \bar{x} \rvert \cdot f_i$	$(x_i - \bar{x})^2 \cdot f_i$
1	1,7	2	3,4	2,4	2,88
2	1,8	1	1,8	1,1	1,21
3	1,9	2	3,8	2,0	2,00
4	2,0	2	4,0	1,8	1,62
5	2,1	2	4,2	1,6	1,28
6	2,2	2	4,4	1,4	0,98
7	2,3	1	2,3	0,6	0,36
8	2,4	1	2,4	0,5	0,25
9	2,5	1	2,5	0,4	0,16
10	2,7	2	5,4	0,4	0,08
11	2,9	3	8,7	0,0	0,00
12	4,0	2	8,0	2,2	2,42
13	5,0	1	5,0	2,1	4,41
14	5,1	1	5,1	2,2	4,84
15	5,5	1	5,5	2,6	6,76
16	5,9	1	5,9	3,0	9,00
Сума		25	72,4	24,3	38,25

Середню арифметичну обчислимо за формулою (3.3):

$$\bar{x} = \frac{\sum_{i=1}^{16}(f_i \cdot x_i)}{\sum_{i=1}^{16} f_i} = \frac{72{,}4}{25} = 2{,}90\,.$$

Розмах варіації $R = x_{max} - x_{min} = 5{,}9 - 1{,}7 = 4{,}2$.

Середнє лінійне відхилення обчислимо за формулою (3.15):

$$\bar{d} = \frac{\sum_{i=1}^{16} \left| x_i - \bar{x} \right| \cdot f_i}{\sum_{i=1}^{16} f_i} = \frac{24{,}3}{25} = 0{,}97.$$

Для визначення дисперсії скористаємося рівнянням (3.17). Однак, оскільки n <30, в знаменнику ставимо (*n*-1):

$$\sigma^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot f_i}{n-1} = \frac{38{,}25}{24} = 1{,}59.$$

Стандартне відхилення дорівнює:

$$\sigma = \sqrt{\sigma^2} = \sqrt{1{,}59} = 1{,}26 \text{ діб.}$$

Таким чином, значення середньої тривалості перебування у колективних засобах розміщення у всіх регіонах України відхиляються від середньої тривалості перебування на 1,26 доби.

Розрахуємо коефіцієнт варіації за формулою (3.27):

$$C_v = \frac{\sigma}{\bar{x}} \cdot 100\% = \frac{1{,}26}{2{,}90} \cdot 100\% = 43{,}5\,\%.$$

Розрахований коефіцієнт свідчить про те, що сукупність регіонів за середньою тривалістю перебування у колективних засобах розміщення є неоднорідною, оскільки C_v > 30 %.

Для порівняльної оцінки коефіцієнтів варіації ми вирахували також їх величину для ряду середньої тривалості перебування у колективних засобах розміщення по регіонах України за 2016 р. Результати аналізу наведені на рис. 3.3, з якого видно, що середня тривалість перебування по регіонах України в 2019 р. варіює відносно сильно (C_v = 43,5 %) у порівнянні з 2016 р. (C_v = 41,7 %). Крім цього, згідно з табл. 3.7, тривалість перебування має високу неоднорідність просторового розподілу, тобто сильно змінюється від регіону до регіону як в 2016 р., так і в 2019 р.

Як було сказано вище, при порівнянні ступеня варіації двох рядів, у яких середні величини різні, коректно використовувати відносний показник варіації, а не абсолютний, інакше можна зробити невірні висновки. Сказане добре видно на розглянутому прикладі. Так, за величиною абсолютного показника варіації, тобто за величиною σ сильніше варіює ряд за 2016 р. ($\sigma_{2016} = 1{,}85$; $\sigma_{2019} = 1{,}26$), а за величиною відносного показника варіації (C_v) сильніше варіює ряд за 2019 р. ($C_{v2016} = 41{,}7$ %; $\sigma_{v2019} = 43{,}5$ %).

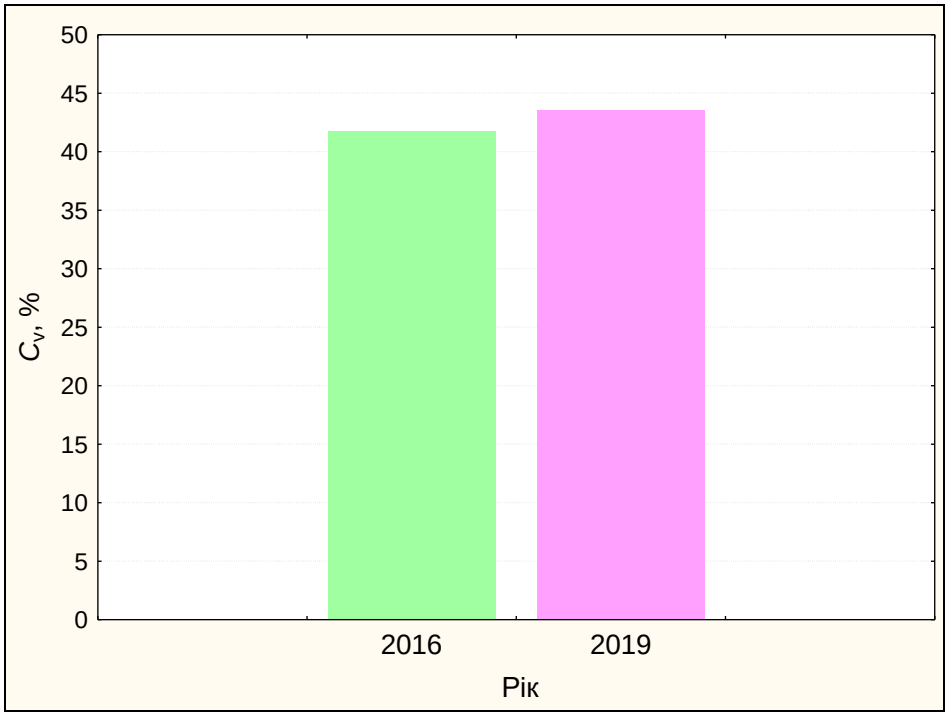


Рис. 3.3. Гістограма розподілу коефіцієнта варіації середньої тривалості перебування туристів у колективних засобах розміщення по регіонах України

Питання для самоконтролю

3.1. Що розуміють під середньою величиною?

3.2. Які існують види середніх величин?

3.3. Якими є види і призначення степеневих середніх величин?

3.4. Якими є види і призначення структурних середніх величин?

3.5. В яких формах обчислюються степеневі середні залежно від представлення вихідних даних?

3.6. Що являє собою проста середня арифметична?

3.7. У чому полягає відмінність середньої арифметичної зваженої від простої середньої арифметичної?
3.8. Перелічить властивості середньої арифметичної величини.
3.9. В яких випадках використовують структурні середні?
3.10. Що характеризують мода і медіана?
3.11. Чому медіану і моду називають структурними середніми?
3.12. Як обчислюють медіану при непарному і парному об'ємі вибірки?
3.13. Вкажіть особливості визначення моди і медіани в дискретному і інтервальному рядах.
3.14. Що таке квартилі, децилі, і якою є методика їх розрахунку? Які ще види структурних середніх існують?
3.15. Що характеризують показники варіації?
3.16. Що характеризує розмах варіації, як він обчислюється?
3.17. Назвіть основні недоліки розмаху варіації.
3.18. Що таке середнє лінійне відхилення, як воно обчислюється?
3.19. Назвіть основні недоліки середнього лінійного відхилення.
3.20. Що являє собою дисперсія і як вона обчислюється?
3.21. Перелічить властивості дисперсії.
3.22. У чому полягає значення стандартного відхилення, як воно обчислюється?
3.23. Який відносний показник варіації використовується найчастіше?
3.24. Що називається коефіцієнтом варіації, як він обчислюється і в яких випадках використовується?

Завдання для самостійного виконання

Завдання 3.1. В табл. 1 наведено кількість іноземних туристів, обслугованих туроператорами та турагентами України за період з 2000 по 2019 рр. Розрахуйте середню кількість іноземних туристів, обслугованих за рік.

Таблиця 1

Кількість іноземних туристів, обслугованих туроператорами та турагентами України за період з 2000 по 2019 рр.

Рік	Кількість іноземних туристів	Рік	Кількість іноземних туристів
2000	377871	2010	335835
2001	416186	2011	234271
2002	417729	2012	270064
2003	590641	2013	232311
2004	436311	2014	17070
2005	326389	2015	15159
2006	299125	2016	35071
2007	372455	2017	39605
2008	372752	2018	75945
2009	282287	2019	86840

Завдання 3.2. На основі пропонованих даних розрахуйте середній прибуток турагентств.

Таблиця 2

Розподіл турагентств за об'ємом річного прибутку

Прибуток, млн. грн.	Кількість турагентств
100-200	90
200-300	30
300-400	40
400-600	50
600-800	60
800-1000	20

Завдання 3.3. За даним про об'єм продажів турпакетів визначте їх середню ціну реалізації.

Таблиця 3

Розподіл турпакетів різної ціни

Порядковий номер	1	2	3	4	5
Ціна турпакету, євро, x_i	230	235	240	245	260
Кількість проданих турпакетів, од., f_i	17	15	12	10	7

Завдання 3.4. За даними про розподіл доходів готелів України та інших місць для тимчасового проживання за 2010 рік визначте: 1) їхній середній дохід на основі моди і медіани; 2) перший і третій квартилі; 3) перший і дев'ятий децилі.

Таблиця 4

Розподіл доходів готелів та інших місць для тимчасового проживання за 2010 рік, млн. грн.

Групи готелів та ін. місць для тимчасового проживання за середнім прибутком, млн. грн.	Відсоток готелів та ін. місць для тимчасового проживання до підсумку
до 12	4
12–33	48
33–54	4
54–75	7
75–96	15
96–138	7
138–223	4
223–391	4
391–1235	7

Завдання 3.5. Визначте середньооблікову кількість штатних працівників готелю за рік, якщо відома їх кількість за певний період: 2012-2018 рр.

Таблиця 5

Середньооблікова кількість штатних працівників готелю за рік

Рік	2012	2013	2014	2015	2016	2017	2018
Кількість працівників	185	1200	1216	1217	1219	1220	1219

Завдання 3.6. Визначте медіанний і модальний вік туристів, які здійснили поїздки з метою відпочинку за рік.

Таблиця 6

Вікові групи туристів і кількість їх поїздок на рік з метою відпочинку

Вікові групи, років, x_i	Частота (кількість поїздок), f_i	Сума накопичених частот, S
до 16	4	
17-20	6	
21-29	4	
30-39	5	
40-49	6	
50-59	11	
60-69	12	
70+	7	

Завдання 3.7. В таблиці 7 наведені витрати суб'єктів туристичної діяльності (за певний рік) на послуги сторонніх організацій, що використовуються при виробленні туристського продукту. Визначте значення першого і третього квартилів, тобто витрати (25 % і 75 %) на послуги сторонніх організацій.

Таблиця 7

Витрати суб'єктів туристичної діяльності на послуги сторонніх організацій, що використовуються при виробленні туристського продукту

Витрати суб'єктів туристичної діяльності на послуги (млн. грн.)	Кількість послуг до підсумку (частота, *f*)	Накопичена частота (*S*)	Накопичений відсоток послуг до підсумку
0,1-0,5	1		
0,5-0,9	2		
0,9-1,3	3		
1,3-2,9	1		
2,9-6,1	1		
6,1-12,5	8		
12,5-18,9			
Разом			

Завдання 3.8. В таблиці 8 наведені дані про середню тривалість перебування (діб) туристів в різних колективних засобах розміщення України за 2018 р. Згрупуйте дані за їхньою частотою і визначте середню арифметичну, середнє лінійне відхилення, дисперсію, стандартне відхилення, коефіцієнт варіації тривалості перебування туристів у колективних засобах розміщення.

Таблиця 8

Розрахунок показників варіації

Номер варіанти, i	Значення (діб), x_i	Частота, f_i	$x_i \cdot f_i$	$\left\|x_i - \bar{x}\right\| \cdot f_i$	$(x_i - \bar{x})^2 \cdot f_i$
1.	1,9				
2.	3,4				
3.	2,5				
4.	2,7				
5.	3				
6.	2,1				
7.	5,8				
8.	2,8				
9.	2,1				
10.	2,2				
11.	2,9				
12.	2,4				
13.	4,5				
14.	3,9				
15.	1,9				
16.	2,8				
17.	1,8				
18.	2,1				
19.	2				
20.	6,9				
21.	2,1				
22.	1,8				
23.	2,2				
24.	1,5				
25.	2,1				
Сума					

РОЗДІЛ 4
ЗАКОНИ РОЗПОДІЛУ

Варіаційний ряд є статистичним аналогом розподілу ознаки (випадкової величини), а його числові характеристики – середня арифметична, дисперсія та ін. – аналогами відповідних числових характеристик випадкової величини: математичного сподівання, дисперсії та ін. Так само поняття частості (відносної частоти) для варіаційного ряду є аналогічним поняттю ймовірності для випадкової величини. При цьому під *розподілом ознак* (випадкових величин) розуміється співвідношення між їхніми значеннями і частотою зустрічаваності.

Одним з найважливіших завдань математичної статистики є встановлення теоретичного закону розподілу випадкової величини, що характеризує досліджувану ознаку за спробним (емпіричним) розподілом, який представляє варіаційний ряд. Для вирішення цього завдання необхідно визначити вид і параметри закону розподілу. Характер, тип розподілу виражає загальну закономірність даного типу розподілу, тобто відображає загальні умови, що випливають із сутності і природи явища, і особливості, які впливають на варіацію досліджуваної ознаки.

Таким чином, при аналізі варіаційних рядів виявляються закономірності розподілу, які можуть бути описані за допомогою деяких кривих, які називаються теоретичними кривими розподілу. Характер розподілу виявляється при великій кількості спостережень і малій довжині інтервалу варіаційного ряду.

Якщо довжина інтервалу варіаційного ряду прямує до нуля, то графічне зображення емпіричного варіаційного ряду (полігон, гістограма) приймають вигляд плавної кривої, яка називається кривою розподілу. Кількість видів кривих і, відповідно, законів розподілу є великою, проте деякі з них мають теоретичне обґрунтування і зустрічаються частіше за інші.

Залежно від того, за якими даними будується розподіл (за

спостереженнями або за обчисленими даними), розрізняють *емпіричні* і *теоретичні* криві розподілу.

Емпірична крива розподілу – це фактична крива розподілу, отримана за даними спостережень, яка відображає як загальні, так і випадкові умови, що визначають розподіл.

Теоретична крива розподілу – це крива, яка виражає функціональний зв'язок між зміною варіювальної ознаки і зміною частот, і характеризує певний тип розподілу. При цьому теоретичний розподіл відіграє роль певної ідеалізованої моделі емпіричного розподілу, а сам аналіз варіаційного ряду зводиться до зіставлення емпіричного і теоретичного розподілів.

В практиці статистичних досліджень зустрічаються, наприклад, такі розподіли: нормальний, біноміальний, розподіл Пуассона та ін. Кожний розподіл має свою специфіку і сферу застосування.

У загальному випадку для визначення виду розподілу досліджуваної ознаки виконують зіставлення емпіричних і теоретичних частот цього розподілу між собою.

Знаходження ряду теоретичних частот для наявного емпіричного розподілу називається вирівнюванням емпіричних кривих за виявленим (наприклад, нормальним) законом. Цей процес має велике теоретичне і практичне значення. Кожному значенню ознаки x_i відповідає при цьому певне значення так званої функції розподілу $F(x_i)$, що показує, наскільки ймовірним є існування варіант, менших за дане значення x_i.

Таким чином, для емпіричного ряду розподілу важливо знайти функцію розподілу, тобто підібрати таку теоретичну криву розподілу, яка найбільш точно відображала б закономірність розподілу. Знаходження функції кривої розподілу називається моделюванням емпіричного ряду розподілу.

4.1. Нормальний закон розподілу

В більшості розподілів, з якими доводиться зустрічатися при вивченні природних і соціально-економічних явищ (в тому числі явищ у сфері туризму), проявляється певна закономірність:

– крайні значення – найменше та найбільше – з'являються рідко;

– чим ближче значення ознаки до середньої арифметичної, тим воно частіше зустрічається;

– в центрі розподілу є такі значення, які зустрічаються найчастіше і утворюють у варіаційному ряду модальний клас.

Такий розподіл значень ознаки настільки часто зустрічається в різних галузях науки і практики, що спочатку він приймався за норму будь-якого масового випадкового прояву ознак, і у зв'язку з цим набув особливої назви – *нормальний розподіл* (або *розподіл Гаусса*). Практика показує, що характеристики розподілу соціально-економічних явищ співпадають з теоретичним розподілом, що називається нормальним розподілом.

Закон нормального розподілу характерний для розподілу подій в разі, коли їх результат являє собою результат спільного впливу великої кількості незалежних чинників, і жоден з цих чинників не робить переважаючого впливу.

Припущення про те, що більшість результатів господарської діяльності (доходи, прибуток і т. п.) як випадкові величини підкоряються закону, близькому до нормального, широко використовується в літературі з проблеми кількісної оцінки економічного ризику.

Насправді нормальний розподіл соціально-економічних явищ (в тому числі туризму) в чистому вигляді зустрічається рідко, однак, якщо однорідність сукупності дотримана, часто фактичні розподіли є близькими до нормального. На практиці для перевірки обґрунтованості прийнятого розподілу використовуються різні критерії згоди (між емпіричним і теоретичним розподілом), які дозволяють прийняти або відкинути висунуту гіпотезу про закон розподілу.

Значимість нормального розподілу визначається тим, що він служить хорошим наближенням для великої кількості наборів випадкових величин, одержуваних при спостереженнях і експериментах.

Теоретичне обґрунтування того, що за допомогою нормального розподілу можна описати переважну більшість процесів, які реально відбуваються, дає закон великих чисел. Зі збільшенням числа спостережень нерідко багато інших видів розподілів приймають вигляд нормального. З урахуванням цих властивостей, доброї розробленості і порівняно простої формальності структури нормальний розподіл частіше за інших застосовується в багатовимірній статистиці.

Як вже зазначалося вище, нормальний розподіл майже завжди має місце, коли спостережувані величини формуються під спільним впливом великої кількості незалежних (випадкових) чинників, при цьому жоден з них істотно не переважає інші. У зв'язку з цим нормальний розподіл застосовується для опису розподілу ознак, на які діє велика кількість незалежних факторів, серед яких немає домінуючих.

Щільність нормального розподілу ймовірностей описується рівнянням:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma}\right)^2\right], \qquad (4.1)$$

де $f(x)$ – щільність розподілу (ордината кривої нормального розподілу, аналог щільності розподілу випадкової величини); x – значення досліджуваної ознаки; $\bar{x}$ – середнє значення ознаки в розподілі (або *математичне сподівання* $M(x)$); σ – стандартне відхилення.

Слід зазначити, що формально математичне сподівання відповідає середній величині емпіричного розподілу, однак, по суті, ці показники ототожнювати не можна. Середню величину визначають як суму всіх членів ряду, віднесену до їх загальної кількості, а

математичне сподівання *M(x)* являє собою суму добутків членів ряду та їх ймовірностей: $M(x) = \sum p_i x_i = p_1 x_1 + p_2 x_2 + ... + p_n x_n$. Емпірична середня прямує до математичного сподівання випадкової величини в міру збільшення кількості випробувань. При невеликій кількості випробувань середня може значно відхилятися від свого математичного сподівання.

Графік зміни щільності нормального розподілу зображений на рис. 4.1*а*. Він являє собою симетричну колоколоподібну криву.

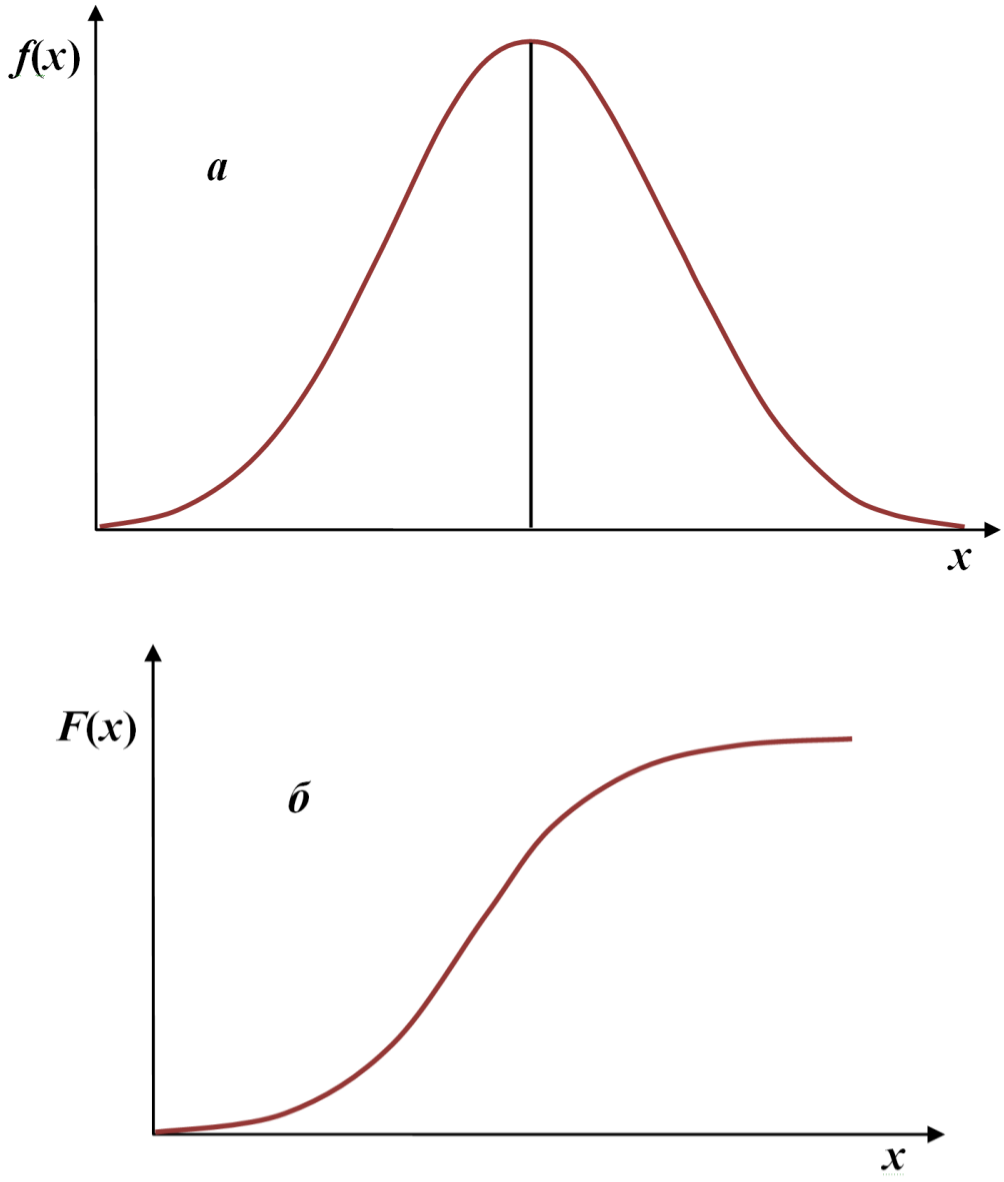


Рис. 4.1. Щільність (*а*) і функція (*б*) нормального розподілу

Інтегральна щільність *F*(*x*), або функція нормального розподілу, наведена на рис. 4.1*б* і визначається як:

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{x}\exp\left[-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma}\right)^2\right]dx. \qquad (4.2)$$

Основні властивості кривої нормального розподілу:

– є сильна тенденція даних групуватися навколо центру (середнього);

– крива розподілу є симетричною відносно осі, яка проходить через центр розподілу $\bar{x}=Mo=Me$, тобто позитивні і негативні відхилення від центру є рівноймовірними*;*

– частота відхилень швидко падає (тобто дві гілки кривої асимптотично наближаються до осі абсцис), коли відхилення від центру стають великими;

– при зміні величини $\bar{x}$ форма нормальної кривої не змінюється, тільки її графік зміщується праворуч або ліворуч;

– зміна величини σ призводить до зміни ширини кривої: при зменшенні σ крива стає більш вузькою за рахунок меншого розсіювання варіант навколо середньої, а при збільшенні σ крива розширюється.

Вся розташована під кривою площа дорівнює одиниці або 100 %, оскільки ця площа дорівнює ймовірності того, що випадкова величина *x* прийме одне зі своїх значень, тобто дорівнює ймовірності достовірної події.

Таким чином, площа, обмежена кривою зверху і віссю абсцис знизу, характеризує ймовірність появи певних значень ознаки. Для симетричного нормального розподілу (рис. 4.2) ординати таких значень *x*, які відрізняються від $\bar{x}$ на $\pm\sigma$, виділяють площу, яка становить ≈ 68 % від загальної площі, тобто в межах $\bar{x}\pm\sigma$ знаходиться 68 % всіх значень ознаки, в межах $\bar{x}\pm 2\sigma$ – 96 % значень, в межах $\bar{x}\pm 3\sigma$ – 99,8 % значень ознаки. Цей висновок називається *правилом* «*трьох сигм*», відповідно до якого можна вважати, що всі можливі значення нормально розподіленої ознаки вміщуються в інтервал $\bar{x}\pm 3\sigma$.

Інтервал зміни величини *x*, заданий у вигляді $\bar{x}\pm n\sigma$,

називається *довірчим*. *Ймовірність* попадання випадкової величини в цей інтервал називається *довірчою* (в теорії надійності вона називається *надійністю*). Вірогідність непопадання в довірчий інтервал називається *рівнем значущості* (в теорії надійності – це *ймовірність ризику*).

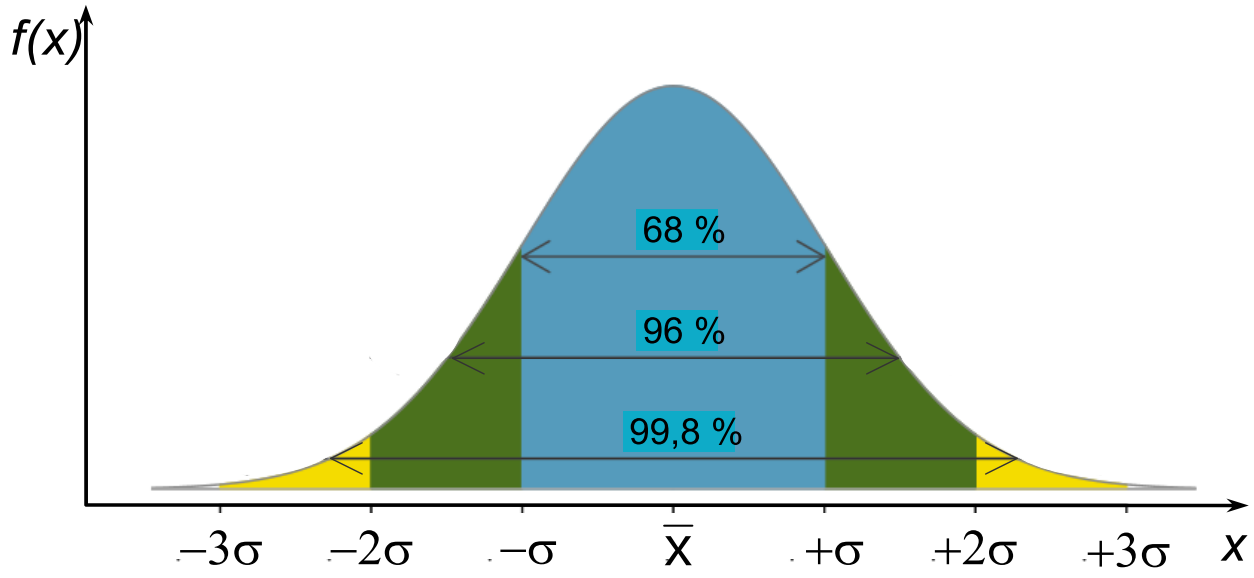


Рис. 4.2. Ілюстрація правила «трьох сигм»

Вважається, що, якщо 75 % варіант вибірки знаходиться в межах $\bar{x} \pm \sigma$, то це відповідає нормі (стандартному відхиленню); якщо в межах $\bar{x} \pm 2\sigma$, – є незначне відхилення від норми; якщо не виходить за межі $\bar{x} \pm 3\sigma$, то можна стверджувати, що має місце нормальний розподіл.

Для симетричних розподілів частоти будь-яких двох варіант, рівновіддалених в обидва боки від центру, є рівними між собою. Розраховані для таких рядів розподілів характеристики, такі як середнє арифметичне, мода і медіана, збігаються: $\bar{x} = Mo = Me$. Якщо зазначені співвідношення порушені, то це свідчить про наявність асиметрії розподілу.

Належність спостережуваних даних до нормального закону є необхідною передумовою для коректного застосування більшості класичних методів математичної статистики. У зв'язку з цим перевірка на відхилення від нормального закону є частою

процедурою в ході проведення спостережень (або вимірювань, випробувань).

Одним із способів перевірки відхилення розподілу щільності ймовірностей від нормального закону є оцінка таких характеристик, як *коефіцієнти асиметрії* (*As*) та *ексцес* (*Es*).

4.2. Асиметрія та ексцес

Ознаки природних і соціально-економічних процесів і явищ іноді характеризуються розподілом, який значно відрізняється від нормального закону. Коли які-небудь чинники сприяють появі значень ознаки, що відрізняються від середньої величини в бік зменшення або в бік збільшення, утворюються асиметричні розподіли.

Асиметричний варіаційний ряд – це ряд, в якому частоти варіант, які однаково розподілені щодо середньої ліворуч і праворуч, не рівні між собою і змінюються по-різному. Ряд з таким асиметричним розподілом частот часто називають *скошеним*. Відповідно до цього розрізняють *ліву і праву асиметрії*.

Для оцінки відхилення емпіричного розподілу від нормального, крім інших показників, використовують *коефіцієнти асиметрії* (*As*) і *ексцесу* (*Es*).

Асиметрія – це ступінь несиметричності розподілу відносно середнього, тобто коефіцієнт асиметрії служить мірою асиметрії і використовується для перевірки розподілу на симетричність.

Коефіцієнти асиметрії (*As*) обчислюються як:

для *незгрупованих даних*
$$As = \left(\frac{\sum (x_i - \bar{x})^3}{n} \right) / \sigma^3, \qquad (4.3)$$

для *згрупованих даних*
$$As = \left(\frac{\sum (x_i - \bar{x})^3 \cdot f_i}{\sum f_i} \right) / \sigma^3. \qquad (4.4)$$

Стандартна помилка для *As* визначається за формулою:

$$S_{As} = \sqrt{\frac{6(n-1)}{(n+1)(n+3)}}. \qquad (4.5)$$

Розрізняють розподіл, *скошений праворуч* (*додатний*) або *ліворуч* (*від'ємний*), рис. 4.3-4.4**.**

При строго симетричних розподілах сума третіх ступенів відхилень варіант x_i від середньої арифметичної дорівнює нулю, тоді і $As = 0$. При правобічній асиметрії As матиме додатну величину ($As > 0$; рис. 4.3); при лівобічній асиметрії – від'ємну величину ($As < 0$; рис. 4.4).

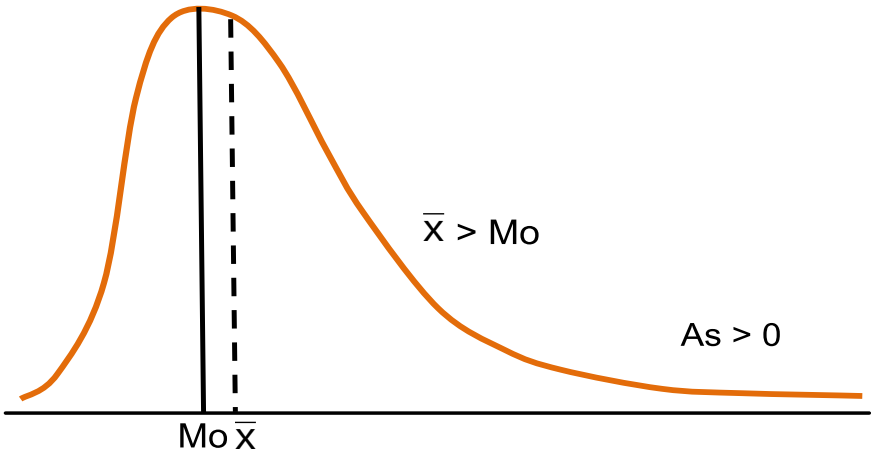


Рис. 4.3. Розподіл, скошений вправоруч (додатна асиметрія)

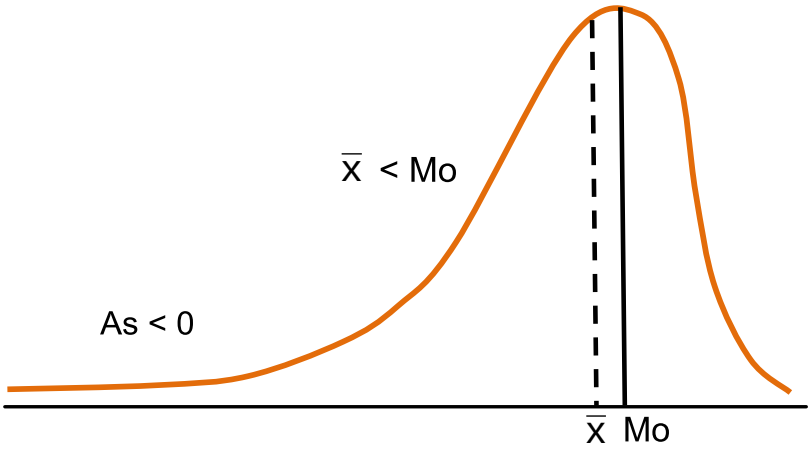


Рис. 4.4. Розподіл, скошений ліворуч (від'ємна асиметрія)

Так, при $\bar{x} > Me > Mo$ різниці між $\bar{x} - Mo$ та $\bar{x} - Me$ є додатними, і асиметрія є правобічною ($As > 0$, рис. 4.3). Це означає, що в розподілі частіше зустрічаються більш високі значення ознаки. При

$\bar{x} < Me < Mo$, навпаки, різниці $\bar{x} - Mo$ та $\bar{x} - Me$ є від'ємними, і асиметрія є лівобічною ($As < 0$, рис. 4.4). Це означає, що в розподілі частіше зустрічаються низькі значення ознаки.

Чим більшою є величина $|As|$, тим більш асиметричним є розподіл.

Іноді використовують таку оціночну шкалу асиметрії:

$|As| \le 0{,}25$ – асиметрія є незначною;

$0{,}25 \le |As| \le 0{,}5$ – асиметрія є помітною (помірною);

$|As| \ge 0{,}5$ – асиметрія є суттєвою.

Прийнято також вважати, що, якщо відношення коефіцієнта асиметрії до величини своєї помилки менше трьох (тобто $|As|/S_{As} < 3$), то асиметрія є несуттєвою, а її наявність пояснюється впливом випадкових факторів. А в іншому разі асиметрія є статистично значущою, і факт її наявності потребує додаткової інтерпретації.

Поряд з симетричними і скошеними розподілами варіаційні ряди можуть мати відхилення по висоті. У зв'язку з цим, крім аналізу симетричності розташування кривої розподілу відносно середньої арифметичної, проводиться також аналіз *гостровершинності розподілу*. Відхилення висоти максимуму вгору або вниз від вершини кривої нормального розподілу називається *ексцесом* (рис. 4.5).

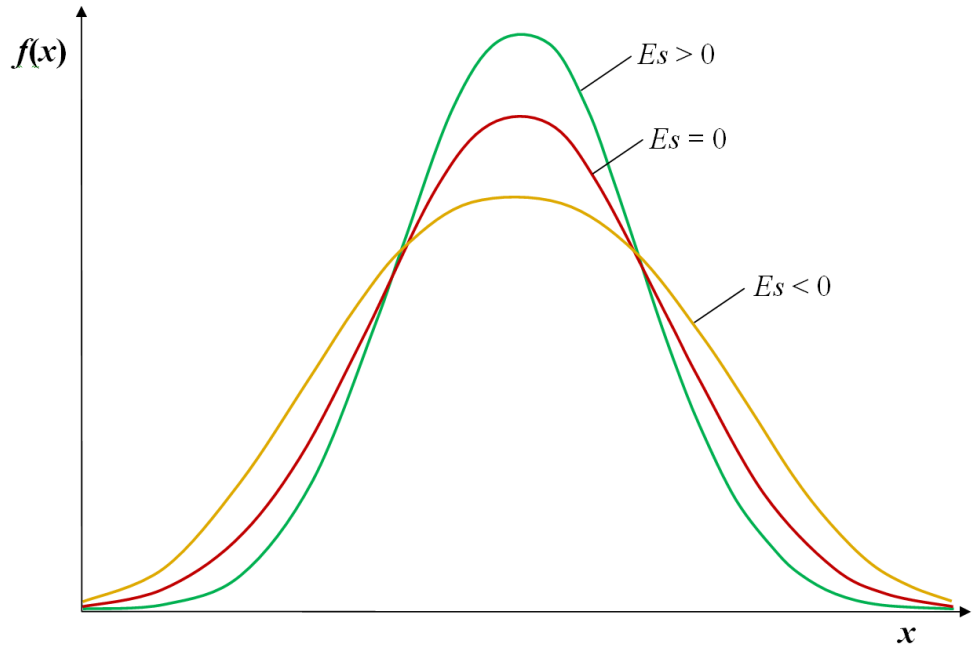


Рис. 4.5. Криві розподілу з різним ступенем крутості (ексцесом)

Показник ексцесу обчислюється за формулою:

для *незгрупованих даних* $$Es = \frac{1}{\sigma^4}\frac{\sum(x_i - \bar{x})^4}{n} - 3; \quad (4.6)$$

для *згрупованих даних* $$Es = \frac{1}{\sigma^4}\frac{\sum(x_i - \bar{x})^4 \cdot f_i}{\sum f_i} - 3. \quad (4.7)$$

Стандартна помилка для *Es* визначається як:

$$S_{Es} = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+5)}}. \quad (4.8)$$

Показник ексцесу Es характеризує крутість кривої розподілу – її загостреність або пологість у порівнянні з нормальною кривою (рис. 4.5).

Нормальний розподіл має нульовий ексцес, $Es = 0$. Крім того, якщо відношення показника ексцесу до величини своєї помилки менше трьох ($|Es|/S_{Es} < 3$), ексцес вважається незначним, і його величиною можна знехтувати. Вважається також, що розподіл з ексцесом в діапазоні від –1 до +1 приблизно відповідає нормальному вигляду.

Якщо крива розподілу характеризується високовершінністю, то ексцес вважається *додатним* ($Es > 0$). В цьому разі хвости розподілу «легше», а пік гостріше, ніж у нормального розподілу (вершина кривої розподілу лежить вище вершини нормальної кривої). Це говорить про скупчення значень ознаки в центральній частині ряду розподілу, тобто про переважну появу в даних значень, близьких до середньої величини.

Крива розподілу, для якої характерною є виражена плосковершинність, свідчить про від'ємний ексцес ($Es < 0$). В цьому разі хвости розподілу «важче», а пік більш «приплющений», ніж у нормального розподілу (вершина кривої розподілу лежить нижче вершини нормальної кривої). Це означає, що значення ознаки не концентруються в центральній частині ряду, а розсіяні по всьому діапазоні від x_{min} до x_{max}.

Показники асиметрії і ексцесу мають велике значення для аналізу статистичної сукупності, оскільки вони не тільки відображають форму, а й дозволяють визначити однорідність досліджуваних соціально-економічних явищ і процесів.

Коефіцієнти асиметрії та ексцесу використовуються при $n \geq 50$ для грубої, попередньої перевірки гіпотези про близькість досліджуваного розподілу частот (ймовірностей) до нормального розподілу. Вони дозволяють відкинути, але не дозволяють прийняти гіпотезу нормальності. Для нормального розподілу *As* та *Es* є близькими до нуля. Крім цього, гіпотеза про нормальність закону розподілу величини *x* висувається, якщо $|As|/S_{As} < 3$ і $|Es|/S_{Es} < 3$. Більш обґрунтовані висновки про відповідність закону розподілу можна зробити, використовуючи критерії згоди (див. Розділ 6).

4.3. Логарифмічно нормальний розподіл

Логарифмічно нормальний розподіл – це розподіл додатної змінної, логарифм якої має гауссовий розподіл. Щільність ймовірності $f(x)$ та інтегральна щільність $F(x)$ мають такий вигляд:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x} \exp\left[-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right], \qquad (4.9)$$

$$\mu \in (-\infty, +\infty), \quad \sigma > 0, \quad x \in (0, +\infty)$$

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^x \frac{1}{t} \exp\left[-\frac{1}{2}\left(\frac{\ln t - \mu}{\sigma}\right)^2\right] dt = \frac{1}{2}\left[1 + erf\left(\frac{ln\, x - \mu}{\sigma\sqrt{2}}\right)\right], \quad (4.10)$$

де $erf(x)$ – функція помилок.

У рівняннях (4.9) і (4.10) μ та σ є середнім значенням і стандартним відхиленням не самої величини *x*, а її логарифма ($\ln x$).

Числові характеристики визначають як:

– найбільш ймовірне значення (мода): $\exp(\mu - \sigma^2)$,

– медіанне значення: $\exp(\mu)$,

– середнє значення: $\exp\left(\mu+\frac{\sigma^2}{2}\right)$,

– стандартне відхилення: $\exp\left(\mu+\frac{\sigma^2}{2}\right)\sqrt{\exp(\sigma^2)-1}$,

– коефіцієнт асиметрії: $\left(\exp(\sigma^2)+2\right)\sqrt{\exp(\sigma^2)-1}$,

– коефіцієнт ексцесу: $\exp(4\sigma^2)+2\exp(3\sigma^2)+3\exp(2\sigma^2)-6$.

На відміну від нормального розподілу (або гауссового розподілу), крива логарифмічно нормального розподілу є асиметричною, має правобічну асиметрію (рис. 4.6). Зокрема, середнє значення, медіанне та найбільш ймовірне значення (тобто мода) не збігаються.

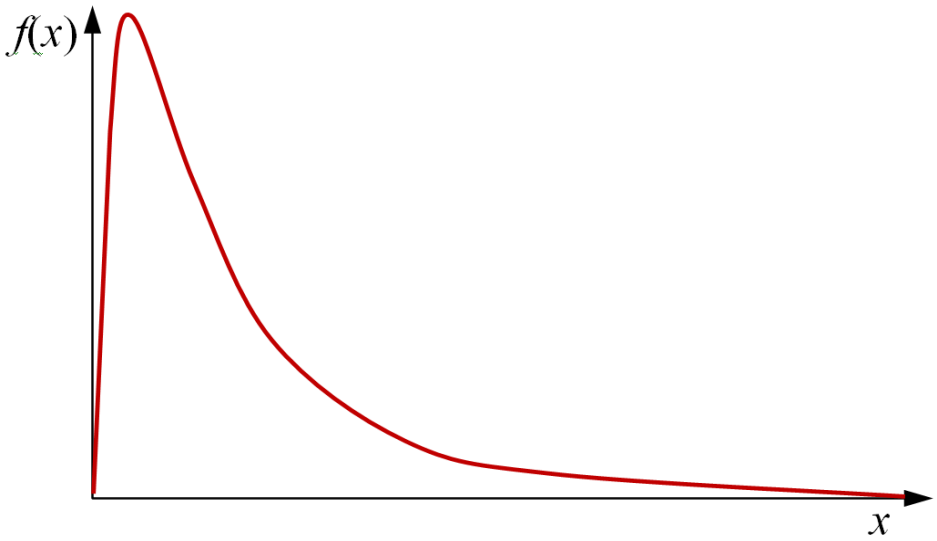


Рис. 4.6. Графік функції щільності логнормального розподілу

Розглянемо нормальний (логнормальний) розподіл на прикладі кількості ночей, проведених туристами за місяць в об’єктах їх розміщення, тобто дані, що характеризують сезонну активність туристів.

***Приклад* 4.1.** На рис. 4.7 наведені емпіричні і розрахункові частоти щомісячного розподілу загальної кількості ночей, проведених громадянами країн-членів Європейського Союзу в об’єктах їх

розміщення за період 2008-2019 рр. (Дані запозичені з сайту: https://ec.europa.eu/eurostat і оброблені авторами навчального посібника). Буквою *N* позначено кількість ночей за місяць.

З рис. 4.7 добре видно асиметрію розподілу: правий «хвіст» кривої розподілу значно довше і ширше за лівий. Більшість місяців характеризуються середньою кількістю ночей, однак, є місяці (наприклад, липень, серпень), які характеризуються значно більшою кількістю ночей порівняно з іншими місяцями (правий «хвіст» гістограми). Загальний характер розподілу частот підтверджує обґрунтованість його відхилення від нормального закону. Про це свідчать також величини коефіцієнтів асиметрії та ексцесу As = 0,66 (S_{As} = 0,20), Es = -0,38 (S_{Es} = 0,40), а також відношення $|As|/S_{As}$ = 3,27. Відношення величини ексцесу до своєї помилки, на відміну від коефіцієнта асиметрії, менше трьох: $|Es|/S_{Es}$ = 0,95.

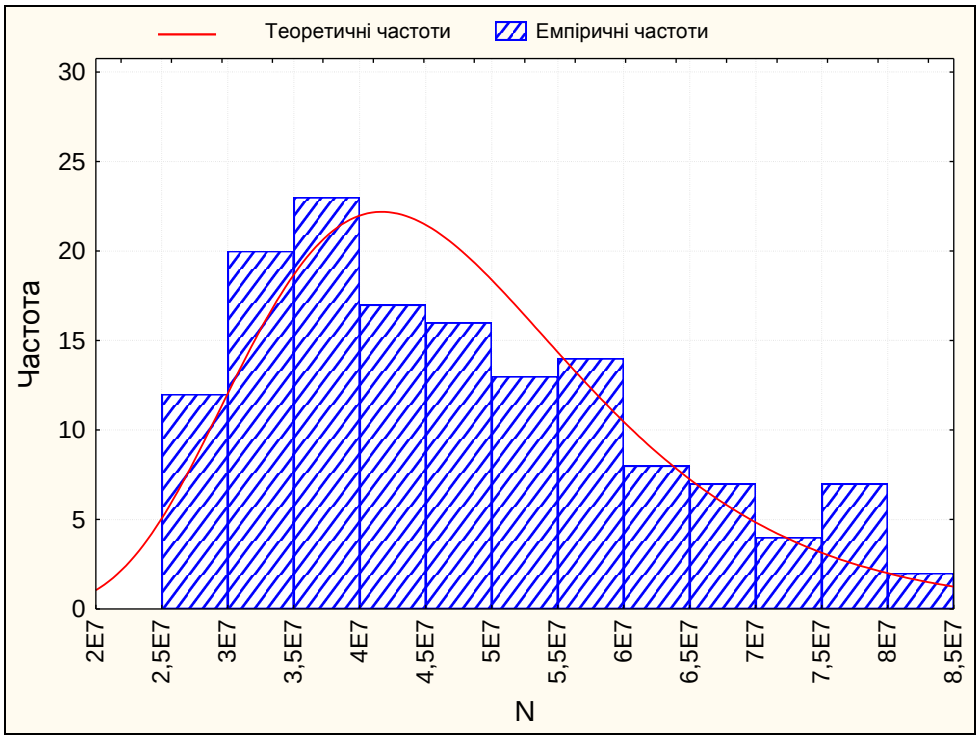


Рис. 4.7. Гістограма розподілу загальної кількості ночей, проведених туристами за місяць в об'єктах їх розміщення за період 2008-2019 рр.

Водночас емпіричні частоти логарифмів загальної кількості ночей ln(*N*), (рис. 4.8) за місяць добре апроксимуються кривою

щільності нормального розподілу з коефіцієнтами асиметрії і ексцесу, відповідно: As = 0,17 ($|As|/S_{As}$ = 0,86 < 3), Es = -0,89 ($|Es|/S_{Es}$ = 2,23 < 3).

Таким чином, загальний характер емпіричного розподілу частот підтверджує обґрунтованість вибору моделі логарифмічно нормального розподілу, наведеного на рис. 4.8.

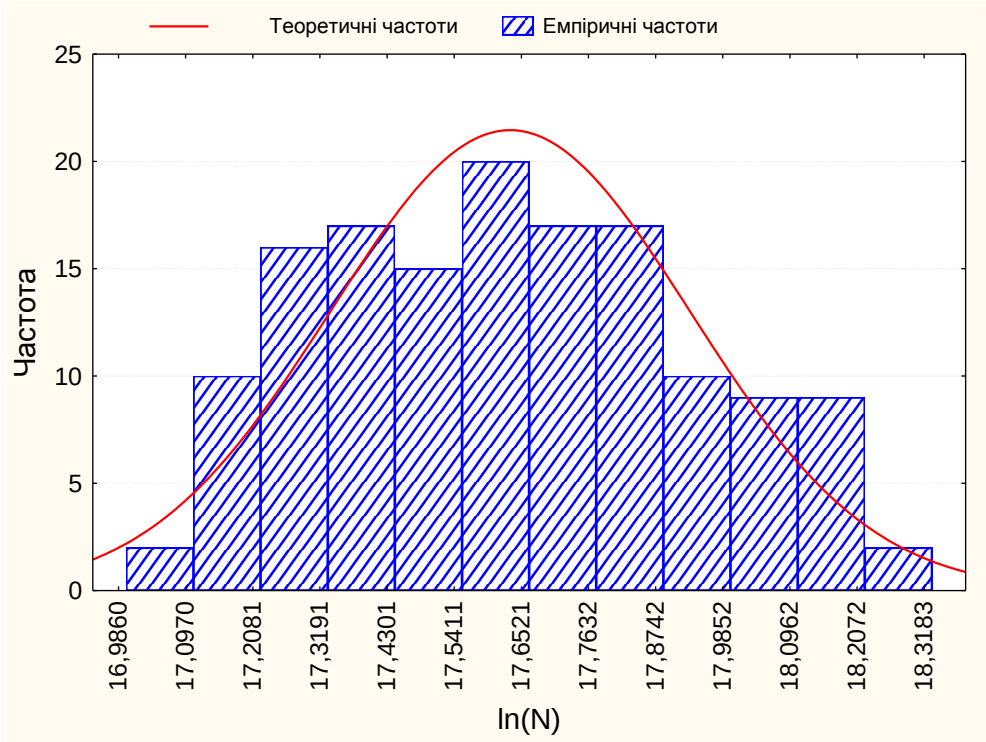


Рис. 4.8. Гістограма розподілу ln(N)частотного

Для перевірки згоди розподілів можна також використовувати Q-Q діаграму – діаграму квантиль-квантиль або нормальних імовірнісних графіків (рис. 4.9). Діаграма показує, наскільки результати вимірювань досліджуваної величини відхиляються від теоретичних значень, отриманих з нормального розподілу з тими ж самими параметрами, що і для емпіричних даних. У разі ідеального збігу діаграма являє собою пряму лінію. Інакше кажучи, якщо квантилі нормального (теоретичного) розподілу і емпіричні квантилі пропорційні між собою, і вибіркові точки вишикуються на «теоретичній» прямій, то апроксимацію можна вважати вдалою.

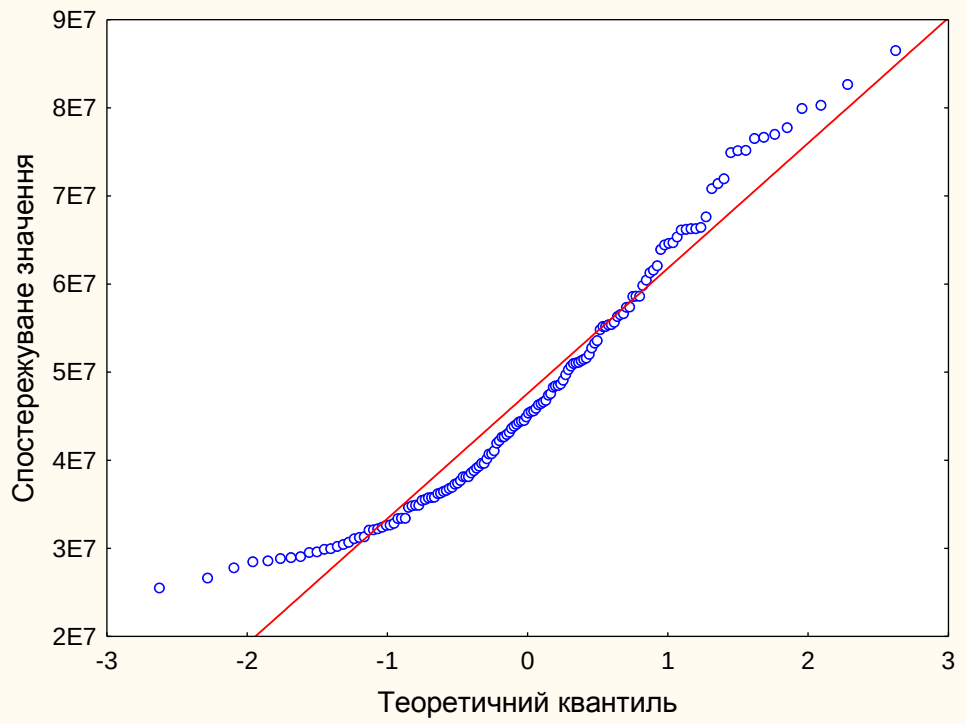


Рис. 4.9. Графік квантиль-квантильного нормального розподілу щомісячної кількості ночей, проведених туристами в об'єктах їх розміщення

Для розглянутого нами прикладу ідеальний збіг не спостерігається (рис. 4.9), і графік свідчить, швидше, про неприйнятність гіпотези щодо нормального розподілу даних, що характеризують щомісячну кількість ночей, проведених туристами в об'єктах розміщення. Проте, для логнормального розподілу дані краще апроксимуються прямою лінією (рис. 4.10).

Для візуального аналізу та попередньої оцінки розподілу даних за нормальним (логнормальним) законом можна також використовувати діаграми розмаху («ящик з вусами», Box and Whisker Plot або Box Plot), про які вже йшлося в Розділі 3.

Рис. 4.11*а* демонструє правобічну асиметрію, що означає відхилення розподілу від нормального закону розподілу. Водночас, на рис. 4.11*б* графік розмаху, побудований на основі логарифмів кількості ночей, є відносно симетричним

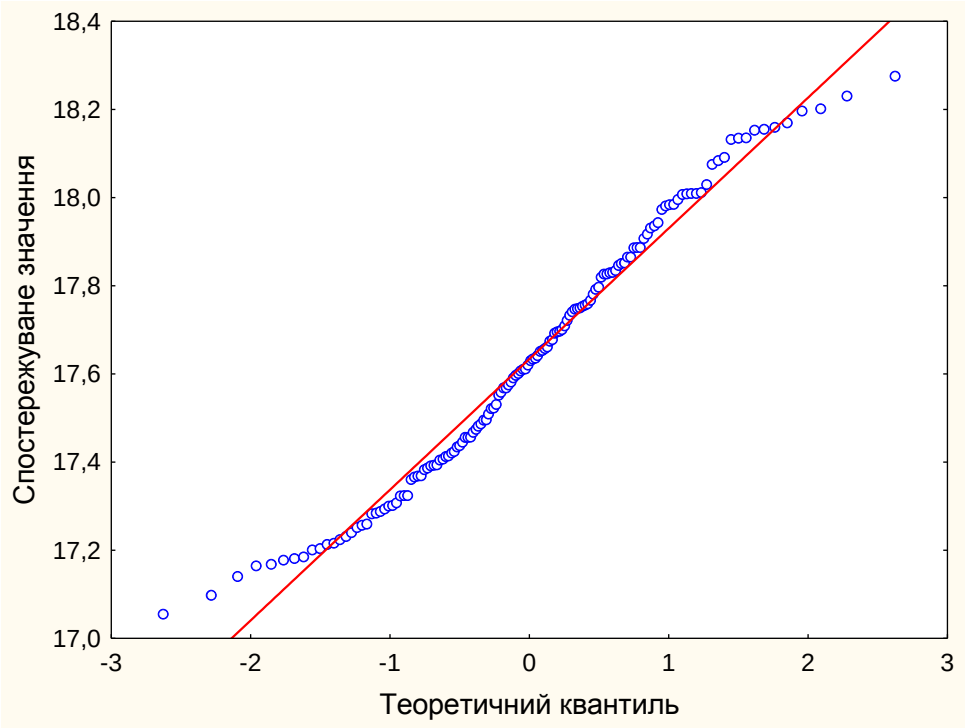
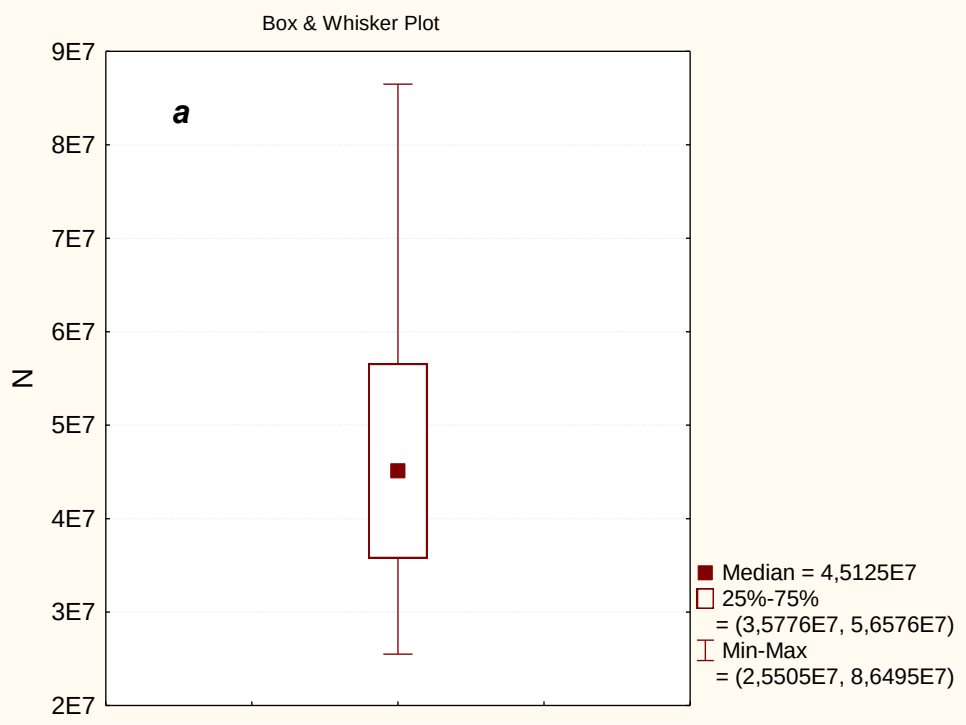


Рис. 4.10. Графік квантиль-квантильного логнормального розподілу щомісячної кількості ночей, проведених туристами в об'єктах їх розміщення



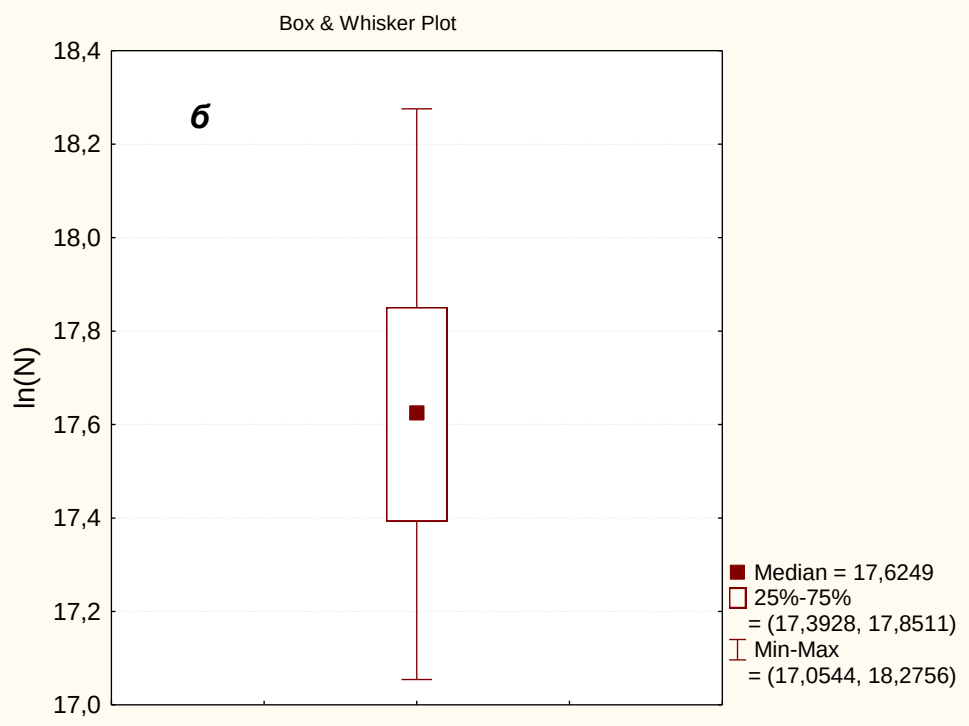


Рис. 4.11. Розподіл щомісячної кількості ночей

Крім цього, з діаграми розмаху (4.11*а*) можна отримати певну інформацію про сезонність туристичного попиту. Наприклад, наявність довгого верхнього «вуса» говорить про те, що кількість місяців з великою кількістю ночей, проведених туристами в об'єктах їх розміщення, які витягали статистику «середньої кількості ночей в місяць» вгору, не є великою (це червень, липень і серпень). Основна маса ночей (три квартилі знизу до $5{,}7\times10^7$) – це 75 % з 12 місяців (тобто 9 місяців, без місяців: червень, липень та серпень).

Таким чином, незважаючи на те, що аналізований розподіл відхилений від нормального закону розподілу, можна спростувати гіпотезу про те, що досліджувані дані розподілені за логарифмічно нормальним законом.

Після попереднього вибору закону розподілу рекомендується застосовувати строгі критерії згоди.

Перевірку розподілу досліджуваної ознаки в сукупності за нормальним законом можна також здійснити за допомогою інших критеріїв, які розглядаються в Розділі 6.

Питання для самоконтролю

4.1. Які розрізняють типи кривих розподілу?

4.2. Чим відрізняються емпіричні та теоретичні розподіли ймовірностей випадкових величин (або частот вибірки)?

4.3. В чому полягає сутність закону нормального розподілу ймовірностей (або частот вибірки)?

4.4. Наведіть характеристики нормального розподілу.

4.5. Як задається нормальний закон розподілу?

4.6. Як називається графік щільності нормального розподілу?

4.7. Як змінюється крива нормального розподілу при зміні її параметрів?

4.8. Які показники використовують для оцінки відхилення емпіричного розподілу від нормального закону розподілу ймовірностей (частот)?

4.9. Що характеризують асиметрія та ексцес розподілу?

4.10. Які значення приймають показники асиметрії та ексцесу для нормального закону розподілу?

4.11. Про що свідчать нерівності: $As|/S_{As} < 3$); $|Es|/S_{Es} < 3$.

4.12. Як використовується правило «трьох сигм» для характеристики розподілу?

4.13. Що таке довірчий інтервал, довірча ймовірність у разі нормального розподілу?

4.14. Що таке рівень значущості?

4.15. Схарактеризуйте логарифмічно нормальний закон розподілу ймовірностей випадкових величин (або частот вибірки). Наведіть приклади логнормального розподілу, які стосуються туризму.

4.16. Чим відрізняється логнормальний розподіл від нормального закону розподілу?

4.17. Які існують графічні методи для візуальної оцінки нормального розподілу?

Завдання для самостійного виконання

Завдання 4.1. Є дані про розподіл загальної кількості ночей, проведених громадянами країн-членів Європейського Союзу в об'єктах їх розміщення за 2018 рік (табл. 1). Дані запозичені з сайту https://ec.europa.eu/eurostat і оброблені авторами навчального посібника.

Побудуйте криві емпіричних і теоретичних частот. Обчисліть відношення величин $|As|/S_{As}$, $|Es|/S_{Es}$ і на їх основі попередньо перевірте гіпотезу про близькість розподілу частот до нормального (логарифмічно нормального) закону.

Таблиця 1

Кількість ночей, проведених громадянами країн-членів Європейського Союзу в об'єктах їх розміщення за 2018 рік

Місяць	Кількість ночей за місяць, N
1	422 184
2	472 082
3	1 568 071
4	4 093 993
5	7 756 692
6	8 682 330
7	15 138 857
8	16 535 354
9	6 087 415
10	2 342 959
11	714 090
12	841 694

Завдання 4.2. Для перевірки згоди розподілів за даними завдання 4.1 (табл. 1) використайте діаграму розмаху і Q-Q діаграму.

РОЗДІЛ 5
КОРЕЛЯЦІЙНИЙ І РЕГРЕСІЙНИЙ АНАЛІЗ

При вивченні соціально-економічних процесів і явищ, в тому числі туризму, часто виникає завдання встановлення зв'язку між досліджуваними змінними. Наприклад, туроператора може цікавити, як збільшується об'єм продажів туристичних пакетів залежно від якості реклами, тривалості туру, курсу національної валюти до долара або євро і т. д.

Статистика має у своєму розпорядженні спеціальні методи, які дозволяють відповісти на ці питання. Це методи *кореляційного* і *регресійного аналізу*.

Кореляційний аналіз встановлює тісноту і напрямок зв'язку між досліджуваними змінними, а *регресійний аналіз* – емпіричні залежності, що зв'язують ці змінні. Практично, обидва види досліджень застосовують для того, щоб передбачити поведінку одних величин на підставі зміни інших.

5.1. Кореляційний аналіз

При визначенні зв'язку між змінними, перш за все, встановлюють, яка змінна є можливою причиною, а яка – наслідком зв'язку. У наведених вище прикладах причина – це якість реклами, тривалість туру, курс національної валюти до долара або євро. Наслідок – об'єм продажів туристичних пакетів. Проте, зв'язок не у всіх випадках може бути причиною.

Для того, щоб абстрагуватися від конкретних об'єктів і формалізувати майбутні дії, в статистиці і, взагалі, в математиці при визначенні взаємозв'язку між змінними вводять поняття незалежних і залежних змінних. Незалежні – це причини або передбачувані причини, залежні – наслідки причинного або передбачуваного зв'язку.

З урахуванням сказаного вище ознаки за їхнім значенням ділять

на два класи. Ознаки, що обумовлюють зміну інших (пов'язаних з ними ознак), називають *факторними* або просто *факторами*. Ознаки, що змінюються під дією факторних ознак, називають *результативними*.

В математиці зв'язок між незалежною і залежною змінними визначається функцією. Кажуть, що якщо кожному елементу x множини $X = \{x\}$ відповідає один і тільки один елемент y множини $Y = \{y\}$, то на множині X визначена функція $y = f(x)$. По суті, функція – це правило (або закон), згідно з яким кожен x відображається в y.

Таким чином, залежність величини y від x називається *функціональною,* якщо кожному значенню величини x, яка називається аргументом (факторною ознакою), відповідає єдине значення y – функції (результативної ознаки).

В статистиці, де вивчаються випадкові події і величини, саме через випадковість вказати таке правило не завжди видається можливим. У зв'язку з цим при визначенні взаємозв'язку між змінними оперують поняттям не функціональної, а *статистичної* (*ймовірнісної*) залежності, тобто залежності не детермінованої, а можливої.

У разі *статистичного* зв'язку кожному значенню однієї величини відповідає певний (умовний) розподіл ймовірності іншої величини. Це пов'язано з тим, що в будь-якій математичній моделі на досліджуваний показник впливають не тільки змінні, що явно входять в модель, а й велика кількість факторів, які існують в дійсності, але не враховуються моделлю, причому частина з цих факторів – це випадкові величини.

Найважливішим окремим випадком *статистичного зв'язку* є ***кореляційний зв'язок,*** коли кожному значенню однієї змінної відповідає певне *умовне середнє значення* (або математичне сподівання) іншої змінної, і при зміні значення однієї величини (факторної ознаки) умовне середнє значення іншої величини (результативної ознаки) змінюється закономірно.

Умовним середнім $\bar{y}_x$ називається середнє арифметичне спостережуваних значень *Y*, які відповідають $X = x$. Наприклад, якщо при $X = 2$ величина *Y* приймає значення $y_1 = 4$, $y_2 = 8$, $y_3 = 6$, то $\bar{y}_x = (4+8+6)/3 = 6$.

Термін *кореляція* (*лат*. correlatio) означає співвідношення, зв'язок, взаємозв'язок.

Кореляційний аналіз слід застосовувати тільки в тому разі, якщо дані спостережень або експерименту можна вважати *випадковими* і вибраними з *нормальної* сукупності.

Зв'язки між явищами та їхніми ознаками класифікуються за напрямком, за рівнем тісноти і за аналітичним вираженням.

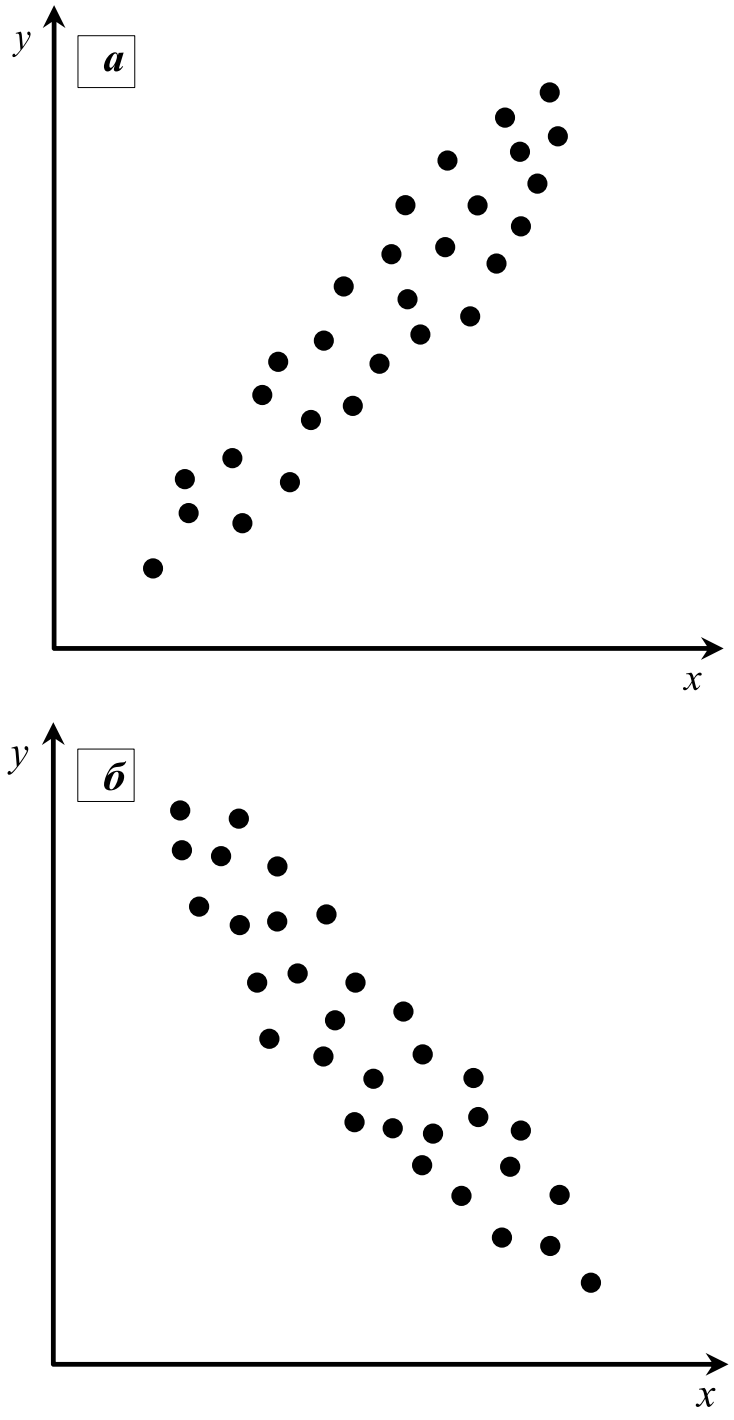
Отже, *завданнями кореляційного аналізу* є:

1. Встановлення напрямку – *позитивного* («прямого») або *негативного* («зворотного») – кореляційного зв'язку. Якщо зі збільшенням величини однієї змінної (наприклад, *x*) зростає в середньому величина іншої (наприклад, *y*), то кореляційний зв'язок є позитивним. Якщо зі збільшенням *x* змінна *y* має в середньому тенденцію до зменшення, то кореляційний зв'язок між цими змінними вважається негативним.
2. Встановлення форми кореляційного зв'язку, тобто визначення виду функції регресії (*лінійна*, *нелінійна*), що характеризує зв'язок між випадковими величинами (ознаками).
3. Оцінка тісноти кореляційного зв'язку за допомогою її кількісної оцінки – *коефіцієнта кореляції* (*r*).
4. Перевірка значущості коефіцієнта кореляції, який виступає мірою зв'язку між випадковими величинами (ознаками).

Для вивчення взаємозв'язку (кореляції) наочним інструментом є *діаграми розсіювання*. На рис. 5.1 наведені різні види цих діаграм у вигляді точок (x_i, y_i) випадкових величин *x* та *y*.

Побудова діаграми розсіювання є першим кроком кореляційного аналізу, оскільки вона наочно показує вид зв'язку між незалежною і залежною змінними. На рис. 5.1*а* зображено позитивний лінійний

зв’язок, оскільки з ростом *x* зростає *y*, причому кожному значенню *x* відповідає цілком певне значення *y*. Тут зростання є приблизно лінійним; очевидною є тенденція до розташування точок певним чином, а саме, смугою, що дає можливість встановити кореляцію *x* та *y*. Рис. 5.1*б* демонструє негативний лінійний зв’язок, оскільки зі збільшенням *x* зменшується *y*. Діаграма на рис. 5.1*в* показує відсутність кореляції між змінними *x* та *y*.



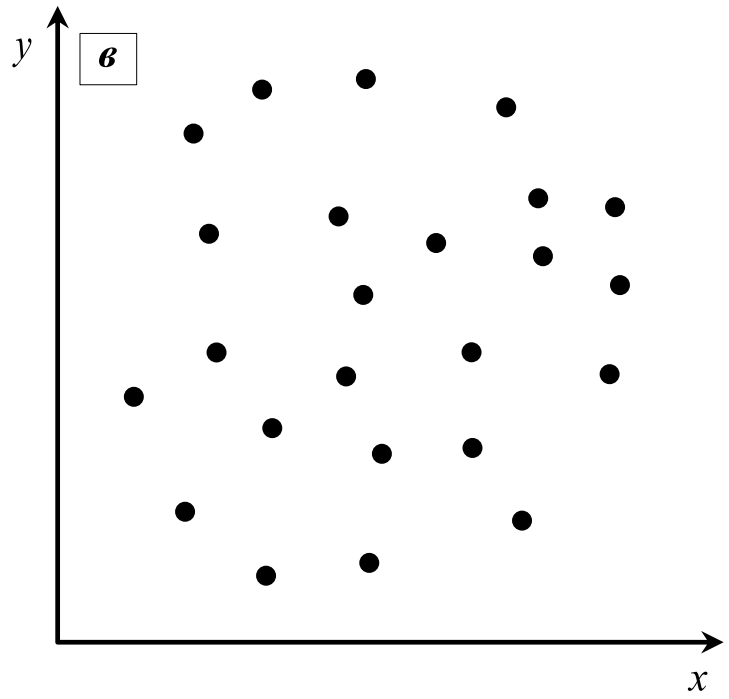


Рис. 5.1. Діаграма розсіювання випадкових величин *x* та *y*:
а) – позитивна кореляція; *б*) – негативна кореляція; *в*) – відсутність кореляції

Графічний спосіб визначення тісноти зв’язку є простим, наочним, але недостатньо точним. Статистична обробка результатів спостереження дозволяє визначити кількісні показники тісноти зв’язку явищ і процесів шляхом обчислення ***коефіцієнта кореляції.***

Якщо зв’язок між ознаками має лінійний характер, то коефіцієнт кореляції називається *коефіцієнтом лінійної кореляції* Пірсона. Ця кореляція найбільш популярна, тому часто, коли говорять про кореляцію, мають на увазі саме кореляцію Пірсона, тобто лінійну кореляцію. Крім цього, в статистичній теорії досить добре розвиненим є апарат оцінки лінійного коефіцієнта кореляції.

В деяких випадках, коли зв’язок є нелінійним або вибірка є неоднорідною, і, отже, для них коефіцієнт кореляції Пірсона є неприйнятним, тоді для перевірки гіпотези про зв’язок двох змінних використовують інші коефіцієнти кореляції. Наприклад, можуть бути застосовані коефіцієнти кореляції Спірмена або Кендалла. Обмежимося розглядом лише лінійних зв'язків.

Для генеральної сукупності коефіцієнт кореляції, як правило,

невідомий, тому він *оцінюється за експериментальними* (спостережуваними) *даними*, що являють собою вибірку об'єму n пар значень (x_i, y_i), отриману при вимірюванні двох ознак. Коефіцієнт кореляції, який визначається за вибірковими даними, називається *вибірковим коефіцієнтом кореляції* і позначається символом r.

Властивості вибіркового коефіцієнта кореляції:

– коефіцієнт кореляції є безрозмірною величиною, і його значення не залежить від одиниць виміру ознак x та y. Коефіцієнт лінійної кореляції Пірсона приймає значення на відрізку від (–1) до (+1), тобто $-1 \le r \le 1$;

– знак r вказує напрямок взаємозв'язку: якщо $r > 0$, то кореляційний зв'язок між змінними є *позитивним*, і r приймає значення на відрізку від 0 до (+1). Якщо $r < 0$, то кореляційний зв'язок між змінними є негативним, і r приймає значення на відрізку від (–1) до 0.

Сила зв'язку між явищами (тобто його тіснота) і спрямованість визначаються величиною коефіцієнта кореляції.

Якщо коефіцієнт лінійної кореляції Пірсона за модулем близький до 1 ($|r| \approx 1$), то це відповідає високому рівню *лінійного зв'язку* (*кореляції*) між змінними; зв'язок є повним, інакше кажучи, функціональним – в цьому разі існує функція $y = f(x)$, яка зв'язує значення y та x. При цьому не обов'язково, щоб зв'язок між y та x був причинно-наслідковим, оскільки обидві ознаки можуть незалежно (але паралельно) змінюватися під дією загальної причини, недоступної для спостереження в даному дослідженні.

Залежно від того, наскільки $|r|$ наближається до 1, розрізняють сильний, помірний і слабкий кореляційний зв'язок.

при $0{,}9 < |r| < 1$ – *зв'язок дуже сильний*;

при $0{,}7 < |r| \le 0{,}9$ – *зв'язок сильний*;

при $0{,}5 < |r| \le 0{,}7$ – *зв'язок значний*;

при $0{,}3 < |r| \le 0{,}5$ – *зв'язок помірний;*

при $0 < |r| \le 0{,}3$ – *зв'язок слабкий*;

при $r = 0$ – *зв'язок відсутній*.

При значенні коефіцієнта кореляції, дуже близькому до 0, лінійний кореляційний зв'язок між двома досліджуваними змінними (ознаками) відсутній, однак може існувати нелінійна форма кореляційного зв'язку. Інакше кажучи, якщо коефіцієнт кореляції дорівнює нулю, то це не свідчить про незалежність відповідних величин. У цьому випадку говорять, що величини є некорельованими.

Важливо зазначити, що величина коефіцієнта лінійної кореляції Пірсона не може перевищувати +1 і бути меншою за –1. Ці два числа (+1 і –1) є межами для коефіцієнта кореляції. Якщо при розрахунку виходить величина, більша за +1 або менша за –1, це означає, що сталася помилка в обчисленнях.

Кореляційні зв'язки можуть бути *однофакторними* і *багатофакторними*.

Коли досліджується зв'язок між однією ознакою-фактором і однією ознакою-наслідком (результативною ознакою), тоді кореляція називається *однофакторною* (*парною кореляцією*).

Коли досліджується вплив багатьох взаємодіючих між собою ознак (факторів) на результативну ознаку (наслідок), тоді кореляція називається *багатофакторною* (*множинною кореляцією*).

Найчастіше в дослідженнях сфери туризму вивчають парну кореляцію і використовують *парний коефіцієнт кореляції*, що, імовірно, пов'язано з простотою його розрахунків.

Парний коефіцієнт кореляції визначається за такою формулою:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \qquad (5.1)$$

де x_i та y_i – конкретне значення ознак; $\bar{x}$ та $\bar{y}$ – середні величини ознак, n – об'єм вибірки.

Коефіцієнт кореляції можна виражати також з урахуванням стандартних відхилень. Наприклад, з рівняння (5.1) можна отримати

такий вираз для r_{xy}:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n\,\sigma_x \sigma_y}, \qquad (5.2)$$

де σ_x та σ_y – стандартні відхилення величин *x* та *y*, відповідно:

$$\sigma_x = \sqrt{\frac{\sum x_i^2}{n} - \bar{x}^2}, \qquad \sigma_y = \sqrt{\frac{\sum y_i^2}{n} - \bar{y}^2}. \qquad (5.3)$$

Коефіцієнт кореляції можна виражати також іншими формулами, які легко виводяться одна з одної. Наприклад, з (5.2) можна отримати такі формули для коефіцієнта кореляції:

$$r_{xy} = \frac{\left[\frac{1}{n}\sum(x_i y_i)\right] - \bar{x}\,\bar{y}}{\sigma_x \sigma_y}; \qquad (5.4)$$

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)}}; \qquad (5.5)$$

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right]\left[n\sum y_i^2 - \left(\sum y_i\right)^2\right]}}. \qquad (5.6)$$

Квадрат коефіцієнта лінійної кореляції називається *коефіцієнтом детермінації* $R = r^2$, який вимірює частку варіації *y*, що пояснюється впливом *x*, і навпаки.

За ступенем (силою) зв'язку можна ввести три градації *R*:

$R < 0{,}3$ – *слабкий зв'язок*;

$0{,}3 < R < 0{,}7$ – *помірний зв'язок*;

$R > 0{,}7$ – *сильний зв'язок*.

Наприклад, якщо $r = 0{,}92$, то $R = r^2 = 0{,}85$ або 85 %. Це означає, що 85 % варіації залежної змінної визначається варіацією незалежної змінної. Решта 15 % складають непоясниму або випадкову варіацію. Так, якщо такий коефіцієнт детермінації отриманий при розгляді залежності «*x* – пора року (сезон)» і «*y* – кількість туристів», то 15 %

можуть бути викликані іншими факторами (економічними, політичними, екологічними та ін.), відмінними від пори року. У міру того, як r наближається до нуля, значення R зменшується ще швидше. Наприклад, якщо $r = 0{,}5$, тоді $R = r^2 = 0{,}25$, тобто тільки 25 % варіації залежної змінної можуть бути пов'язані з варіацією незалежної змінної. Якщо коефіцієнт детермінації занадто малий, тоді потрібно шукати інші чинники – змінні, які причинно зумовлюють залежну змінну.

Коефіцієнт кореляції, так само, як і інші параметри вибірки, змінюється від вибірки до вибірки при повторних дослідженнях. Мірою мінливості коефіцієнта кореляції служить помилка коефіцієнта кореляції (σ_r), яка визначається за такою формулою:

$$\sigma_r = \sqrt{\frac{1-r^2}{n-2}}\,. \qquad (5.7)$$

Показники тісноти зв'язку, обчислені за даними порівняно невеликої статистичної сукупності, можуть спотворюватися дією випадкових причин. Це викликає необхідність перевірки їх суттєвості, що дає можливість поширювати висновки за результатами вибірки на генеральну сукупність. У зв'язку з цим, наступним кроком кореляційного аналізу повинна стати оцінка статистичної достовірності кореляції. Це завдання розглянуто у Розділі 6 «Статистична перевірка гіпотез».

Як вже зазначалося, кореляційний аналіз має своїм завданням кількісне визначення тісноти зв'язку між двома ознаками (при *парній кореляції*) і між результативною і кількома факторними ознаками (при *багатофакторній кореляції*).

Розглянуті вище коефіцієнти кореляції характеризують ступінь впливу факторної ознаки x на результативну ознаку y, і тому вони належать до *парної кореляції*.

Крім парних коефіцієнтів кореляції, кількісною мірою інтенсивності зв'язку є також *часткові коефіцієнти кореляції* і сукупний *коефіцієнт множинної кореляції*.

Часткова кореляція характеризує залежність між результативною і однією факторною ознаками при фіксованому значенні (при виключенні дії інших факторних ознак).

Множинна кореляція характеризує залежність результативної і двох (або більше) факторних ознак, охоплених дослідженням.

Ми обмежилися розглядом лише парної кореляції.

5.2. Регресійний аналіз

Регресійний аналіз тісно пов'язаний з методами кореляційного аналізу. На відміну від кореляційного аналізу, який вивчає напрямок і силу кореляційного зв'язку ознак, *регресійний аналіз* вивчає залежність кількісної або якісної ознаки від однієї або декількох кількісних ознак. У цьому разі стає більш ясним вплив (вага) окремих факторів на результативну ознаку, краще розуміється природа досліджуваного явища.

Зміна функції (результативної ознаки) залежно від змін одного або кількох аргументів (факторних ознак) називається *регресією.*

Завданнями регресійного аналізу є:

1. Встановлення форми залежності між змінними (ознаками), тобто визначення форми рівняння регресії (лінійна, нелінійна). Це завдання вирішується шляхом аналізу досліджуваного взаємозв'язку.

2. Оцінка рівняння регресії, оцінка невідомих значень (прогноз) залежної змінної.

Отже, регресійний аналіз полягає у визначенні аналітичного вираження зв'язку, в якому зміна однієї величини (яка називається залежною або результативною ознакою) обумовлена впливом однієї або декількох незалежних величин (факторів).

Для вираження регресії використовують кореляційні рівняння, або рівняння регресії, емпірично і теоретично обчислені ряди регресії, їхні графіки, які називаються лініями регресії, а також коефіцієнти регресії (лінійної і нелінійної).

Показники регресії виражають кореляційний зв'язок

двосторонньо, враховуючи зміну усереднених значень $\bar{y}$ змінної y при зміні значень x_i змінної x, і, навпаки, показують зміну середніх значень $\bar{x}$ змінної x за зміненими значеннями y_i змінної y. Виняток становлять часові ряди, що показують зміну змінних в часі.

Існує багато різних форм і видів кореляційних зв'язків. Завдання регресійного аналізу зводиться до того, щоб в кожному конкретному випадку виявити форму (*лінійну* або *нелінійну*) зв'язку і виразити її відповідним кореляційним рівнянням (тобто *рівнянням регресії*), яке дозволяє передбачати можливі зміни змінної y на підставі відомих змін змінної x, пов'язаної з y кореляційно.

Рівняння регресії – це *рівняння лінії регресії*, навколо якої групуються точки кореляційного поля. Рівняння вказує основний напрямок і тенденцію зв'язку.

Як вже було зазначено, залежно від кількості факторів (незалежних ознак), включених у рівняння регресії, прийнято розрізняти *однофакторну* (*парну*) і *багатофакторну* (*множинну*) регресії.

Найчастіше використовується *лінійне рівняння парної регресії*, коли результуюча ознака y залежить від однієї факторної x.

У загальному вигляді рівняння регресії можна представити так:

$$\bar{y} = f(x). \qquad (5.8)$$

У разі, якщо $f(x)$ – лінійна функція, тобто якщо між величинами x та y встановлена лінійна кореляційна залежність, *регресію* називають *лінійною*. Тоді становить інтерес встановлення цієї залежності у вигляді рівняння прямої лінії $y = ax + b$ (де a та b – коефіцієнти). Таке рівняння називається *рівнянням регресії*.

Графічно взаємозв'язок двох ознак зображується за допомогою поля кореляції. В системі координат на осі абсцис відкладаються значення факторної ознаки, а на осі ординат – результативної. При відсутності тісних зв'язків має місце безладне розташування точок на графіку. Чим сильніше зв'язок між ознаками, тим тісніше будуть групуватися точки навколо певної лінії, яка виражає форму зв'язку.

На рис. 5.2 наведені парні значення величин x та y, які відповідають одна одній: (x_1, y_1), (x_2, y_2),…, (x_n, y_n). Якщо провести пряму AB, яка «найкращим» чином вирівнює систему середніх значень $\bar{y}$, то буде отримана функціональна залежність:

$$\bar{y} = a\,x + b, \qquad (5.9)$$

яка є рівнянням прямої AB і наближено відображає зв'язок між x та $\bar{y}$.

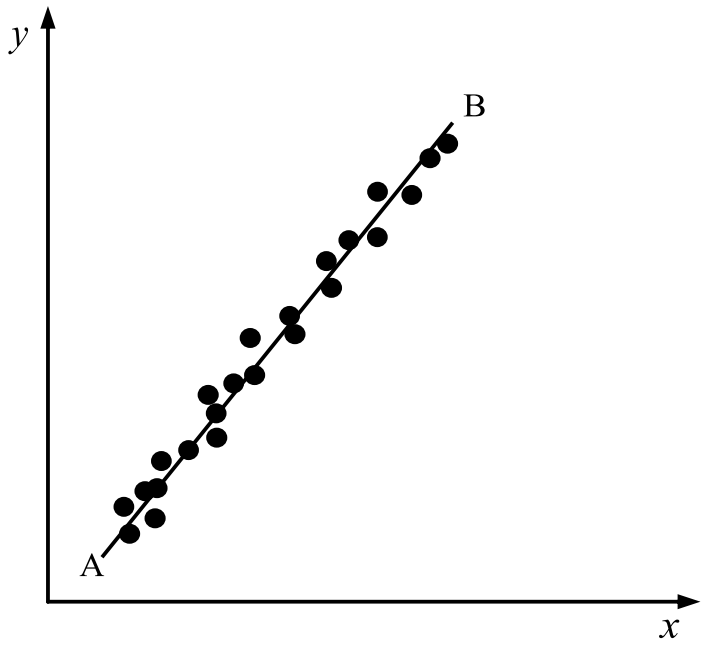


Рис. 5.2. Графік кореляційного поля

Рівняння (5.9) називається *рівнянням парного лінійного кореляційного зв'язку* (*рівнянням парної лінійної регресії*). Лінія AB називається *лінією регресії y* по *x*.

Для того, щоб пряма AB «найкращим» чином вирівнювала середні значення, її необхідно провести так, щоб сума квадратів відстаней від неї (виміряних паралельно осі y) всіх точок була найменшою, тобто менше, ніж від будь-якої іншої прямої. Інакше кажучи, для оцінки параметрів лінійної регресії застосовують метод найменших квадратів (МНК). Це класичний підхід, що дозволяє отримувати такі оцінки параметрів a і b, при яких сума квадратів відхилень фактичних значень результативної ознаки y від

теоретичних $\hat{y}$ (розрахункових) стає мінімальною: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \to \min$. Всі значення y, отримані з проведеної таким способом лінії регресії, мають найбільшу кореляцію із значеннями, які дійсно спостерігалися.

У рівнянні (5.9) параметр a – *коефіцієнт регресії*, який показує, наскільки в середньому величина однієї ознаки змінюється при зміні на одиницю міри іншої ознаки, кореляційно пов'язаної з першою; параметр b – вільний член, який показує усереднений вплив на результативну ознаку неврахованих (не виділених для дослідження) чинників.

Рівняння лінійної залежності (5.9) можна представити за допомогою статистичних характеристик:

$$\hat{y} = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) = a\,x + b, \qquad (5.10)$$

де стандартні відхилення σ_x та σ_y визначаються за формулою (5.3).

У рівнянні (5.10) кутовий коефіцієнт $a = r\frac{\sigma_y}{\sigma_x}$ визначає тангенс кута нахилу лінії регресії на діаграмі в координатах X – Y. Він називається *коефіцієнтом парної регресії y* по *x*.

Знак при коефіцієнті регресії a свідчить про напрямок залежності y від x: при $a \geq 1$ залежність є прямою; при $a \leq 0$ залежність є зворотною.

Параметри a і b рівняння лінійної регресії (5.10) можна визначити за допомогою методу найменших квадратів, однак існують і інші методи знаходження оцінок лінійної регресії, наприклад, метод максимальної правдоподібності. Критеріями кращого способу оцінювання є вимоги змістовності, незміщеності і ефективності оцінок, знайдених таким способом. Оцінки, отримані за методом найменших квадратів, задовольняють всім цим вимогам, тобто є «найкращими».

Система нормальних рівнянь для знаходження параметрів лінійної парної регресії методом найменших квадратів має такий вигляд:

$$\begin{cases} nb + a\sum x = \sum y; \\ b\sum x + a\sum x^2 = \sum x \cdot y, \end{cases} \tag{5.11}$$

де n – об'єм досліджуваної сукупності.

Параметри рівняння парної лінійної регресії (5.10) визначаються як:

$$a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}, \qquad b = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum y_i x_i}{n\sum x_i^2 - (\sum x_i)^2}, \tag{5.12}$$

або:

$$a = r\frac{\sigma_y}{\sigma_x} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}. \quad b = \frac{1}{n}\sum y_i - r\frac{\sigma_y}{\sigma_x} \cdot \frac{1}{n}\sum x_i = \bar{y} - a\,\bar{x}. \tag{5.13}$$

Таким чином, *рівняння парної лінійної регресії* показує зміну середнього значення результативної ознаки при зміні факторної ознаки на одну одиницю її виміру. Визначивши значення параметрів a і b та підставивши їх у рівняння зв'язку (5.10), можна визначити теоретичні значення y, що залежать тільки від заданого значення x.

Для оцінки точності *регресії* використовується *коефіцієнт детермінації*, що являє собою квадрат лінійного коефіцієнта кореляції r_{xy}^2. Коефіцієнт детермінації характеризує частку дисперсії результативної ознаки y під впливом варіації ознаки-фактора і визначається за формулою:

$$r_{xy}^2 = 1 - \frac{\sigma_{зал.}^2}{\sigma_{заг.}^2}. \tag{5.14}$$

Величина $\sigma_{заг.}^2$ – загальна дисперсія, яка характеризує розкид спостережуваних величин y_i відносно середнього значення $\bar{y}$; $\sigma_{зал.}^2$ – залишкова дисперсія, яка характеризує відхилення спостережуваних величин y_i від теоретичних (розрахункових) значень $\hat{y}_i$, отриманих за регресійною залежністю, тобто за рівнянням регресії:

$$\sigma^2_{заг.} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \overline{y^2} - \bar{y}^2, \qquad \sigma^2_{зал.} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \qquad (5.15)$$

Величина $1 - r^2_{xy}$ дає характеристику частки дисперсії результативної ознаки *y*, викликаної впливом інших, не врахованих факторів.

Поряд з перевіркою значущості окремих параметрів здійснюється перевірка значущості рівняння регресії. У зв'язку з цим, після побудови рівняння регресії потрібно перевірити адекватність цього рівняння. Це завдання розглядається в Розділі 6 «Статистична перевірка гіпотез».

Проведемо однофакторний кореляційний і регресійний аналіз на основі даних з галузі туризму.

***Приклад* 5.1.** Є дані про кількість українських туристів, які виїжджали до Німеччини за період 2006-2017 рр. (табл. 5.1) Дані запозичені з сайту Державної служби статистики України. Необхідно перевірити, чи існує взаємозв'язок між кількістю туристів, які виїжджали до Німеччини, і курсом гривні до євро. Інакше кажучи, необхідно виявити, як змінюється кількість туристів при зміні курсу гривні відносно євро. У стовпчиках 2-5 (табл. 5.1) числа представлені в експоненційному форматі. Наприклад, запис 2,07E+09 в загальному вигляді представляється так: $2{,}07 \cdot 10^9$.

Для розв'язання цієї задачі необхідно обчислити коефіцієнт кореляції, що характеризує тісноту зв'язку між цими ознаками.

Таблиця 5.1

Кількість українських туристів, які виїжджали до Німеччини за період 2000-2017 рр.

Роки	Курс гривні до євро, x_i	Кількість туристів y_i	x_i^2	y_i^2	$x_i \cdot y_i$
1	2	3	4	5	6
2006	6,34	45 525	40,25	2,07E+09	288628,5
2007	6,92	59 216	47,91	3,51E+09	409774,7
2008	7,72	57 295	59,63	3,28E+09	442317,4

1	2	3	4	5	6
2009	10,89	48 207	118,5	2,32E+09	524974,2
2010	10,52	37 418	110,73	1,40E+09	393637,4
2011	11,06	45 755	122,22	2,09E+09	506050,3
2012	10,27	15 063	105,5	2,27E+08	154697,0
2013	10,61	5 366	112,64	2,88E+07	56933,3
2014	15,61	5 842	243,69	3,41E+07	91193,6
2015	24,22	10 610	586,77	1,13E+08	256974,2
2016	28,27	8 999	799,27	8,10E+07	254401,7
2017	30,00	4 827	900,50	2,33E+07	144810,0
Сума	172,43	344 123	3247,61	1,52E+10	3524392,3

Для наочності мінливість курсу гривні до євро і кількості туристів наведено на рис. 5.3.

Рис. 5.3. Мінливість курсу гривні до євро і кількості українських туристів, які виїжджали до Німеччини за період 2006-2017 рр.

Для визначення r_{xy} обчислимо вибіркові середні:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{172{,}43}{12} = 14{,}37, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{344123}{12} = 28677;$$

$$\frac{1}{n}\sum_{i=1}^{n}(x_i \cdot y_i) = \frac{3524392{,}3}{12} = 293699{,}4.$$

Тепер обчислимо значення вибіркових стандартних відхилень:

$$\sigma_x = \sqrt{\frac{\sum x_i^2}{n} - (\bar{x})^2} = \sqrt{\frac{3247{,}61}{12} - 206{,}50} = 8{,}0;$$

$$\sigma_y = \sqrt{\frac{\sum y_i^2}{n} - (\bar{y})^2} = \sqrt{\frac{1{,}52\text{E}+10}{12} - 8{,}22\text{E}+08} = 21087{,}0.$$

Підставимо отримані значення у формулу для обчислення коефіцієнта кореляції:

$$r_{xy} = \frac{\left[\frac{1}{n}\sum(x_i y_i)\right] - \bar{x}\,\bar{y}}{\sigma_x \sigma_y} = \frac{293699{,}4 - 412088{,}5}{8{,}0 \cdot 21087{,}0} = -\frac{118389}{168696} = -0{,}7.$$

Зв'язок між результативною ознакою *Y* і фактором *X* є значним і негативним. Рівняння регресії:

$$\hat{y} = \bar{y} + r_{xy}\frac{\sigma_y}{\sigma_x}(x - \bar{x}) = 28677 - 0{,}7\frac{21087}{8}(x - 14{,}37) = -1846{,}8x + 55213.$$

Таким чином, кореляційний зв'язок між кількістю українських туристів, які виїжджали до Німеччини, і курсом гривні до євро близький до лінійного негативного. Чим вищим є курс гривні до євро, тим меншою є кількість туристів, і навпаки (рис. 5.4).

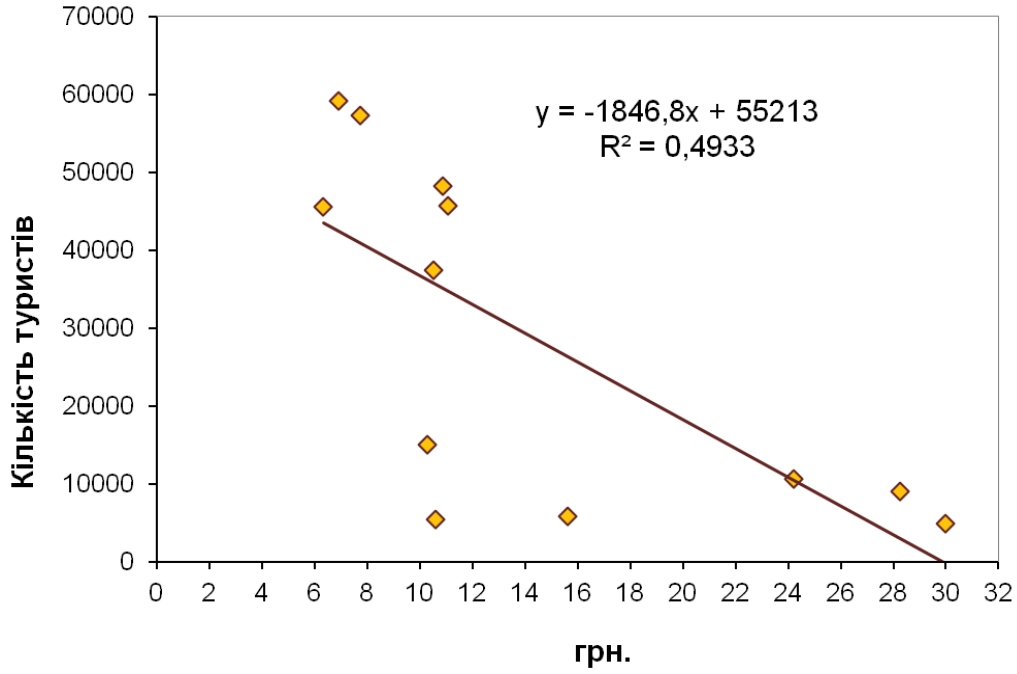


Рис. 5.4. Мінливість кількості українських туристів залежно від курсу гривні до євро

***Приклад* 5.2.** На основі вибіркових даних про ділову активність туроператорів 9-ти регіонів України необхідно оцінити тісноту зв'язку між доходами (млн. грн.) від наданих послуг і матеріальними витратами (млн. грн.) на вироблення туристичного продукту за 2017 р. Для визначення параметрів рівняння регресії побудована розрахункова таблиця 5.2. Дані запозичені з сайту Державної служби статистики України і оброблені авторами навчального посібника.

Таблиця 5.2

Показники роботи туроператорів у 2017 році за регіонами

№	Регіони	Матеріальні витрати, млн. грн. x_i	Доходи від надання туристичних послуг, млн. грн. y_i	x_i^2	$x_i \cdot y_i$	$\bar{y}$
1.	Вінницька	4,218	12,762	17,792	53,830	32,755
2.	Волинська	0,578	16,470	0,334	9,520	17,977
3.	Дніпропетровська	0,310	3,541	0,096	1,098	16,888
4.	Закарпатська	2,497	7,680	6,235	19,177	25,768
5.	Запорізька	0,122	1,777	0,015	0,217	16,125
6.	Івано-Франківська	26,182	267,619	685,497	7006,801	121,929
7.	Київська	0,728	4,273	0,530	3,111	18,586
8.	Львівська	63,996	357,634	4095,488	22887,145	275,454
9.	Одеська	61,205	117,819	3746,052	7211,112	264,122
	Сума	159,836	789,575	8552,039	37192,011	789,604

Обчислення параметрів рівняння лінійної регресії здійснюємо за методом найменших квадратів. Для цього використовуємо систему нормальних рівнянь (5.11), яка для цього прикладу має такий вигляд:

$$\begin{cases} nb + a\sum x = \sum y; \\ b\sum x + a\sum x^2 = \sum xy. \end{cases} \qquad \begin{cases} 9b + 159{,}836a = 789{,}575; \\ 159{,}836b + 8552{,}039a = 37192{,}011. \end{cases}$$

Звідси: $a = 4{,}06$; $b = 15{,}63$, отже, $y = 4{,}06x + 15{,}63$.

Таким чином, при збільшенні матеріальних витрат туроператора

на вироблення туристичного продукту на 1 млн. грн. його сукупний дохід збільшиться в середньому на 4,06 млн. грн.

Питання для самоконтролю

5.1. Який зв’язок називається кореляційним?
5.2. Які ознаки називаються факторними? результативними?
5.3. У чому полягає відмінність кореляційного зв’язку від функціонального?
5.4. Що являє собою умовне середнє значення?
5.5. До яких даних спостережень (експерименту) можна застосувати кореляційний аналіз?
5.6. Сформулюйте основні завдання кореляційного аналізу.
5.7. Що являють собою діаграми розсіювання?
5.8. Поясніть сенс коефіцієнта кореляції Пірсона і назвіть його властивості.
5.9. Які градації має лінійний коефіцієнт кореляції?
5.10. Що являють собою однофакторна (парна) і багатофакторна (множинна) кореляції?
5.11. Який висновок робить дослідник, якщо вибірковий коефіцієнт кореляції Пірсона дорівнює: $r = 0{,}72$; $r = 0{,}92$; $r = 0{,}20$?
5.12. Який висновок можна зробити про залежність двох випадкових величин, якщо коефіцієнт кореляції $r = 0$; $r = 1$; $r = -1$?
5.13. Які показники характеризують форму і тісноту кореляційного зв’язку?
5.14. Що являє собою коефіцієнт детермінації, і які його градації існують?
5.15. Що являє собою помилка коефіцієнта кореляції?
5.16. Сформулюйте основні завдання регресійного аналізу.
5.17. Яку кореляційну залежність називають лінійною?
5.18. Який вигляд має рівняння парної лінійної регресії?
5.19. Поясніть сенс параметрів рівняння парної лінійної регресії.
5.20. Що показує кутовий коефіцієнт в рівнянні регресії?
5.21. У чому полягає метод найменших квадратів?
5.22. Наведіть приклад лінійного позитивного та негативного кореляційного зв’язку між ознаками зі сфери туризму.

Завдання для самостійного виконання

Завдання 5.1. Заповніть таблицю 1, використовуючи дані сайту Державної служби статистики України про кількість українських туристів, які виїжджали до Польщі за період з 2006 по 2017 рр.

Оцініть тісноту зв'язку між кількістю туристів і курсом гривні до євро.

Знайдіть рівняння регресії.

Побудуйте: 1) графіки мінливості кількості туристів і курсу гривні до євро за вказаний період; 2) лінію регресії, що характеризує кореляційний зв'язок між кількістю туристів і курсом гривні до євро.

Таблиця 1

Кількість українських туристів, які виїжджали до Польщі за період 2006-2017 рр.

Роки	Курс гривні до євро, x_i	Кількість туристів y_i	x_i^2	y_i^2	$x_i \cdot y_i$
2006	6,34				
2007	6,92				
2008	7,72				
2009	10,89				
2010	10,52				
2011	11,06				
2012	10,27				
2013	10,61				
2014	15,61				
2015	24,22				
2016	28,27				
2017	30,00				
Сума					

Завдання 5.2. На основі вибіркових даних про ділову активність туроператорів 7-ми регіонів України, оцініть тісноту зв'язку між доходами (млн. грн.) від наданих послуг і матеріальними витратами

(млн. грн.) на вироблення туристичного продукту за 2017 р. Заповніть таблицю 2 і побудуйте лінійну парну регресію. Обчислення параметрів рівняння регресії виконайте методом найменших квадратів.

Таблиця 2

Показники роботи туроператорів у 2017 році за регіонами

№	Регіони	Матеріальні витрати, млн. грн. x_i	Доходи від надання туристичних послуг, млн. грн. y_i	x_i^2	$x_i \cdot y_i$	$\bar{y}$
1.	Рівненська	0,641	5,100			
2.	Сумська	0,039	1,994			
3.	Тернопільська	0,712	2,150			
4.	Харківська	18,962	21,644			
5.	Херсонська	6,415	7,743			
6.	Черкаська	2,509	3,706			
7.	Чернівецька	0,289	1,545			
	Сума					

РОЗДІЛ 6
СТАТИСТИЧНА ПЕРЕВІРКА ГІПОТЕЗ

На різних стадіях статистичного дослідження іноді виникає необхідність в перевірці деяких *припущень* (*гіпотез*) щодо природи і величини певних параметрів аналізованої генеральної сукупності. Наприклад, можна висунути припущення про рівність параметрів (середніх) ознак двох різних розподілів; про незалежність вибірок; про те, що дана ознака в генеральній сукупності розподілена за нормальним законом, тобто вибірка витягнута з нормальної генеральної сукупності і т. д. Оскільки вибіркові дані становлять лише частину статистичних даних всієї генеральної сукупності, твердження, що висуваються для перевірки, називають гіпотезами.

6.1. Основні поняття і визначення

Будь-яке припущення про параметри і властивості всієї генеральної сукупності, що розглядається, за статистичними даними вибіркової сукупності називається *статистичною гіпотезою*.

Процес встановлення істинності або хибності гіпотези є процесом її емпіричного обґрунтування. Процедура обґрунтованого зіставлення висунутої гіпотези з вибірковими даними називається *статистичною перевіркою гіпотези*. Іншими словами, не маючи відомостей про всю генеральну сукупність, висловлену гіпотезу зіставляють за певними правилами з вибірковими даними і роблять висновок про те, чи є гіпотеза вірною, чи ні. Наприклад, зіставляють деякі статистичні показники, обчислені за даними вибірки, зі значеннями цих самих показників, визначених теоретично. Ця процедура зіставлення і називається перевіркою гіпотези.

Розв’язання кожної задачі статистичної перевірки гіпотез зазвичай починається з формулювання так званої *нульової (основної) гіпотези* H_0. Вона називається так тому, що передбачає, що між порівнюваними параметрами різниця є несуттєвою, дорівнює нулю.

Решта гіпотез, що відрізняються від H_0 і протиставляються їй, називаються *альтернативними (конкуруючими)* і позначаються H_1. Текст гіпотези записується після двокрапки.

Нехай перевіряється гіпотеза про рівність деякого параметра μ значенню μ_0, тобто гіпотеза H_0: $\mu = \mu_0$. В цьому разі альтернативна гіпотеза може полягати, наприклад, в припущенні, що: H_1: $\mu \neq \mu_0$; H_1: $\mu < \mu_0$ або H_1: $\mu > \mu_0$.

Таким чином, поряд з висунутою гіпотезою H_0 розглядають і суперечну їй гіпотезу H_1. Якщо висунута, тобто нульова, гіпотеза відкинута, то має місце суперечна, тобто альтернативна, гіпотеза, яка є несумісною з нульовою гіпотезою.

Розрізняють гіпотези *прості,* що містять тільки одне припущення (наприклад, H_0: $\mu = 5$), і *складні*, які містять більше одного припущення, що вказують область можливих значень параметра (наприклад, H_1: $\mu < 5$; H_1: $\mu > 5$).

При перевірці гіпотези можлива помилка, яка полягає в тому, що буде відкинута правильна нульова гіпотеза, тобто існує ризик прийняти помилкове рішення. Оскільки вибірка – це лише частина генеральної сукупності, то, наприклад, в якийсь невеликій частці випадків нульова гіпотеза H_0 може виявитися відкинутою, тоді як у дійсності в генеральній сукупності вона є справедливою. Таку помилку називають *помилкою першого роду*, тобто *помилка першого роду має місце тоді, коли відкидається вірна гіпотеза* H_0. Імовірність допуститися помилки першого роду ($\alpha = P(H_1/H_0)$) позначається α і називається вона *рівнем значущості*. У статистичних дослідженнях використовують такі його значення: 0,001; 0,005; 0,01; 0,05. Найчастіше рівень значущості приймають рівним $\alpha = 0{,}05$. Якщо висновки, які мають бути зроблені за результатами перевірки гіпотез, пов'язані з великою відповідальністю (наприклад, при дослідженні об'єктів підвищеної небезпеки та відповідальності або процесів, які є загрозою для навколишнього середовища і здоров'я людини), то рекомендується вибирати $\alpha = 0{,}01$ або $\alpha = 0{,}001$.

Якщо, наприклад, взяти $\alpha = 0{,}05$, то це означає, що ми дозволяємо собі помилитися з 5 %-вим ризиком, якщо приймаємо гіпотезу H_0. Інакше кажучи, в п'яти випадках зі ста є ризик допуститися помилки першого роду, а саме – відкинути правильну гіпотезу H_0.

Якщо при перевірці гіпотези H_0 не відкидається, то даний факт не означає, що висловлене в нульовій гіпотезі твердження є єдино вірним. Просто твердження нульової гіпотези не суперечить наявним вибірковим даним.

Якщо в якійсь невеликій частці випадків нульова гіпотеза H_0 приймається, тоді як насправді в генеральній сукупності вона є хибною, тобто справедливою є альтернативна гіпотеза H_1, тоді таку помилку називають *помилкою другого роду*. Інакше кажучи, *помилка другого роду має місце тоді, коли приймається помилкова гіпотеза* H_0. Імовірність помилки другого роду (надійність оцінки) прийнято позначати β, тобто $(\beta = P(H_0/H_1))$. Імовірність $1-\beta$ називають *потужністю критерію*. Чим ближче потужність критерію до одиниці, тим більш ефективним є критерій.

Наприклад, висувається гіпотеза про те, що розвиток туризму в деякому регіоні країни є неприбутковим, а у дійсності він є прибутковим. Тоді відбувається помилка першого роду. А припущення про те, що в цьому ж самому регіоні розвиток туризму є прибутковим, коли насправді він є неприбутковим, призводить до помилки другого роду.

Наслідки помилок першого і другого роду нерівнозначні. Вважається, що одна з помилок (першого роду) веде до більш консервативного або більш обережного рішення, а друга (другого роду), навпаки, веде до ризику, іноді невиправданого.

Імовірності помилок α та β є взаємопов'язаними; спроба знизити одну з них призведе до збільшення другої. Крім цього, чим меншими будуть помилки першого і другого роду, тим точнішим буде статистичний висновок. При заданому об'ємі вибірки одночасно зменшити α і β і, взагалі, повністю виключити помилки неможливо.

Проте виникає питання про зменшення ймовірності появи цих помилок. Одночасне зменшення цих ймовірностей є можливим лише при збільшенні, іноді дуже істотному, об'єму вибірки, що, звичайно, не завжди вдається зробити. При незмінному фіксованому об'ємі вибірки зменшення ймовірності появи помилки одного виду незмінно призводить до збільшення значення ймовірності появи помилки іншого виду. Яка помилка є більш значущою – залежить від постановки завдання і мети дослідження. У загальній схемі статистичної перевірки гіпотез завжди задається ймовірність здійснення помилки першого роду, тобто рівень значимості. При цьому вважається, що дослідник, як правило, висуває досить імовірну гіпотезу, для спростування якої потрібні вагомі аргументи. У зв'язку з цим рівень значущості вибирається досить малим – найчастіше $\alpha = 0{,}05$.

Перевірка гіпотез здійснюється за допомогою *статистичних критеріїв*. Ці величини, що також називаються *критеріями достовірності*, – спеціально розроблені статистичні показники з відомими функціями розподілу, що дозволяють із заданою довірчою ймовірністю (інакше – рівнем довіри) перевірити істинність нульової або альтернативної гіпотез. Функції розподілу цих величин табульовані, тобто зведені в спеціальні таблиці, де містяться значення функції для різних чисел ступенів вільності k (іноді її позначають df – *degrees of freedom*) або об'єму вибірки n і рівнів значущості α. Ці розрахункові функції є також в програмах обробки статистичних даних, в тому числі в табличному процесорі *Microsoft Excel*. За визначенням, *ступінь вільності* – число вільно варіювальних одиниць в складі чисельно обмеженої статистичної сукупності.

Таким чином, статистичний критерій встановлює, при яких значеннях цієї статистики гіпотеза, що перевіряється, не відкидається, а при яких вона відкидається.

Статистичні критерії поділяються на *параметричні* і *непараметричні*.

Критерії, які служать для перевірки гіпотез про параметри

розподілу генеральної сукупності (найчастіше – нормального розподілу), називаються *параметричними*. Застосування параметричних критеріїв потребує обов'язкового знання закону розподілу досліджуваних ознак в сукупності і обчислення їх основних параметрів, таких як середнє або стандартне відхилення. Наприклад, відомо, що вибірки витягнуті з генеральних сукупностей з нормальним законом розподілу і однаковими дисперсіями.

Критерії, які для перевірки гіпотез не потребують знання параметрів розподілу і не використовують припущень про розподіл генеральної сукупності, називаються *непараметричними*. Застосування непараметричних критеріїв не потребує знання закону розподілу досліджуваних ознак в сукупності і обчислення їх основних параметрів.

При нормальному розподілі ознаки параметричні критерії мають більшу потужність, ніж непараметричні. Вони здатні з меншою ймовірністю помилки відкинути нульову гіпотезу, якщо вона не є вірною. У зв'язку з цим у всіх випадках, коли порівнювані вибірки взяті із сукупностей, що розподіляються нормально, слід віддавати перевагу параметричним критеріям.

У разі дуже великих відмінностей розподілів ознаки від нормального вигляду слід застосовувати непараметричні критерії, які в цій ситуації часто виявляються більш потужними. У ситуаціях, коли варіювальні ознаки виражаються не числами, а умовними знаками, застосування непараметричних критеріїв виявляється єдино можливим.

Перевірка гіпотези за своєю сутністю є виявленням попадання певної спостережуваної величини, обчисленої за вибірковими даними, в проміжок, який визначається значеннями конкретної випадкової величини, яка теоретично визначена для даної гіпотези і яка називається критерієм K перевірки основної гіпотези H_0. При цьому, закон розподілу критерію повинен бути завжди відомим.

З визначення критерію ясно, що він є одновимірною випадковою величиною, отже, його значення розташовані на дійсній

прямій. Після вибору певного критерію множина всіх його можливих значень розбивається на дві підмножини: одна з них містить значення критерію, при яких H_0 відкидається і називається *критичною областю*; інша – значення критерію, при яких H_0 не відкидається і називається *областю невідхилення* гіпотези (областю допустимих значень).

Точки, які поділяють ці дві області, називаються *критичними* ($k_{кр.}$) і знаходяться по таблиці розподілу обраного критерію.

Відповідно до обраного критерію, використовуючи вибіркові дані, знаходять спостережуване (вибіркове) значення критерію $K_{спост.}$, яке називають *статистикою критерію*.

Отже, для отримання висновку про вірність або хибність певної гіпотези необхідно перевірити потрапляння статистики критерію в так звану критичну область, яка визначається також за допомогою критерію.

Основне правило або *основний принцип перевірки* будь-якої статистичної гіпотези формулюється так:

- якщо вибіркове (спостережуване) значення критерію ($K_{спост.}$) потрапляє в критичну область, то основна гіпотеза H_0 відкидається;
- якщо вибіркове (спостережуване) значення критерію ($K_{спост.}$) не потрапляє в критичну область, гіпотеза H_0 не відкидається.

Позначимо критичну область Ω. Якщо обчислене за вибіркою значення критерію $K_{спост.}$ потрапляє в критичну область Ω, то гіпотеза H_0 відкидається і приймається гіпотеза H_1. В цьому разі можна припуститися помилки першого роду, ймовірність якої дорівнює α. Інакше кажучи, ймовірність того, що критерій $K_{спост.}$ прийме значення з критичної області Ω, повинна дорівнювати заданому значенню α, тобто $P(K_{спост.} \in \Omega) = \alpha$.

Критична область Ω визначається неоднозначно. Можливі три випадки розташування Ω.

Теоретично доведено, що критична область може бути

однобічною (лівобічною або правобічною) або двобічною. Правобічна і лівобічна області визначаються нерівностями $K > k_{кр.}$ і $K < k_{кр.}$, відповідно, двобічна – двома нерівностями: $K < k^1_{кр.}$ і $K > k^2_{кр.}$. При цьому двобічна область може бути симетричною, якщо $k^2_{кр.} = - k^1_{кр.}$. Тоді вона буде визначатися нерівністю $|K| > k_{кр.}$.

Правобічна критична область (рис. 6.1*а*) складається з інтервалу $(k^{кр}_{пр.\alpha}; +\infty)$, де $k^{кр}_{пр.\alpha}$ визначається з умови $P(K_{спост.} > k^{кр}_{пр.\alpha}) = \alpha$ і називається правобічною точкою, яка відповідає рівню значущості α. Рівень значущості α буде відповідати площі критичної області (наприклад, заштрихована область на рис. 6.1*а*).

Лівобічна критична область (рис. 6.1*б*) складається з інтервалу $(-\infty; k^{кр}_{лів.\alpha})$, де $k^{кр}_{лів.\alpha}$ визначається з умови $P(K_{спост.} < k^{кр}_{лів.\alpha}) = \alpha$ і називається лівобічною точкою, яка відповідає рівню значущості α.

Двобічна критична область (рис. 6.1*в*) складається з таких двох інтервалів: $(-\infty; k^{кр}_{лів.\alpha/2})$ та $(k^{кр}_{пр.\alpha/2}; +\infty)$, де точки $k^{кр}_{лів.\alpha/2}$ і $k^{кр}_{пр.\alpha/2}$ визначаються з умов $P\left(K_{спост.} < k^{кр}_{лів.\alpha/2}\right) = \frac{\alpha}{2}$ та $P\left(K_{спост.} > k^{кр}_{пр.\alpha/2}\right) = \frac{\alpha}{2}$ і називаються двобічними критичними точками.

Отже, якщо обчислений за вибіркою $K_{спост.}$ потрапляє в критичну область, нульова гіпотеза H_0 відкидається, якщо ні, то немає підстав її відхилити.

В сучасних статистичних пакетах зазвичай порівнюються не тільки $K_{спост.}$ та $k_{кр.}$, але й заданий рівень значущості α і ймовірність того P (*p-value*), що, наприклад, для правобічної критичної області $K > K_{спост.}$. Наприклад, на рис. 6.2*а* ця ймовірність P дорівнює площі під кривою розподілу критерію, розташованої праворуч від $K_{спост.}$.

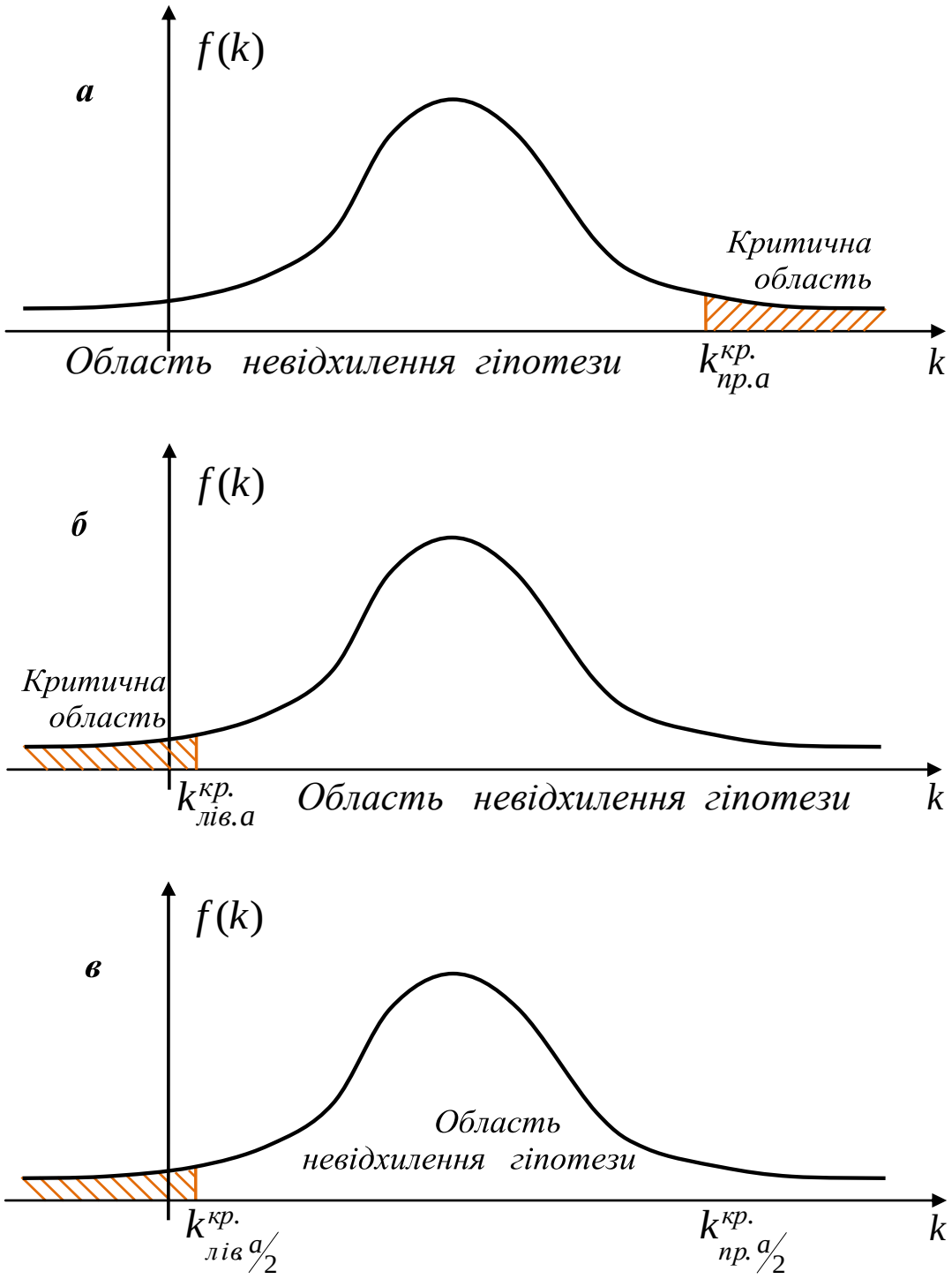


Рис. 6.1. Види критичної області

Якщо ймовірність P виявляється більше заданого рівня значущості α ($P > \alpha$), то гіпотеза H_0 не відхиляється (рис. 6.2*а*), в іншому випадку гіпотеза відхиляється (рис. 6.2*б*, $P < \alpha$).

Цей підхід також є вірним для лівобічної і двобічної симетричної критичної області.

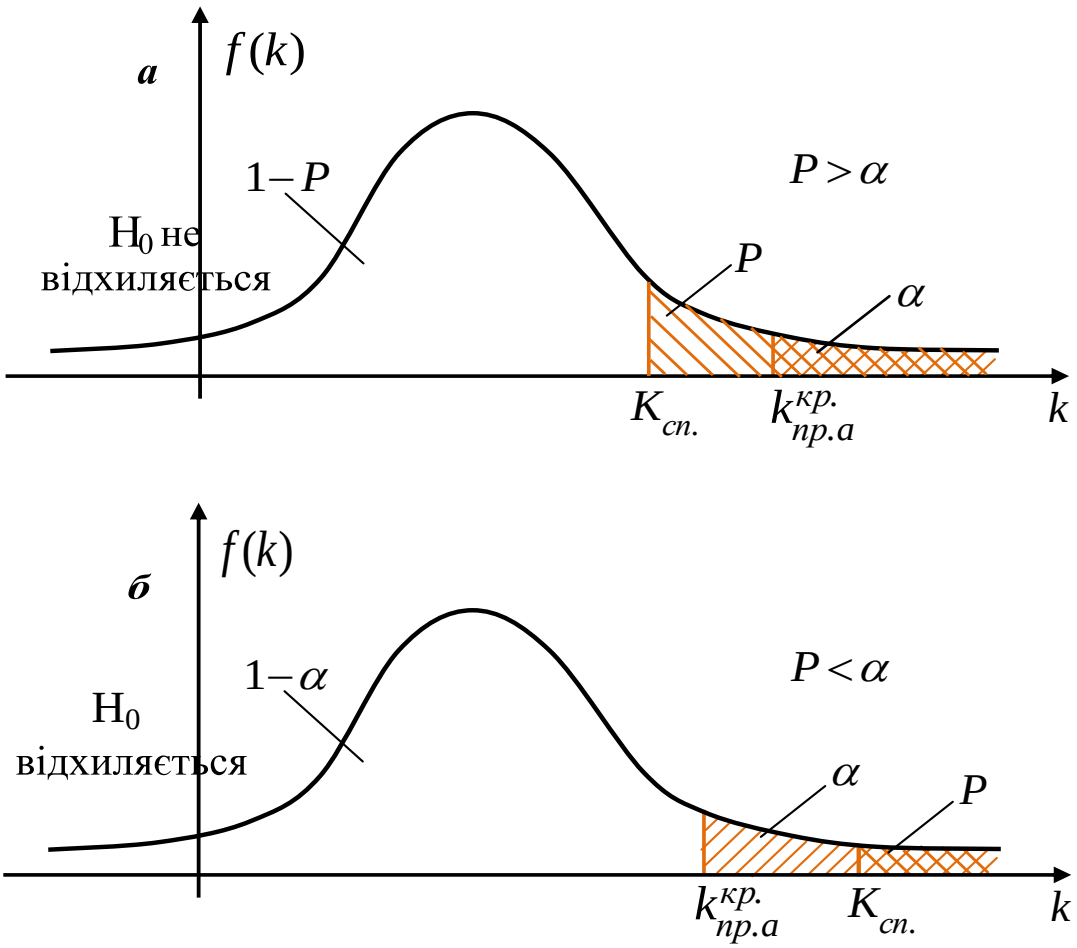


Рис. 6.2. Перевірка гіпотези за допомогою порівняння *P* (*p-value*) з рівнем значущості α

Алгоритм використання будь-якого критерію містить:

- вибір відповідного статистичного методу;
- формулювання нульової H_0 та альтернативної H_1 гіпотез;
- вибір статистичного критерію і задання довірчої ймовірності (рівня значущості);
- обчислення емпіричного (за вибіркою) значення статистичного критерію;
- знаходження критичного значення критерію за допомогою таблиць;
- прийняття рішення на основі порівняння емпіричного (спостережуваного) і критичного значень критерію.

Статистичні гіпотези можна умовно об’єднати в декілька груп:

Група 1 охоплює гіпотези, в яких передбачається можливе значення одного з основних параметрів генеральної сукупності, при цьому закон розподілу самої генеральної сукупності повинен бути відомий.

Група 2 містить гіпотези, в яких передбачається рівність двох або кількох значень параметрів або ознак генеральних сукупностей.

Група 3 охоплює гіпотези, в яких передбачається конкретний вид закону розподілу даної генеральної сукупності, якщо він невідомий.

Група 4 містить гіпотези, в яких передбачається дослідити значущість деяких коефіцієнтів, що характеризують можливу залежність між двома або кількома ознаками даної генеральної сукупності.

Нижче наведено приклади статистичної перевірки гіпотез, які належать до 3-ї та 4-ї групи.

6.2. Перевірка гіпотези про вид закону розподілу

Розглянемо приклад перевірки гіпотези за критерієм *Хі-квадрат* (χ^2) щодо законів розподілу частот випадкових величин. Ця задача стає практично важливою тоді, коли йдеться про заміну того чи іншого закону нормальним розподілом, оскільки більшість практичних прийомів аналітичної статистики спирається на цей закон.

Розподіл *Хі-квадрат Пірсона* являє собою розподіл ймовірностей квадратів k незалежних випадкових величин, кожна з яких розподілена за нормальним законом з нульовим середнім арифметичним і дисперсією, яка дорівнює одиниці. Геометрично розподіл являє собою сім'ю кривих для $k = 1, 2, 3, \ldots$. Кожна крива асиметрична, з позитивною асиметрією, і, тільки починаючи з $k = 10$, відповідна крива відображає закон розподілу, близький до нормального закону. За своєю природою розподіл χ^2 є розподілом дискретних величин.

Та обставина, що, з одного боку, розподіл Пірсона є сумою квадратів випадкових величин, а з іншого – за допомогою суми квадратів вимірюються відхилення, помилки і, зокрема, дисперсія, дає підставу використовувати цей розподіл для перевірки відхилення емпіричних і теоретичних частот і, таким чином, зіставляти закони розподілу.

В обчислювальному відношенні перевірка за критерієм χ^2 дозволяє встановити значущість різниць між спостережуваними та прогнозованими частотами:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_{in} - f_{ip})^2}{f_{ip}}, \qquad (6.1)$$

де k – кількість груп частот; f_{in}, f_{ip}, (i = 1, 2, 3,..., k) – спостережувані та прогнозовані частоти.

У практичних умовах виникають два випадки: 1) відомий деякий розподіл частот і потрібно емпірично (з певною ймовірністю) підтвердити або відхилити закон розподілу; 2) потрібно підтвердити або відкинути з наперед заданою ймовірністю, що отриманий в результаті експерименту або спостереження розподіл відповідає відомому закону. Таким чином, можливі випадки: коли теоретичні частоти заздалегідь відомі, і коли вони невідомі; і обидві ці задачі називаються перевіркою на згоду або адекватність розподілів.

При перевірці гіпотези про відповідність емпіричного розподілу теоретичному порівнюють фактичне значення критерію χ^2 з табличним $\chi^2_{кр.}$. Якщо $\chi^2 < \chi^2_{кр.}$, то емпіричний розподіл відповідає теоретичному. В іншому випадку емпіричний розподіл не відповідає теоретичному, розподіл частот в ньому носить інший характер.

***Приклад* 6.1.** Припустимо, що новий вид туризму, який ще недостатньо розвинений в регіоні, оцінюється фахівцями за п’ятьма категоріями: 1 – дуже привабливий; 2 – привабливий; 3 – відносно привабливий; 4 – мало-привабливий; 5 – непривабливий. Відповідні

показники відносних частот, отриманих на підставі опитування 100 фахівців, наведені в табл. 6.1 (стовпець 2). Крім того, було опитано 300 туристів, які користувалися цим видом туризму. Їхня кількість за категоріями наведена в стовпці 3. По суті, вони являють собою частоти, отримані на підставі вибірки $n = 300$, тобто спостережувані частоти. Необхідно визначити, наскільки відрізняється фактичний розподіл частот оцінки нового виду туризму від прогнозованого фахівцями.

Таблиця 6.1

Оцінка привабливості нового виду туризму

Категорії	Відсоток	Кількість опитаних, f_{in}	Об'єм вибірки	Прогнозовані частоти, f_{ip}	$f_{in} - f_{ip}$	$(f_{in} - f_{ip})^2$	$\frac{(f_{in} - f_{ip})^2}{f_{ip}}$
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	40	130	300	0,4· 300 = 120	10	100	0,833
2	30	80	300	0,3·300 = 90	-10	100	1,111
3	20	50	300	0,2·300 = 60	-10	100	1,667
4	5	25	300	0,05·300 = 15	10	100	6,667
5	5	15	300	0,05·300 = 15	0	0	0
Сума	100	300					10,278

Для того, щоб отримати прогнозовані частоти, приведені до об'єму вибірки, необхідно прогнозовані частоти кожної категорії, тобто величини стовпчика 2, помножити на $n = 300$ і розділити на 100. Отримані дані наведені в стовпці 5.

Тепер можна перейти до обчислення значення χ^2. Для цього використовується формула (6.1). Подальший розрахунок χ^2 представлений стовпцями 6-8. Отримане значення $\chi^2 = 10{,}278$ – сума чисел останнього 8-го стовпця табл. 6.1:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_{in} - f_{ip})^2}{f_{ip}} = 10{,}278.$$

Щоб використовувати це значення для оцінки законів збігу розподілу частот, необхідно сформулювати нульову і альтернативну

гіпотези. Отримані в результаті опитування туристів частоти при заданому рівні значущості відповідають прогнозованим частотам, або зазначені частоти далекі від спостережуваних. Крім того, необхідно обчислити число ступенів вільності $df = k - 1$ для визначення відповідної кривої розподілу χ^2. У нашому випадку k дорівнює числу груп частот. Отже, $df = 5 - 1 = 4$. Після цього для заданого рівня значущості, наприклад, $\alpha = 0{,}05$, тобто значення ймовірності, при якій нульова гіпотеза відкидається, необхідно за спеціальною таблицею (або за допомогою спеціальної програми) знайти критичне значення $\chi^2_{кр.}$, яке відповідає цій ймовірності. В даному випадку для $df = 4$ і $\alpha = 0{,}05$ воно дорівнює 9,49 (див. табл. 6.2). Якщо знайдене (тобто розрахункове) значення $\chi^2 \geq \chi^2_{кр.}$, нульова гіпотеза відхиляється. В іншому випадку гіпотеза не відхиляється.

Таблиця 6.2

Фрагмент таблиці критичних значень *Xi*-квадрат ($\chi^2_{кр.}$) розподілу

df	α=0,05	α=0,01	α=0,001
1	3,84	6,63	10.83
2	5,99	9,21	13,82
3	7,81	11,07	16,27
4	9,49	13,28	18,47
5	11,07	15,09	20,51
6	12,59	16,81	22.46
7	14,07	18,48	24,32
8	15,51	20,09	26,12
9	16,92	21,67	27,88
10	18,31	33,21	29,59
11	19,68	34,73	31,26
12	21,03	26,22	32,91

Оскільки в нашому випадку при $\alpha = 0{,}05$, $\chi^2 \approx 10{,}28 > 9{,}49 = \chi^2_{кр.}$

(рис. 6.3), то нульова гіпотеза відхиляється, і можна зробити висновок про те, що при рівні значущості 0,05 фактичний розподіл частот оцінки нового виду туризму відрізняється від прогнозованого спеціалістами.

Таким чином, нульова гіпотеза завжди буде відкинута, якщо $\chi^2 \geq \chi^2_{кр.}$ для будь-якого рівня значущості.

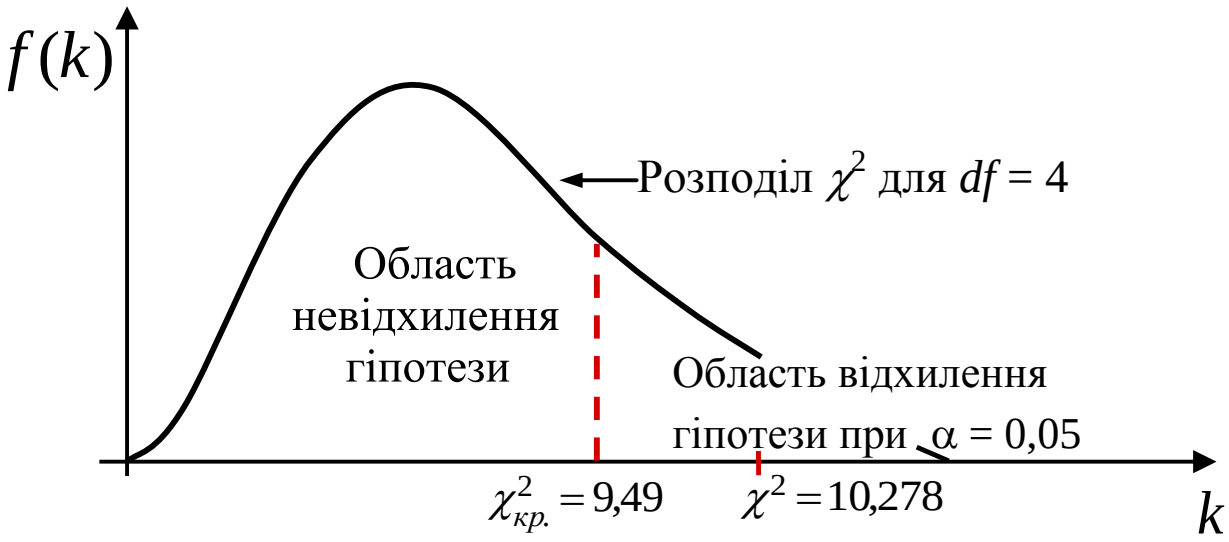


Рис. 6.3. Перевірка гіпотези за критерієм *Xi*-квадрат (χ^2)

Пояснимо деякі моменти, пов'язані зі ступенем вільності (df) у статистиці. Як вже було зазначено вище, за визначенням ступінь вільності – це число вільно варіювальних одиниць у статистичній сукупності, тобто кількість елементів ряду, які можуть вільно змінюватися.

Наприклад, якщо сукупність складається з n-ї кількості членів і характеризується середньою величиною, то будь-який член цієї сукупності може мати яке завгодно значення, не змінюючи при цьому середню $\bar{x}$, крім однієї варіанти, значення якої визначається різницею між сумою значень усіх інших варіант і величиною $n\bar{x}$. Наприклад, якщо $\bar{x} = \frac{x_1 + x_2 + x_3}{3}$, то звідси x_1 (або x_2, x_3) можна визначити як: $x_1 = (3 \cdot \bar{x}) - (x_2 + x_3)$. Отже, одна варіанта кількісно обмеженої

статистичної сукупності не має свободи варіації. Звідси число ступенів вільності для такої сукупності дорівнюватиме її об'єму n без одиниці, тобто $df = n - 1$.

При наявності не одного, а декількох обмежень свободи варіації число ступенів вільності варіації дорівнюватиме $df = n - v$, де v позначає число обмежень свободи варіації.

Приклад 6.2. Є дані про ділові подорожі громадян Великої Британії різного віку протягом одного року, табл. 6.3. Дані запозичені з сайту National Travel Survey (https://www.gov.uk/government/collections/national-travel-survey-statistics) і оброблені авторами навчального посібника. У стовпчиках 2-4 наведені вік туристів за групами, спостережувана частота в групі f_{in}, центри (середини) інтервалів x_i. Необхідно підтвердити або спростувати гіпотезу про те, що при рівні значущості $\alpha = 0{,}05$ емпіричний розподіл частот (кількості туристів за віковими групами) відповідає нормальному закону розподілу.

Таблиця 6.3

Перевірка гіпотези про нормальний розподіл частот

№	Вік туристів за групами	Кількість туристів, f_{in}	Середина інтервалу, x_i	$(x_i - \bar{x})^2$	$f(x_i, \bar{x}, \sigma)$	Прогнозо-вані частоти, f_{ip}	$\frac{(f_{in} - f_{ip})^2}{f_{ip}}$
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	17-19	10	18,5	660,84	0,00514	3,36	15,014
2	20-29	28	25,0	368,91	0,01101	19,2	3,825
3	30-39	44	34,5	94,22	0,02253	44,21	0,001
4	40-49	51	44,5	0,09	0,02880	56,51	0,606
5	50-59	56	54,5	105,95	0,02185	42,88	3,798
6	60-69	24	64,5	411,81	0,00985	19,32	1,232
7	70-79	5	74,5	917,67	0,00263	5,17	0,005
	Сума	218					24, 481

Для того, щоб розв'язати поставлену задачу, необхідно

визначити значення χ^2, використовуючи рівняння (6.1). У свою чергу, для цього потрібно по кожній групі обчислити прогнозовану частоту f_{ip}. Такі частоти обчислюються на основі множення значень щільності ймовірності нормального розподілу $f(x_i, \bar{x}, \sigma)$, відповідно, на довжину інтервалу h і об'єм вибірки $n = 218$.

Таким чином, процедура зводиться до обчислення значень щільності ймовірності $f(x_i, \bar{x}, \sigma)$, де x_i, $\bar{x}$ і σ – середина інтервалу, середня арифметична вибірки і стандартне відхилення, відповідно. Для цього використовується формула (4.1) (див. Розділ 4):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma}\right)^2\right].$$

У свою чергу, для обчислення значень $\bar{x}$ і σ застосовуються вирази (3.3) і (3.23) (див. Розділ 3):

$$\bar{x} = \frac{\sum(f_i \cdot x_i)}{\sum f_i}, \qquad \sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2 \cdot f_i}{\sum f_i}}.$$

В результаті, на основі цих формул отримуємо: $\bar{x} = 44$, $\sigma = 13{,}85$.

Використовуючи ці значення, послідовно заповнюємо стовпці 5 і 6 таблиці 6.3. Потім, за виразом $f(x_i, \bar{x}, \sigma) \cdot h \cdot n$ знаходимо значення прогнозованої частоти f_{ip} і далі – відношення $\frac{(f_{in} - f_{ip})^2}{f_{ip}}$, представлені стовп-цем 8. Підсумовуючи значення стовпця 8, отримуємо $\chi^2 = 24{,}481 \approx 24{,}48$. Криві емпіричних і теоретичних частот наведені на рис. 6.4.

При числі ступенів вільності $df = n - 1 = 7 - 1 = 6$ і рівні значущості $\alpha = 0{,}05$ критичне значення $\chi^2_{кр.} = 12{,}59$ (див. табл. 6.2). Оскільки $\chi^2 = 24{,}48 > 12{,}59 = \chi^2_{кр.}$, гіпотеза відкидається, тобто розходження між емпіричним і теоретичним законами розподілу частот при $\alpha = 0{,}05$ є значущим. Отже, вік туристів, які здійснюють

ділові поїздки, не підкоряється нормальному закону розподілу.

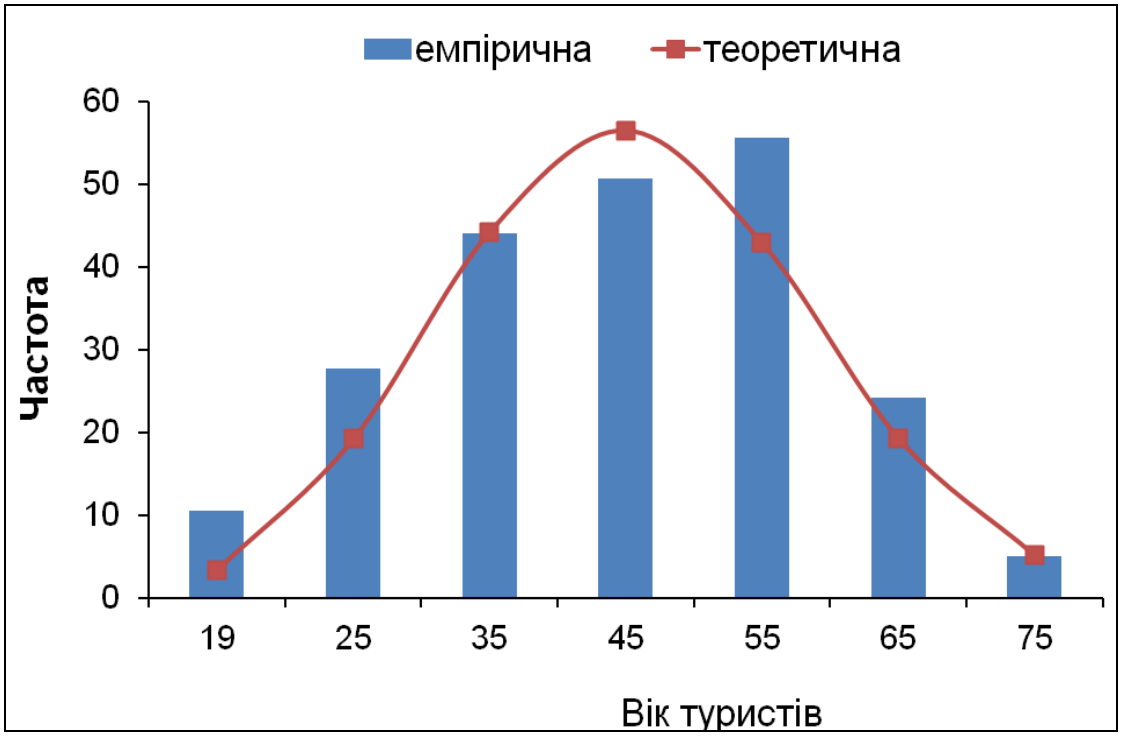


Рис. 6.4. Емпіричний і теоретичний розподіл кількості туристів за віковими групами

Повертаючись до прикладу 4.1 в розділі 4, зазначимо, що для натурального логарифма даних, що характеризують розподіл загальної кількості ночей, проведених туристами за місяць в об'єктах їх розміщення за період 2008-2019 роки, для $\bar{x} = 17{,}63$ і $\sigma = 0{,}27$, при рівні значущості $\alpha = 0{,}05$, $df = 7$ отримуємо $\chi^2 = 12{,}106$ (табл. 6.4), що менше $\chi^2_{кр.} = 14{,}07$ (табл. 6.2). Отже, припущення про те, що дані, які характеризують щомісячну кількість ночей, проведених туристами в об'єктах їх розміщення за вказаний період, розподілені за логарифмічно нормальним законом, є вірним.

Таблиця 6.4

Перевірка гіпотези про нормальний розподіл частот

№	Інтервали $y_i = \ln x_i$	Частота, f_{in}	Середина інтервалів, y_i	$(y_i - \mu)^2$	$f(y_i, \mu, \sigma)$	Прогнозовані частоти, f_{ip}	Спостережувана частота,	$\frac{(f_{in} - f_{ip})^2}{f_{ip}}$

							f_{in}	
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	17,05-17,20	9	17,125	0,255	0,072	6	9	1,500
2	17,20-17,35	19	17,275	0,126	0,171	14	19	1,786
3	17,35-17,50	28	17,425	0,042	0,301	24	28	0,667
4	17,50-17,65	21	17,575	0,003	0,391	31	21	3,226
5	17,65-17,80	23	17,725	0,009	0,375	30	23	1,633
6	17,80-17,95	19	17,875	0,060	0,266	21	19	0,19
7	17,95-18,10	14	18,025	0,156	0,139	11	14	0,818
8	18,10-18,25	10	18,175	0,297	0,054	7	11	2,286
9	18,25-18,40	1	18,325	0,483	0,015			
	Сума	218						12,106

Завершуючи цей параграф, слід зазначити, що застосування розподілу χ^2 для виявлення згоди між емпіричними і прогнозованими розподілами частот має деякі обмеження. Вони починають відігравати суттєву роль, коли емпіричні частоти груп даних дуже низькі, а саме, менше 5. У цьому разі дискретність χ^2 розподілу вносить помилки при обчисленні конкретного його значення. У зв'язку з цим на практиці групи з низькою частотою рекомендується об'єднати в одну групу, наприклад, частоти останніх рядків стовпців 7 та 8, табл. 6.4.

Крім цього, при об'ємі вибірки $n < 30$ критерій χ^2 дає вельми наближені значення. Точність підвищується з ростом n.

6.3. Перевірка значущості коефіцієнта кореляції

Оскільки вибірковий коефіцієнт кореляції r обчислюється за вибірковими даними, він є випадковою величиною. Якщо $r \neq 0$, то постає питання: чи пояснюється це дійсно існуючим лінійним зв'язком між, наприклад, x та y або викликано випадковими

факторами?

Коефіцієнт кореляції, як і будь-який інший вибірковий коефіцієнт, служить оцінкою свого генерального параметра (тобто істинного, невипадкового коефіцієнта кореляції) і, як величина випадкова, супроводжується помилкою:

$$\sigma_r = \sqrt{\frac{1-r^2}{n-2}} \ . \qquad (6.2)$$

Для перевірки суттєвості коефіцієнта кореляції обчислюють спостережуване (тобто розрахункове) значення *t-критерію* ($t_р$), яке являє собою відношення вибіркового коефіцієнта кореляції (r) до своєї помилки (σ_r):

$$t_{спост.} = \frac{r}{\sigma_r} = |r| \cdot \sqrt{\frac{n-2}{1-r^2}} \ , \qquad (6.3)$$

де n – об’єм вибірки.

Відношення (6.3) служить критерієм для перевірки *нульової гіпотези* – припущення про те, що в генеральній сукупності істинний коефіцієнт кореляції дорівнює нулю.

Далі спостережуване значення $t_{спост.}$ порівнюють з його критичною (тобто табличною) величиною $t_{кр.}$, яка отримується з таблиці значень *t-критерію* Стьюдента при заданому рівні значущості α і числі ступенів вільності $df = n - 2$.

Якщо для прийнятого рівня значущості α значення $t_{спост.} > t_{кр.}$, то нульова гіпотеза (тобто гіпотеза про відсутність зв’язку) відкидається і вважається, що між досліджуваними ознаками, наприклад, x та y існує лінійний зв’язок з достовірністю $(1-\alpha)$. І, навпаки, в разі $t_{спост.} < t_{кр.}$ нульова гіпотеза приймається, що означає відсутність лінійного зв’язку між x та y.

Табличні значення $t_{кр.}$ для різних рівнів значущості α і чисел ступенів вільності $df = n - 2$ наведені в табл. 6.5 (повна таблиця наведена у *Додатку* Б).

Таблиця 6.5

Фрагмент таблиці критичних значень *t*-критерію Стьюдента

df	α		
	0,10	0,05	0,01
1	6,3138	12,7062	63,6567
2	2,9200	4,3027	9,9248
3	2,3534	3,1824	5,8409
4	2,1318	2,7764	4,6041
5	2,0150	2,5706	4,0321
6	1,9432	2,4469	3,7074
7	1,8946	2,3646	3,4995
8	1,8595	2,3060	3,3554
9	1,8331	2,2622	3,2498
10	1,8125	2,2281	3,1693
11	1,7959	2,2010	3,1058
12	1,7823	2,1788	3,0545

Можна також скористатися таблицею критичних значень коефіцієнта кореляції, з якої знаходять величину критичного значення коефіцієнта кореляції $r_{кр.}$ за числом ступенів вільності $df = n - 2$ і рівнем значущості α. Якщо $r_{спост.} < r_{кр.}$, то в генеральній сукупності відсутня значуща кореляція між досліджуваними ознаками, а відмінність від нуля вибіркового коефіцієнта кореляції пояснюється лише випадковістю вибірки або тим, що об'єм вибірки є недостатнім для виявлення лінійного зв'язку. Якщо ж $r_{спост.} > r_{кр.}$, то робиться висновок, що коефіцієнт кореляції значно відрізняється від нуля, і існує статистично значуща кореляція.

Зауважимо, що чим меншим є об'єм вибірки, тим більшим має бути розрахункове значення коефіцієнта кореляції для прийняття гіпотези про лінійну залежність між величинами *x* та *y*. Проте, як

завгодно близьке до одиниці значення *r* не гарантує їх причинно-наслідкової обумовленості, оскільки можливим є інший характер їхнього взаємозв'язку. Так, одні явища можуть одночасно, але незалежно одне від одного (спільні події) відбуватися або змінюватися (*помилкова* регресія). Інші – бути в причинній залежності не одне від одного, а в більш складному причинно-наслідковому зв'язку (*непряма* регресія). Таким чином, при значущому коефіцієнті кореляції остаточний висновок про наявність причинно-наслідкового зв'язку можна зробити лише з урахуванням специфіки досліджуваної проблеми.

Слід зазначити, що правильне застосування коефіцієнта кореляції передбачає нормальний розподіл (або розподіл Гауса) двовимірної сукупності значень випадкових величин *x* та *y*. Якщо коефіцієнт кореляції має значущу відмінність від нуля, то його розподіл тим сильніше відрізняється від нормального, чим меншим є кількість спостережень *n* і чим більше його абсолютне значення. Отже, коефіцієнт кореляції не буде точною оцінкою генерального параметра, якщо він обчислений на нечисленній вибірці, і його абсолютне значення перевищує 0,5.

Зазначимо, що *t*-критерій Стьюдента застосовується для об'єму вибірки
$n < 50$.

Для об'єму вибірки $n \geq 50$ можна обчислити значення *Z-критерію* за формулою: $Z_{спост.} = r \cdot \sqrt{n-1}$. Отримане значення порівнюється з табличним критичним $Z_{кр.}$. За умови $Z_{спост.} \geq Z_{кр.}$ кореляція визнається достовірною.

Значущість коефіцієнта кореляції також можна перевірити за допомогою *F-критерію* (критерію Фішера). Без обчислення будь-яких критеріальних коефіцієнтів можна скористатися таблицею Р. Фішера. У ній наведені значення лінійних коефіцієнтів кореляції, при яких кореляція може бути визнана достовірною. Значення $r_{кр.}$ вибираються для заданого рівня значущості α і об'єму вибірки *n*

(таблиця побудована для числа ступенів вільності, пов'язаного з n співвідношенням: $df = n - 2$).

***Приклад* 6.3.** Необхідно визначити значущість вибіркового коефіцієнта кореляції, обчисленого в прикладі 5.1, розділ 5.

Оскільки величина коефіцієнта кореляції в результаті розв'язання прикладу 5.1 визначена – $r_{xy} = -0{,}7$, знайдемо емпіричне значення t-критерію:

$$t_{спост.} = |r| \cdot \sqrt{\frac{n-2}{1-r^2}} = 0{,}7 \cdot \sqrt{\frac{12-2}{1-0{,}7^2}} = 3{,}10.$$

Число ступенів вільності дорівнює $df = n - 2 = 12 - 2 = 10$, рівень значущості виберемо таким, що дорівнює $\alpha = 0{,}05$. За таблицею «Критичні значення t-критерію Стьюдента» при різних рівнях значущості (табл. 6.5, *Додаток Б*) знаходимо критичне значення $t_{кр.}(df = 10;\ \alpha = 0{,}05) \approx 2{,}23$.

Оскільки $t_{спост.} > t_{кр.}$ $(3{,}10 > 2{,}23)$, нульова гіпотеза відкидається на 5 %-му рівні значущості. Отже, з імовірністю $P > 0{,}95$ можна стверджувати, що між курсом гривні до євро і кількістю туристів, що виїжджають за кордон, існує статистично значущий (негативний) кореляційний зв'язок.

Наведений спосіб оцінки значущості вибіркового коефіцієнта кореляції не є єдиним.

Перевіримо значущість коефіцієнта кореляції також за допомогою F-критерію (критерію Фішера):

$$F = \frac{r^2}{1-r^2} \cdot \frac{k_2}{k_1}, \qquad (6.4)$$

де k_1, k_2 є числами ступенів вільності дисперсій і розраховуються за формулою: $k_1 = m - 1 = 2 - 1 = 1$, де m – для лінійної функції дорівнює 2; $k_2 = n - m = 12 - 2 = 10$, де n – об'єм вибірки; в прикладі, що розглядається, $n = 12$.

Розрахуємо значення критерію Фішера:

$$F = \frac{r^2}{1-r^2} \cdot \frac{k_2}{k_1} = \frac{0{,}49}{1-0{,}49} \cdot \frac{12-2}{2-1} = 9{,}61.$$

Фактичне значення критерію Фішера порівнюють з критичним значенням. Якщо $F_{факт.} > F_{кр.}$, то зв'язок між ознаками є істотним, в іншому випадку (якщо $F_{факт.} < F_{кр.}$) зв'язок є несуттєвим. Табличне значення критерію Фішера для $k_1 = 1$, $k_2 = 10$, згідно з табл. 6.6 (повна таблиця *F*- критерію Фішера наведена у *Додатку* В), становить: $F_{кр.} = 4{,}96$, що менше фактичного значення, тобто $F_{факт.}(9{,}61) > F_{кр.}(4{,}96)$. Отже, зв'язок між досліджуваними ознаками, а саме між курсом гривні до євро і кількістю українських туристів, що виїжджають до Німеччини, при рівні значущості $\alpha = 0{,}05$ є істотним.

Таблиця 6.6

Фрагмент таблиці значень критерію Фішера (*F*-критерію) для рівня значущості α =0,05

k_1 / k_1	1	2	3	4	5	6	8
1	161,4	199,5	215,7	224,5	230,1	233,9	238,8
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59

6.4. Перевірка значущості рівняння регресії

Оскільки параметри рівняння регресії a і b розглядають як випадкові величини, рівняння регресії $\bar{y} = ax + b$ також розглядається як випадковий об'єкт, тобто використання рівнянь регресії (регресійних моделей) для вирішення певних практичних завдань можливо лише в тому випадку, якщо вони відбивають реальні істотні зв'язки. У зв'язку з цим для того, щоб правомірно використовувати рівняння регресії, тобто поширити висновки на генеральну сукупність, перевіряють його значущість.

Отже, важливим завданням регресійного аналізу, поряд з перевіркою значущості окремих параметрів, є перевірка значущості рівняння регресії (або адекватності регресійних моделей), мета якої – з'ясувати, чи не є параметри отриманого рівняння регресії результатом впливу випадкових причин.

Для цього рівняння регресії піддають загальному статистичному аналізу: *перевіряють нульові гіпотези про значущість рівняння* регресії і *його коефіцієнтів* a і b, а також *визначають довірчі інтервали*.

Якість моделі (рівняння регресії) з відносних відхилень по кожному спостереженню визначають на основі розрахунку величини *середньої помилки апроксимації* $\bar{\varepsilon}$:

$$\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\ \%, \qquad (6.5)$$

де y_i – емпіричні значення результативної ознаки; $\hat{y}_i$ – теоретичні значення результативної ознаки. Значення середньої помилки апроксимації не повинно перевищувати 10-15 %.

Перевірка значущості рівняння регресії в цілому проводиться на основі F-критерію Фішера, величину якого отримують, зіставляючи факторну і залишкову дисперсії в розрахунку на одну ступінь вільності:

$$F_{розрах.} = \frac{\sigma^2_{факт.}}{\sigma^2_{зал.}}, \qquad (6.6)$$

де

$$\sigma^2_{факт.} = \frac{\sum(\hat{y}_i - \bar{y})^2}{m} \qquad \sigma^2_{зал.} = \frac{\sum(y_i - \hat{y}_i)^2}{n-m-1} \tag{6.7}$$

Розрахункове значення *F*-критерію Фішера (6.6) порівнюють з критичним (табличним) $F_{кр.}(a; k_1; k_2)$ при рівні значущості α та ступенях вільності $k_1 = m$ і $k_2 = n-m-1$. При цьому, якщо фактичне значення *F*-критерію виявиться більшим за відповідне табличне значення, то визнається статистична значущість рівняння в цілому, тобто рівняння регресії адекватно репрезентує дані генеральної сукупності.

Для парної лінійної регресії $m = 1$, тому:

$$F_{розрах.} = \frac{\sigma^2_{факт.}}{\sigma^2_{зал.}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \hat{y}_i)^2} \cdot (n-2). \tag{6.8}$$

На практиці, зокрема, в комп'ютерних технологіях регресійного аналізу, застосовують інший спосіб перевірки значущості рівняння регресії за емпіричними даними. Він полягає в тому, що замість співвідношення між факторною і залишковою дисперсіями використовують *коефіцієнт детермінації* r^2_{xy}. Величина *F*-критерію пов'язана з коефіцієнтом детермінації за такою формулою:

$$F_{розрах.} = \frac{r^2_{xy}}{1-r^2_{xy}} \cdot (n-2). \tag{6.9}$$

Якщо при заданому рівні значущості α і заданому *n*, $F_{розрах.} > F_{кр.}$, нульова гіпотеза про те, що $r^2_{xy} = 0$ відкидається, тобто рівняння $\hat{y} = ax + b$ вважається адекватним емпіричним даним генеральної сукупності.

У парній лінійній регресії оцінюється значущість також окремих параметрів рівняння. Для оцінки статистичної значущості параметрів рівняння регресії визначають *t*-критерій Стьюдента і довірчі інтервали кожного з показників.

Оцінку значущості параметрів регресії a і b за допомогою t-критерію Стьюдента t_a і t_b здійснюють шляхом зіставлення їх значень з величиною їх помилки m_a і m_b, відповідно:

$$t_a = \frac{a}{m_a}; \qquad t_b = \frac{b}{m_b}. \tag{6.10}$$

Помилки параметрів лінійної регресії a і b визначають на основі формул:

$$m_a = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{(n-2)\sum(x_i - \bar{x})^2}} = \sqrt{\frac{\sigma^2_{зал.}}{\sum(x_i - \bar{x})^2}} = \frac{\sigma_{зал.}}{\sigma_x\sqrt{n}}; \tag{6.11}$$

$$m_b = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2 \sum x_i^2}{(n-2)n\sum(x_i - \bar{x})^2}} = \sqrt{\sigma^2_{зал.} \cdot \frac{\sum x_i^2}{n^2\sigma_x^2}} = \sigma_{зал.} \cdot \frac{\sqrt{\sum x_i^2}}{n\sigma_x}. \tag{6.12}$$

В тому разі, якщо $t_{кр.} > t_{фак.}$ для заданого рівня значущості α і числа ступенів вільності $(n-2)$, то визнається випадкова природа формування a і b. Якщо $t_{кр.} < t_{фак.}$, гіпотеза α-значущості про те, що $a = 0$ ($b = 0$), відкидається, тобто a і b не випадково відрізняються від нуля і сформувалися під впливом систематично діючого фактора x. Тоді для параметрів a і b можна знайти довірчі інтервали.

Довірчий інтервал $\hat{y}$ для будь-якого значення x_i (i = 1, 2, ..., n) визначається на основі співвідношення:

$$a \pm t_{кр.} m_a; \quad b \pm t_{кр.} m_b, \tag{6.13}$$

де $t_{кр.}$ – критичне значення t-розподілу Стьюдента, знайдене для прийнятого рівня значущості α і числа ступенів вільності $(n-2)$.

***Приклад* 6.4.** Перевіримо адекватність рівняння лінійної регресії y на x, отриманого при розв'язанні прикладу 5.1 (розділ 5): $y = -1846{,}8x + 55213$.

Для цього прикладу значення $F_{розрах.}$ обчислимо за формулою (6.8):

$$F_{розрах.} = \frac{\sigma^2_{факт.}}{\sigma^2_{зал.}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \hat{y}_i)^2} \cdot (n-2) = \frac{2{,}62 \cdot 10^9}{2{,}69 \cdot 10^9} \cdot (12-2) = 9{,}74.$$

За таблицею критичних точок розподілу *F*-Фішера (табл. 6.6 або *Додаток В*) $F_{кр.}$ при рівні значущості $\alpha = 0{,}05$ і числах ступенів вільності $k_1 = 1$, $k_2 = 10$ складає $F_{кр.} = 4{,}96$.

Оскільки фактичне значення критерію Фішера більше табличного $F_{розрах.} > F_{кр.}$ (9,74 > 4,96), то нульова гіпотеза відкидається, тобто рівняння лінійної регресії $\hat{y} = ax + b$ ($y = -1846{,}8x + 55213$) статистично значуще описує результати експерименту.

Для цього прикладу значення $F_{розрах.}$ обчислимо також за формулою (6.9): $F_{розрах.} = \frac{r_{xy}^2}{1 - r_{xy}^2}(n-2) = 9{,}61$.

Невеликі розбіжності результатів розрахунків $F_{розрах.}$ за формулою (6.8) і (6.9) пов'язані з округленням.

Питання для самоконтролю

6.1. Що в статистиці розуміють під гіпотезою та її перевіркою? Наведіть приклад статистичної гіпотези з галузі туризму.

6.2. В чому полягає мета перевірки статистичних гіпотез?

6.3. Яка статистична гіпотеза називається нульовою? альтернативною? Наведіть приклади.

6.4. Назвіть типи гіпотез.

6.5. Яка гіпотеза називається простою? складною? Наведіть приклади.

6.6. Поясніть сенс помилок першого і другого роду, що виникають під час перевірки гіпотез.

6.7. Що являє собою рівень значущості? Як він пов'язаний з довірчою ймовірністю?

6.8. Які величини рівня значущості найчастіше зустрічаються в соціально-економічних дослідженнях?

6.9. Що мають на увазі, коли говорять, що рівень значущості дорівнює 0,05?

6.10. Що являє собою потужність критерію?

6.11. Що називається статистичним критерієм?

6.12. Які критерії називаються параметричними, а які –

непараметричними? Назвіть умови їх застосування.

6.13. Що таке критичний рівень значущості? Чим він відрізняється від рівня значущості?

6.14. Що являють собою області відхилення і невідхилення гіпотез?

6.15. Що являють собою критичні точки?

6.16. Як в статистиці інтерпретується поняття «число ступенів вільності»?

6.17. Сформулюйте основний принцип перевірки статистичної гіпотези.

6.18. Що являють собою критична і довірча області критерію?

6.19. В чому полягає різниця між двосторонньою та односторонньою перевірками гіпотез?

6.20. Що називається спостережуваним (або емпіричним) значенням критерію?

6.21. Сформулюйте алгоритм використання статистичного критерію.

6.22. Як перевірити значущість коефіцієнта кореляції?

6.23. Як перевіряється значущість рівняння регресії і окремих його параметрів?

6.24. Як перевіряється гіпотеза про вид розподілу за допомогою критерію *Xi*-квадрат? Як обчислюється $\chi^2_{кр.}$?

6.25. Що таке критичне значення *t*-критерію Стьюдента? Назвіть умови його застосування.

6.26. Що таке критичне значення *F*-критерію Фішера? Назвіть умови його застосування.

6.27. Як будується довірчий інтервал прогнозу в разі лінійної регресії?

Завдання для самостійного виконання

Завдання 6.1. Є дані про туристичні поїздки громадян Великої Британії різного віку протягом одного року (табл. 1). Дані запозичені з сайту *National Travel Survey*. Потрібно підтвердити або спростувати гіпотезу про те, що при рівні значущості $\alpha = 0{,}05$ емпіричний розподіл частот відповідає нормальному закону розподілу.

Таблиця 1

Перевірка гіпотези про нормальний розподіл частот

№	Вік туристів за групами	Кількість туристів, f_{in}	Середина інтервалу, x_i	$(x_i - \bar{x})^2$	$f(x_i, \bar{x}, \sigma)$	Прогнозовані частоти, f_{ip}	$\frac{(f_{in} - f_{ip})^2}{f_{ip}}$
1	0-16	9					
2	17-20	11					
3	21-29	9					
4	30-39	10					
5	40-49	11					
6	50-59	17					
7	60-69	18					
8	70-79	12					
	Сума						

Завдання 6.2. Визначте значущість вибіркового коефіцієнта кореляції, обчисленого в результаті вирішення *завдання* 5.1 (розділ 5) для самостійного виконання.

РОЗДІЛ 7
ЗАСТОСУВАННЯ MS EXCEL ДЛЯ СТАТИСТИЧНОЇ ОБРОБКИ ДАНИХ В СФЕРІ ТУРИЗМУ

Потужним інструментальним засобом при виконанні статистичних досліджень є використання комп'ютерної техніки. У зв'язку з цим великого поширення набули спеціальні пакети прикладних програм. Вони дозволяють забезпечити швидкість статистичних розрахунків, достовірність результатів, можливість легко представляти дані в графічній, табличній і аналітичній формах.

Серед подібних програм великою популярністю користуються Microsoft Excel, Statistica та ін.

У цьому розділі описано можливості Microsoft Excel (MS Excel) на прикладі розв’язання задач сфери туризму. На наочних прикладах розглянуто принципи розрахунків статистичних показників та прийоми статистичної обробки даних за допомогою програмного пакета MS Excel. У посібнику описується російськомовна версія програми MS Excel, тому пункти меню програми наведено російською мовою.

7.1. Основи роботи з електронною таблицею MS Excel

MS Excel – це програма управління електронними таблицями, що дозволяє обробляти числові дані, створювати графіки і т. д.

Математичні обчислення в **MS Excel.** Особливість електронних таблиць полягає в можливості застосування формул для опису зв’язку між значеннями різних комірок. При введенні формул використовуються константи, адреси комірок, оператори і функції. Якщо формула містить посилання на комірки, а значення в цих комірках змінюються, то Excel автоматично обчислює формули і оновлює значення, використовуючи нові дані.

Введення формули повинно завжди починатися зі знака рівності

«=» або зі знака «+». Формула може містити різні математичні оператори. Сама формула відображається в *рядку формул*, а безпосередньо в комірці відображається результат обчислень (рис. 7.1).

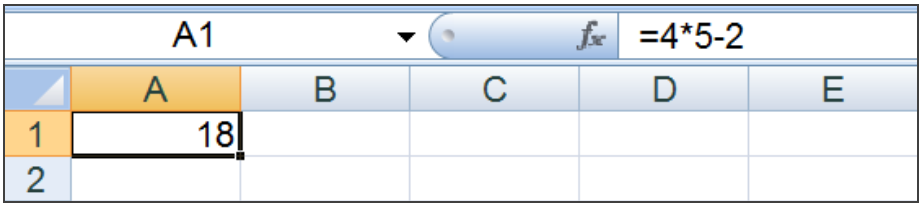


Рис. 7.1. Приклад формули у комірці

Варіант введення формул, показаний на рис. 7.1, використовується рідко. В формулах здебільшого використовують адреси комірок для підстановки значень (рис. 7.2). Щоб ввести адресу комірки необхідно клацнути мишею на потрібній комірці.

Інші приклади запису формул: =A1^2+(B2+10)^3; =(A5/B3)*100%.

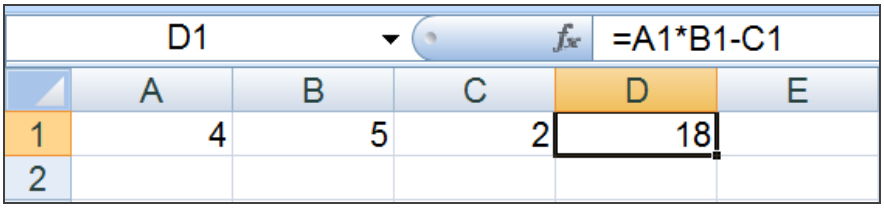
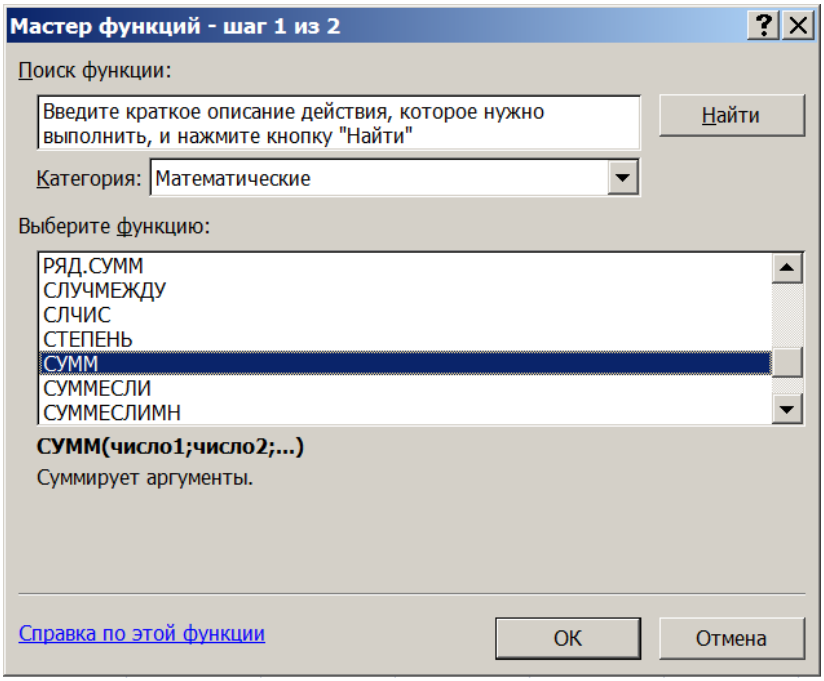


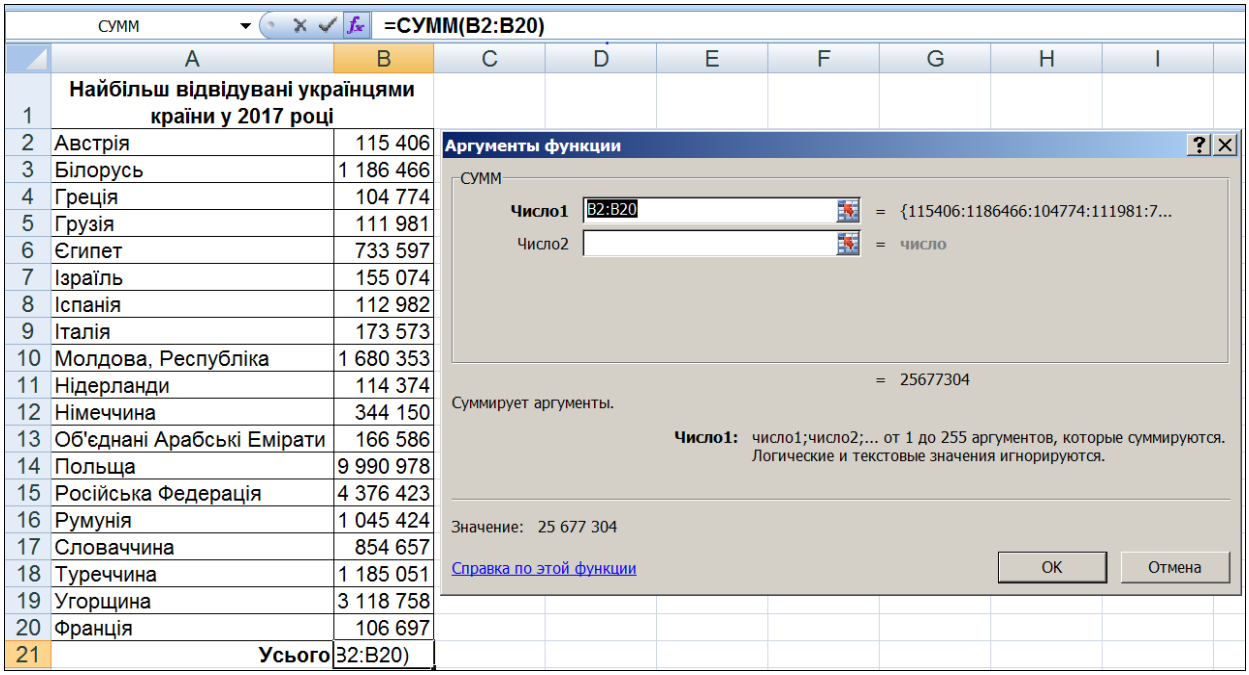
Рис. 7.2. Введення формули із зазначенням адрес комірок

Однією з найбільш затребуваних стандартних функцій MS Excel є функція підсумовування. Часто необхідно скласти ряд значень, що містяться в суміжних комірках. Можна, звичайно, написати у формулі адреси всіх комірок, які треба враховувати, наприклад, =A2+B2+C2. Однак, простіше вказати діапазон комірок і застосувати до них вбудовану функцію підсумовування СУММ(). Цю функцію особливо корисно використовувати при великому наборі даних. Для виклику функції СУММ() потрібно використати кнопку **«Мастер функций»** на панелі інструментів Excel, на якій зображено символ «f_x» (). Відразу ж після його запуску відкривається перше діалогове вікно, в якому можна вибрати потрібну опцію. У полі «Категория»

вибирається потрібна категорія функцій, а в полі «Выберите Функцию» – назва самої функції. Після вибору функції СУММ(), вибираємо категорію «Математические», рис. 7.3*а*. Крім цього, на стандартній панелі інструментів передбачена кнопка Σ, яка дозволяє реалізувати функцію підсумовування чисел, розташованих в рядку або в стовпці.



а)



б)

Рис. 7.3. Виклик математичної функції СУММ()

Друге вікно ***Мастер функций*** (рис. 7.3*б*) містить поля для введення аргументів обраної функції. Праворуч від кожного поля аргументу відображається поточне значення аргументу. Для закінчення діалогу слід натиснути кнопку «ОК», і створена функція з'явиться в рядку формул

Наведемо деякі інші корисні функції: ЕСЛИ(), СЧЕТЕСЛИ(), СУММЕСЛИ(). Їх назви показують, що вони повертають, рахують і підсумовують не всі значення, а тільки задовольняють деякій умові. Розглянемо застосування цих функцій на конкретному прикладі.

***Приклад* 7.1.** Є дані про доходи від Надання туристичних послуг по регіонах України за 2019 р. За даними Державної служби статистики України ці показники роботи суб'єктів туристичної діяльності (юридичні особи) за регіонами мають вигляд як на рис. 7.4 (стовпці А і В). Потрібно: 1) встановити ті регіони України, доходи яких від надання туристичних послуг перевищують 10 млн. грн.; 2) визначити кількість регіонів, доходи яких від туристичних послуг перевищують 10 млн. грн.; 3) обчислити суму доходів, що перевищують 10 млн. грн.

Для розв'язання першої задачі використовуємо функцію ЕСЛИ(умова; повертане значення). Ця функція використовується при перевірці умов для значень і формул. Повертає одне значення, яке задовольняє умові.

В комірці С4 введемо формулу = ЕСЛИ(B4>10000000; B4; ""). Скопіюємо цю формулу в комірки С5:С28. Якщо число > 10 млн., в комірках С5:С28 повертаються значення комірок В5:В28. Інакше повертається порожній текстовий рядок (""). З рис. 7.4 видно, що доходи ≤ 10 млн. грн. не відображені в комірках С8, С13, С14 і т. д. Це, наприклад, Житомирська, Кіровоградська, Луганська області та ін.

C4 | =ЕСЛИ(B4>10000000; B4; "")

	A	B	C	D	E
1	**Показники роботи суб'єктів туристичної діяльності у 2019 р. за регіонами (юридичні особи)**				
2					
3	Регіон	Дохід від надання туристичних послуг (без ПДВ, акцизів і аналогічних обов'язкових платежів), тис. грн.	Регіони з доходами більше 10 млн. грн.		
4	Вінницька	24 710 600	24 710 600		
5	Волинська	19 476 600	19 476 600		
6	Дніпропетровська	38 778 900	38 778 900		
7	Донецька	14 196 500	14 196 500		
8	Житомирська	6 544 600			
9	Закарпатська	20 751 900	20 751 900		
10	Запорізька	25 389 300	25 389 300		
11	Івано-Франківська	314 013 300	314 013 300		
12	Київська	67 510 200	67 510 200		
13	Кіровоградська	5 710 300			
14	Луганська	1 733 100			
15	Львівська	564 885 500	564 885 500		
16	Миколаївська	4 241 800			
17	Одеська	176 352 700	176 352 700		
18	Полтавська	5 953 600			
19	Рівненська	13 650 400	13 650 400		
20	Сумська	6 218 100			
21	Тернопільська	6 251 200			
22	Харківська	56 494 000	56 494 000		
23	Херсонська	39 200 700	39 200 700		
24	Хмельницька	6 877 900			
25	Черкаська	15 759 400	15 759 400		
26	Чернівецька	19 255 900	19 255 900		
27	Чернігівська	3 483 200			
28	м.Київ	30 491 261 500	30 491 261 500		

Рис. 7.4. Приклад застосування функції ЕСЛИ()

Для розв’язання другої задачі**,** тобто для обчислення кількості регіонів, доходи яких становлять більше 10 млн. грн., застосуємо функцію СЧЕТЕСЛИ(інтервал, умова). Ця функція підраховує у

виділеному інтервалі кількість значень, що задовольняють умові. Введемо в комірку B30 функцію: =СЧЁТЕСЛИ(B4:B28; ">10000000"). Результат показує (рис. 7.5, комірка B30), що лише 16 регіонів заробили за 2019 р. більше 10 млн. грн.

B30 =СЧЁТЕСЛИ(B4:B28; ">10000000")

	A	B	C	D	E	F
4	Вінницька	24 710 600	24 710 600			
5	Волинська	19 476 600	19 476 600			
6	Дніпропетровська	38 778 900	38 778 900			
7	Донецька	14 196 500	14 196 500			
8	Житомирська	6 544 600				
9	Закарпатська	20 751 900	20 751 900			
.....						
25	Черкаська	15 759 400	15 759 400			
26	Чернівецька	19 255 900	19 255 900			
27	Чернігівська	3 483 200				
28	м.Київ	30 491 261 500	30 491 261 500			
29						
30		16				

Рис. 7.5. Приклад застосування функції СЧЕТЕСЛИ()

Третю задачу розв'яжемо за допомогою функції СУММЕСЛИ(інтервал, умова), рис. 7.6. Значення, що задовольняють умові, вибираються з діапазону даних, заданого першим аргументом, і підсумовуються. В комірку B31 введемо формулу =СУММЕСЛИ(B4:B28, ">10000000"). В результаті отримаємо суму тільки тих доходів, які перевищують 10 млн. грн. Для достовірності отриманих результатів можна підсумувати числа діапазону C4:C28.

Фільтрацію даних, які не задовольняють критерію користувача, можна також здійснити за допомогою інструменту Excel – «**Фильтрация**». Інструмент можна запустити, виконавши послідовність команд ***Данные→Фильтр***.

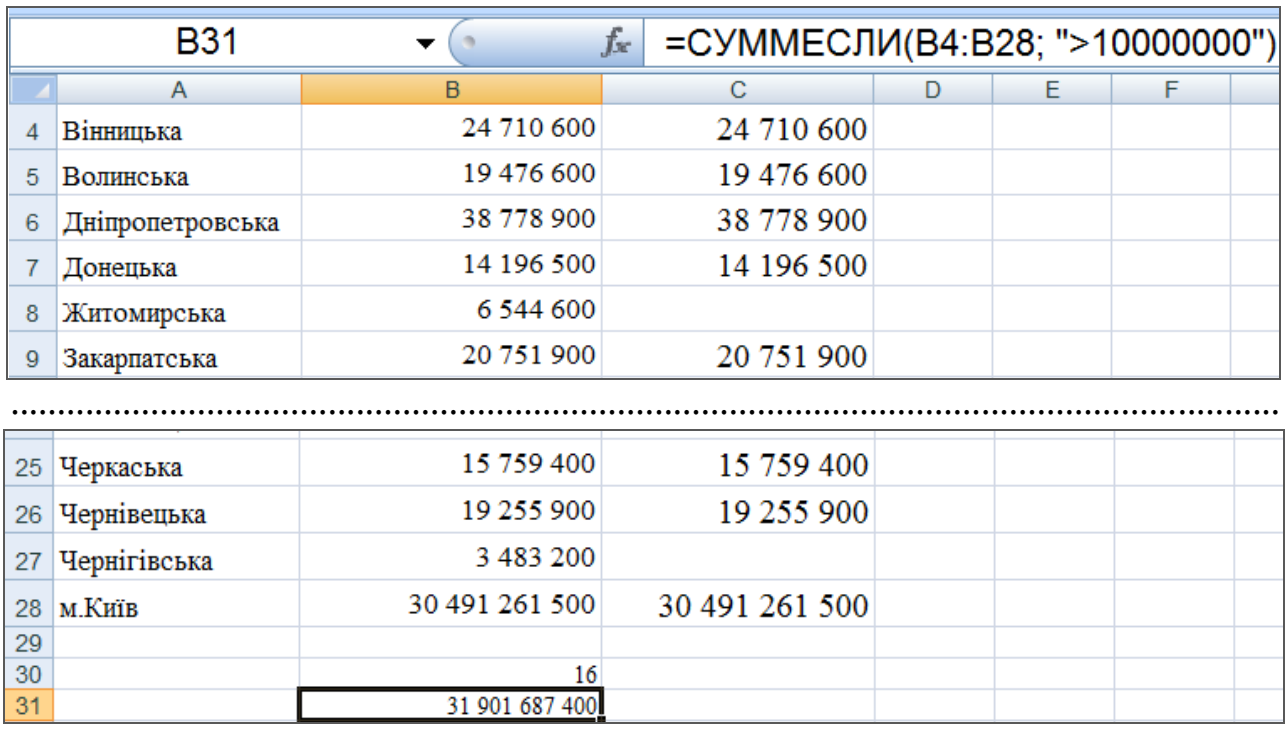
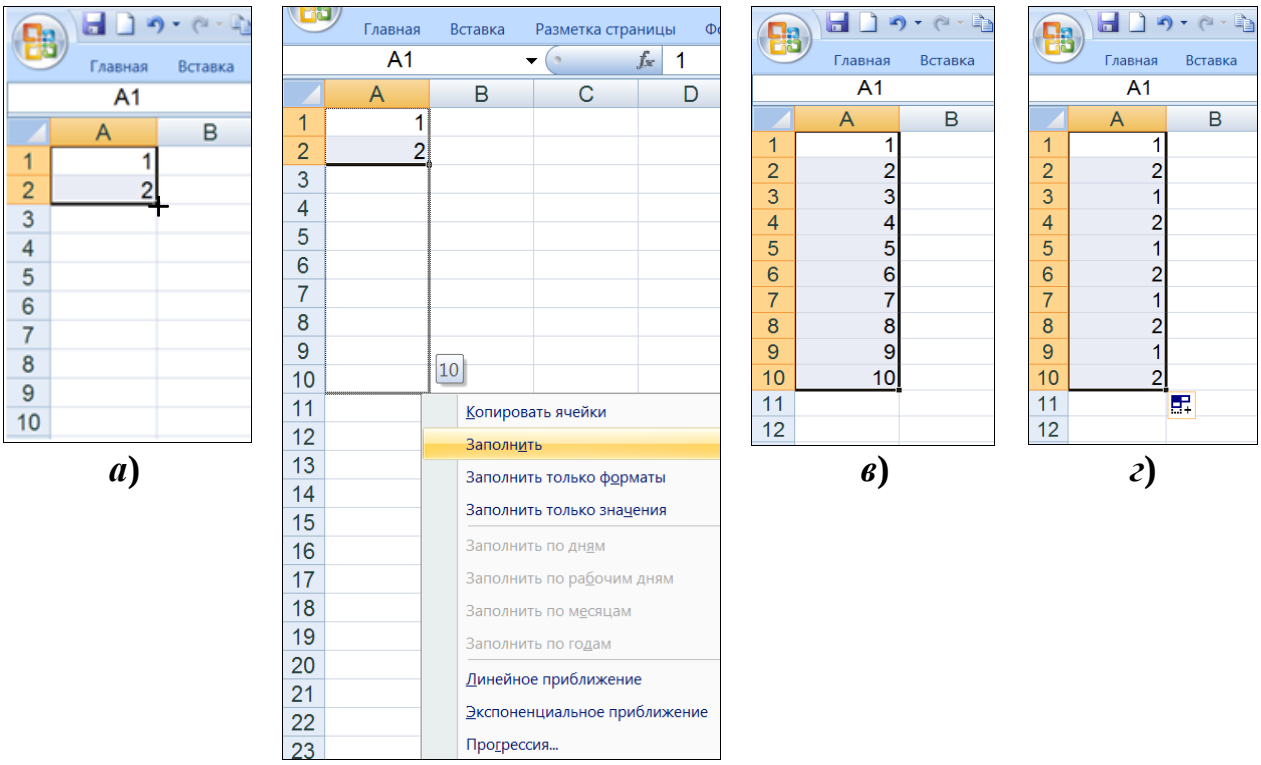


Рис. 7.6. Приклад застосування функції СУММЕСЛИ()

Можна виділити два способи фільтрації: *автофільтр* та *розширений фільтр*. *Автофільтр* дозволяє вибирати окремі записи безпосередньо в робочому аркуші. *Розширена фільтрація* дозволяє використовувати більше критеріїв для відбору необхідної інформації.

Автозаповнення комірок. *Автозаповнення комірок числами* – найбільш показова операція в Excel, яка належить до можливостей функції *перетягування комірок*. Як приклад розглянемо стовпець А, у дві комірки якого А1 і А2 занесені числа 1 і 2 (рис. 7.7*а*). Якщо за допомогою миші виділити ці дві комірки і навести курсор на маленький квадрат в нижньому правому куті виділеної ділянки, курсор перетвориться в маленький чорний хрестик (рис. 7.7*а*). Тоді, натиснувши на ліву кнопку миші, потягнемо курсор вниз і відпустимо кнопку миші в комірці А10. Або, натиснувши на праву кнопку миші, потягнемо курсор вниз і відпустимо кнопку миші в комірці А10. У вікні, що відкриється, вибираємо команду «заполнить» (рис. 7.7*б*). У виділених комірках з’явиться послідовний ряд чисел, як це показано на рис. 7.7*в*.



а) *б*) *в*) *г*)

Рис. 7.7. Автозаповнення комірок

Іншим способом автозаповнення комірок є використання клавіші <Ctrl>. Натискаємо на клавішу <Ctrl>, встановлюємо курсор на маленький квадрат в нижньому правому куті виділеної зони. Після появи хрестика, натиснувши на ліву кнопку миші, тягнемо курсор вниз і відпускаємо кнопку в останній потрібній комірці. Тоді в виділених комірках з’явиться вже не послідовний ряд чисел, як на рис. 7.7*в*, а повторюваний набір значень, які були в перших двох комірках (рис. 7.7*г*). Для редагування (зміни) формули і, взагалі, даних в комірці потрібно виділити комірку і натиснути функціональну клавішу F2 (редагування). Редагування завершується так само, як і введення. Замість натискання F2 можна також двічі клацнути по редагованій комірці мишею.

Абсолютні і відносні адресації (***посилання***). За замовчуванням посилання на комірки в формулах розглядаються як *відносні.* У цьому разі місце розташування активної комірки є початковим, адреси інших комірок вказуються відносно активних. Це означає, що при копіюванні формули, адреси в посиланнях автоматично змінюються

відповідно до відносного розташування вихідної комірки і створюваної копії. Нехай, наприклад, в комірці B1 є посилання на комірку A1. У відносному представленні можна сказати, що посилання вказує на комірку, яка розташовується на один стовпець лівіше даної. Якщо формула буде скопійована в іншу комірку, наприклад, в B2, то таке відносне вказання посилання збережеться, тобто при копіюванні формули в комірку B2 посилання буде продовжувати вказувати на комірку, що розташовується лівіше, в даному разі на комірку A2.

При *абсолютній адресації* адреси посилань при копіюванні не змінюються, тобто абсолютний спосіб адресації дозволяє однозначно визначати в формулах адреси комірок. Абсолютна адреса завжди позначається знаком долара «$». Наприклад, запис $A$8 незмінно означає адресу комірки, що стоїть на перетині стовпця A і рядка 8. Для автоматичної абсолютної адресації, наприклад, комірки A8, необхідно виділити посилання на цю комірку і натиснути клавішу F4.

Таким чином, при копіюванні формул відносні адреси комірок змінюються, тоді як абсолютні адреси не змінюються.

7.2. Побудова діаграм в MS Excel

В MS Excel є набір стандартних типів і видів діаграм. Діаграми можна створювати одним із таких способів: як рисунок на одному робочому аркуші з даними (рис. 7.8), тоді діаграма називається *вбудованою*; на окремому аркуші робочої книги без даних, і цей аркуш називається *аркушем діаграми*.

Для створення діаграм в MS Excel необхідно запустити ***Мастер диаграмм***, натиснувши кнопку  на стандартній панелі інструментів, або виконати послідовність команд ***Вставка*** → ***Диаграмма***. Після цього необхідно вибрати відповідний «Тип» діаграми. Кожен тип діаграми має ще кілька видів.

Наприклад, на рис. 7.8 за допомогою *Майстра діаграми* як тип діаграми вибрана «Гистограмма с группировкой», що характеризує

кількість громадян України, які виїжджали за кордон (з метою туризму) за період 2006-2017 рр.

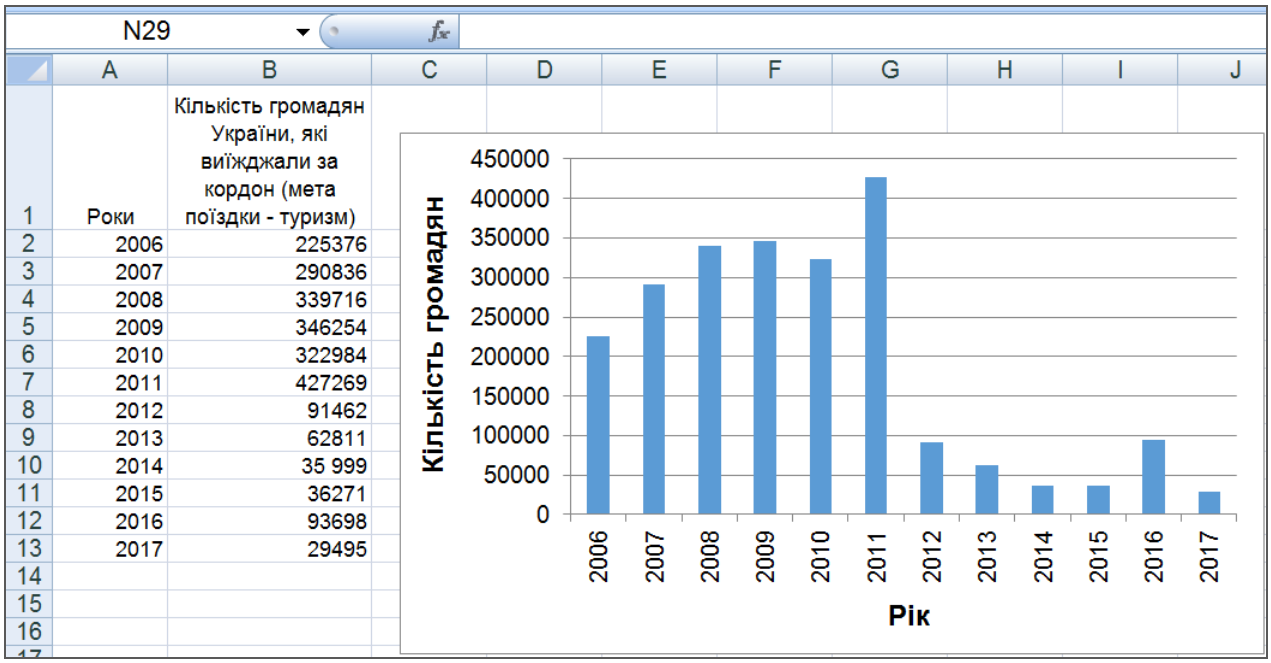
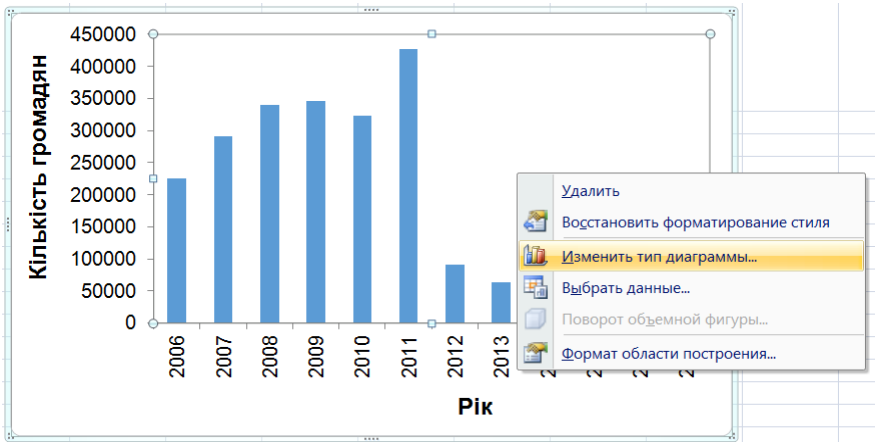


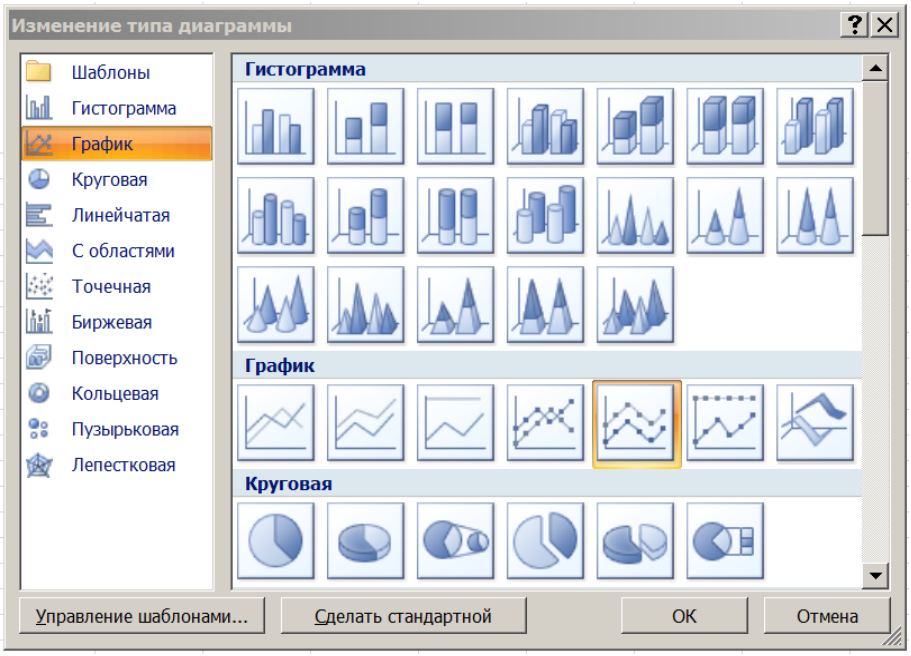
Рис. 7.8. Діаграма на одному робочому аркуші з даними

Другим кроком у створенні діаграми є задання діапазону, що містить дані, які будуть представлені на діаграмі. Якщо в діаграмі передбачається використовувати заголовки рядків і стовпців, то їх слід також включити в виділений діапазон. Якщо цього не було зроблено, або потрібно змінити зроблений вибір, то можна скористатися командою «Выбрать данные», попередньо натиснувши на праву кнопку миші в полі діаграми (рис. 7.9). Тут же можна «*Изменить тип диаграммы*» (рис. 7.9*а*) і діапазон даних. Наприклад, на рис. 7.9 *Гістограму* замінили на «График с маркерами и накоплением».

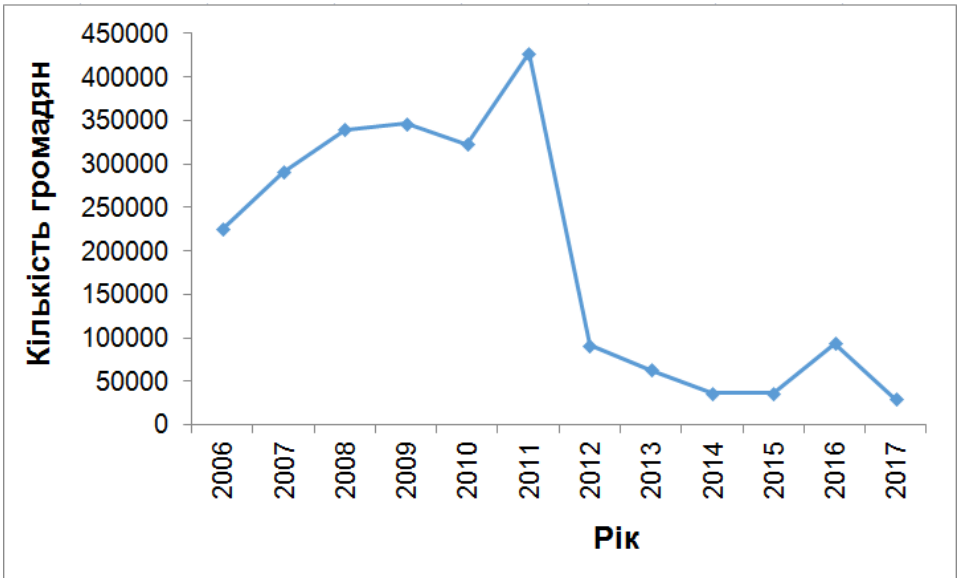
Далі можна додати легенду, якщо *Майстер* цього не зробив; можна підписати осі, дати діаграмі назву, підписати дані і виконати деякі інші операції. Для реалізації цих операцій у вікні *Майстра* є вкладки: «Название диаграммы», «Название осей», «Легенда», «Подписи данных», «Оси» та ін.



а)



б)



в)

Рис. 7.9. Зміна типу діаграми

Вибір варіанта розташування діаграми (*на окремому аркуші* або *на робочому аркуші з таблицею*) здійснюється на останньому кроці роботи *Майстра діаграм* у вікні «Размещение диаграммы» (рис. 7.10).

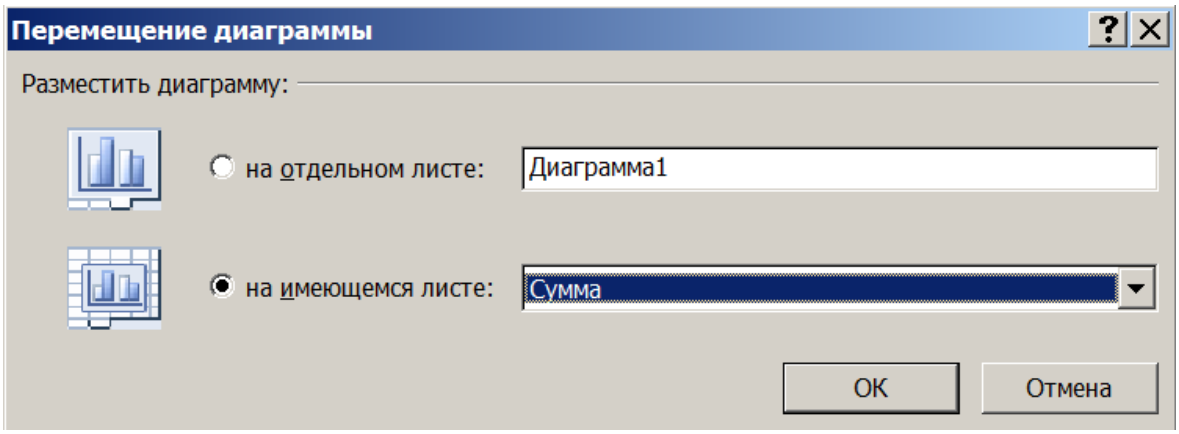


Рис. 7.10. Вікно розташування діаграми

7.3. Засоби статистичної обробки в MS Excel

Середовище Microsoft Excel має вбудовані засоби (інструменти) із статистичної обробки різних даних. Зокрема, в ньому є велика кількість *вбудованих статистичних функцій* і програмна надбудова **«Пакет анализа»**, призначена для обробки даних і розв'язання задач математичної статистики. Коротко розглянемо ці засоби.

Перш ніж перейти до інструментів статистичного аналізу, розглянемо деякі *статистичні функції* MS Excel.

Вбудовані статистичні функції **MS Excel.** Перше, що завжди робиться при статистичній обробці даних, це обчислення елементарних статистичних характеристик вибірок (як мінімум: середнього, стандартного відхилення) по кожній групі. В MS Excel є низка спеціальних функцій, які призначені для обчислення цих вибіркових характеристик. Перш за все, це функції, що характеризують центр розподілу: СРЗНАЧ() – обчислює середнє арифметичне; МЕДИАНА() – дозволяє отримати медіану досліджуваної вибірки; МОДА() – обчислює значення, що найбільш часто зустрічається у вибірці.

До функцій, які обчислюють показники, що характеризують розсіювання варіант, зараховуються: ДИСП() – дозволяє оцінити дисперсію за вибірковими даними; СТАНДОТКЛОН() – обчислює стандартне відхилення.

Форму емпіричного розподілу дозволяють оцінити функції: СКОС() – оцінює ступінь несиметричності розподілу щодо середнього вправо і вліво; ЭКСЦЕСС() – обчислює величину ексцесу за вибірковими даними, що характеризує ступінь гостровершинності кривої розподілу.

Використовуючи статистичні функції, обчислимо деякі статистичні характеристики вибірки, наведеної в таблиці на рис. 7.11.

	A	B	
1	**Витрати суб'єктів туристичної діяльності (юридичні особи), у 2019 р. за деякими регіонами**		
2	на розміщення і проживання у готелях		
3	Вінницька	1 101	
4	Волинська	1 657	
5	Дніпропетровська	15 692	
6	Донецька	8 139	
7	Закарпатська	7 529	
8	Запорізька	1 108	
9	Київська	34 916	
10	Одеська	52 324	
11	Рівненська	3 829	
12	Харківська	18 827	
13	Черкаська	11 969	
14	Чернівецька	14 250	
15			
16	**Мін**	1 101	=МИН(B3:B14)
17	**Макс**	52 324	=МАКС(B4:B15)
18	**Середнє**	14278,3	=СРЗНАЧ(B3:B14)
19	**Станд. відхилення**	15374,8	=СТАНДОТКЛОН(B3:B14)

Рис. 7.11. Розрахунок деяких характеристик вибірки за допомогою статистичних функцій МИН(), МАКС(), СРЗНАЧ(), СТАНДОТКЛОН()

Надбудова «Анализ данных». Список статистичних функцій та

їхній опис можна переглянути через **«Мастер функций»**, вибравши в ньому категорію «*Статистические*». *Майстер функцій* відкривається при натисканні на кнопку f_x у рядку формул.

Для активації надбудови **«Пакет анализа»** (модуля), наприклад, в MS Excel 2007 слід натиснути кнопку «Microsoft Office» (в лівому верхньому кутку головного вікна), а потім кнопку **«Параметры Excel»** (рис. 7.12).

У вікні **«Параметры Excel»**, у списку команд лівої частини вікна потрібно вибрати команду «Надстройки», у списку «Управление надстройками» вибрати позицію «Надстройки Excel» і натиснути кнопку «Перейти», рис. 7.13*а*. Після цього з'явиться вікно «Надстройки», в якому необхідно встановити прапорець **«Пакет анализа»** (рис. 7.13*б*) і натиснути кнопку «ОК».

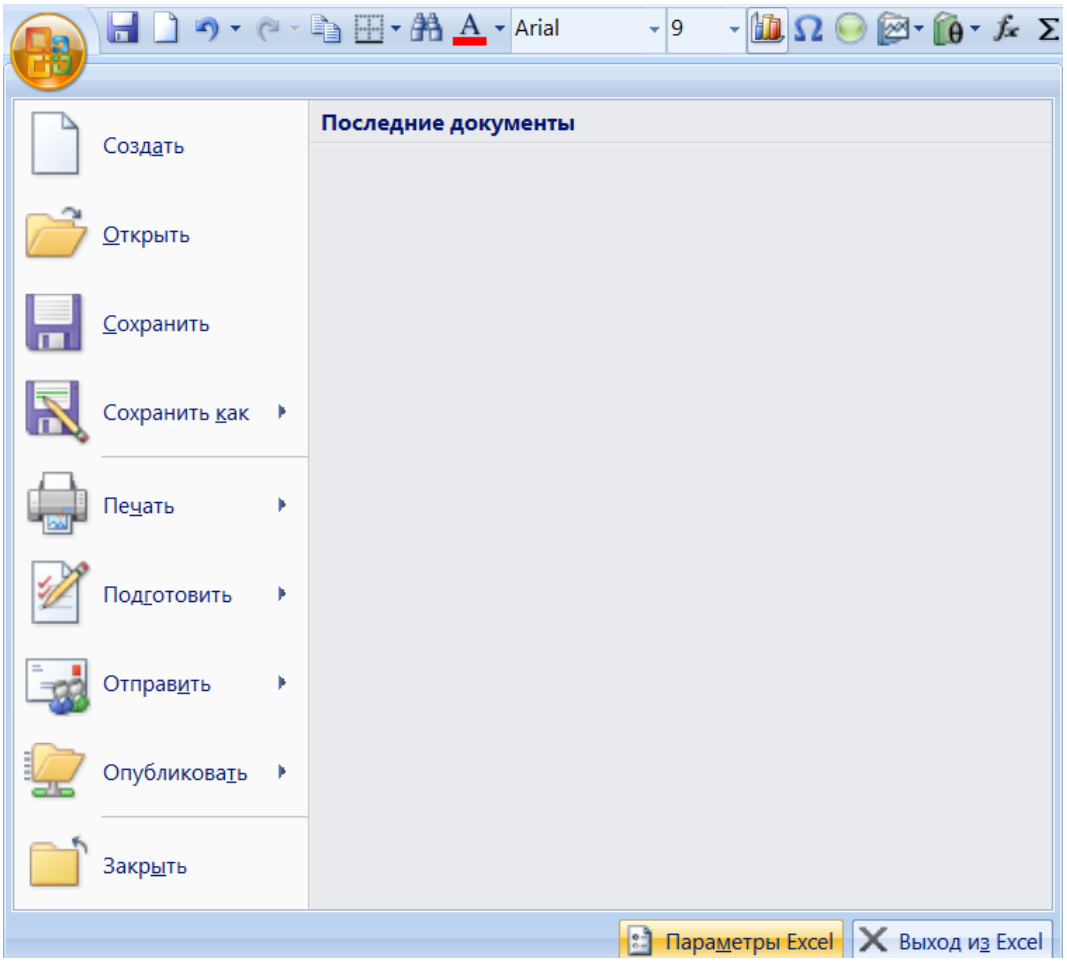
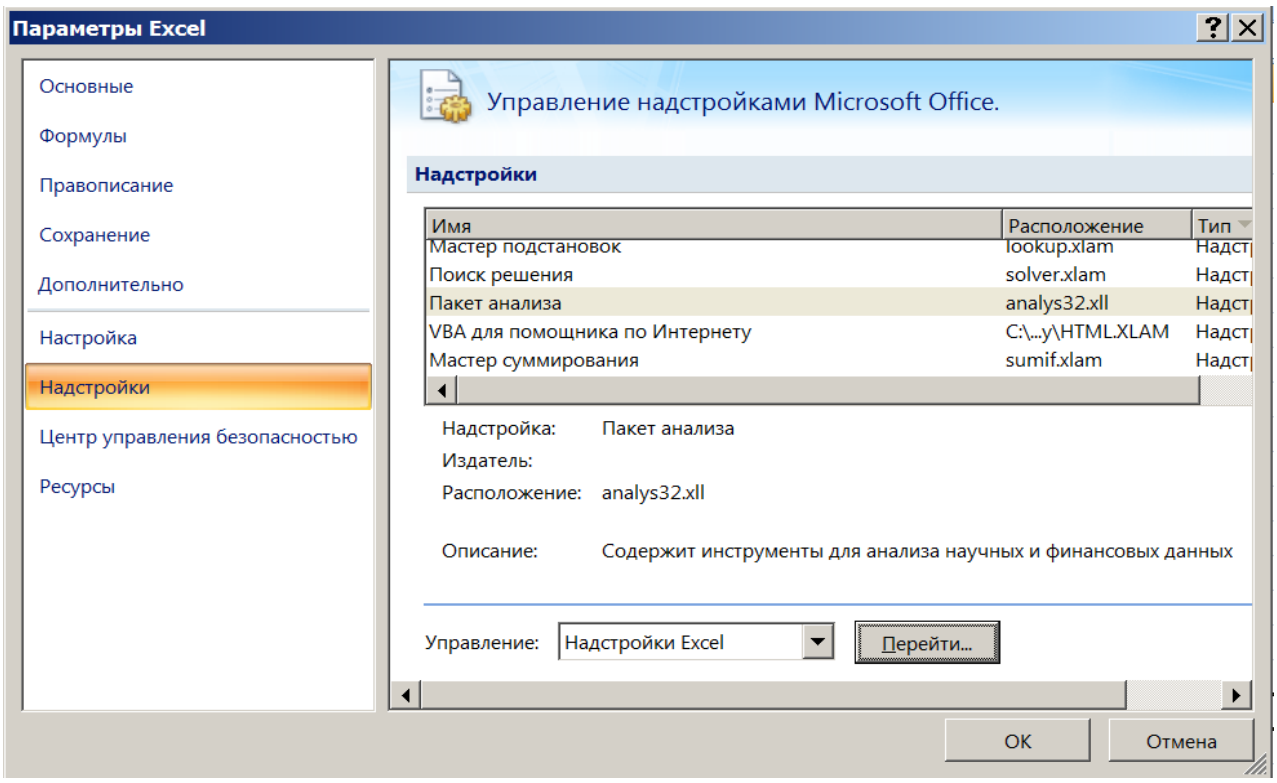
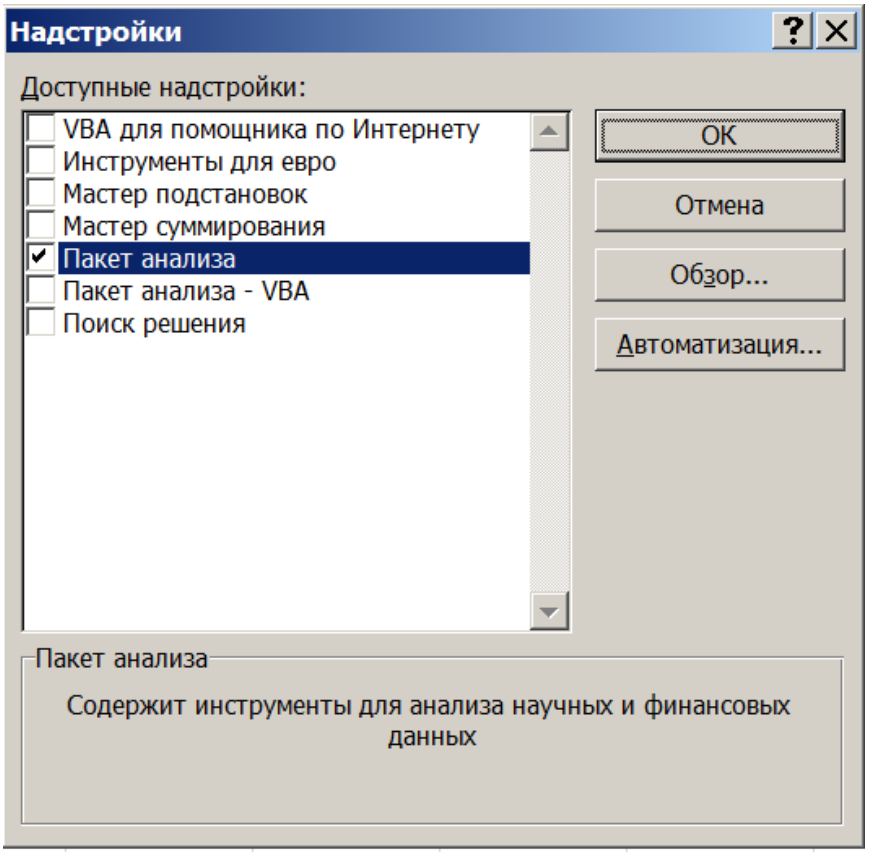


Рис. 7.12. Початок надбудови «Пакет анализа»



а)



б)

Рис. 7.13. Надбудови MS Excel і активація надбудови «Пакет анализа»

В результаті підключення «Надстройки» на вкладці «Данные» у групі «Анализ» стане доступною команда «Анализ данных», рис. 7.14. При виборі цієї команди буде відкриватися вікно «Анализ данных» і **«Инструменты анализа».**

Таким чином, для використання статистичного пакету аналізу даних необхідно: виконати команду «Анализ данных»; вибрати необхідний рядок в списку «Инструменты анализа»; ввести вхідний та вихідний діапазони і вибрати необхідні параметри.

Описова статистика. Крім перерахованих вище функцій, для роботи з декількома вибірками і обчислення їх статистичних характеристик Excel містить інструмент «Описательная статистика» з «Пакета анализа».

Інструмент «Описательная статистика» створює звіт, що містить статистичні характеристики досліджуваної вибірки. Після відкриття вікна «Анализ данных», виберемо рядок «Описательная статистика» і натиснемо кнопку «ОК» (рис. 7.15).

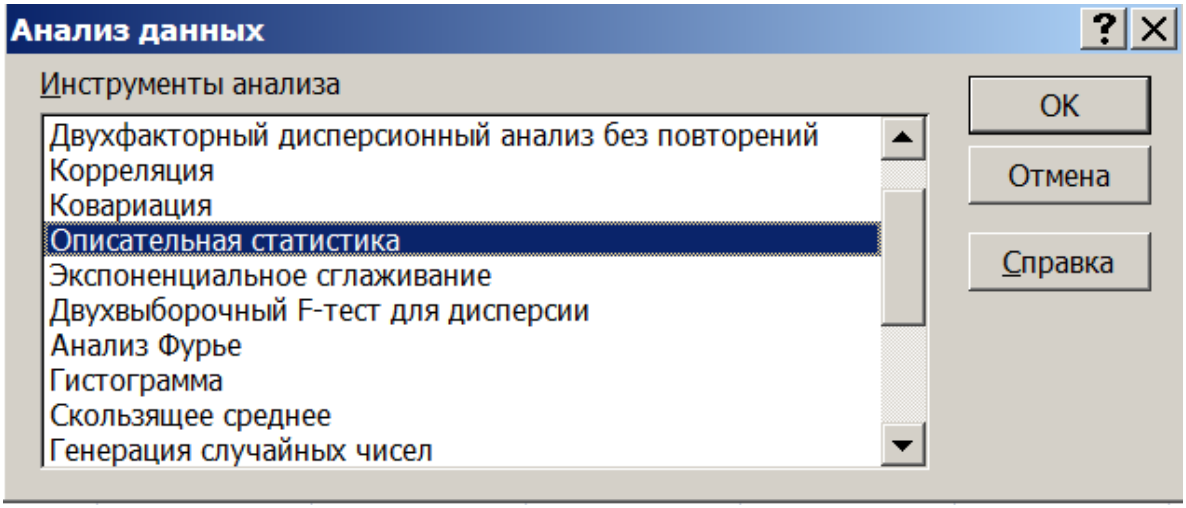


Рис. 7.14. Інструменти статистичного аналізу

У діалоговому вікні, що з’явилося, необхідно вказати: «Входной диапазон», тобто посилання на комірки, що містять аналізовані дані; «Выходной диапазон», тобто посилання на комірки, в які будуть виведені результати аналізу. У розділі «Группирование» необхідно

встановити прапорець в положення «за стовпцями». Крім цього, потрібно встановити прапорець в полі «Итоговая статистика» і натиснути на кнопку «ОК».

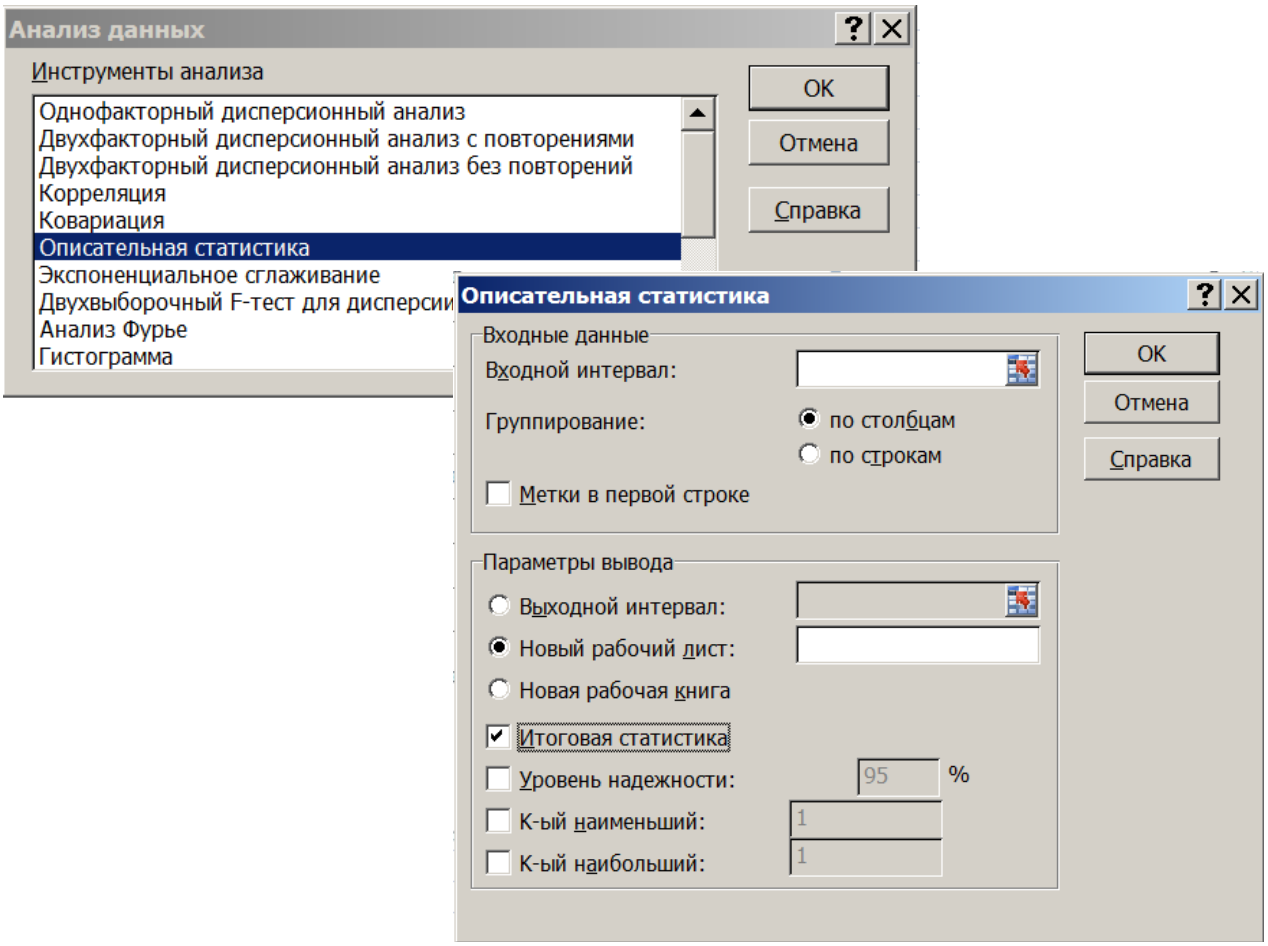


Рис. 7.15. Вибір «Инструмента анализа» даних і вікно «Описательная статистика»

В результаті аналізу в зазначеному вихідному діапазоні для кожного стовпця даних виводяться такі статистичні характеристики як: середнє, стандартна помилка (середнього), медіана, мода, стандартне відхилення, ексцес, асиметрія, рівень надійності та ін.

Крім зазначених засобів описової статистики у «Пакет анализа» MS Excel включені також засоби проведення кореляційно-регресійного аналізу, критерії відмінності та інші засоби, що дозволяють проводити статистичний аналіз різних типів даних.

Обмежимося найпростішими і найбільш часто використовуваними засобами, реалізованими в «Мастер функций» і

«Пакет анализа» MS Excel.

Для побудови, наприклад, вибіркових функцій розподілу використовуються спеціальна функція ЧАСТОТА() і процедура «Пакет анализа» «Гистограмма». При цьому весь діапазон зміни випадкової величини розбивається на інтервали рівної ширини, які називають *кишенями*. За кількістю попадань значень випадкової величини в кожну кишеню обчислюються частоти, за якими і будується «Гистограмма» вибіркової функції розподілу статистичних ймовірностей.

Функція ЧАСТОТА() задається як формула масиву: ЧАСТОТА(массив_данных; массив_карманов), де *масив_даних* – це масив або посилання на множину даних, для яких обчислюються частоти; *масив_кишень* – це масив або посилання на множину інтервалів, в які групуються значення аргументу масиву даних.

Розглянемо приклад побудови вибіркового розподілу за даними про щоденні продажі турпакетів деяких туроператорів.

***Приклад* 7.2.** Є дані про щоденні продажі турпакетів деяких туроператорів за два місяці. На рис. 7.16 наведена вибірка за два місяці і діапазон кишень – граничних значень. Потрібно побудувати вибірковий розподіл денних продажів турпакетів і визначити основні статистичні характеристики для представлених даних. Дані будуть групуватися в інтервали 0-96, 97-101, 102-106 і т. д. При підрахунку в кишеню зараховуються значення на нижній межі і не зараховуються значення на верхній межі.

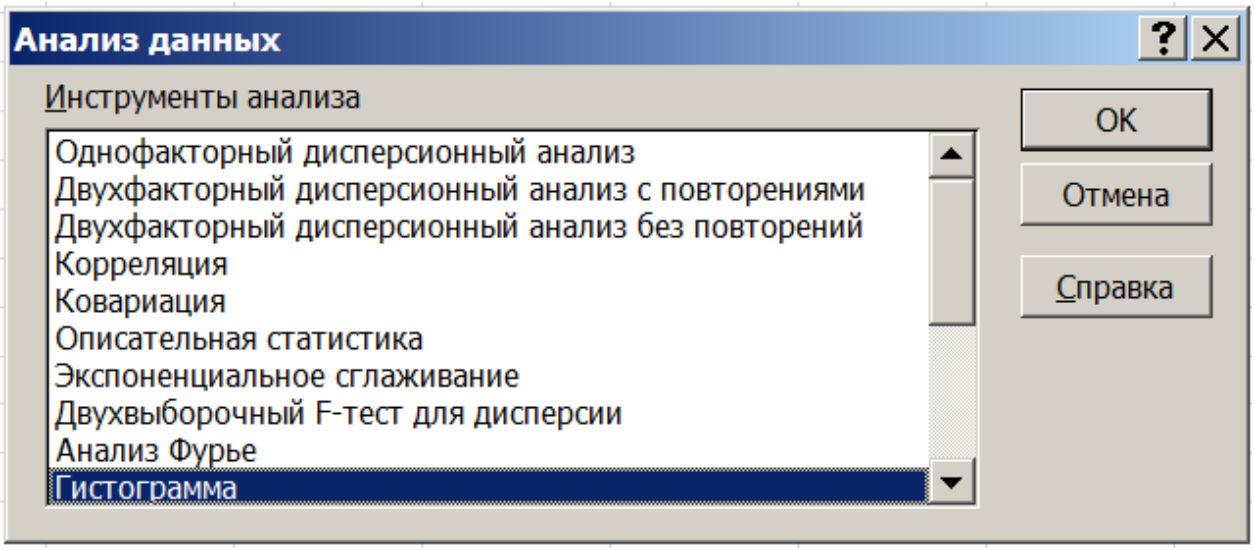
Побудуємо вибірковий розподіл денних продажів турпакетів, використовуючи інструмент «Гистограмма». Виклик інструменту – через меню ***Данные»→Анализ данных.***

Параметри діалогового вікна «Гистограмма» наведені на рис. 7.17*а*. В полі «Входной интервал» вводиться діапазон даних спостережень (A3:D17), рис. 7.17*б*. В полі «Интервал карманов» (необов’язковий параметр) вводиться діапазон комірок, які визначають вибрані інтервали (*кишені*); в нашому випадку це

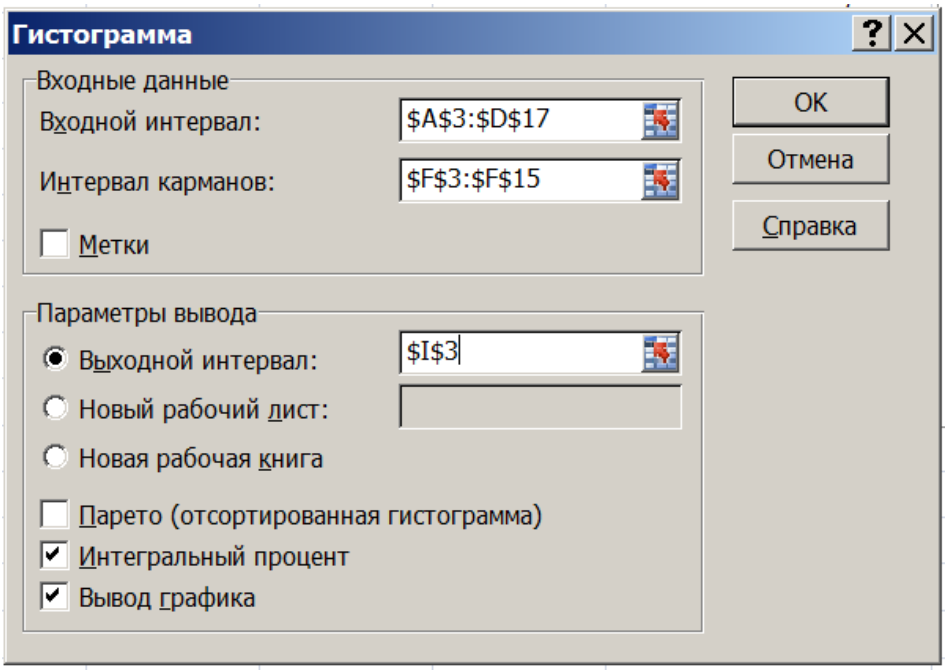
діапазон даних (F3:F15), рис. 7.17*б*. Якщо діапазон кишень не був введений, то набір інтервалів, рівномірно розподілених між мінімальним і максимальним значеннями даних, буде створено автоматично.

	A	B	C	D	E	F
1		Денні продажи				Кишені
2						
3	100	104	116	115		96
4	128	113	130	125		101
5	117	123	125	119		106
6	135	116	122	137		111
7	128	116	110	118		116
8	125	149	129	134		121
9	122	125	134	136		126
10	138	114	135	153		131
11	135	96	126	156		136
12	133	102	121	129		141
13	117	123	138	123		146
14	137	127	132	120		151
15	132	128	143	121		156
16	126	129	140	136		
17	121	129	127	106		

Рис. 7.16. Кількість турпакетів, проданих за два місяці



***а*)**



***б*)**

Рис. 7.17. Виклик і заповнення діалогового вікна «Гистограмма»

Робоче поле «Выходной интервал» призначене для введення посилання на ліву верхню комірку вихідного діапазону. Розмір вихідного діапазону буде визначено автоматично. «Интегральный процент» (рис. 7.17*б*) дозволяє встановити режим генерації інтегральних відсоткових відношень і включення в гістограму графіка інтегральних відсотків. «Вывод графика» (рис. 7.17*б*) дозволяє встановити режим автоматичного створення вбудованої діаграми на аркуші (рис. 7.18), що містить вихідний діапазон.

Кишеня	*Частота*	*Інтегральний %*
96	1	1,67%
101	1	3,33%
106	3	8,33%
111	1	10,00%
116	6	20,00%
121	8	33,33%
126	11	51,67%
131	10	68,33%
136	10	85,00%
141	5	93,33%
146	1	95,00%
151	1	96,67%
156	2	100,00%
Ще	0	100,00%

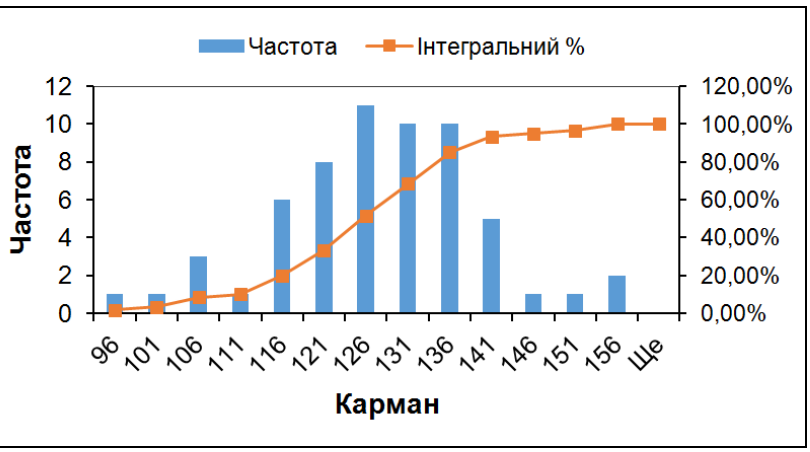


Рис. 7.18. Результат обчислення абсолютних, накопичених частот та їх діаграми

Для визначення основних статистичних характеристик в двох групах даних потрібно в «Инструменты анализа» вибрати рядок «Описательная статистика» (рис. 7.17*а*). У діалоговому вікні, що з'явиться (рис 7.19), в робочому полі «Входной интервал» зазначимо вхідний діапазон даних – A2:B31. В робочому полі «Выходной интервал» зазначимо вихідний діапазон – комірку C2. В розділі «Группирование» вибираємо положення «за стовпцями». Встановлюємо прапорець в полі «Итоговая статистика» і натискаємо кнопку «ОК».

В результаті аналізу у зазначеному вихідному діапазоні для кожного стовпця даних отримаємо відповідні характеристики (рис. 7.20), серед яких найважливішими є *Середнє*, *Стандартна помилка* (середнього) і *Стандартне відхилення*.

Слід зазначити, що інструмент «Описательная статистика» обчислює повний довірчий інтервал вибірки. На рис. 7.20, наприклад, для одного місяця він дорівнює 53, для іншого – 60. Оскільки середнє арифметичне вибірки зазвичай не співпадає із середнім арифметичним генеральної сукупності, то актуальним є визначення прийнятних меж зміни середнього арифметичного вибірок – *довірчого інтервалу середнього*.

Відомо, що в межі $[\bar{x} \pm \sigma]$ нормально розподілена випадкова величина потрапляє з довірчою ймовірністю 0,68 (68 %), в межі $[\bar{x} \pm 2\sigma]$ – з імовірністю 0,96 (96 %), в межі $[\bar{x} \pm 3\sigma]$ – з імовірністю 0,998 (99,8 %), де σ – стандартне відхилення від середнього.

Отже, можна стверджувати, що в 95 % випадків значення вибірки потраплять в довірчий інтервал [122,9-26,5; 122,9+26,5] (для стовпця 1) і [124,1-30; 124,1+30] (для стовпця 2), рис. 7.20.

Для обчислення напівширини довірчого інтервалу середнього за заданим рівнем значущості, стандартним відхиленням і числом значень у вибірці потрібно використати функцію ДОВЕРИТ().

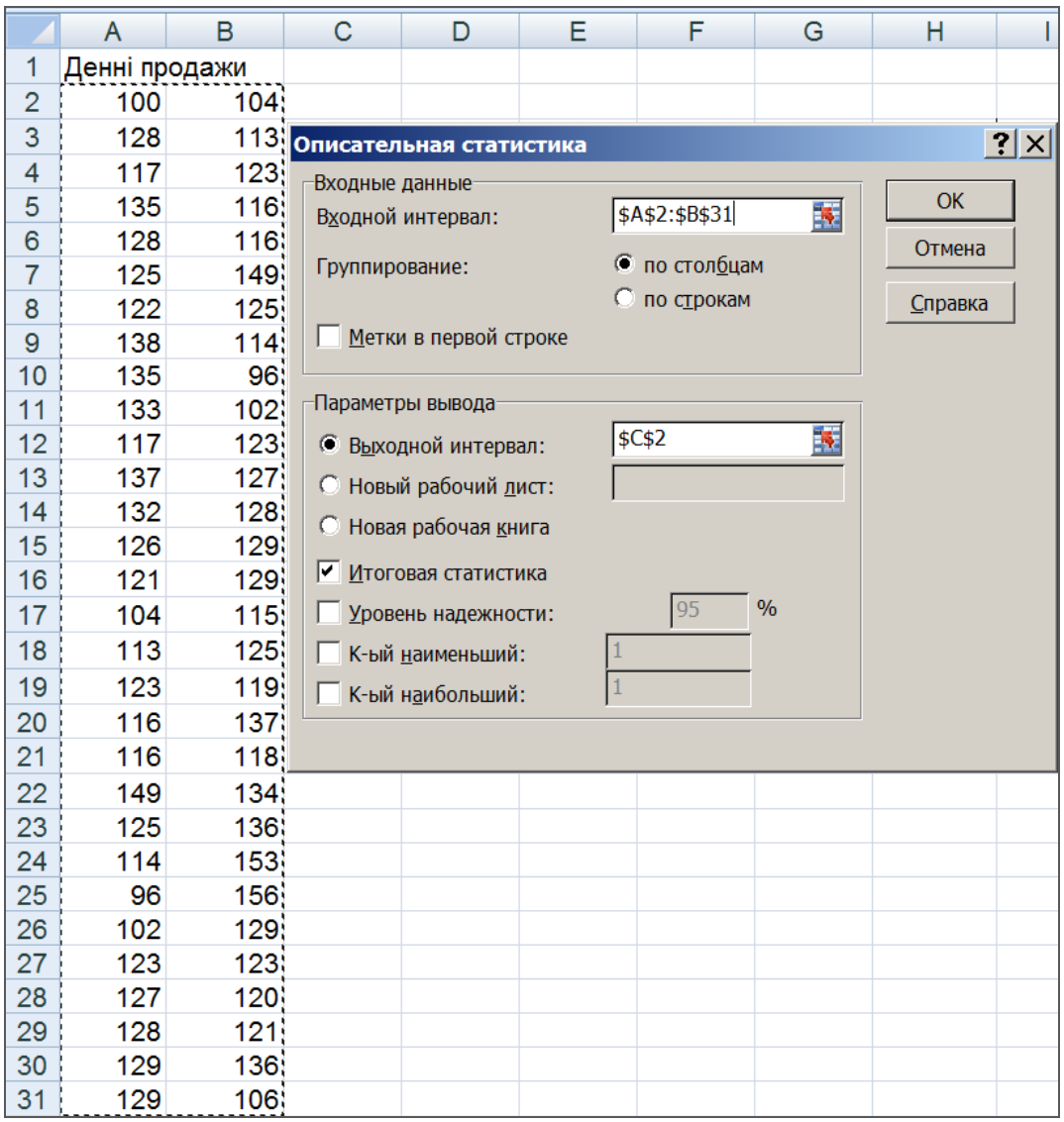


Рис. 7.19. Таблиця з прикладу 7.2 і застосування інструменту «Описательная статистика»

H	I	J	K
Столбец1		*Столбец2*	
Среднее	122,9	Среднее	124,1
Стандартная ошибка	2,187	Стандартная ошибка	2,551
Медиана	125	Медиана	123
Мода	128	Мода	123
Стандартное отклонение	11,98	Стандартное отклонение	13,97
Дисперсия выборки	143,5	Дисперсия выборки	195,2
Эксцесс	0,295	Эксцесс	0,423
Асимметричность	-0,434	Асимметричность	0,377
Интервал	53	Интервал	60
Минимум	96	Минимум	96
Максимум	149	Максимум	156
Сумма	3688	Сумма	3722
Счет	30	Счет	30

Рис. 7.20. Результати роботи інструменту «Описательная статистика»

7.4. Прийняття статистичних рішень

Перевірка відповідності теоретичному розподілу. Одним з завдань статистичного аналізу є оцінка ступеня відповідності (розбіжності) отриманих емпіричних даних (вибірки) відомому теоретичному розподілу, зокрема, нормальному розподілу. Це пов'язано з тим, що при вирішенні реальних завдань закон розподілу і його параметри здебільшого невідомі. Водночас застосовувані статистичні методи як передумови часто вимагають визначеного закону розподілу.

Найчастіше перевіряється припущення про нормальний розподіл генеральної сукупності, оскільки більшість статистичних процедур орієнтована на вибірки, отримані з нормально розподіленої генеральної сукупності.

Для оцінки відповідності наявних експериментальних даних нормальному закону розподілу зазвичай використовують графічний метод, вибіркові параметри форми розподілу і критерії згоди.

Графічний метод дозволяє давати орієнтовну оцінку розбіжності або збігу розподілів.

При великій кількості спостережень ($n > 100$) непогані результати дає обчислення вибіркових параметрів форми розподілу: ексцесу і асиметрії. Прийнято говорити, що припущення про нормальність розподілу суперечить наявним даним, якщо асиметрія близька до нуля, тобто лежить в діапазоні від –0,2 до 0,2, а ексцес – від –1 до +1.

Розглянемо приклад оцінки відповідності експериментальних даних нормальному закону розподілу.

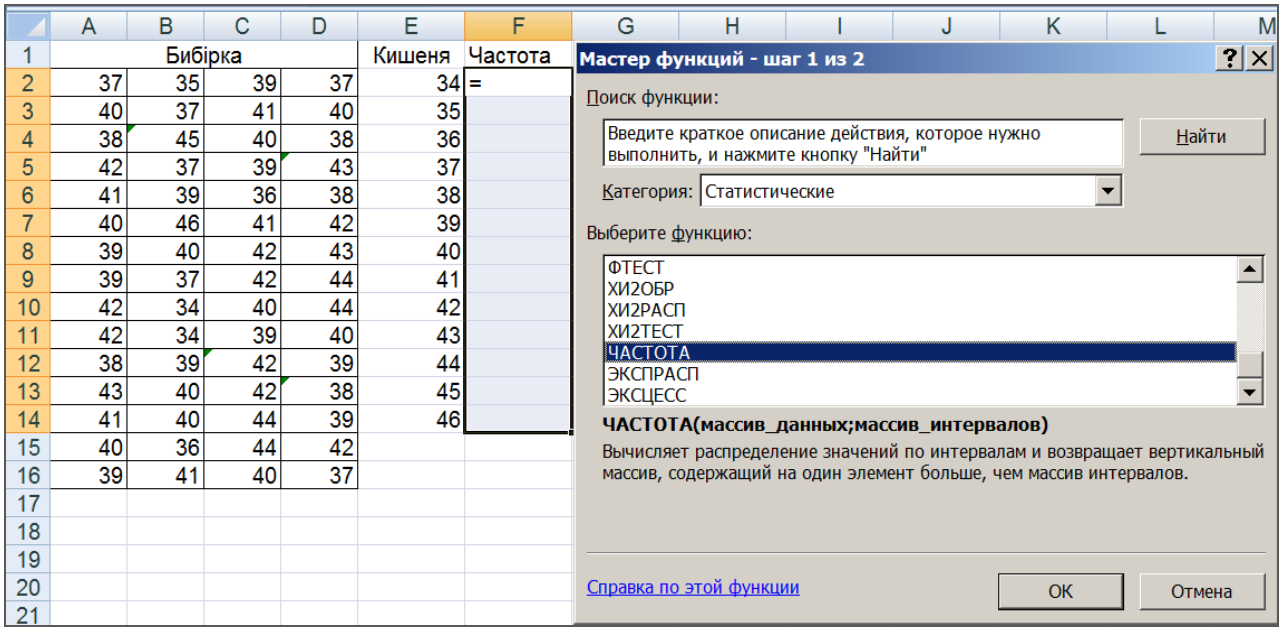
***Приклад* 7.3.** Нехай є деяка вибірка, для якої необхідно візуально оцінити міру її відповідності нормальному розподілу. Для цього потрібно побудувати графіки вибіркового і теоретичного розподілу ймовірностей (частот).

Вводимо вихідні дані в комірки A2:D16 (рис. 7.21*а*). У стовпці E введемо інтервали кишень і за допомогою функції ЧАСТОТА() у

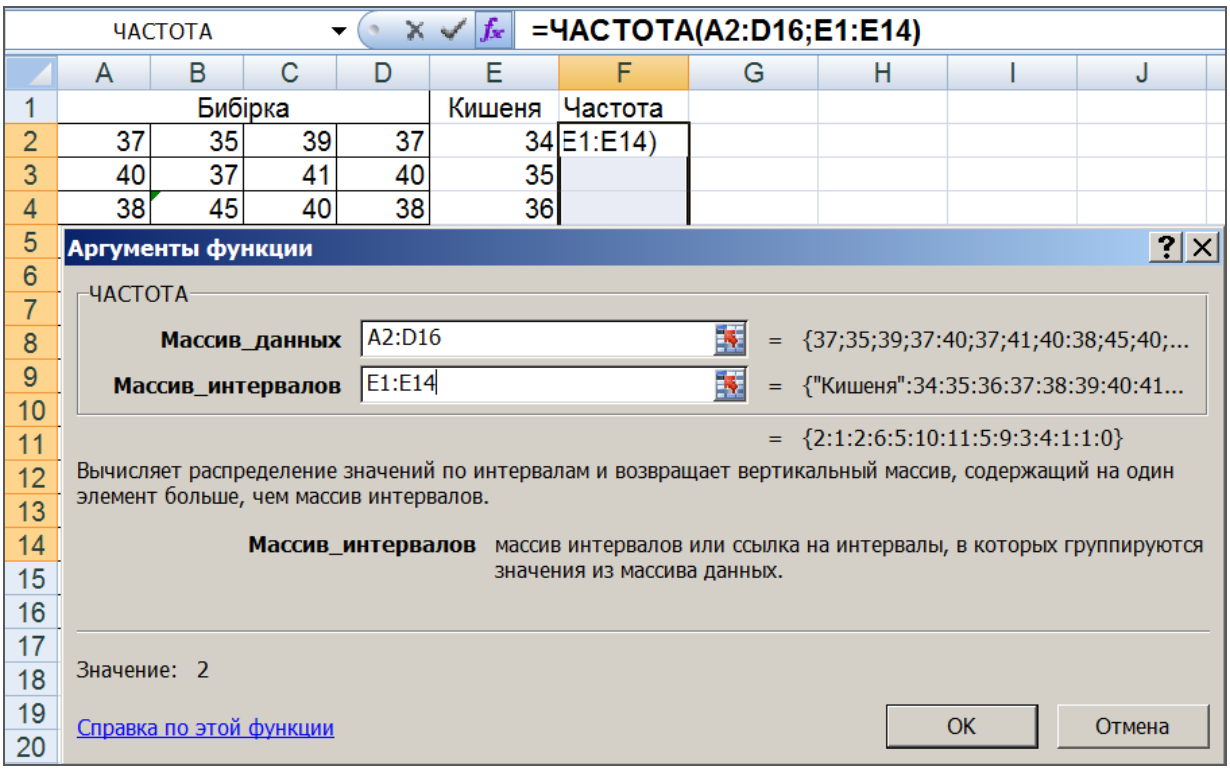
стовпці F обчислимо абсолютні частоти значень вибірки.

Після виклику функції ЧАСТОТА() у вікні, що відкрилося, в полі «Массив_данных» виберемо діапазон A2:D16 (7.21*б*), а в полі «Массив_интервалов» – діапазон кишень E2:E14. Після цього необхідно завершити дію не натисканням клавіші <Enter>, а натисканням комбінації клавіш <Ctrl + Shift + Enter>. Таким чином ми повідомимо Excel, що потрібно виконати операцію над масивом, тобто створити формулу масиву. У відповідь Excel автоматично візьме формулу в фігурні дужки: {=ЧАСТОТА(A2:D16;E1:E14)} і зробить необхідні обчислення з елементами масиву.

Для обчислення теоретичних частот і побудови теоретичного нормального розподілу ймовірностей в комірку H1 запишемо функцію =НОРМРАСП(E2;СРЗНАЧ($E$2:$E$14);СТАНДОТКЛОН($E$2:$E$14);0), рис. 7.22. Потім копіюємо її на діапазон E3:E14.



***а*)**



б)

Рис. 7.21. Застосування функції ЧАСТОТА()

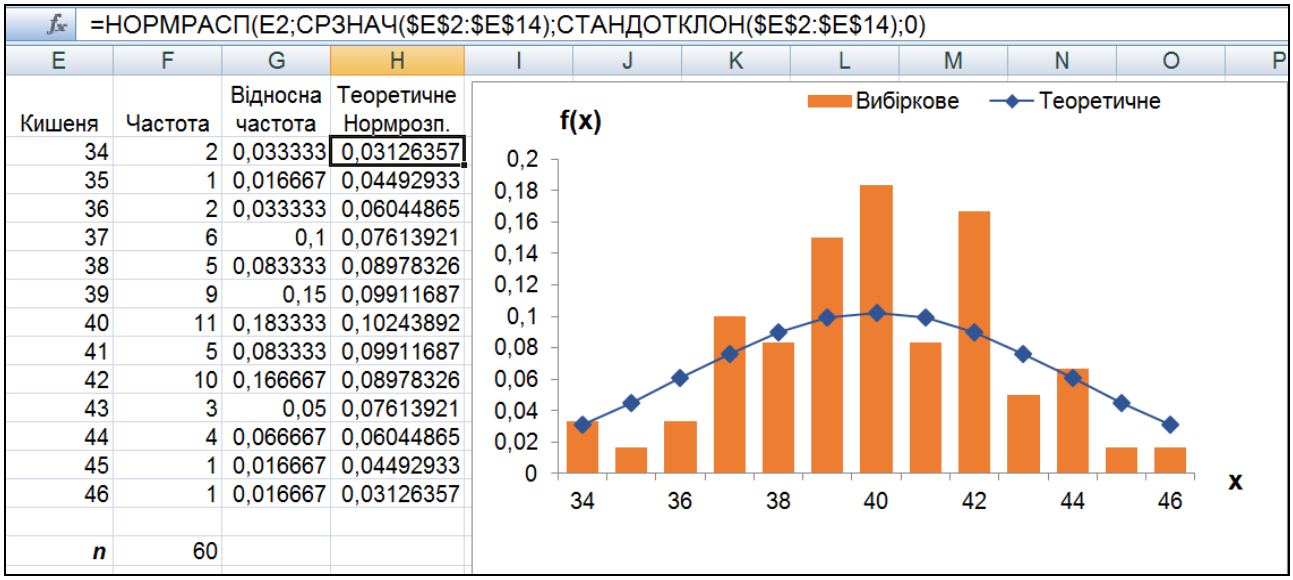


Рис. 7.22. Зіставлення вибіркового розподілу даних і кривої нормального розподілу

Функцію НОРМРАСП(), як і інші функції, можна ввести як вручну, так і викликати з групи «Статистические». Стандартна форма функції має такий вигляд: НОРМРАСП(*X*; *Среднее*; *Стандартное_откл.*; *Интегральная*), рис. 7.23. Тут *X* – значення

аргументу, на основі якого обчислюється нормальний розподіл; *Среднее* – середнє значення розподілу; *Стандартное_откл.* – стандартне відхилення розподілу.

При *Интегральной* = 0 розраховується щільність $f(x)$, а при =1 – функція $F(x)$ нормального розподілу.

Далі потрібно протягуванням скопіювати вміст комірки С2 у діапазон комірок С3:С14 (рис. 7.23). Перед копіюванням, необхідно створити абсолютну адресацію для комірок В16 і В17. Для цього в рядку формул потрібно виділити В16 і натиснути на кнопку F4. В результаті В16 перетворюється в $B$16. Ті ж самі дії потрібно виконати для В17.

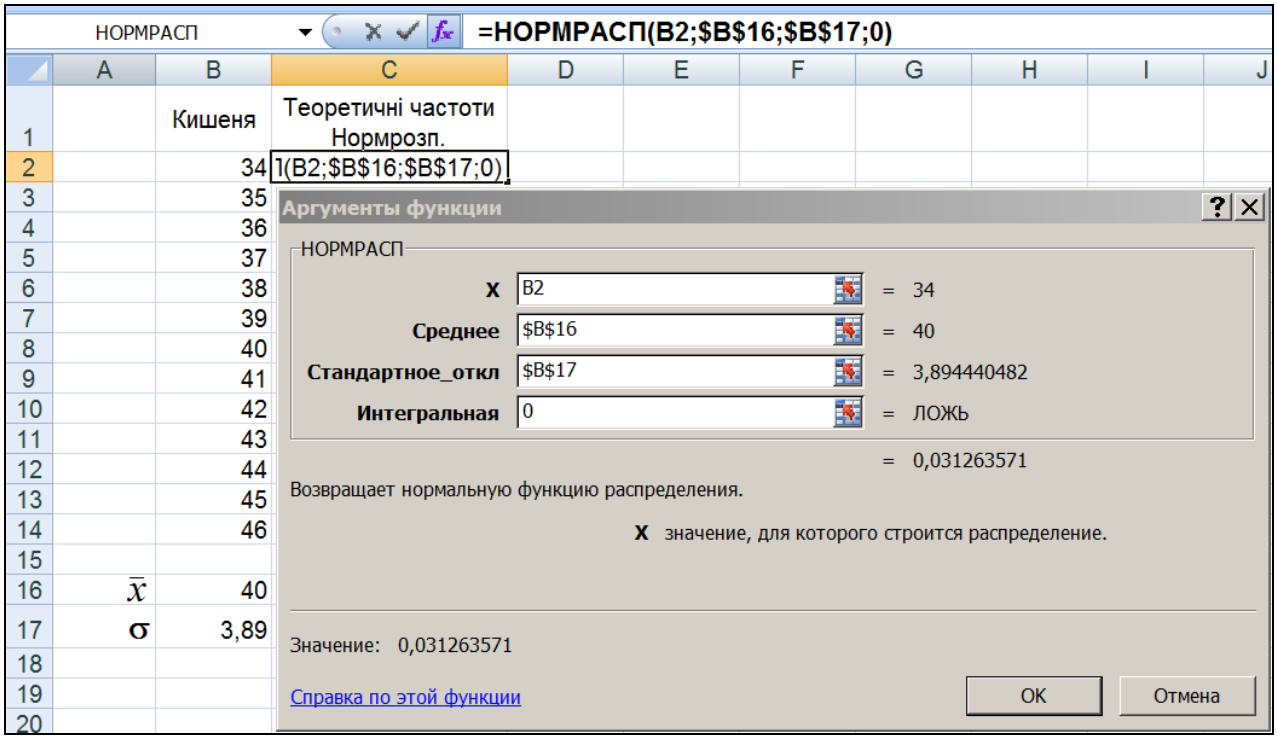


Рис. 7.23. Вбудована функція НОРМРАСП()

Для визначення відповідності вибірки нормальному розподілу, найбільш переконливі результати дає використання критеріїв згоди, серед яких великого поширення набув непараметричний критерій χ^2 (*Хі-квадрат*). Він ґрунтується на порівнянні емпіричних частот інтервалів групування з теоретичними (очікуваними) частотами, розрахованими за формулами нормального розподілу.

Зазначимо, що скільки-небудь впевнено про нормальність закону розподілу можна судити, якщо є не менше 50 результатів спостережень. У разі меншої кількості даних можна говорити лише про те, що дані не суперечать нормальному закону, і в цьому разі зазвичай використовують графічні методи оцінки відповідності. При більшій кількості спостережень доцільним є спільне використання графічних і статистичних (наприклад, тест *Хі*-квадрат або аналогічні) методів оцінки, які доповнюють один одного.

Використання критерію згоди Хі-квадрат. Для застосування критерію бажано, щоб вибіркові дані були згруповані в інтервальний ряд, а в кожному інтервалі знаходилося не менше 5 спостережень (частот).

Зауважимо, що порівнюватися повинні саме абсолютні частоти, а не відносні. При цьому, як і будь-який інший статистичний критерій, критерій *Хі-квадрат* не доводить справедливість нульової гіпотези (відповідність емпіричного розподілу нормальному), а лише може дозволити її відкинути з певною ймовірністю (рівнем значущості).

В MS Excel критерій *Хі-квадрат* (χ^2) реалізований у функції ХИ2ТЕСТ(). Функція обчислює ймовірність збігу спостережуваних (фактичних) значень і теоретичних (очікуваних) значень. Якщо обчислена ймовірність нижче рівня значущості (0,05), то нульова гіпотеза відкидається і стверджується, що спостережувані значення не відповідають нормальному закону розподілу.

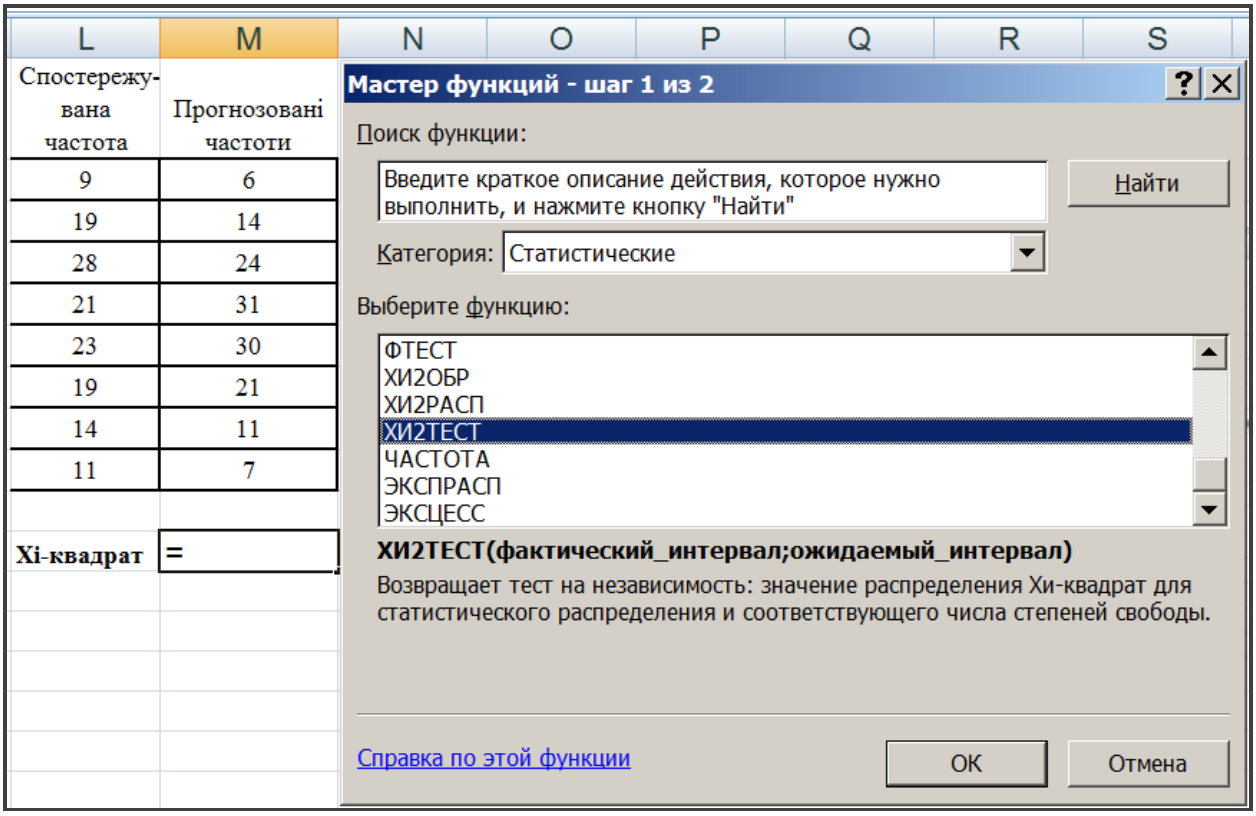
Функція ХИ2ТЕСТ() має таку структуру:

ХИ2ТЕСТ(*фактический интервал*; *ожидаемый интервал*).

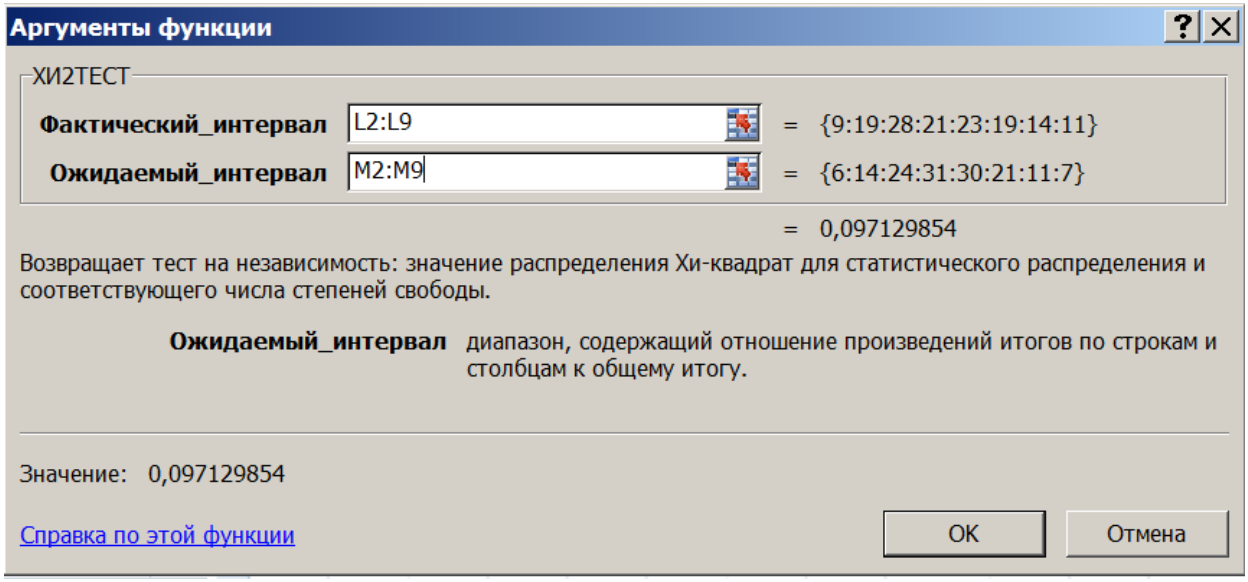
Перевіримо відповідність вибіркових даних з прикладу 6.2 (табл. 6.4, див. розділ 6) нормальному закону розподілу. Там само наведена процедура обчислення теоретичних частот, тому тут ми будемо лише використовувати отримані частоти.

Спостережувані і теоретичні частоти для прикладу 6.2 наведені на рис. 7.24*а*. Завдання полягає в тому, щоб визначити, наскільки відрізняється фактичний розподіл частот оцінки нового виду туризму

від прогнозованого фахівцями.



а)



б)

Рис. 7.24. Приклад заповнення робочих полів функції ХИ2ТЕСТ()

Встановимо курсор миші у вільну комірку М11 і введемо функцію ХИ2ТЕСТ(), рис. 7.24*а*. Як фактичний інтервал задамо

діапазон L2:L9, а як очікуваний інтервал – діапазон M2:M9 (рис. 7.24*б*). Після натискання кнопки «ОК» в комірці M11 з'явиться значення ймовірності того, що вибіркові дані відповідають нормальному закону розподілу – 0,0971. Це значення відображається також на рис. 7.24*б* праворуч під діапазоном чисел.

Оскільки отримана ймовірність відповідності експериментальних даних $p = 0{,}097$ більше, ніж рівень значущості $\alpha = 0{,}05$, то можна стверджувати, що нульова гіпотеза не може бути відкинута, і, отже, дані не суперечать нормальному закону розподілу.

7.5. Кореляційно-регресійний аналіз в MS Excel

Як відомо, для оцінки взаємозв'язку між вибірками (змінними *X* та *Y*) застосовують *регресійний* і *кореляційний аналіз*. Перший встановлює форму взаємозалежності, другий – ступінь зв'язку вибірок.

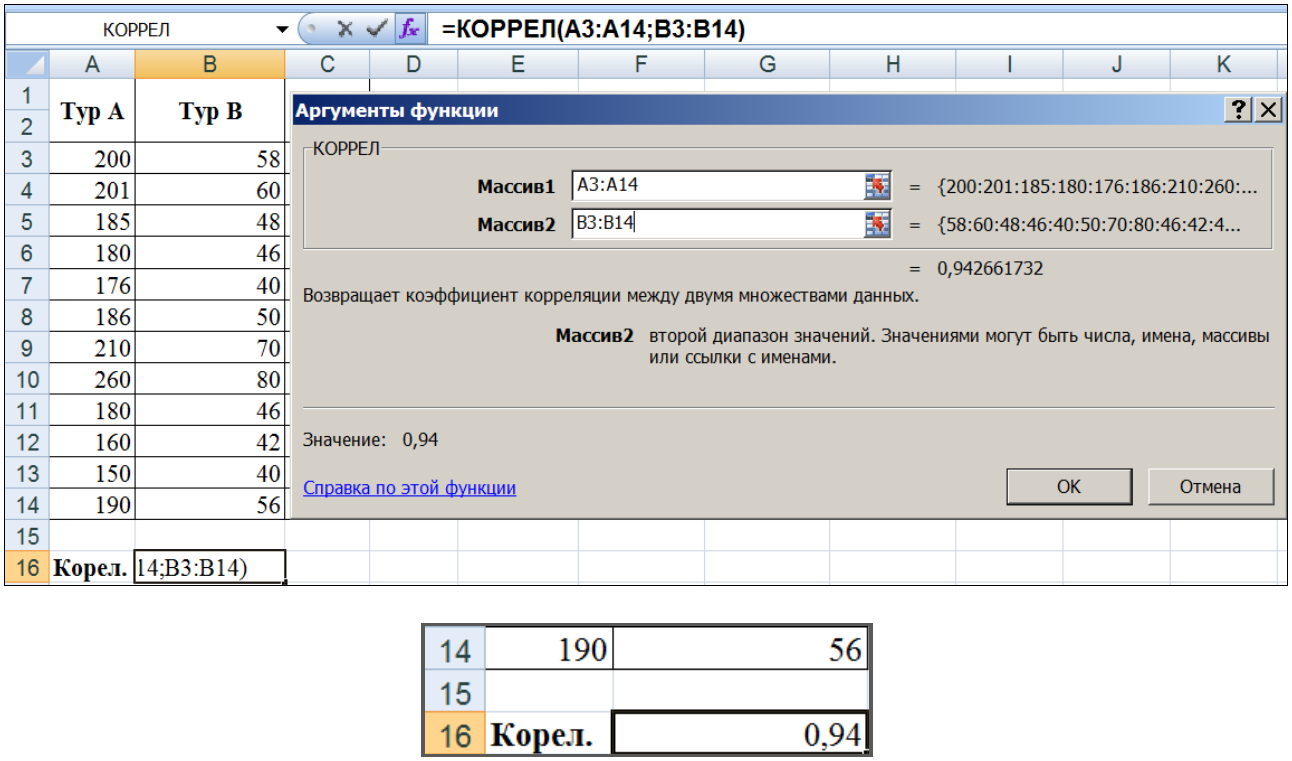
Кореляція. Ступінь зв'язку двох вибірок (випадкових величин *X* та *Y*) оцінюється *коефіцієнтом кореляції*. В MS Excel для обчислення парних коефіцієнтів лінійної кореляції використовується спеціальна функція КОРРЕЛ(). Функція має таку структуру: КОРРЕЛ(массив1; массив2), де «массив1» – це діапазон комірок першої випадкової величини; «массив2» – це другий інтервал комірок зі значеннями другої випадкової величини.

Розглянемо приклад застосування функції КОРРЕЛ().

***Приклад* 7.4.** Є результати річного спостереження реалізації турпакетів двох туристських маршрутів *A* і *B* (рис. 7.25). Потрібно визначити, чи є взаємозв'язок між кількістю продажів турпакетів обох маршрутів.

Для обчислення значення коефіцієнта кореляції між вибірками встановимо курсор миші у вільну комірку (наприклад, B16). Після виклику функції КОРРЕЛ() введемо в полі «Массив1» діапазон даних A3:A14; в полі «Массив2» – діапазон даних B3:B14. Після натискання кнопки «ОК» в комірці B16 з'явиться значення коефіцієнта кореляції

– 0,94. Таке значення коефіцієнта кореляції свідчить про те, що протягом року була висока ступінь прямого лінійного взаємозв'язку між кількостями проданих турпакетів обох маршрутів (r_{xy} = 0,94).



	A	B
1		
2	**Тур А**	**Тур В**
3	200	58
4	201	60
5	185	48
6	180	46
7	176	40
8	186	50
9	210	70
10	260	80
11	180	46
12	160	42
13	150	40
14	190	56
15		
16	**Корел.**	14;B3:B14)

14	190	56
15		
16	**Корел.**	0,94

Рис. 7.25. Результат застосування функції КОРРЕЛ()

Кореляційна матриця. При великій кількості спостережень, коли коефіцієнти кореляції необхідно послідовно обчислювати з декількох рядів числових даних, для зручності одержувані коефіцієнти зводять у таблиці, які називаються кореляційними матрицями.

Кореляційна матриця – це таблиця, в якій на перетині відповідних рядка та стовпця знаходиться коефіцієнт кореляції між відповідними параметрами.

В MS Excel для обчислення кореляційних матриць використовується інструмент **«Корреляция»**, для активації якого після запуску «**Анализа данных**» в списку «Инструменты анализа» потрібно вибрати рядок **«Корреляция»** і натиснути на кнопку «ОК».

У діалоговому вікні, яке з'явиться, в полі «Входной интервал» слід ввести посилання на комірки, що містять аналізовані дані.

Вхідний інтервал повинен містити не менше двох стовпців.

В розділі «Группировка» прапорець встановлюємо відповідно до введених даних (наприклад, «за стовпцями»).

В полі вікна «Выходной диапазон» потрібно ввести посилання на комірки, в які будуть виведені результати аналізу. Розмір вихідного діапазону визначається автоматично. Тут можна також вибрати новий робочий аркуш для виведення даних або нову робочу книгу.

У вихідний діапазон буде виведена кореляційна матриця, в якій на перетині кожних рядка і стовпця знаходиться коефіцієнт кореляції між відповідними параметрами. Комірки вихідного діапазону, що мають збіжні координати рядків і стовпців, містять значення 1, оскільки кожен стовпець у вхідному діапазоні повністю корелює з самим собою.

***Приклад* 7.5.** Додамо третій параметр в *Прикладі* 7.4 – реалізація турпакету третього туристського маршруту *C* і визначимо взаємозв'язок між кількістю продажів турпакетів трьох маршрутів.

Для цього виберемо інструмент **«Корреляция»** в меню **«Анализ данных»**. У діалоговому вікні, що з'явиться, вкажемо «Входной интервал» A3:C14 та «Выходной интервал» – E3 (рис. 7.26).

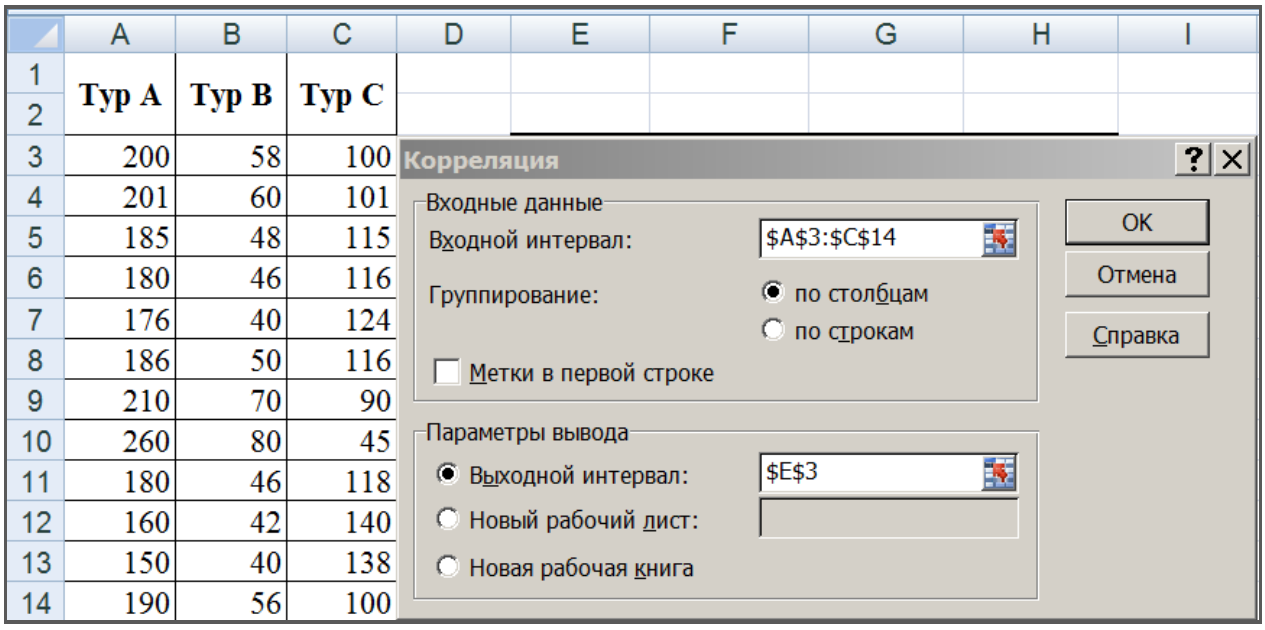


Рис. 7.26. Установка параметрів кореляційного аналізу

У вихідному діапазоні отримуємо *кореляційну матрицю* (рис. 7.27), в комірках якої на перетині стовпців і рядків записані коефіцієнти кореляції між турами A, B і C.

	Столбец 1	*Столбец 2*	*Столбец 3*
Столбец 1	1		
Столбец 2	0,94	1	
Столбец 3	-0,99	-0,95	1

Рис. 7.27. Результати обчислення кореляційної матриці

Таким чином, в результаті аналізу виявлено: *зворотний* лінійний взаємозв'язок дуже сильного ступеня між кількістю продажів турпакетів маршрутів A і C (r = –0,99 – практично лінійний зв'язок), B і C (r = –0,95) та лінійний дуже сильний прямий зв'язок між кількістю продажів турпакетів маршрутів A і B (r = 0,94).

***Регресія*.** Мірою ефективності регресійної моделі є коефіцієнт детермінації R^2 (R-квадрат). Коефіцієнт детермінації визначає, з яким ступенем точності отримане регресійне рівняння описує (апроксимує) вихідні дані.

Досліджується також значущість регресійної моделі за допомогою *F-критерію* (Фішера). Якщо величина *F-критерію* (p < 0,05), то регресійна модель є значущою.

Достовірність відмінності параметрів регресії (коефіцієнтів) від нуля перевіряється за допомогою критерію *Стьюдента*. У випадках, коли p > 0,05, коефіцієнт може вважатися нульовим, а це означає, що вплив відповідної незалежної змінної на залежну змінну є недостовірним, і ця незалежна змінна може бути виключена з рівняння.

В MS Excel для отримання коефіцієнтів регресії використовується процедура **«Регрессия»** із **«Пакет анализа»**. Крім того, можуть бути використані: функція ЛИНЕЙН() для отримання параметрів регресійного рівняння і функція ТЕНДЕНЦИЯ() для отримання передбачених (теоретичних) значень залежної змінної за

потрібними значеннями незалежної змінної.

***Приклад* 7.6.** Розглянемо застосування інструменту **«Регрессия»** на *прикладі* 5.1 (розділ 5), в якому перевіряється взаємозв'язок між кількістю туристів, які виїжджали до Німеччини, і курсом гривні до євро. Потрібно на основі цих даних побудувати регресійне рівняння.

Якщо вихідні дані вже внесені, то для реалізації інструменту **«Регрессия»** вибираємо ***Анализ данных→Регрессия***. У діалоговому вікні «Входной интервал» *Y* необхідно ввести посилання на діапазон залежних даних, що містить один стовпець даних. У вікні «Входной интервал» *X* потрібно ввести посилання на діапазон незалежних даних. У вікні «Выходной интервал» потрібно ввести посилання на комірки, в які будуть виведені результати аналізу (рис. 7.28).

Для візуальної перевірки відмінності експериментальних точок від передбачених за регресійною моделлю слід встановити прапорець в полі «График подбора» і натиснути кнопку «ОК».

Вихідний діапазон буде містити результати дисперсійного аналізу, коефіцієнти регресії, стандартну похибку обчислення *Y*, стандартні відхилення, кількість спостережень та ін.

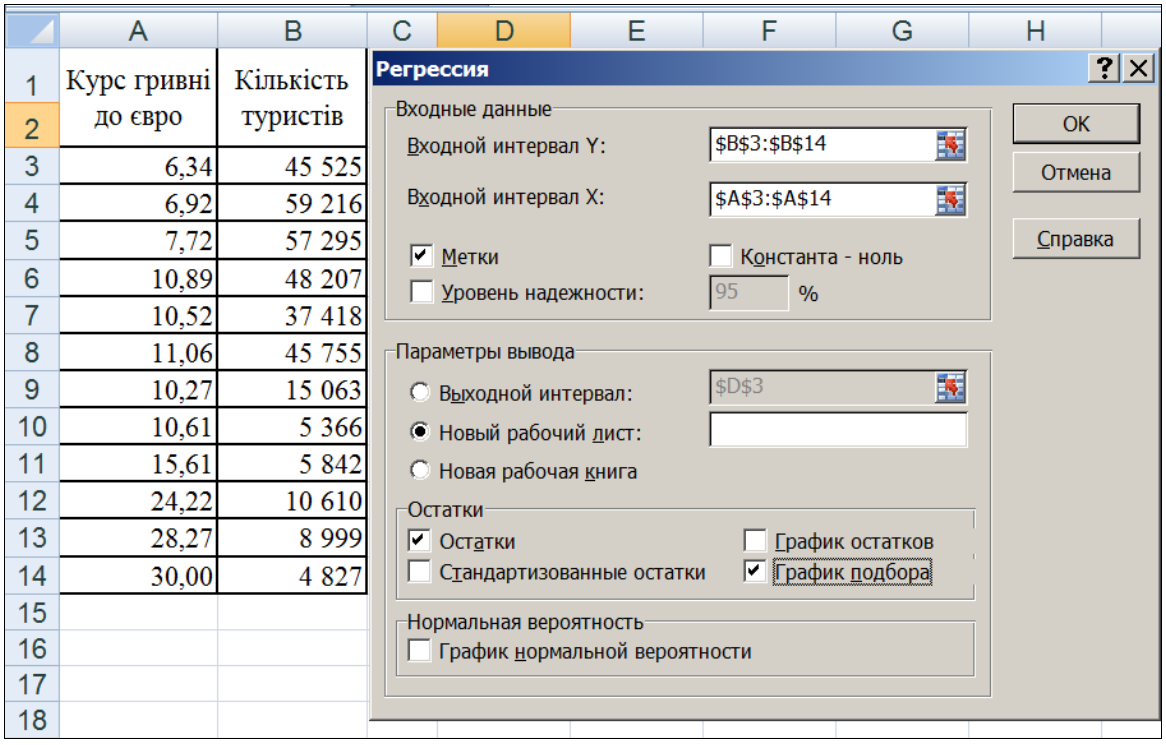


Рис. 7.28. Приклад заповнення діалогового вікна «Регрессия»

Результати для розглянутого прикладу наведені на рис. 7.29.

З отриманих результатів можна встановити (округлюючи деякі значення) величини багатьох характеристик регресійної моделі.

За розділом таблиці *Регрессионная статистика* можна встановити:

Коефіцієнт кореляції: $r_{xy} = 0{,}70$.

Коефіцієнт детермінації: $R = r_{xy}^2 = 0{,}49$. Значення коефіцієнта детермінації, що наводиться, визначає, з яким ступенем точності отримане регресійне рівняння апроксимує вихідні дані.

Корінь квадратний із залишкової дисперсії (*стандартна помилка*): $S_{зал.}$. = 16414,96.

	A	B	C	D	E	F	G	H	I	J	K
1	ВЫВОД ИТОГОВ										
2											
3	*Регрессионная статистика*										
4	Множественный	0,702343									
5	R-квадрат	0,493285									
6	Нормированный	0,442613									
7	Стандартная ош	16414,96									
8	Наблюдения	12									
9											
10	Дисперсионный анализ										
11		*df*	*SS*	*MS*	*F*	*Значимость F*					
12	Регрессия	1	2623093908	2623093908	9,7349595	0,01087232					
13	Остаток	10	2694509321	269450932							
14	Итого	11	5317603229								
15											
16		*Коэффици*	*Стандартна*	*t-статисти*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,*	*Верхние 95,0%*		
17	Y-пересечение	55213,46	9736,02982	5,67104508	0,0002064	33520,2378	76906,69	33520,238	76906,69		
18	Переменная X 1	-1846,77	591,896483	-3,1200897	0,0108723	-3165,59764	-527,94255	-3165,598	-527,9426		
19											
20	ВЫВОД ОСТАТКА										
21											
22	*Наблюдение*	*едсказанно*	*Остатки*								
23	1	43504,94	2020,05843								
24	2	42433,81	16782,1851								
25	3	40956,4	16338,6012								
26	4	35102,14	13104,8624								
27	5	35785,44	1632,55743								
28	6	34788,19	10966,8133								
29	7	36247,14	-21184,135								
30	8	35619,23	-30253,233								
31	9	26385,38	-20543,383								
32	10	10484,69	125,307768								
33	11	3005,273	5993,72666								
34	12	-189,639	5016,63893								
35											

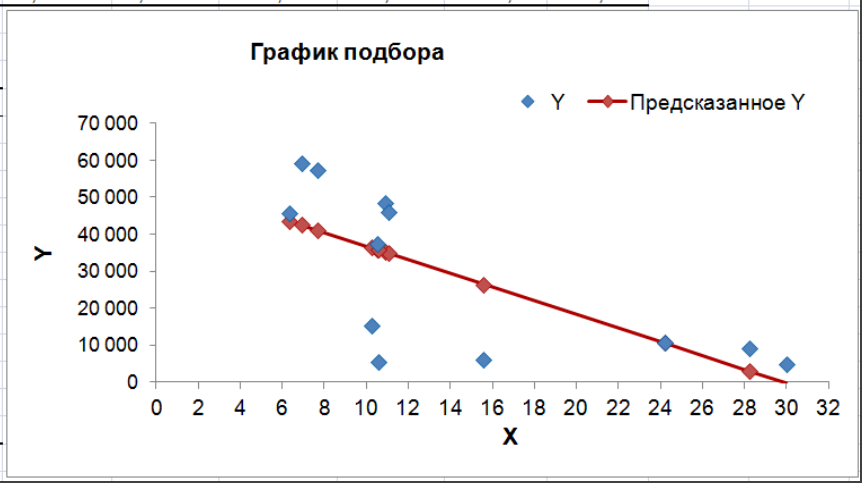


Рис. 7.29. Результати аналізу і графік відповідності фактичних і передбачених (за регресійною моделлю) точок

В розділі *Дисперсионный анализ* оцінюється достовірність отриманої моделі за рівнем значущості *F-критерію* Фішера – *p* (рядок *Регрессия*, стовпець *Значимость F*), який повинен бути < 0,05. У розглянутому прикладі значимість критерію становить 0,01087232, тобто $p < 0{,}05$. Отже, модель може вважатися адекватною з ймовірністю 0,95.

Далі необхідно визначити значення коефіцієнтів моделі, тобто коефіцієнти рівняння регресії. Ці коефіцієнти відображені в стовпці *Коэффициенты*, який містить значення *b* в рядку *Y-пересечение*, *a* – в рядку *Переменная* X1: $a = -1846{,}77$, $b = 55213{,}46$ (*рівняння регресії* з округленими значеннями *a* і *б*: $\hat{y}_x = 55213{,}5 - 1846{,}8x$).

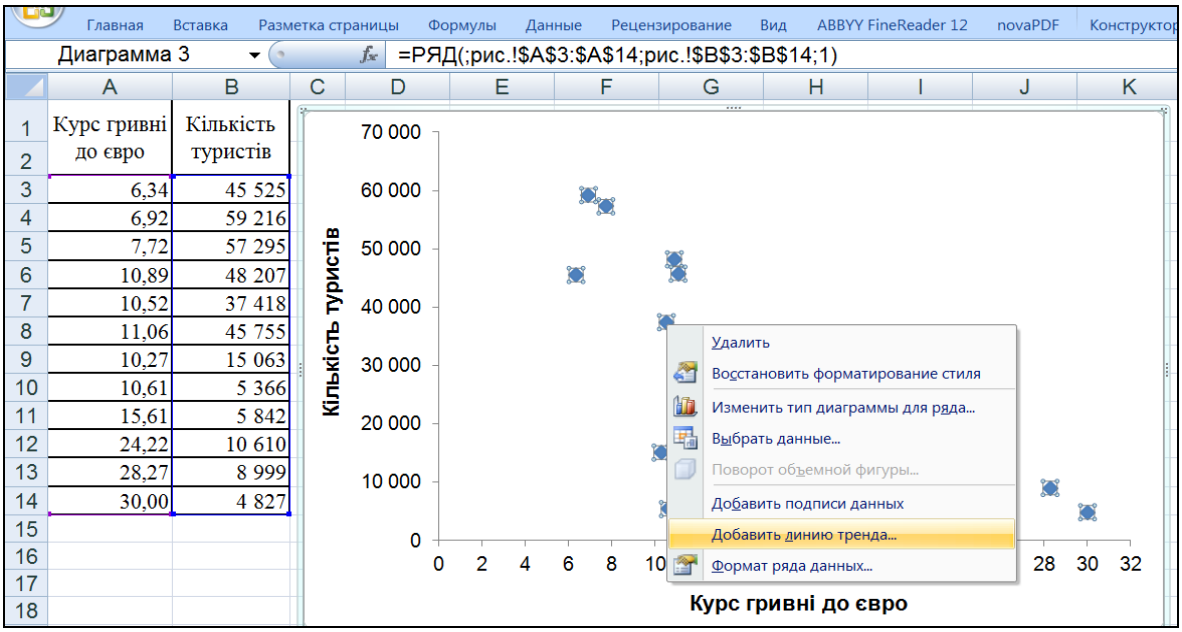
Для визначення достовірності коефіцієнтів моделі, тобто рівняння регресії, в стовпці *p-значення* наводиться достовірність відмінності відповідних коефіцієнтів від нуля. У разі, якщо $p > 0{,}05$, коефіцієнт може вважатися нульовим; це означає, що незалежна змінна практично не впливає на залежну змінну. У розглянутому прикладі $p = 0{,}0002064$ і $p = 0{,}0108723$, тобто значення *p*-рівня ≤ 0,05, і параметри є достовірними.

Проте, оскільки $R = 0{,}49 < 0{,}6$), можна зробити висновок, що точність апроксимації є недостатньою, і модель потребує покращення (введення нових незалежних змінних і т. д.).

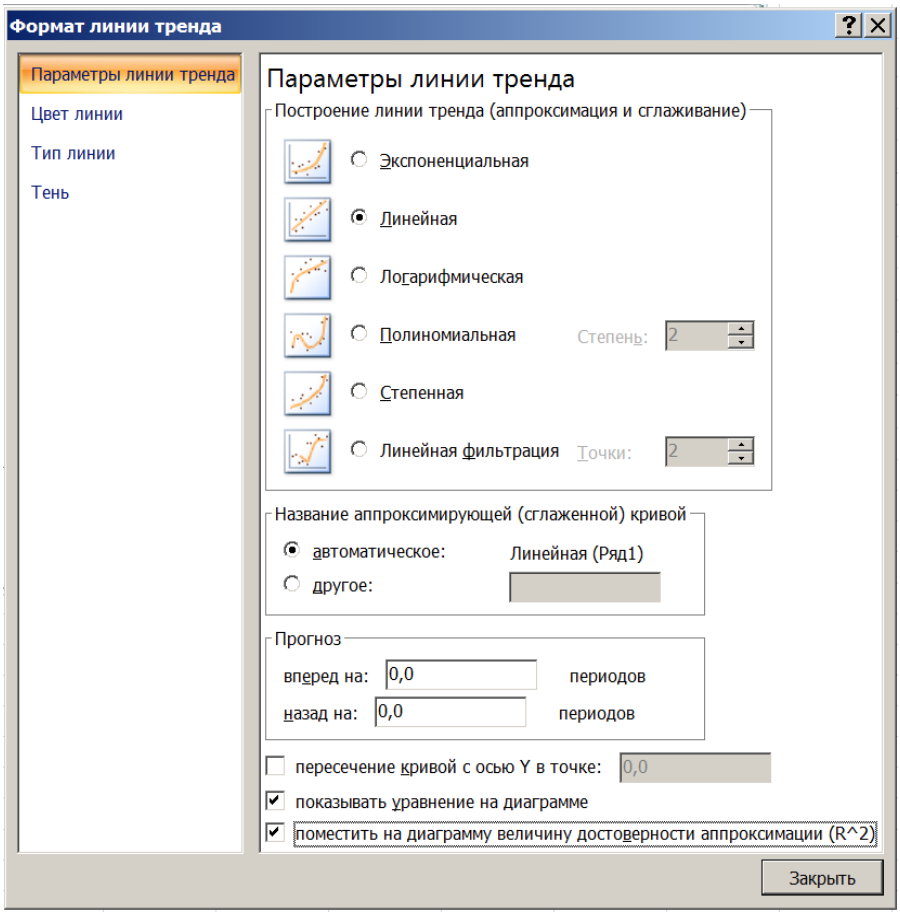
В останньому розділі таблиці (рис. 7.29) наводяться також: с*тандартні помилки* для параметрів регресії: $m_a = 9736{,}03$; $m_b = 591{,}90$; фактичні значення *t-критерію Стьюдента*: $t_a = 5{,}67$, $t_b = -3{,}12$. Знак «–» перед значенням *t-критерію* до ($t_b = -3{,}12$) означає, що емпіричне значення нижче рівня стандартного значення; д*овірчі інтервали*: $33520{,}2 \le b^* \le 76906{,}7$; $-3165{,}6 \le a^* \le -527{,}9$.

Рівняння регресії. Зазначимо, що рівняння регресії можна встановити і без запуску **«Анализа данных»** та інструменту **«Регрессия»**. Після введення даних (*X* та *Y*) у відповідні поля і запуску **«Мастера диаграмм»** вибираємо тип діаграми «Точечные». Далі, клацнувши правою кнопкою миші в будь-якій точці діаграми, в меню вибираємо «Добавить линию тренда» (рис. 7.30*а*). Призначаємо

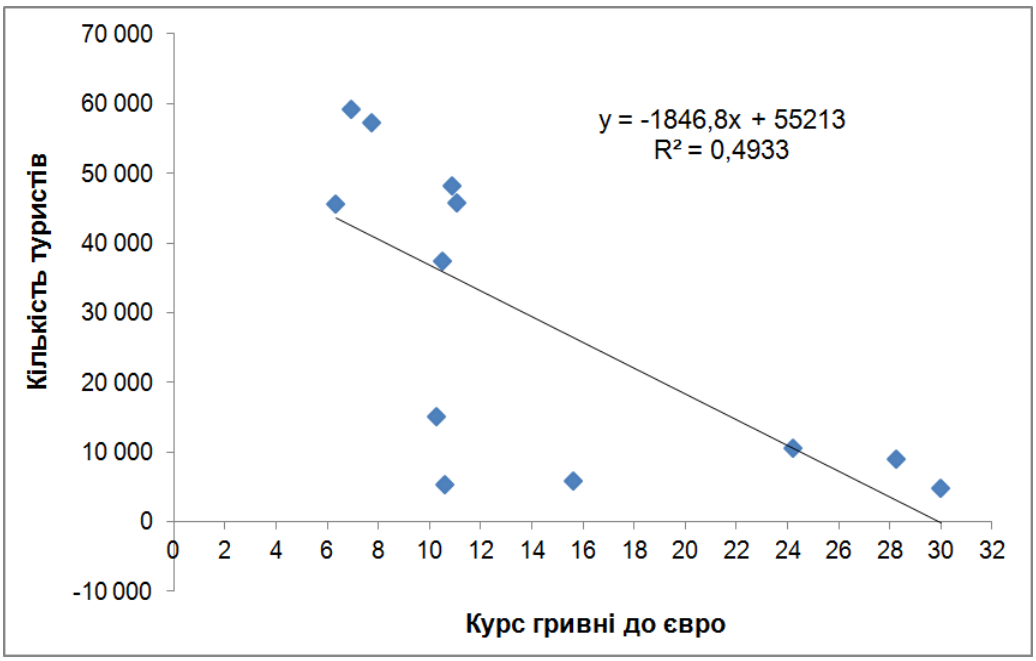
параметри для лінії: тип – «Линейная» (рис. 7.30*б*). Внизу – «Показать уравнение на диаграмме». Там само, для отримання величини коефіцієнта детермінації можна встановити – «Поместить на диаграмму величину достоверности аппроксимации (R^2)».



а)



б)



в)

Рис. 7.30. Приклад побудови рівняння і лінії регресії

Завдання для самостійного виконання

Завдання 7.1. Є дані про щомісячну кількість реалізованих туристичною фірмою турпакетів за періоди до і після початку активної рекламної компанії. У табл. 1 наведена кількість реалізованих турпакетів по місяцях. Потрібно знайти середні значення, стандартні відхилення і коефіцієнти варіації цих рядів, а також побудувати гістограми абсолютних і накопичених частот до і після рекламної компанії.

Таблиця 1

Кількість реалізованих туристичною фірмою турпакетів

Місяці	Без реклами	З рекламою
1	73	100
2	64	94
3	53	82
4	78	75
5	59	63

6	50	83
7	68	89
8	78	98
9	70	90
10	65	85
11	58	78
12	79	99

Завдання 7.2. Є дані про зарплати двох груп працівників готельно-ресторанного бізнесу: *готель* і *ресторан*. Необхідно визначити основні статистичні характеристики в групах даних, використовуючи інструмент *Описательная статистика* зі списку *Инструменты Анализа*. Результати роботи інструменту подайте у вигляді рисунку.

Таблиця 2

Заробітна плата працівників готелю та ресторану

Персонал готелю, грн.	Персонал ресторану, грн.
14000	20000
12500	15000
11000	13500
10600	11600
10000	12800
9900	12500
9000	11000
8500	10000
7000	8500
6000	8000

Завдання 7.3. Визначте ступінь взаємозв’язку між валовим доходом і витратами туроператора, заданого вибірками в табл. 3. Використовуючи інструменти **«Анализа данных»**, проведіть кореляційно-регресійний аналіз. У вигляді рисунку представте

«Вывод Итогов», в тому числі графік, який представляє лінію регресії з рівнянням регресії.

Таблиця 3

Валовий дохід і витрати туроператора

Валовий дохід, грн.	Витрати, грн.
450 000	75 000
560 000	80 000
520 000	75 000
785 000	93 000
635 000	86 000
465 000	76 500
745 000	82 000
780 000	90 000

СПИСОК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ

Основна

1. Барчуков И. С. Методы научных исследований в туризме: учеб. пособие. Москва: Издательский центр «Академия», 2008. 224 с.
2. Кущенко О. І. Статистика туризму: Економічна статистика: навчально-методичний посібник. Харків: ХНУ імені В. Н. Каразіна, 2014. 74 с.
3. Математична статистика: навч. посіб. / С. М. Григулич, В. П. Лісовська, О. І. Макаренко та ін. Київ: КНЕУ, 2015. 203 с.
4. Соболева Е. А. Статистика туризма: Статистическое наблюдение. Москва: Финансы и статистика, 2004. 160 с.
5. Ткач Є. І., Сторожук В. П. Загальна теорія статики: підручник. 3-тє вид. Київ: Центр учбової літератури, 2009. 442 с.
6. Єріна А. М., Пальян З. О. Теорія статистики: Практикум. 7-ме вид., стер. Київ: Знання, 2009. 255 с.

Додаткова

1. Барковський В. В., Барковська Н. В., Лопатін О. К. Теорія ймовірностей та математична статистика. Київ: ЦНЛ, 2002. 448 с.
2. Бережной В. И., Бигдай О. Б., Бережная О. В., Киселева О. А. Статистика в примерах и задачах: учеб. пособие. Москва: ИНФРА-М, 2016. 287 с.
3. Гельман В. Я. Решение математических задач средствами EXCEL: Практикум. Санкт-Петербург: Питер, 2003. 240 с.
4. Годин А. М. Статистика: учебник. 11-е изд. Москва: «Дашков и К°», 2014. 412 с.
5. Статистика: учебник / Глинский В. В., Ионин В. Г., Серга Л. К. и др.; под ред. В. Г. Ионина. 4-е изд., доп. перераб. Москва: ИНФРА-М, 2017. 655 с.
6. Гмурман В. Е. Руководство к решению задач по теории вероятностей и математической статистике : учеб. пособие. Москва: Высш. шк., 2005. 404 с.
7. Статистика: учебник / Дарда Е. С., Дианов Д. В., Садовникова Н. А., Шмойлова Р. А. и др.; под ред. В. Г. Минашкина. Москва: Издательство Юрайт, 2014. 448 с.
8. Дегтярева И. Н. Статистика: учеб. пособие для СПО. Саратов: Профобразование, 2017. 181 с.
9. Замедлина Е. А. Статистика: учеб. пособие. Москва: РИОР: ИНФРА-М, 2014. 160 с.
10. Канцедал С. А. Основы статистики: учеб. пособие. Москва: ФОРУМ:

ИНФРА-М, 2011. 192 с.
11. Кармелюк Г. І. Теорія ймовірностей та математична статистика. Посібник з розв'язування задач: навч. посіб. Київ: Центр учбової літератури, 2007. 576 с.
12. Методологічні положення зі статистики туризму. Наказ Державної служби статистики України 23.12.2011 р. № 372. URL: http://www.ukrstat.gov.ua/
13. Міжнародні рекомендації зі статистики туризму, 2008 рік. ООН. URL: http://unstats.un.org/unsd/publication/SeriesM/Seriesm_83rev1r.pdf
14. Морозова С. В. Статистика предприятий отрасли: учеб.-метод. пособие. Москва: ИНФРА-М; Минск: Нов. Знание, 2014. 271 с.
15. Статистика: учебник / Мхитарян В. С., Дуброва Т. А., Минашкин В. Г. и др. Москва Академия, 2011. 271 с.
16. Полякова В. В., Шаброва Н. В. Основы теории статистики: учеб. пособие. Екатеринбург: Изд-во Урал. ун-та, 2015. 148 с.
17. Сергеева И. И., Чекулина Т. А., Тимофеева С. А. Статистика: учебник. Москва: ФОРУМ: ИНФРА-М, 2016. 303 с.
18. Статистика туризма: учебник / под ред. А. Ю. Александровой. Москва: Федеральное агентство по туризму, 2014. 464 с.
19. Щурик М. В. Статистика: навч. посібн. 2-ге видання, оновлене і доповнене. Львів: «Магнолія-2006», 2009. 545 с.
20. Mining trips from location-based social networks for clustering travelers and destinations. Linus W. Dietz, et. all, *Information Technology & Tourism* 2020. Vol. 22. P. 131–166.

Електронні інформаційні ресурси

1. Державна служба статистики України. URL: http://www.ukrstat.gov.ua
2. Львівська міська рада. URL: https://www.city-adm.lviv.ua
3. Міжнародна система бронювання готелей. URL: https://www.booking.com
4. Спілка Зеленого туризму. URL: https://www.greentour.com.ua
5. Statistics/Eurostat. URL: https://ec.europa.eu/eurostat/web/tourism/
6. UK National Travel Survey. URL: https://www.gov.uk/government/collections/national-travel-survey-statistics
7. UNWTO World Tourizm. URL: http://www.unwto.org/facts/eng/barometer.htm

Додаток А

Критичні значення критерію відповідності χ^2 (*Xi*-квадрат)

df	α=0,05	α=0,01	α=0,001	*df*	α=0,05	α=0,01	α =0,001
1	3,84	6,63	10.83	21	32,67	38,93	46,80
2	5,99	9,21	13,82	22	33,92	40,29	48,27
3	7,81	11,07	16,27	23	35,17	41,64	49,73
4	9,49	13,28	18,47	24	36,42	42,98	51,18
5	11,07	15,09	20,51	25	37,65	44,31	52.62
6	12,59	16,81	22.46	26	38,89	45,64	54,05
7	14,07	18,48	24,32	27	40,11	46,96	55,48
8	15,51	20,09	26,12	28	41,34	48,28	56,89
9	16,92	21,67	27,88	29	42,56	49,59	58,30
10	18,31	33,21	29,59	30	43,77	50,89	59,70
11	19,68	34,73	31,26	31	44,99	52.19	61,10
12	21,03	26,22	32,91	32	46,19	53,49	62.49
13	22,36	27,69	34,53	33	47,40	54,78	63,87
14	23,68	29,14	36,12	34	48,60	56,06	65,25
15	25,00	30,58	37,70	35	49,80	57,34	66,62
16	26,30	32,00	39,25	36	51,00	58,62	67,98
17	27,59	33,41	40,79	37	52,19	59,89	69,35
18	28,87	34,81	42,31	38	53,38	61,16	70,70
19	30,14	36,19	43,82	39	54,57	62,43	72,06
20	31,41	37,57	45,31	40	55,76	63,69	73,40

Додаток Б

Критичні значення *t*-критерію Стьюдента

df	α			*df*	α		
	0,10	0,05	0,01		0,10	0,05	0,01
1	6,3138	12,7062	63,6567	18	1,7341	2,1009	2,8784
2	2,9200	4,3027	9,9248	19	1,7291	2,0930	2,8609
3	2,3534	3,1824	5,8409	20	1,7247	2,0860	2,8453
4	2,1318	2,7764	4,6041	21	1,7207	2,0796	2,8314
5	2,0150	2,5706	4,0321	22	1,7171	2,0739	2,8188
6	1,9432	2,4469	3,7074	23	1,7139	2,0687	2,8073
7	1,8946	2,3646	3,4995	24	1,7109	2,0639	2,7969
8	1,8595	2,3060	3,3554	25	1,7081	2,0595	2,7874
9	1,8331	2,2622	3,2498	26	1,7056	2,0555	2,7787
10	1,8125	2,2281	3,1693	27	1,7033	2,0518	2,7707
11	1,7959	2,2010	3,1058	28	1,7011	2,0484	2,7633
12	1,7823	2,1788	3,0545	29	1,6991	2,0452	2,7564
13	1,7709	2,1604	3,0123	30	1,6973	2,0423	2,7500
14	1,7613	2,1448	2,9768	40	1,6839	2,0211	2,7045
15	1,7531	2,1314	2,9467	60	1,6707	2,0003	2,6603
16	1,7459	2,1199	2,9208	120	1,6577	1,9799	2,6174
17	1,7396	2,1098	2,8982	∞	1,6448	1,9600	2,5758

Додаток В

Критичні значення *F*-критерію Фішера при рівні значущості α=0,05

k_2 \ k_1	1	2	3	4	5	6	8	12	24	∞
1	161,5	199,5	215,7	224,6	230,2	233,9	238,9	243,9	249,0	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,74	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31

90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,63	1,26
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,60	1,21
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,55	1,10
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,54	1,06
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,53	1,03
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1

Навчальне видання

Мелконян Джема Варанцівна

Яворська Вікторія Володимирівна

СТАТИСТИКА В ТУРИЗМІ

НАВЧАЛЬНИЙ ПОСІБНИК

В авторській редакції

Підп. до друку 30.08.2021. Формат 60x84/16.
Умов.-друк. арк. 11,39. Тираж 37 пр.
Зам. № 2807.

Видавець і виготовлювач
Одеський національний університет
імені І. І. Мечникова

Україна, 65082, м. Одеса, вул. Єлісаветинська, 12
Тел.: (048) 723 28 39. E-mail: druk@onu.edu.ua
Свідоцтво суб'єкта видавничої справи ДК № 4215 від 22.11.2011 р.