КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

М.В. Грисенко
О.А. Чугаєв

КІЛЬКІСНІ МЕТОДИ АНАЛІЗУ МІЖНАРОДНИХ ЕКОНОМІЧНИХ ВІДНОСИН

НАВЧАЛЬНИЙ ПОСІБНИК

КИЇВСЬКИЙ УНІВЕРСИТЕТ 2023

УДК 339.9:330.47:311:004.9(075.8)

Рецензенти:
д-р екон. наук, проф. П. Дзюба
канд. фіз.-мат. наук, доц. А. Рижов

Рекомендовано до друку вченою радою ННІ міжнародних відносин
(*протокол* № 13 *від* 29 *червня* 2022 *року*)

Ухвалено науково-методичною радою
Київського національного університету імені Тараса Шевченка
(*протокол* № 6-22 *від* 30 *серпня* 2022 *року*)

Грисенко М.В.

Г85 Кількісні методи аналізу міжнародних економічних відносин : навч. посіб. / М.В. Грисенко, О.А. Чугаєв. – ВПЦ "Київський університет", 2023. – 364 с.

ISBN 978-966-933-221-9

Видання побудовано на міждисциплінарній основі. Подано теоретичні основи базових кількісних методів і реальних моделей, що застосовують при аналізі міжнародних економічних відносин. Поєднано елементи таких дисциплін, як "Міжнародні економічні відносини", "Економіко-математичний і статистичний аналіз", "Інформатика", "Іноземна мова". Представлено джерела міжнародної економічної статистики, основи використання спеціалізованого програмного забезпечення, теоретичні аспекти окремих кількісних методів і приклади їх застосування.

Призначено для студентів, які навчаються за освітньо-професійною програмою "Міжнародні економічні відносини" освітнього рівня "Магістр". Може бути корисним для аспірантів, викладачів і науковців, економістів-аналітиків і практиків, фінансистів та актуаріїв, бізнесменів і менеджерів не тільки як практичний посібник, але й як довідник.

УДК 339.9:330.47:311:004.9(075.8)

Навчальне видання

ГРИСЕНКО Марина Віталіївна
ЧУГАЄВ Олексій Анатолійович

КІЛЬКІСНІ МЕТОДИ АНАЛІЗУ
МІЖНАРОДНИХ ЕКОНОМІЧНИХ ВІДНОСИН

Навчальний посібник

Оригінал-макет виготовлено ВПЦ "Київський університет"

Формат 60 × 84$^{1/16}$. Ум. друк. арк. 21,2. Наклад 100. Зам. № 223-10607. Гарнітура Cambria.
Папір офсетний. Друк офсетний. Вид. № МВ 9. Підписано до друку 10.05.23.

Видавець і виготовлювач – ВПЦ "Київський університет"
б-р Т. Шевченка, 14, м. Київ, кімн. 01601, ☎ (044) 239 32 22; (044) 239 31 72;
факс (044) 239 31 28; e-mail: vpc_div.chiev@univ.net.ua; http: vpc.knu.ua
Свідоцтво внесено до Державного реєстру ДК № 1103 від 31.10.02

ISBN 978-966-933-221-9

ПЕРЕДМОВА

Навчальний посібник орієнтує читача на активне формування навичок ознайомлення з базами статистичних даних та самостійне використання методів кількісного аналізу. Автори мали на меті продемонструвати можливості різних кількісних методів, проблеми їх застосування на практиці та способи розв'язання цих проблем. Саме тому значна частина видання присвячена вивченню реальних прикладних моделей як наукових результатів самих авторів, так й інших науковців. Для прикладів використано реальні або умовні (якщо не вказано джерело) дані.

Навчальний посібник написано спеціально для навчальної дисципліни "Кількісні методи аналізу міжнародних економічних відносин", що не виключає можливості використання окремих його складових при викладанні курсів: "Світова економіка", "Теорія і форми міжнародних економічних відносин", "Міжнародні фінанси", "Математика для економістів", "Регулювання міжнародних економічних відносин", "Міжнародна статистика", "Економіко-математичне моделювання світогосподарських процесів", "Іноземна мова (англійська)".

Дисципліна "Кількісні методи аналізу міжнародних економічних відносин" ставить перед собою такі завдання:

- ознайомити студентів з основними кількісними методами, які використовують у сфері МЕВ;
- розширити систему знань щодо статистичних показників у сфері МЕВ, теоретичні знання у сфері методики кількісного аналізу;
- сформувати навички пошуку статистичної інформації для аналізу МЕВ, підготовки та здійснення кількісного аналізу дослідження в МЕВ з використанням спеціального програмного забезпечення;
- надати понятійно-термінологічний апарат у сфері соціально-економічних статистичних показників і методів кількісного аналізу українською та англійською мовами.

Під час підготовки навчального посібника було використано матеріали підручників, навчальних посібників, наукові публікації, статистичну інформацію офіційних веб-сайтів міжнародних організацій, державних органів і недержавних структур, інформаційні сайти, присвячені спеціалізованому програмному забезпеченню.

Кількісні методи надають потужний інструментарій для проведення досліджень у сфері міжнародних економічних відносин. Вони дозволяють оцінювати силу впливу факторів на процеси у світовій економіці, наслідки реалізації міжнародних економічних зв'язків на внутрішній соціально-економічний розвиток країн, здійснювати прогнози на майбутнє, забезпечувати підтримку при прийнятті рішень у сфері зовнішньої і внутрішньої економічної політики.

Кількісні методи є потужним математичним інструментом дослідження, але кожний предмет дослідження має свою специфіку використання цих методів. Саме тому авторами використано *міждисциплінарний підхід, що поєднує такі складові*:

1. Міжнародні економічні відносини як предмет дослідження. Використання універсальних кількісних методів ілюструють приклади аналізу різних форм міжнародних відносин. Частина навчального посібника присвячена джерелам статистичної інформації щодо показників міжнародних економічних відносин і суміжних сфер.

2. Практичні аспекти використання кількісних методів: можливості цих методів, етапи аналізу, приклади застосування, застереження, інтерпретація результатів, способи розв'язання проблем, поєднання різних методів.

3. Програмне забезпечення, що дозволяє автоматизувати розрахунки й представляти їх результати у зручній формі. Існує велика кількість програмного забезпечення, яке значно полегшує використання кількісних методів, особливо при дослідженні великого масиву даних. Зокрема, у посібнику подано приклади для Microsoft Excel як найбільш розповсюдженого програмного продукту, спеціалізованої програ-

ми для кількісного аналізу, що надає можливість використання ширшого спектру методів у зручний спосіб.

4. Іноземна мова. Ключові терміни подано як українською, так і англійською мовами. Володіння іноземною мовою у цій сфері є беззаперечним, адже міжнародні економічні статистичні бази даних і програмне забезпечення представлені та доступні переважно англійською мовою.

У навчальному посібнику розкрито теоретичні та практичні аспекти використання різноманітних кількісних методів. Особливу вагу приділено підготовчому етапу проведення аналізу: пошуку джерел статистичних даних, ознайомленню з базовими функціями спеціалізованого програмного забезпечення, формуванню бази вхідних даних і первинному аналізу їх структури. Подано алгоритми та приклади застосування методів аналізу середніх, дисперсійного, кореляційного, регресійного, кластерного, частотного аналізу, а також сигнального методу. Як приклади також наведено реальні моделі, методики та результати низки наукових досліджень, показано можливості їх використання для аналізу сучасних міжнародних економічних відносин.

Автори висловлюють щиру подяку колегам за поради та цінні зауваження, які були враховані при підготовці посібника.

Розділ 1
РЕАЛЬНІ МОДЕЛІ У СФЕРІ МІЖНАРОДНИХ ЕКОНОМІЧНИХ ВІДНОСИН

1.1. Економіко-математичне моделювання в міжнародних економічних відносинах

Міжнародні економічні відносини є складно організованою динамічною підсистемою функціонування світової економіки, що віддзеркалює зовнішньоекономічну діяльність держав і структурних підрозділів їхніх господарств, які базуються на міжнародному поділі праці. Їх характеризує динамізм у часі та просторі, галузева й територіальна структура. Ці сторони зовнішньої економічної діяльності легко піддаються математичній формалізації, що зберігає ідентичність реальних співвідношень, масштабність складних явищ і процесів, які досліджують у просторовому та часовому аспектах. На цій основі за допомогою кількісних методів можлива побудова економіко-математичних моделей, які досліджують і встановлюють конкретні економічні закономірності та взаємозалежності, що спостерігаються у міжнародних економічних відносинах.

Метою будь-якого дослідження зазвичай є визначення параметрів досліджуваного соціально-економічного об'єкта, які задовольняють певні вимоги та критерії. Виконуючи дослідження за допомогою штучних або мислених моделей, що віддзеркалюють об'єктивну реальність, дослідник прагне одержати додаткову інформацію про об'єкт дослідження. У процесі дослідження виникає необхідність зміни значень параметрів об'єкта, отже і зміни значень показника, який відповідає критеріям. Процес дослідження закінчується, коли дослідник знаходить сукупність значень параметрів об'єкта, які задовольняють заданим критеріям із заданою точністю та достовірністю.

На практиці експериментування з реальними об'єктами зазвичай обходиться дуже дорого, тому частіше за все для проведення наукових експериментів реальні об'єкти замінюють відповідними простішими та доступними, властивості

яких подібні до властивостей реальних об'єктів у певній суттєвій частині.

Уявлення про об'єкт або явище, яке досліджують, можна виразити у формі опису, рисунку, схеми, формули, штучно зробленої конструкції тощо. Можна використовувати й окремі явища або об'єкти соціально-економічної природи.

Методологія моделювання довгий час розвивалась у надрах переважно природничих наук. Із часом моделювання набуло статусу універсального методу наукового пізнання.

Об'єкт, з метою вивчення якого проводять дослідження, називають *оригіналом*, а об'єкт, що досліджують замість оригіналу для вивчення певних властивостей, – *моделлю.* Моделями можуть бути природні об'єкти з властивостями, що подібні до властивостей оригіналу або створені спеціальні штучні об'єкти із потрібними властивостями.

Модель – це матеріально або уявно поданий об'єкт, що являє собою образ оригіналу та відображає найбільш важливі для певного дослідження риси та властивості оригіналу, при безпосередньому вивченні якої можна дістати нову інформацію про об'єкт-оригінал.

Математичну модель визначають як внутрішньо несуперечливу замкнену систему сукупності математичних і логічних співвідношень, що описують структуру та функції реального об'єкта дослідження. Якщо об'єкт дослідження – зі сфери економіки, а модель побудовано на основі математичного апарату, то її називають економіко-математичною моделлю (ЕММ). Отже, ЕММ – це математичний опис економічного процесу або явища з метою його дослідження та управління.

Моделювання – це метод або процес вивчення властивостей об'єктів-оригіналів шляхом дослідження відповідних властивостей їх моделей, коли замість одного явища (об'єкта досліджень – "оригіналу"), яке вивчають, розглядають інше, подібне до нього у деяких суттєвих аспектах явище, що має назву "модель-образ". Під моделюванням розуміють процес побудови, вивчення та використання моделі-образу. Моделювання тісно пов'язане з такими науковими категоріями, як абстракція, аналогія, гіпотеза тощо.

Пізнавальні можливості моделі зумовлені тим, що вона імітує певні істотні риси об'єкта-оригіналу. При побудові моделі зазвичай поза увагою залишаються певні аспекти об'єкта моделювання. Цей процес (за суттю суб'єктивний) називають *абстрагуванням*. Від цього залежить *адекватність* моделі, тобто її відповідність об'єкту моделювання з погляду можливості перенесення на останній висновків, що одержані за допомогою моделі. Здобуті знання щодо моделі переносять на оригінал та одночасно переходять з мови моделі на мову оригіналу, вважаючи побудовану модель такою, що адекватна оригіналу. Для створення узагальнюючої теорії про модельований об'єкт здійснюють практичну перевірку результатів, що здобуті шляхом виконання модельних експериментів та їх використання.

Припустимо, що деякий об'єкт Q має деяку властивість C_0, яка нас цікавить. Для отримання ЕММ, яка описує цю властивість, користуються таким **алгоритмом**:

Крок 1. Визначити показник властивості (тобто міру властивості в деякій системі вимірювання).

Крок 2. Установити перелік властивостей C_1, C_2,… C_m, з якими властивість C_0 пов'язана деякими відношеннями (це можуть бути внутрішні властивості об'єкта та властивості зовнішнього середовища).

Крок 3. Описати у вибраній форматній системі властивостей зовнішнього середовища зовнішні фактори x_1, x_2,… x_n, які впливають на шуканий показник Y, внутрішні властивості z_1, z_2,… z_l, а невраховані властивості зарахувати до групи неврахованих факторів w_1, w_2,… w_s.

Крок 4. З'ясувати, за можливості, закономірні відношення між Y та всіма врахованими факторами й параметрами, скласти математичний опис (модель) економічного явища чи процесу.

Основним питанням математичного моделювання є таке: наскільки точно складена математична модель відображає відношення між врахованими факторами, параметрами та показником, що оцінюють властивості реального об'єкта.

Етапи побудови
економіко-математичних моделей

І етап. **Визначення економічної проблеми та проблемної ситуації** (напр., великі витрати на обсяги експорту, державний борг, збитки у міжнародній торгівлі, спад обсягу продаж, інфляція тощо). На цьому етапі формулюють предмет і мету дослідження.

ІІ етап. **Постановка задачі**. У досліджуваній економічній системі виділяють структурні та функціональні елементи, що відповідають меті; виявляють найважливіші якісні характеристики цих елементів. На цьому етапі відбуваються:

- визначення об'єкта моделювання (група країн, країна, міжнародний фінансовий ринок, світовий регіон, ТНК тощо);
- дослідження зовнішнього середовища об'єкта;
- якісний опис взаємозв'язків між елементами моделі;
- визначення мети моделювання й критеріїв її досягнення.

ІІІ етап. **Формалізації задачі**. На цьому етапі формулюють математичну модель:

- вводять позначення для характеристик економічного об'єкта, які враховують при дослідженні;
- формалізують, наскільки можливо, взаємодію між характеристиками економічного об'єкта;
- вводять до змістовного опису математичні символи й позначення, а також математичний запис мети моделювання.

ІV етап. **Вибір методу моделювання**.

V етап. **Процес моделювання**. На цьому етапі відбуваються:

- підготовка плану числових та імітаційних експериментів;
- підготовка вихідних даних;
- комп'ютерні розрахунки за математичною моделлю;
- аналіз здобутих розв'язків;
- оцінювання результатів моделювання та покращення моделі.

VI етап. **Практичне застосування**, на якому відбувається практична перевірка здобутих за допомогою моделі знань і використання їх для побудови узагальнюючої теорії щодо функціонування об'єкта чи керування ним.

Розглянемо моделі, змінними в яких виступають показники міжнародних економічних відносин: функції експорту

та імпорту; фактори цін експорту та імпорту; поточний рахунок і ВВП; фінансовий рахунок валютний курс і відсоткові ставки; міграція робочої сили.

1.2. Моделі міжнародної торгівлі

Функція експорту в неявному вигляді:

$$X = F\left(ER \cdot \frac{P_F}{P_D}; YWR; Y_P; HD \right), \quad (1.1)$$

де X – реальний експорт; ER – валютний курс; P_D – внутрішні ціни; P_F – експортні ціни або ціни за кордоном; YWR – реальний світовий попит; Y_P – виробничий потенціал сектору, що орієнтований на експорт; HD – реальний внутрішній попит.

Функцію експорту також можна подати у такому вигляді:

$$X = F(GDP; GNPN; DEFX; DEFGNPW), \quad (1.2)$$

де X – реальний (у постійних цінах) експорт; GDP – реальний ВВП; $GNPN$ – реальний світовий ВНП; $DEFX$ – індекс експортних цін (дефлятор експорту); $DEFGNPW$ – індекс цін світового ВНП.

Експорт туристичних послуг розраховують за формулою:

$$T_e = a_0 + a_1 y_r^f + a_2 \left(P_d / E_{d/f} \cdot P_f \right), \quad (1.3)$$

де T_e – експорт туристичних послуг; y_r^f – реальний дохід у країнах, звідки прибувають туристи (зважений, відповідно до часток країн у сумарній кількості туристів, що прибувають); P_d – індекс внутрішніх цін (індекс цін послуг готелів і ресторанів або дефлятор ВВП); $E_{d/f}$ – валютний курс; P_f – індекс цін у країнах, звідки прибувають туристи.

У моделі зазвичай передбачено достатні лаги, наприклад рішення про поїздку часто формують заздалегідь, виходячи із зароблених раніше доходу.

Функція імпорту в неявному вигляді:

$$M = F\left(ER \cdot P_f / P_d; Y_d \right), \quad (1.4)$$

де M – реальний імпорт; ER – валютний курс; P_d – внутрішні ціни; P_f – експортні ціни або ціни за кордоном; Y_d – реальний попит у країні (або реальний ВВП).

Функцію імпорту також можна подати у такому вигляді:

$$M = F\left(GDP; DEFM/DEFGDP\right), \quad (1.5)$$

де M – реальний імпорт; DGP – реальний ВВП; $DEFM$ – індекс імпортних цін; $DEFGDP$ – індекс цін (дефлятор) ВВП.

Можна побудувати функцій експорту та імпорту у розрізі окремих галузей (напр., машинобудування).

Імпорт туристичних послуг розраховують за формулою:

$$im_r^d = a_0 + a_1 y_r^f + a_2\left(E_{d/f}(P_{im}/P_d)\right) + a_3 D_e, \quad (1.6)$$

де im_r^d – попит на реальний імпорт; y_r^f – реальний ВВП; $E_{d/f}$ – номінальний валютний курс (од. нац. вал./дол.), середній за період; P_{im} – дефлятор ВВП, або індекс споживчих цін у країнах-контрагентах; P_d – індекс внутрішніх цін; D_e – надлишковий попит (реальний ВВП мінус трендовий реальний ВВП).

1.3. Моделі торгівельного та платіжного балансу

Розглянемо модель ВВП:

$$BIP = c\cdot(BIP - T) + G + I + X - m\cdot BIP, \quad (1.7)$$

де BIP – ВВП; $BIP = c$ – гранична схильність до споживання; T – чисті податки; G – державні витрати; I – приватні інвестиції; X – експорт; m – частка імпорту у ВВП.

Після модифікації, дістанемо:

$$BIP = -c\cdot T/(1-c+m) + G/(1-c+m) + (I+X)/(1-c+m). \quad (1.8)$$

Валовий національний використовуваний дохід і ВВП пов'язані співвідношенням:

$$BII = BIP + Yf + Trf, \quad (1.9)$$

де BII – валовий національний використовуваний дохід; BIP – ВВП; Yf – чисте надходження доходів; Trf – чисте надходження зовнішніх трансфертів.

Поточний рахунок платіжного балансу моделюють за формулою:

$$\frac{CA_t}{Y_t} = \frac{TB_t}{Y_t} + r_t\cdot\frac{NFA_{t-1}}{Y_t}, \quad (1.10)$$

де CA – поточний рахунок платіжного балансу; TB – торговельний баланс; r – номінальна відсоткова ставка; NFA – чисті закордонні активи; Y – номінальний ВВП.

Поточний рахунок платіжного балансу віддзеркалює чисті внутрішні заощадження:

$$(SPR - IANPR) + (T - G) = X - M + Yf + Trf = CA, \quad (1.11)$$

де *SPR* – приватні заощадження; *IANPR* – приватні інвестиції; *T* – чисті податки; *G* – державні витрати; *X*– експорт; *M* – імпорт; *Yf* – чисте надходження доходів; *Trf* – чисте надходження зовнішніх трансфертів; *CA* – поточний рахунок.

Модель впливу платіжного рахунку на організовані заощадження населення має вигляд:

$$Sorg = a_0 + a_1 \cdot \pi + a_2 \cdot Sunorg + a_3 \cdot curr\_acc + a_4 \cdot Sorg_{t-1}, \quad (1.12)$$

де *Sorg* – схильність до організованих заощаджень населення, тобто відношення таких заощаджень до доходів населення (вклади до банків і небанківських депозитних установ, придбання цінних паперів, вкладення у страхові поліси); *Sunorg* – схильність до неорганізованих заощаджень населення (готівка, тобто грошовий агрегат М0); π – інфляція; *curr_acc* – сальдо рахунку операцій з капіталом і фінансових операцій (% від ВВП).

1.4. Моделі міжнародних фінансів

Грошова маса та валютні резерви пов'язані співвідношенням

$$dDCg + dCp + dR = dM, \quad (1.13)$$

де *dDCg* – зміна кредиту державному сектору; *dCp* – зміна кредиту приватному сектору; *dR* – зміна валютних резервів; *dM* – зміна обсягу грошей.

Залежність фондового індексу ПФТС від відпливу капіталу за кордон:

$$PFTS = a_1 \cdot GDP_{t-1} + a_2 \cdot OUTFLOW + a_3 \cdot DEPOSIT_2 + a_0, \quad (1.14)$$

де *PFTS* – зміна індексу ПФТС за квартал (%); GDP_{t-1} – реальний ВВП у попередньому кварталі (млн грн); *OUTFLOW* – приріст фінансових активів за прямими, портфельними та іншими інвестиціями за квартал (% ВВП); $DEPOSIT_2$ – приріст строкових депозитів у банках України за квартал (% ВВП).

Визначення рівноважного валютного курсу відбувається згідно з монетарною моделлю:

$$s_t = Ms_t^* - ms_t + \varphi y_t - \varphi^* y_t^* - \lambda r_t + \lambda^* r_t^*, \quad (1.15)$$

де *S* – номінальний валютний курс; *ms* – пропозиція грошей; *y* – ВВП; *r* – номінальна відсоткова ставка; φ, λ – регресійні коефіцієнти; * – позначає іноземні показники. Усі змінні, крім відсоткової ставки, – натуральні логарифми.

За режиму таргетування грошової маси та валютного курсу відсоткову ставку визначають так:

$$rs_t = rs_{t-1} + a(m_t - m_t^*) + b(er_t - er_t^*), \quad (1.16)$$

де *rs* – короткострокова номінальна відсоткова ставка; *m* – логарифм грошових балансів; *er* – номінальний валютний курс; *a*, *b* – коефіцієнти для цілей таргетування грошової маси та валютного курсу; * – позначає цільовий рівень, а не фактичний.

За режиму таргетування інфляції відсоткова ставка:

$$rs_t = rr_t^* + \pi_t^e + m(\pi_t - \pi_t^*) - \nu(y_t - y_t^*), \quad (1.17)$$

де *rs* – короткострокова номінальна відсоткова ставка; *rr* – реальна відсоткова ставка; π^e – очікувана інфляція; π – інфляція; *y* – розрив ВВП (*output gap*); *m*, *v* – коефіцієнти для цілей таргетування інфляції і розриву ВВП; * – позначає цільовий рівень, а не фактичний.

1.5. Моделі міжнародної міграції

Залежність трудової еміграції від показників ринку праці:

$$EM = a_0 + a_1 \cdot WAGE - a_2 \cdot POPUL + a_3 \cdot UNEMPL, \quad (1.18)$$

де *EM* – річна кількість емігрантів; *WAGE* – номінальна заробітна плата; *POPUL* – економічно активне населення; *UNEMPL* – навантаження незайнятих трудовою діяльністю громадян на одне вільне робоче місце.

Було здійснено моделювання факторів впливу на міграційні процеси в країнах Європейського Союзу та визначено вплив окремих показників соціально-економічного розвитку країни на міграційні показники[1]. Моделювання проводи-

[1] Див. : Hrysenko M., & Pryiatelchuk O. Modelling the factors influencing migration processes in the European Union. Economic Annals-XXI, 183(5-6), 26-42. doi: https://doi.org/10.21003/ea.V183-03. (2020).

ли окремо для показників M_1 – імміграційні та M_2 – еміграційні потоки. Моделі мають наступний вигляд:

$$M_1 = -103300 + 0,1X_1 + 198200X_2 + 460,8X_3 + \\ + 10240X_4 + 1650X_5. \quad (1.19)$$

$$M_2 = 26590 - 0,4X_1 + 150600X_2 + 30,9X_3 + \\ + 1211,0X_4 - 189,4X_5. \quad (1.20)$$

Основні досліджувані фактори для аналізу:

X_0 – індекс конкурентоспроможності;

X_1 – ВВП на душу населення;

X_2 – ВВП у % світового ВВП;

X_3 – прямі іноземні інвестиції;

X_4 – рівень безробіття;

X_5 – рівень оподаткування заробітної плати.

Розділ 2
ДЖЕРЕЛА МІЖНАРОДНОЇ ЕКОНОМІЧНОЇ СТАТИСТИКИ

Розглянемо широкий спектр джерел міжнародної статистики (які умовно згруповано за сферами) – переважно ті, які надають порівнювану інформацію за різними країнами на безкоштовній основі. Але є й джерела, доступ до яких обмежено. Додаткова інформація міститься на сайтах статистичних органів окремих країн, центральних банків, інтеграційних об'єднань, недержавних установ тощо.

Під час перегляду статистичної інформації важливо розуміти точні визначення показників. Зокрема, якщо йдеться про ВВП, то потрібно розуміти: за який період, у якій валюті, яким є спосіб перерахунку, у яких цінах тощо. Якщо йдеться про короткостроковий борг, то потрібно розуміти: чи включає він частину довгострокового боргу, до погашення якої залишається менше року.

Більшість джерел статистичної інформації містять інформацію про показники та методику їх розрахунку, зокрема у розділі *Метадані*/*Metadata*. Іноді опції щодо способу вимірювання показника обирають у процесі формування запиту. Іншим варіантом є деталізація інформації в окремих частинах електронної таблиці (аркуші, рядки, стовпчики) або у додатках до текстового файлу.

2.1. Формат даних

Доступ до даних може бути організований у різні способи:

- доступні для скачування файли з електронними таблицями у форматах xls, xlsx, csv тощо;
- доступні для скачування текстові файли з таблицями, але формату pdf;
- онлайн-таблиці;

- онлайн-таблиці, що сформовані на основі багатокрокового *запиту*/*query*, в якому користувач указує параметри даних, що бажає одержати; результати у такому випадку часто можна скачати у формі електронних таблиць;
- карти або діаграми та графіки – деякі сайти мають інструменти для візуалізації статистичних даних;
- автоматичне оновлення інформації у власній базі на основі API (*application programing interfaces*).

Якщо електронні таблиці зберігаються в іншому форматі, ніж xls або xlsx, наприклад csv, то у Microsoft Excel вони можуть не відкритися належним чином. Тоді необхідно:

- перш за все, змінити розширення файлу на txt у файловому менеджері;
- у полі тип файлу обрати *Всі файли*/*All files* для того, щоб побачити файли всіх форматів;
- у Microsoft Office Excel у меню *Файл*/*File* обрати *Відкрити*/*Open*, знайти у діалоговому вікні потрібний файл і відкрити його.

Далі у новому діалоговому вікні (рис. 2.1) у полі *Вкажіть тип даних*/*Choose the best type that describes your data* необхідно обрати опцію *з розділювачами*/*Delimited* (у більшості випадків) і натиснути кнопку *Далі*/*Next*; обрати у полі *символом-розділювачем є*/*Delimiters* опцію *Кома*/*Comma* (рис. 2.2). Внизу вікна текст і цифри тепер буде розподілено правильно за стовпчиками. Тепер залишилося натиснути *Готово*/*Finish*, і таблиця відобразиться правильно.

Можна також зберегти файл у форматі xls або xlsx. Якщо після вказаних дій таблиця не відображається правильно, то спробуйте інші варіанти (*Фіксованої ширини*/*Fixed width* або інші знаки в полі *символом-разділювачем є*/*Delimiters*).

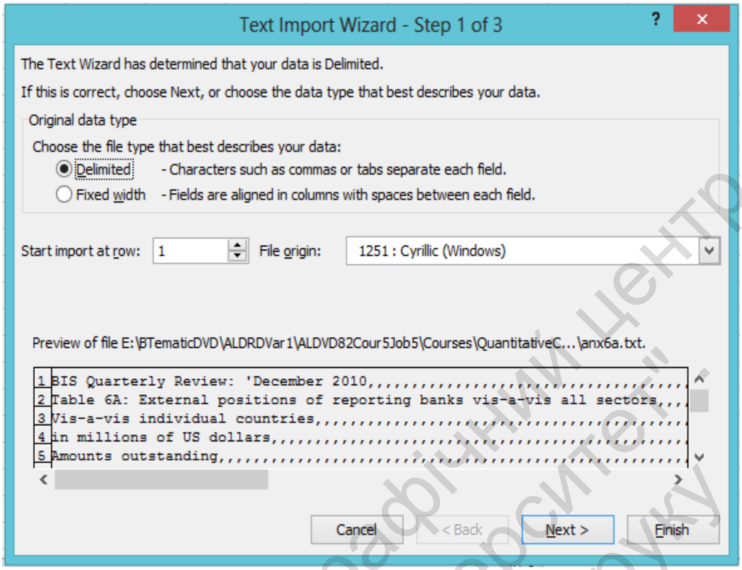


Рис. 2.1

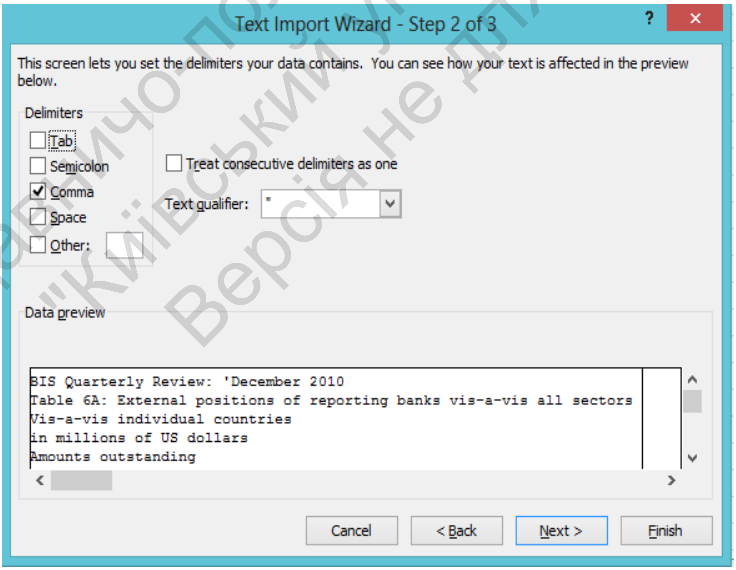


Рис. 2.2

2.2. Статистика Світового банку. Світові індикатори розвитку

Світовий банк на своєму сайті дає можливість скачати велику кількість міжнародних економічних статистичних даних[2]. Низку статистичних баз даних Світового банку розглянемо далі, а поки сконцентруємося на основній базі даних **Світові індикатори розвитку/World Development Indicators** – *WDI*[3], де можна сформувати запит для отримання інформації за цією або іншими *базами даних*/*Database* (разом доступно 86 станом на 2022 р.) щодо необхідних *показників*/*Series* (1443), *країн або їх груп*/*Country* (266) та *часових періодів*/*Time* (частина показників доступна з 1960 р.). Повний файл усієї бази даних у форматі xlsx або csv доступний для скачування з цієї сторінки[4].

Пріоритет щодо одиниць виміру надають одиницям, а не тисячам, мільйонам чи мільярдам (доларів або інших одиниць виміру) або відсоткам (зокрема, від іншого показника, напр., від ВВП тощо). Хоча це незручно візуально (можна побачити великі числа типу 2354168908 або 2.35E+09), але з такими одиницями виміру зручніше проводити подальші розрахунки, не переймаючись тим, що, наприклад, перед тим, як поділити мільярди на мільйони, мільярди потрібно перетворити на мільйони, помноживши на 1000. Розділ *Метадані*/*Metadata* містить також пояснення щодо країн і показників. У табл. 2.1 подано класифікацію показників для кращого орієнтування у базі даних, а у табл. 2.2 – корисні поради щодо термінології і скорочень.

[2] http://data.worldbank.org/data-catalog або https://databank.worldbank.org/home.aspx

[3] https://databank.worldbank.org/source/world-development-indicators

[4] https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators

Таблиця 2.1

Economic Policy & Debt – економічна політика і борг	Balance of payments – платіжний баланс	Current account – поточний рахунок	Balances – баланси
			Goods, services & income – товари, послуги та доходи
			Transfers – поточні трансферти
		Capital & financial account – рахунок операцій з капіталом і фінансовий рахунок	
		Reserves & other items – резервні активи та інші статті	
	External debt – зовнішній борг	Debt outstanding – непогашений борг	
		Debt ratios & other items – коефіцієнти заборгованості та інші показники	
		Debt service – обслуговування боргу	
		Net flows – чисті потоки	
	National accounts – національні рахунки – ВВП та його складові	Adjusted savings & income – скориговані заощадження і дохід	
		Atlas GNI & GNI per capita – валовий національний дохід за методом атлас (у цілому і на душу населення)	
		Growth rates – темпи зростання	
		Local currency at constant prices – у національній валюті в постійних цінах	Aggregate indicators –узагальнені показники
			Expenditure on GDP – ВВП за витратами
			Other items – інші статті
			Value added – у розрізі секторів економіки
		Local currency at current prices – те саме у національній валюті у поточних цінах	
		US$ at constant 2000 prices – те саме у постійних цінах у доларах у цінах 2000 року	
		US$ at current prices – те саме у доларах у поточних цінах	
		Shares of GDP & other – частки ВВП	
	Official development assistance – офіційна допомога у розвитку		
	Purchasing power parity – паритет купівельній спроможності		

Продовження табл. 2.1

Financial Sector – фінансовий сектор	Access – доступність	
	Assets – активи	
	Capital markets – ринки капіталу	
	Exchange rates & prices – валютні курси та ціни	
	Interest rates – процентні ставки	
	Monetary holdings (liabilities) – монетарні активи та зобов'язання	
Private Sector & Trade – приватний сектор і торгівля	Business environment – регуляторне середовище для бізнесу	
	Exports – експорт	
	Imports – імпорт	
	Private infrastructure investment – приватні інвестиції в інфраструктуру	
	Tariffs – тарифи	
	Total merchandise trade – торгівля товарами	
	Trade facilitation – стимулювання торгівлі	
	Trade indexes – торговельні індекси	
	Travel & tourism – подорожі та туризм	
Public Sector – державний сектор	Conflict & fragility – конфлікти і вразливість	
	Defense & arms trade – оборона і торгівля зброєю	
	Government finance – державні фінанси	Deficit & financing – дефіцит бюджету та його фінансування
		Expense – видатки
		Revenue – доходи
	Policy & institutions – політика та інституції	

Продовження табл. 2.1

Social Protection & Labor – соціальний захист і праця	Economic activity – економічна активність
	Labor force structure – структура робочої сили
	Migration – міграція
	Unemployment – безробіття
	Performance – ефективність соціального захисту
Poverty – бідність	Shared prosperity – спільний добробут
	Income distribution – розподіл доходу
	Poverty rates – коефіцієнти бідності
Infrastructure – інфраструктура	Communications – комунікації
	Technology – технологія
	Transportation – транспорт
Education – освіта	Efficiency – ефективність
	Inputs – показники на вході
	Outcomes – результати
	Participation – участь
Environment – навколишнє середовище	Agricultural production – сільськогосподарське виробництво
	Biodiversity & protected areas – біорізноманіття та заповідники
	Density & urbanization – щільність розселення та урбанізація
	Emissions – викиди газів
	Energy production & use – виробництво та споживання енергії
	Freshwater – прісна вода
	Land use – використання землі
	Natural resources contribution to GDP – внесок природніх ресурсів до ВВП

Закінчення табл. 2.1

Gender – гендерні показники	Health – охорона здоров'я	
	Participation & access – участь і доступ	
	Public life & decision making – громадське життя та участь у прийнятті рішень	
Health – охорона здоров'я	Disease prevention – попередження хвороб	
	Health systems – системи охорони здоров'я	
	Mortality – смертність	
	Nutrition – харчування	
	Population – населення	Dynamics – динаміка
		Structure – структура
	Reproductive health – репродуктивне здоров'я	
	Risk factors – фактори ризику	
Social: health/Universal Health Care – вплив медичних витрат осіб на їх економічний добробут		

Таблиця 2.2

Net	чистий (надходження мінус витрати, або активи мінус зобов'язання)
% of GDP	у % від ВВП
BoP	платіжний баланс
Payments	виплати нерезидентам
Receipts	надходження від нерезидентів
Personal remittances	грошові перекази осіб і оплата праці тимчасових мігрантів
Secondary income/Current transfers	вторинний дохід/поточні трансферти
Outflows	відплив
Inflows	надходження
Foreign direct investment	прямі іноземні інвестиції

Продовження табл. 2.2

Portfolio investment	портфельні інвестиції
Reserves and related items	зміна валютних резервів (мінус означає збільшення валютних резервів країни)
Total reserves	валютні резерви
Stocks	показник станом на певну дату, наприклад, прямі інвестицій, накопичені до вказаного моменту за всі попередні роки
Flows	за період, наприклад прямі інвестиції, які надійшли протягом року
External debt stocks, total	загальний зовнішній борг (короткостроковий + державний + гарантований державою + приватний негарантований + використання кредитів МВФ)
Public and publicly guaranteed (PPG)	державний або гарантований державою (борг)
Private nonguaranteed (PNG)	приватний негарантований державою (борг)
External debt stocks, short-term	короткостроковий зовнішній борг, який має строк менше року з самого початку. Не включає частину довгострокового боргу, яка має погашатися протягом року, тому дані щодо короткострокового боргу є фактично заниженими
Debt service	обслуговування боргу – виплата основної частини боргу та відсотків
Interest	процентні платежі
GNI	валовий національний дохід
IMF repurchases and charges	повернення кредитів МВФ і процентів
IFC	Міжнародна фінансова корпорація
Atlas method	метод Атлас передбачає перерахунок у долари не за поточним курсом, а за середнім курсом останнім часом з поправкою на різницю в інфляції з США, Єврозоною, Великою Британією і Японією
Per capita	на душу населення
Annual % growth	приріст у % на рік
General government	загальний уряд (включає не тільки центральний уряд)
Fixed capital	основні засоби
Gross capital formation	інвестиції в основні засоби та збільшення запасів
Manufacturing	обробна промисловість
Industry	промисловість у цілому
LCU	в одиницях національної валюти

Продовження табл. 2.2

Constant	у постійних цінах
Gross domestic savings	валові внутрішні збереження
Changes in inventories	зміна запасів
Current	у поточних цінах
Trade	сума експорту та імпорту (зовнішньоторговельний оборот)
2015 US$	у доларах у цінах 2015 року
External balance on goods and services	зовнішньоторговельний баланс (експорт мінус імпорт)
PPP	у перерахунку за коефіцієнтом паритету купівельної спроможності (ПКС)
PPP conversion factor	коефіцієнт ПКС (співвідношення цін у даній країні та США)
Price level ratio of PPP conversion factor (GDP) to market exchange rate	відношення коефіцієнту ПКС до валютного курсу
Foreign assets	закордонні активи
Claims on central government	чисті вимоги до центрального уряду (за мінусом депозитів державного сектору)
Adjusted savings	скориговані заощадження (з поправкою на зміни в навколишньому середовищі та витрати на освіту)
S&P Global Equity Indices	ціни на акції в доларах
Market capitalization of listed companies	ринкова капіталізація (кількість акції всіх компаній (крім інституціональних інвесторів) на біржах помножена на їх ціну)
Stocks traded	обсяги торгівлі акціями
Stocks traded, turnover ratio	відношення обсягів торгівлі акціями до ринкової капіталізації
Consumer price index (2010 = 100)	індекс споживчих цін (базовий рік 2010)
Period average	середній за період
Real interest rate	реальна відсоткова ставка по кредитах (номінальна відсоткова ставка мінус інфляція за дефлятором ВВП)
Risk premium on lending(prime rate minus treasury bill rate, %)	премія за ризик (ставка за кредитами першокласним позичальниками мінус ставка за короткостроковими облігаціями уряду)

Закінчення табл. 2.2

Broad money	широкі гроші (готівкові кошти, депозити крім урядових, чеки, депозитні сертифікати і комерційні папери)
Age dependency ratio (% of working-age population)	коефіцієнт вікової залежності (відношення кількості людей до 15 і з 64 років до кількості працюючого населення)
High-technology exports	високотехнологічний експорт (аерокосмічний, фармацевтичний, наукових приладів, електротехніки, комп'ютерів)
Freight	вантажні перевезення
International migrant stock	кількість іммігрантів у країні (які народилися за кордоном, для СРСР – також в іншій республіці)
Income share held by lowest 10%	частка доходу або споживання, яке належить найбіднішим 10% населення
GINI index	індекс Джині (0 – повна рівність у доходах населення, 100 – повна нерівність)
Average time to clear exports through customs (days)	середня кількість днів для розмитнення експорту
Cost to export (US$ per container)	вартість експортних митних процедур для 1 контейнера (за винятком мита та податків на торгівлю)
Merchandise exports	експорт товарів
Bound rate, simple mean, all products	максимальні ставки митного тарифу, погоджені внаслідок переговорів – середня незважена по всіх товарах
Share of tariff lines with international peaks	частка продуктових ліній, за якими ставка митного тарифу більше 15 %
Tariff rate, applied, simple mean	середня незважена ставок митного тарифу, які фактично застосовуються
Most favored nation	для країн, які користуються режимом найбільшого сприяння
Weighted mean	зважена середня
Logistics performance index: Overall (1=low to 5=high)	індекс ефективності логістики: загальний (1 – низький, 5 – високий)
Net barter terms of trade index (2000 = 100)	умови торгівлі (відношення експортних цін до імпортних – щодо базового 2000 р.)
International tourism, expenditures	імпорт туристичних послуг
International tourism, receipts	експорт туристичних послуг
Arms imports	імпорт зброї
Taxes on international trade	податки на міжнародну торгівлю (експортні та імпортні мита, прибутки експортних та імпортних монополій, валютні прибутки та податки)

2.3. Джерела комплексної статистики

Система доступу до баз даних ООН **UNData**[5] дає можливість доступу принаймні до частини показників із більш ніж 30 баз даних або продивитися основну статистику щодо окремих країн або регіонів світу.

Euromonitor International Passport GMID[6] є джерелом комплексної інформації за країнами, компаніями та ринками. Доступ є платним. Ця база даних більш зорієнтована на аналітику для бізнесу у сфері маркетингу, а не тільки на макроекономічний аналіз. За плату можна також одержати огляди ринків конкретних груп товарів за окремою країною. Крім звичайних макроекономічних даних база має інформацію за такими показниками:

- витрати споживачів у розрізі груп товарів або послуг;
- індекси цін за окремими групами товарів;
- ціни на конкретні товари;
- більш деталізована за галузями структура ВВП;
- детальна статистика зі споживання, виробництва, запасів енергії за видами палива чи джерелами електроенергії;
- детальні екологічні показники (зокрема, обсяги використання вторинної сировини з алюмінію, групи тварин під загрозою зникнення тощо);
- експорт та імпорт у розрізі країн-контрагентів і груп товарів;
- детальні показники охорони здоров'я;
- види домогосподарств (зокрема, із власним житлом, кредитним, орендованим, з кухнею тощо);
- популяція домашніх тварин;
- кількість людей, які мають ті чи інші товари (зокрема, комп'ютер, піаніно, кондиціонер, мікрохвильова піч тощо);
- дані з розподілу доходу, такі як доходи людей у розрізі вікових груп, кількість населення з відповідним доходом (зокрема, від 10 до 20 тис. дол.);
- виробництво конкретних видів товарів;

[5] http://data.un.org

[6] https://www.euromonitor.com/our-expertise/passport

- сфера розваг (зокрема, відвідування кінотеатрів, кількість музеїв за видами тощо);
- тренди щодо споживачів (зокрема, індивідуалізація; сегментація споживачів за різноманітними параметрами: студенти, молоді фахівці, батьки, геймери, туристи тощо);
- демографія;
- зайнятість;
- освіта;
- транспорт (зокрема, ціни на пальне, нещодавно зареєстровані двоколісні транспортні засоби, кількість човнів за видами, автобани тощо).

Статиста/Statista[7] також надає доступ до комплексних статистичних даних за країнами та ринками, але різної деталізації, залежно від рівня оплати послуг доступу чи вибору безкоштовного аккаунту.

Світові економічні перспективи МВФ/IMF World Economic Outlook[8] надає історичні, поточні та прогнозні дані за низкою показників у розрізі країн і груп країн. Показники включають ВВП за різними способами вимірювання, інфляцію, доходи та витрати державного бюджету, поточний рахунок платіжного балансу та деякі інші показники, залежно від країни.

Конференція ООН з торгівлі та розвитку ЮНКТАД /UNCTAD[9] надає інформацію за різними сферами, такими як міжнародна торгівля товарами та послугами (зокрема, торгівельна матриця, в якій можна знайти продуктову структуру двосторонньої торгівлі між конкретними країнами), зовнішні фінансові ресурси, ВВП, населення, інформаційна економіка, креативна економіка, морський транспорт тощо.

Світова книга фактів ЦРУ/CIA The World Factbook[10] дає можливість одержати статистичну та фактологічну інформацію щодо кожної країни в найрізноманітніших сферах (історія, люди, економіка, географія, комунікації, транспорт, збройні сили, транснаціональні проблеми); містить як фактичні дані,

[7] https://www.statista.com
[8] http://www.imf.org/external/ns/cs.aspx?id=28
[9] http://unctadstat.unctad.org/EN/
[10] https://www.cia.gov/the-world-factbook

так й оціночні; надає можливість ранжування країн за конкретним показником.

Nation Master[11] містить інформацію з широкого кола статистичних показників і фактологічної інформації; має орієнтацію на ілюстративність даних. Наприклад, надає можливість одержати дані щодо країни, порівнювати країни за різними показниками, продивлятися рейтинги країн за обраним показником (зокрема, за секторами економіки) та ілюстративні карти.

Gap Minder[12] містить інформацію майже за 500 показниками країн у різних сферах із різних джерел. Gap Minder надає широкі можливості для візуалізації, зокрема, побудови чотиривимірного графіка, де за осями подано дві змінні (напр., тривалість життя та ВВП на душу населення), величиною бульбашок – третя, зважуюча, змінна (напр., кількість населення), а четвертий вимір забезпечує кнопка *Play* для анімації змін у часі.

Trading Economics Overview[13] надає найновішу доступну інформацію у розрізі країн, макроекономічних показників або фінансових чи товарних ринків. Тут подано також прогнози на наступні квартали, графіки історичної динаміки показників, рейтинги країн за обраним показником.

Ресурс **Countryeconomy**[14] є інструментом для порівняння країн за різними групами статистичних показників.

Організація економічного співробітництва та розвитку OECD[15] є джерелом широкого спектру статистичних даних щодо країн – членів цієї організації.

Ресурс **Worldometer**[16] показує оцінені зміни різних показників у режимі реального часу (залежно від наявних тенденцій шляхом екстраполяції).

В ЄС комплексним джерелом статистичних даних є **Євростат/Eurostat**[17]. Його база даних дає можливість знайти ве-

[11] http://www.nationmaster.com
[12] http://www.gapminder.org
[13] http://www.tradingeconomics.com
[14] https://countryeconomy.com
[15] http://www.oecd-ilibrary.org/statistics
[16] https://www.worldometers.info

лику кількість показників, але переважно лише за країнами ЄС, країнами – кандидатами на членство, країнами Європейської асоціації вільної торгівлі та іноді – найбільшими економіками світу.

Основна форма доступу – *Запит/Query*. Для здійснення запиту в *Ієрархічному дереві/Data navigation tree* показників обирають потрібний. Наприклад, у *Database by Themes* обирають *Economy and finance* → *Balance of payments – International transactions* → *Balance of payments statistics and International investment positions* (*bop_q6*) → *Balance of payments by country – annual data* (*BPM6*) (*bop_c6_a*).

У новому діалоговому вікні натискають плюс біля одного із параметрів запиту. У вкладці *Currency* обирають одиницю виміру (валюту), наприклад, *Millions of euro*; у вкладці *Flow* – вид потоку: за *Кредитом/Credit* (надходження із закордону), за *Дебетом/Debit* (виплати за кордон) або *Чисті/Net* (надходження мінус виплати). Далі у вкладці *Geo* обирають країну або групу країн; у вкладці *Partner* – країну або групу країн-партнерів; у вкладці *BOP_item* – конкретний вид показника, наприклад, поточний рахунок; у вкладці *Time* – періоди часу. Тепер натискають кнопку *Update*. Таблиця, яка розташована праворуч, буде трансформована, відповідно до умов запиту.

Виміри таблиці можна змінювати, перетаскуючи іконку хрестика зі стрілками навколо параметрів запиту до першої лівої верхньої комірки таблиці. Натиснення кнопки *Download* дозволить закачати файл із цією таблицею. Залишилося тільки обрати формат (напр., xls) і за потреби – додаткові опції файлу.

В Україні комплексним джерелом статистичних даних є **Державна служба статистики України**[18], де дані структуровано за групами показників у розділі "Статистична інформація". Розділ "Експрес-випуски" містить найновішу оприлюднену статистичну інформацію.

---

[17] https://ec.europa.eu/eurostat/web/main/data/database

[18] http://www.ukrstat.gov.ua

2.4. Статистика конкурентоспроможності та середовища для бізнесу

Глобальний звіт з конкурентоспроможності Світового економічного форуму/World Economic Forum Global Competitiveness Report[19] оцінює рейтинги конкурентоспроможності країн за таким складовими: інституції, інфраструктура, адаптація інформаційно-телекомунікаційних технологій, макроекономічне середовище, охорона здоров'я, уміння, ринок товарів, ринок праці, фінансова система, розмір ринку, динаміка бізнесу, інноваційна спроможність.

Міжнародний інститут розвитку менеджменту/International Institute for Management Development – IMD[20] публікує свої рейтинги конкурентоспроможності країн, цифрової конкурентоспроможності, талантів і смарт-міст.

Індекс європейської конкурентоспроможності регіонів /European Regional Competitiveness Index[21] містить інформацію про конкурентоспроможність на рівні регіонів усередині країн ЄС.

Індекс економічної складності/Economic complexity Index[22] фактично використовують як один з індикаторів успішності економіки.

Індекс економічної свободи Інституту Фрейзера/Fraser Institute[23] містить оцінку економічної свободи для країн за такими складовими: величина уряду; правова структура та захист прав власності; доступ до надійних грошей; свобода міжнародної торгівлі; регулювання. Кожна зі складових поділена детальніше. Більшість показників є не точними кількісними показниками, а бальними оцінками якісних показників.

Фундація Спадщина/Heritage Foundation[24] також розраховує індекс економічної свободи, але за іншою методикою.

---

[19] https://www3.weforum.org/docs/WEF_TheGlobalCompetitivenessReport2019.pdf
[20] http://www.imd.org/research/ publications/wcy/index.cfm
[21] https://ec.europa.eu/regional_policy/en/information/maps/regional_competitiveness/
[22] https://oec.world
[23] https://www.fraserinstitute.org/economic-freedom/map?geozone=world&year=2019&page=map
[24] https://www.heritage.org/index/

Здійснення бізнесу/World Bank Doing Business[25] фактично є альтернативним індексом економічної свободи, який ураховує складність бюрократичних процедур та інших аспектів у сфері заснування бізнесу, реєстрації власності, одержання кредиту, захисту інвесторів, сплати податків, транскордонної торгівлі, забезпечення виконання контрактів, розв'язання проблем неплатоспроможності. Проте оновлення рейтингу було припинено. На зміну йому має прийти новий рейтинг **Business Enabling Environment.**

Огляд підприємств/World Bank Enterprise Surveys[26] містить дані про компанії країн у таких сферах: бюрократичні складнощі, податковий тягар, корупційний тягар, втрати від кримінальних злочинів, конкуренція з боку неформального сектору, відключення або складність підключення електроенергії або водопостачання, способи фінансування, напрямки капіталовкладень, способи комунікації, завантаженість виробничих потужностей, важливість експорту та імпорту, структура персоналу, характер власності, розмір тощо.

Міжнародний гід з країнового ризику/PRS Group International Country Risk Guide[27] містить інформацію щодо політичних, фінансових та економічних ризиків за країнами.

Рейтинг країнового ризику Євромані/Euromoney Country Risk Ratings[28] ураховує такі фактори, як політичний ризик, економічна ефективність, структурні оцінки, доступ до банківського фінансування та ринків капіталу, індикатори боргу, кредитні рейтинги.

Компанія E.T. Кірні публікує **Індекс глобальних міст/The A.T. Kearney Global Cities Index**[29], що показує ступінь інтеграції провідних міст із рештою світу за такими складовими: активність бізнесу, людський капітал, інформаційний обмін, культурний досвід, політичне залучення.

[25] https://www.worldbank.org/en/programs/business-enabling-environment
[26] https://datacatalog.worldbank.org/search/dataset/0037947
[27] https://www.prsgroup.com/explore-our-products/icrg/
[28] https://www.euromoneycountryrisk.com
[29] https://www.kearney.com/global-cities/2021

2.5. Статистика міжнародної торгівлі

Світова організація торгівлі/World Trade Organization[30] має кілька баз даних, зокрема:

- Регіональні торговельні угоди.
- Торгівля товарами.
- Торгівля послугами.
- Нетарифні заходи.
- Тарифи.
- Глобальні ланцюги вартості.

Інформаційна система регіональних торговельних угод/ Regional Trade Agreements Information System[31] дає можливість знайти інформацію про кількість різних типів таких угод, а також перейти до їх текстів (додатки містять, зокрема, тарифні графіки, за якими сторони поступово знижують або скасовують імпортні тарифи у взаємній торгівлі). Наявна й інформація щодо **Торговельних спорів/Dispute statistics**[32].

Узагальнена інформація щодо імпортних тарифів за країнами доступна у **Тарифних профілях/Tariff Profiles**[33]. Вони включають:

- основні дані щодо митних тарифів у країні, *розподіл Ставок імпортних тарифів за величиною/Tariffs and imports: Summary and duty ranges*);
- *розподіл Ставок імпортних тарифів та імпорту за групами товарів/Tariffs and imports by product groups*;
- *розподіл Експорту та ставок імпортних тарифів, з якими стикається експорт країни за країнами-контрагентами /Exports to major trading partners and duties faced*).

Для тарифних профілів, зокрема, застосовують такі позначення:

- Ag – для сільського господарства;
- Non-Ag – дня секторів, крім сільського господарства;
- Final bound – зв'язаний (рівень, якій країна зобов'язалася не перевищувати);

[30] https://www.wto.org/english/res_e/statis_e/statis_e.htm
[31] http://rtais.wto.org/UI/PublicMaintainRTAHome.aspx
[32] https://www.wto.org/english/tratop_e/dispu_e/dispustats_e.htm
[33] https://www.wto.org/english/res_e/statis_e/tariff_profiles_list_e.htm

- MFN applied – у рамках режиму найбільшого сприяння – РНС (Most-Favoured-Nation);
- AVG – середній;
- Max – максимальний;
- Duty-free in % – частка імпорту без мита;
- Pref. Margin – преференційна маржа як різниця між звичайною ставкою РНС імпортного тарифу та ставкою преференційного тарифу.

ООН публікує **Статистичну базу даних торгівлі товарами /United Nations Commodity Trade Statistics Database** – UN Comtrade [34].

МВФ має **Статистику напрямів торгівлі/IMF Directions Of Trade Statistics**[35], що містить дані про обсяг двосторонньої торгівлі між країнами: експорт та імпорт. Тут подано географічну структуру зовнішньої торгівлі для кожної країни. МВФ також публікує **Ціни на сировину/IMF Primary Commodity Prices**[36], яка містить тижневі, місячні, квартальні або річні дані щодо цін сировинних товарів.

Світовий Банк має базу **Рішення світової інтегрованої торгівлі/World Bank World Integrated Trade Solution**[37], що є програмним продуктом, який дозволяє доступ до баз даних UN Comtrade, СОТ та ЮНКТАД. Крім доступу до баз даних, WITS є аналітичним інструментом для оцінювання наслідків зміни ставок митних тарифів. Інша база даних – **Тимчасові бар'єри для торгівлі/World Bank Temporary Trade Barriers Database**[38] містить інформацію щодо випадків антидемпінгових процедур, установлення компенсаційного мита, застосування захисних заходів.

Міжнародний торговельний центр/International Trade Center[39] дає можливість одержати для кожної країни та галузі дані з міжнародної торгівлі (обсяги, темпи зростання, рівень диверсифікації, спеціалізації), прямих інвестицій,

[34] https://comtrade.un.org
[35] https://data.imf.org/?sk=9D6028D4-F14A-464C-A2F2-59B2CD424B85
[36] https://www.imf.org/en/Research/commodity-prices
[37] https://wits.worldbank.org/Default.aspx?lang=en
[38] http://data.worldbank.org/data-catalog/temporary-trade-barriers-database
[39] http://www.intracen.org

митних тарифів тощо. Інформація центру може бути корисною при прийнятті рішень про вихід на зовнішні ринки за допомогою міжнародної торгівлі або прямих інвестицій, а також містить інформацію про двосторонню торгівлю різними видами товарів і послуг[40]. Одним із інструментів є **Карта експортного потенціалу/Export Potential Map**[41] у розрізі країн-експортерів, країн-партнерів і груп товарів.

Компанія E.T. Кірні випускає **Індекс глобального розвитку роздрібної торгівлі/The A.T. Kearney Global Retail Development Index**[42], що містить рейтинги країн за їх комерційною привабливістю для розвитку роздрібної торгівлі.

Світова організація туризму/World Tourism Organization[43] містить інформацію про розвиток міжнародного туризму.

В Україні створено базу даних **Державної митної служби**[44], де наявна інформація про торгівлю України у розрізі країн-партнерів, митниць, видів транспорту, груп товарів, а також про митну вартість товарів.

2.6. Статистика міжнародних фінансів

Міжнародний валютний фонд/International Monetary Fund[45] надає можливість знайти широкий спектр показників за країнами, переважно у фінансовій сфері. Тут є посилання на кілька баз даних, основною серед яких є **Міжнародна фінансова статистика/International Financial Statistics** – IFS[46], де наявні річні, квартальні та місячні дані.

У табл. 2.3 подано деякі корисні пояснення щодо термінології і скорочень.

[40] https://intracen.org/resources/trade-statistics
[41] https://exportpotential.intracen.org/en/
[42] https://www.kearney.com/global-retail-development-index
[43] https://www.e-unwto.org/toc/unwtotfb/current
[44] https://bi.customs.gov.ua
[45] https://www.imf.org/en/Data
[46] https://data.imf.org/?sk=4c514d48-b6ba-49ed-8ab9-52b0c1a0179b

Таблиця 2.3

Exchange Rates	офіційний або ринковий валютний курс, валютні курси на кінець періоду та середній за період, до долара та до СДР (СПЗ), крім звичайних курсів вказано ефективний валютний курс, номінальний і реальний
SDR	спеціальні права запозичення (СПЗ або СДР)
International Liquidity	дані про валютні резерви у розрізі складових, також інші закордонні активи та зобов'язання центрального банку та інших банків
Central Bank	структура активів і зобов’язань центрального банку
Depository Corporations	активи та зобов'язання комерційних банків і подібних установ, які приймають гроші на депозити до запиту
Foreign Assets	закордонні активи
Claims on General Government	внутрішні вимоги до загального уряду
Claims on Public Nonfinancial Enterprises	внутрішні вимоги до нефінансових державних підприємств
Claims on Private Sector	внутрішні вимоги до приватного сектору
Reserve Money	грошова база
Time, Savings & Foreign Currency Deposits	строкові, ощадні та валютні депозити
Demand Deposits	депозити до запиту
Foreign Liabilities	зобов'язання перед нерезидентами
General Government Deposits	депозити загального уряду
Capital Accounts	власний капітал
Other Financial Corporations	активи та зобов'язання інших банків та подібних фінансових установ
Financial Corporations Survey	консолідовані дані з активів і зобов'язань усієї банківської системи (включаючи центральний банк, депозитні корпорації, та інші фінансові корпорації)
Interest Rates	відсоткові ставки
Refinancing Rate	ставка рефінансування комерційних банків центральним банком
Money Market Rate	відсоткова ставка грошового (міжбанківського) ринку
Balance of Payments	структура платіжного балансу
Credit	надходження за відповідною статтею платіжного балансу (експорт, надходження доходів і трансферти від нерезидентів)

Закінчення табл. 2.3

Debit	виплати за відповідною статтею платіжного балансу (імпорт, виплати доходів і трансферти нерезидентам)
Assets	закордонні активи за відповідною статтею (плюс означає зменшення активів (приплив капіталу), мінус – збільшення (відплив капіталу))
Liabilities	зобов'язання з відповідної статті (плюс означає – збільшення зобов'язань (приплив капіталу), мінус – зменшення (відплив капіталу))
Reserves and Related Items	Резерви та подібні активи (валютні резерви, використання кредитів МВФ, надзвичайне фінансування)
International Investment Position	міжнародна інвестиційна позиція: закордонні активи резидентів та їх зобов'язання перед нерезидентами у розрізі видів інвестицій
National Accounts	національні рахунки – складові ВВП
Euro-area-wide Residency	метод, за яким для країн Єврозони закордонними активами вважають вимоги до нерезидентів Єврозони, зовнішніми зобов'язаннями вважають зобов'язання перед нерезидентами Єврозони

Міжнародний валютний фонд також оприлюднює кілька інших джерел у сфері міжнародних і внутрішніх фінансів:

- **Статистика платіжного балансу та міжнародна інвестиційна позиція/Balance Of Payments and International Investment Position Statistics**[47] містить детальну інформацію про структуру платіжних балансів кожної країни:

- **Погоджений огляд прямих інвестицій/Coordinated Direct Investment Survey**[48] надає інформацію за країнами у розрізі географічної структури прямих інвестицій і структури за інструментами.

- **Погоджений огляд портфельних інвестицій/Coordinated Portfolio Investment Survey**[49] надає таку саму інформацію з портфельних інвестицій.

- **Валютна структура офіційних валютних резервів/Currency Composition of Official Foreign Exchange Reserves**[50], де подано

[47] https://data.imf.org/?sk=7A51304B-6426-40C0-83DD-CA473CA1FD52
[48] https://data.imf.org/?sk=40313609-F037-48C1-84B1-E1F1CE54D6D5
[49] https://data.imf.org/?sk=B981B4E3-4E58-467E-9B90-9DE0C3367363
[50] https://data.imf.org/?sk=E6A5F467-C14B-4AA8-9F6D-5A09EC4E62A4

інформацію для світу у цілому, але у розрізі окремих країн такої інформації не вказано.

- **Огляд доступу до фінансування/Financial Access Survey**[51] містить інформацію про показники доступності фінансових послуг для клієнтів (зокрема, кількість людей, які мають кредити або депозити на 1000 дорослих; кількість філій банків або банкоматів на 1000 км^2).
- **Індикатори фінансової міцності/Financial Soundness Indicators**[52] надають інформацію про такі показники фінансової системи, як відношення регуляторного капіталу до активів з урахуванням ризику, частка непрацюючих кредитів, прибутковість активів, частка кредитів на нерухомість у всіх кредитах тощо.
- **Статистика урядових фінансів/Government Finance Statistics**[53] містить детальні статистичні дані щодо доходів, витрат, активів і зобов'язань загального уряду та його складових.

Світовий банк також пропонує кілька джерел статистичної інформації у цій сфері:

- **Банківське регулювання та нагляд/Bank Regulation and Supervision**[54] надає інформацію переважно з якісних параметрів банківського регулювання та нагляду за такими складовими: вхід на ринок, власність, капітал, діяльність, зовнішній аудит, внутрішній менеджмент/організаційні вимоги, ліквідність і диверсифікація, захист вкладників, забезпечення, бухгалтерській облік і розкриття інформації, дисципліна/проблемні інституції/вихід з ринку, нагляд.
- **Фінансовий розвиток і структура/Financial Development and Structure**[55] містить інформацію про різноманітні відносні показники фінансової системи.
- **Квартальні дані боргу державного сектору/Quarterly Public Sector Debt**[56] є спільною з МВФ базою даних щодо боргу загального уряду, центрального уряду, нефінансових дер-

---

[51] https://data.imf.org/?sk=E5DCAB7E-A5CA-4892-A6EA-598B5463A34C
[52] https://data.imf.org/?sk=51B096FA-2CD2-40C2-8D09-0699CC1764DA
[53] https://data.imf.org/?sk=a0867067-d23c-4ebc-ad23-d3b015045405
[54] https://www.worldbank.org/en/research/brief/BRSS
[55] https://www.worldbank.org/en/publication/gfdr/data/financial-structure-database
[56] https://datacatalog.worldbank.org/search/dataset/0037906

жавних корпорацій, фінансових державних корпорацій у розрізі за інструментами, термінами та валютами.

- **Квартальні дані зовнішнього боргу/Quarterly External Debt Statistics**[57] містить інформацію про зовнішній борг у розрізі секторів економіки, інструментів, валют; короткострокового зовнішній борг, обслуговування боргу, прострочені виплати.

- **Ціни на грошові перекази/Remittance Prices**[58] містить дані про ціни на грошові перекази невеликих сум грошей у розрізі країни-платника та країни-одержувача.

Банк міжнародних розрахунків/Bank for International Settlements охоплює кілька складових статистики міжнародних фінансів:

- **Платіжні системи/BIS Payment Systems**[59] містить дані щодо платіжних систем, торговельних платформ, клірингових домів і систем розрахунків за цінними паперами.

- **Статистика цін на нерухомість/BIS Property Price Statistics**[60] указує ціни за видами нерухомості (місячні, квартальні, піврічні або річні дані).

- **Статистика боргових цінних паперів/Debt Securities Statistics**[61] містить інформацію щодо міжнародних і внутрішніх боргових цінних паперів.

- **Міжнародна банківська статистика/International Banking Statistics**[62] надає інформацію про зовнішні фінансові вимоги (активи) та зобов'язання банків, зокрема, у розрізі валют, країн і країн-контрагентів.

- **Кредитна статистика/Credit Statistics**[63] надає інформацію про кредитування нефінансового сектору.

- **Статистика деривативів/Derivatives Statistics**[64] містить дані щодо ринку деривативів країн Великої десятки та Швейцарії.

[57] https://www.worldbank.org/en/programs/debt-statistics/qeds
[58] https://databank.worldbank.org/source/remittance-prices-worldwide-(corridors)
[59] https://www.bis.org/statistics/payment_stats.htm?m=1036
[60] https://www.bis.org/statistics/pp.htm?m=2640
[61] https://www.bis.org/statistics/about_securities_stats.htm?m=2638
[62] https://www.bis.org/statistics/about_banking_stats.htm?m=2637
[63] https://www.bis.org/statistics/about_credit_stats.htm?m=2673

▪ **Статистика валютного ринку/BIS Foreign Exchange Statistics**[65] містить інформацію про валютні курси щодо долара, ефективні валютні курси та обсяги торгівлі валютою.

Спільний хаб зовнішнього боргу/Joint External Debt Hub[66] є спільною базою даних Світового банку, Банку міжнародних розрахунків, Міжнародного валютного фонду та Організації економічного розвитку та співробітництва. Статистика представлена у подвійному вимірі: за даними країни-боржника та за даними кредиторів або ринків.

Світова федерація бірж/Worlds Federation Of Exchanges[67] надає інформацію про ринкову капіталізацію акцій, облігацій; оборот торгівлі акціями, облігаціями; капітал, залучений за допомогою випуску акцій; кількість компаній у лістингу; індекси цін на акції тощо.

Суверенний рейтинг Фітч/Fitch Sovereign Rating[68] публікує рейтинги боргу урядів у національній або іноземній валюті.

Інший приклад кредитної рейтингової інформації – **Суверенний рейтинг Стендард енд Пуерс/Standard And Poors Sovereign Rating List**[69].

Індекс впевненості компанії Е.Т. Кірні для прямих іноземних інвестицій/The A.T. Kearney Foreign Direct Investment Confidence Index[70] містить дані, які характеризують умови, що впливають на рішення про здійснення прямих іноземних інвестицій.

В Україні основним джерелом фінансової статистичної інформації є **Національний банк України**[71]. Дані про державний бюджет і державний борг подано на сайті **Міністерства фінансів України**[72].

---

64 https://www.bis.org/statistics/about_derivatives_stats.htm?m=2639
65 https://www.bis.org/statistics/about_fx_stats.htm?m=2674
66 http://www.jedh.org/data.html
67 https://statistics.world-exchanges.org/Account/Login
68 https://www.fitchratings.com/site/home
69 https://www.spglobal.com/ratings/en/ або https://disclosure.spglobal.com/ratings/en/regulatory/ratings-actions
70 https://www.kearney.com/foreign-direct-investment-confidence-index
71 https://bank.gov.ua
72 https://mof.gov.ua/uk

2.7. Статистика компаній

Компанія **Bureau van Dijk** має базу **Orbis**[73] представляє опис можливостей[74]. База даних містить інформацію про більш ніж 300 млн компаній і банків – приватних і державних:

- *загальні відомості*: кількість років функціонування, менеджмент, кількість працівників, юридична форма, номери компанії у реєстрах, продукція та послуги, статут, новини;
- *характер власності*: акціонери, дочірні компанії, рівень незалежності компанії, кінцевий власник, інші компанії у групі, менеджмент, злиття та поглинання, чутки про їхню можливість;
- *фінансова інформація*: показники балансу, звіти про фінансові результати, фінансові коефіцієнти, оцінки фінансової стійкості, ціни на акції;
- *пов'язані особи*: директори, аудитори, радники тощо, їх попередні посади, зв'язок з політично значущими особами;
- *оцінки ризиків*: екологічного, соціального, політичного характеру; участь у публічних тендерах, зв'язок з політично значущими особами.

Міжнародний бізнес-центр/International Business Center і **Бізнес-коледж імені Елі Броуда Мічиганського публічного університету/The Eli Broad College of Business at Michigan State University** розміщують базу даних **globalEDGE**[75]; у розділі Global insights – інформацію за країнами, інтеграційними блоками та галузями (макропоказники міжнародної торгівлі та інформацію щодо найбільших компаній: обсяги продажів, прибуток, активи, ринкова вартість).

База даних Організації економічного співробітництва і розвитку – AMNE[76] містить інформацію про транснаціональні операції багатонаціональних підприємств у розрізі країн і галузей: кількість працівників, торговельний оборот, інвестиції, внутрішній корпоративний експорт та імпорт, додана вартість, витрати на дослідження та розробки тощо.

[73] https://www.bvdinfo.com/en-gb/our-products/data/international/orbis
[74] https://www.bvdinfo.com/en-gb/-/media/brochure-library/orbis.pdf
[75] https://globaledge.msu.edu
[76] https://www.oecd.org/sti/ind/amne.htm

2.8. Статистика інфраструктури

Міжнародна організація цивільної авіації/International Civil Aviation Organization[77] містить дані щодо комерційних авіаперевізників (перевезення, фінансові дані, персонал тощо), аеропортів (перевезення, фінансові дані), провайдерів авіанавігаційних послуг, зареєстрованих цивільних літаків.

Міжнародна дорожня федерація/International Road Federation[78] оприлюднює дані щодо довжини, якості та щільності мережі доріг; обсягів перевезень за видами автотранспорту; забезпеченості автотранспортом; дорожньо-транспортних пригод; виробництва, реєстрації, експорту та імпорту автотранспорту; витрат на дороги; цін і споживання автомобільного пального.

Міжнародний транспортний форум/International Transport Forum[79] надає інформацію про обсяги перевезень, інвестиції до інфраструктури, транспортні податки, викиди вуглекислого газу транспортом тощо.

Індекс ефективності логістики Світового банку/Logistic Performance Index[80] включає такі складові, як ефективність розмитнення, якість торговельної і транспортної інфраструктури, легкість організації відвантаження товарів за прийнятними цінами, якість логістичних послуг, здатність відслідковувати переміщення вантажів, вчасність доставки вантажу.

Енергетична статистика Статистичного відділу ООН/ UNSD Energy Statistics[81] містить дані про виробництво, міжнародну торгівлю, зміни запасів, перетворення, кінцеве споживання (промисловістю, транспортом, домогосподарствами) енергоносіїв за їх видами; виробництво, торгівлю та споживання електроенергії, потужність електростанцій.

Міжнародний телекомунікаційний союз/International Telecommunication Union[82] оприлюднює дані щодо фіксованих

[77] https://data.icao.int/newdataplus
[78] https://www.irf.global/statistics/
[79] https://www.itf-oecd.org/
[80] https://datacatalog.worldbank.org/search/dataset/0038649
[81] https://unstats.un.org/unsd/energystats/
[82] https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

телефонних ліній, мобільного зв'язку, телебачення, інтернету та комп'ютерів, персоналу, тарифів, якості, доходів та інвестицій у сфері телекомунікацій.

База даних Світового банку **Приватна участь в інфраструктурі/World Bank Private Participation in Infrastructure**[83] містить дані про кількість інфраструктурних проєктів та обсяги інвестицій у розрізі секторів (енергетика, транспорт, телекомунікації, водопостачання та водовідведення).

2.9. Демографічна, соціальна, науково-технічна та екологічна статистика

База даних Світового банку **Прогнози населення/World Bank Population Projections**[84] містить довгострокові прогнозні дані щодо кількості населення за країнами, його вікової і статевої структури, показників народжуваності, міграції та смертності.

База даних Світового банку **Міграція та грошові перекази/World Bank Migration And Remittances**[85] містить інформацію про кількість мігрантів, їх склад, основні напрями міграції, доходи та перекази осіб, що працюють за кордоном.

Міжнародна організація праці/International Labour Organization[86] оприлюднює інформацію про показники зайнятості, безробіття, робочі години, зарплати, вартість робочої сили, індекси споживчих цін, травми на виробництві, страйки, доходи та витрати домогосподарств, трудову міграцію тощо.

Міжнародні показники людського розвитку/International Human Development Indicators[87] містить інформацію про широкий спектр показників соціального спрямування: ступінь зручності для людини умов у тій чи іншій країні. Основну інформацію акцентовано на доходах населення, освіті, розвитку охорони здоров'я, екологічних аспектах.

[83] https://ppi.worldbank.org/en/ppi
[84] https://datacatalog.worldbank.org/search/dataset/0037655
[85] https://www.worldbank.org/en/topic/labormarkets/brief/migration-and-remittances
[86] https://www.ilo.org/global/statistics-and-databases/lang--en/index.htm
[87] https://hdr.undp.org

Статистика освіти Світового банку/World Bank Education Statistics[88] охоплює показники доступності та результативності освіти, забезпеченості вчителями, витрат на освіту.

Статистика здоров'я, харчування та населення Світового банку/World Bank Health Nutrition and Population[89] охоплює показники динаміки кількості населення, харчування, репродуктивного здоров'я, фінансування охорони здоров'я, ресурсів медицини, рівня імунізації, поширеності хвороб тощо.

Гендерна статистика Світового банку/World Bank Gender Statistics[90] містить дані про вікові групи населення, показники зайнятості, тривалості життя, доступності освіти, доступ до ресурсів, зарплати, участь в управлінні тощо, залежно від статевої належності.

База даних нерівності доходу Університету об'єднаних націй/UNU-WIDER Income Inequality Database (*United Nations University*) містить дані про коефіцієнт Джині та частки доходу за верствами населення.

Інститут статистики ЮНЕСКО/UNESCO Institute of Statistics[91] надає інформацію у сфері освіти, науки і технології, культури та комунікацій, а також про основні демографічні та соціально-економічні показники. Наприклад, можна знайти інформацію про обсяги витрат на дослідження та розробки у розрізі типів інституцій (університети, державні установи, бізнес-сектор, недержавні організації), видів витрат, сфер науки тощо[92].

Світова організація інтелектуальної власності/World Intellectual Property Organization[93] надає інформацію щодо показників, які пов'язані з патентами, корисними моделями (малими винаходами), торговельними марками, промисловими зразками, сортами рослин, мікроорганізмами.

Світовий звіт про щастя/World Happiness Report[94] демонструє рейтинг країн за суб'єктивними оцінками щастя, а та-

[88] https://datacatalog.worldbank.org/search/dataset/0038480
[89] https://datacatalog.worldbank.org/search/dataset/0037652
[90] https://datacatalog.worldbank.org/search/dataset/0037654
[91] http://www.uis.unesco.org/ Pages/default.aspx
[92] http://data.uis.unesco.org
[93] https://www.wipo.int/ipstats/en/
[94] https://worldhappiness.report/

кож його факторами (ВВП на душу населення, соціальна підтримка, очікувана тривалість здорового життя, свобода вибору у житті, щедрість, свобода від корупції).

Індекс екологічної ефективності/Environmental Performance Index[95] ураховує такі складові: втрати здоров'я від несприятливого довкілля, вплив забруднення на людину та довкілля, біорізноманіття, ліси, рибне господарство, сільське господарство, кліматичні зміни.

База даних Світового банку **Скориговані чисті заощадження/World Bank Adjusted Net Savings**[96] надає оцінку справжнього рівня заощаджень в економіці з урахуванням поправок на інвестиції до людського капіталу (+), вичерпання природних ресурсів і втрат від забруднення довкілля (–).

2.10. Статистика політичної сфери та державного управління

База даних Світового банку **Політичні інституції/World Bank Political Institutions**[97] є дослідницькою базою даних, яка містить інформацію про особливості політичної системи країни, термін перебування президента/прем'єр-міністра та правлячої партії при владі, рівень підтримки на виборах голови держави/уряду, характер правлячої партії, розподіл місць у парламенті за партіями, особливості виборчого законодавства, стабільність, систему взаємозалежності та взаємних обмежень інституцій влади, рівень федералізації.

База даних Світового банку **Загальносвітові індикатори управління/World Bank Worldwide Governance Indicators**[98] охоплюють такі сфери: голосування та підзвітність, політична стабільність і відсутність насилля, ефективність уряду, якість регулювання, верховенство права, контроль за корупцією.

[95] https://sedac.ciesin.columbia.edu/data/collection/epi/sets/browse
[96] https://databank.worldbank.org/source/adjusted-net-savings/preview/on
[97] https://datacatalog.worldbank.org/search/dataset/0039819
[98] https://datacatalog.worldbank.org/search/dataset/0038026

Організація Freedom House[99] публікує рейтинг країн за ступенем дотримання політичних прав і громадянських свобод, включаючи свободу інтернету.

Інститут дослідження миру в Осло/PRIO[100] містить дані про дати та характеристики збройних конфліктів та інші показники, що пов'язані із конфліктами.

Організація Transparency International[101] здійснює оцінювання ступеню поширення корупції, зокрема й у розрізі окремих сфер: політичні партії, законодавчий орган, правоохоронні органи, бізнес, засоби масової інформації, державні службовці, система судочинства, недержавні організації, релігійні організації, збройні сили, система освіти.

2.11. Статистика багатства

База даних Світового банку **Багатство націй/World Bank Wealth of Nations**[102] містить оцінки багатства країни та її складових: чисті закордонні активи, вироблений капітал, людський капітал, природній капітал. Останній поділяють на активи надр, ресурси деревини, інші ресурси лісу, землі під захистом, орні землі, пасовища.

Фінансова установа **Credit Suisse Group** публікує **Звіт про багатство /Credit Suisse Wealth Report**[103], що містить інформацію про середнє багатство домогосподарств у різних країнах. Точнішу інформацію подано за розвиненими країнами; за іншими країнами оцінки подано на основі моделювання.

Звіт про світове багатство/Capgemini World Wealth Report[104] містить дані про кількість і статки осіб, що мають інвестиційні активи більше 1 млн дол. (крім основного житла, предметів колекціонування та споживчих товарів).

Журнал **Форбс/Forbs**[105] публікує рейтинг мільярдерів, обсяг їх статків, країну та сферу.

[99] https://freedomhouse.org/explore-the-map?type=fiw&year=2022
[100] https://www.prio.org/data
[101] https://www.transparency.org/en/cpi/2020/index/nzl
[102] https://datacatalog.worldbank.org/search/dataset/0042066
[103] https://www.credit-suisse.com/about-us/en/reports-research/global-wealth-report.html
[104] https://worldwealthreport.com
[105] https://www.forbes.com/billionaires/#1949f9d7251c

2.12. Статистика брендів та м'якої сили

Індекс країнового бренду/Country Brand Index[106] компанії **FutureBrand** ураховує такі аспекти: ступінь відомості країни (відомість не гарантує сильного бренду); якості, з якими її пов'язують; ступінь поваги до неї; бажання її відвідати/інвестувати до неї/купляти її товари; можливість реалізації цих бажань, рекомендація країни членам сім'ї (друзям, колегам). Асоціації з країною об'єднано у дві групи: статус і досвід.

Дослідницький центр Пью/Pew Research Center[107] проводить опитування громадськості у десятках країн світу з таких питань: економічна ситуація у країні респондента, чи покращиться економічна ситуація в країні, чи буде фінансово краще жити наступне покоління у країні, яким є ставлення до одного з економічних або політичних центрів світу (США, ЄС, Китай, Росія), яка країна є економічним лідером світу. В окремих публікаціях уміщено й інформацію щодо ставлення до інших актуальних проблем сучасності.

Програма опитувань громадської думки ЄС **Євробарометр/Eurobarometer**[108] публікує тематичні випуски, частина з яких має стосунок до вимірювання м'якої економічної сили (ставлення до ЄС, економічного та монетарного союзу, економічних проблем в ЄС, інтеграції з іншими країнами, основних досягнень ЄС; наявність економічного прогресу; країна, в якій планують відпочивати респонденти).

Індекс доброї країни/Good Country Index[109] показує внесок країни (позитивний або негативний) до загального блага для людства. Ураховано такі компоненти: наука й технології, культура, міжнародний мир і безпека, світовий порядок, планета й клімат, процвітання та рівність, здоров'я та добробут. Багато індикаторів взято щодо розміру економіки, тому в лідерах часто опиняються невеликі країни.

[106] https://www.futurebrand.com/uploads/FCI/FutureBrand-Country-Index-2019.pdf
[107] https://www.pewresearch.org/global/database/
[108] https://europa.eu/eurobarometer/screen/home
[109] https://index.goodcountry.org

Компанія **Brand Finance** оцінює вартість національних брендів країн на основі ВВП та якісних показників[110]. Окремо у галузевих звітах опібліковано рейтинги і вартість корпоративних брендів[111].

Компанія **Bloom Consulting**[112] публікує три індекси:

- **Індекс цифрової країни/the Digital Country Index**[113] побудовано на основі частоти згадування країн у пошукових запитах громадян світу дев'ятьма мовами. Виокремлено п'ять вимірів пошуку: туризм, талант, інвестиції, експорт, національна видатність, але лише останній вимір розрізняє позитивну та негативну інформацію.
- **Рейтинг брендів країн: торговельне видання/Country Brand Ranking: Trade Edition**[114] ураховує чисті статистичні дані про прямі іноземні інвестиції та різні інфометричні показники у сфері інвестування.
- **Рейтинг брендів країн: туристичне видання/Country Brand Ranking: Tourism Edition** ураховує: чисті надходження від туризму та різні інфометричні показники у цій сфері[115].

110 https://brandirectory.com/rankings/nation-brands
111 https://brandirectory.com/rankings
112 https://www.bloom-consulting.com/en/country-brand-ranking
113 https://www.digitalcountryindex.com
114 https://www.bloom-consulting.com/en/pdf/rankings/Bloom_Consulting_Country_Brand_Ranking_Trade.pdf
115 https://www.bloom-consulting.com/en/pdf/rankings/Bloom_Consulting_Country_Brand_Ranking_Tourism.pdf

Розділ 3
ОРГАНІЗАЦІЯ ДАНИХ

3.1. Класифікація даних у статистичному аналізі

Початком інтелектуального аналізу даних у міжнародних економічних відносинах є визначення типу даних та їх організація. Успішність застосування будь-якого методу аналізу даних залежить від відповідності аналізованих даних його вихідним припущенням. Методи, що придатні для одного типу даних, можуть призводити до серйозних помилок при їх використанні для інших типів даних. Першим етапом аналізу будь-яких даних зазвичай є визначення їх типу.

Залежно від типу *спостережень*/*cases* дані поділяють на:

- *просторові дані*/*cross-sectional data*, коли кожне спостереження – це різні об'єкти в один період часу або на один й той самий момент часу, наприклад різні фірми у 2022 р.;
- *часові ряди*/*time series*, коли кожне спостереження – це той самий об'єкт у різні періоди або моменти часу, наприклад, Україна в 2000, 2001, 2002, ... 2022 р.;
- панельні дані, коли кожне спостереження – це комбінація об'єкта та періоду часу, наприклад, країна у певний рік.

Розглянемо типи ознак або змінних, які характеризують властивості спостережень.

Нехай вивчають деяку *генеральну сукупність*/*population*, яку описує випадкова величина *X*. Кожній одиниці (елементу) сукупності може відповідати значення деякої ознаки, яку ще називають *змінною*/*variable* або *варіантом* чи *показником*, оскільки вона може набувати різних значень у різних елементів. Значення ознаки утворюють набори даних у математичній статистиці. Наприклад, досліджують фондовий ринок. Статистична сукупність – це множина акцій різних компаній, якими торгують на фондовій біржі. Основною ознакою кожної акції є її ціна. Інформація про ціни утворюватиме статистичні дані для дослідження. Інколи елементи статистичної сукупності можуть характеризувати кілька ознак. Наприклад, якщо *генеральна сукупність*/*population* – це всі сім'ї, що

проживають у певному регіоні, то, залежно від мети, ознаками можуть бути: річний сукупний дохід, кількість дітей дошкільного або шкільного віку, загальна площа квартири або будинку, наявність автомобіля, задоволення роботою комунальних служб тощо. У загальному випадку для кожної генеральної сукупності може існувати багато змінних ознак, які мають кількісний або якісний вираз.

Якісна/qualitative ознака характеризує належність елемента статистичної сукупності до якої-небудь якісної категорії. Якісні дані тільки реєструють певну якість, що має елемент, проте не вимірюють її. Вони лише вказують, до якої з якісних категорій належить елемент статистичної сукупності. Наприклад, при виявленні рейтингу кандидата на посаду президента країни кожен потенційний виборець може бути зарахований до однієї із трьох якісних категорій: "за", "проти", "не визначився". Іншим прикладом може бути кваліфікаційна або посадова структура розподілу працівників фірми, що є розподілом персоналу за певними якісними категоріями.

Кількісну/quantitative ознаку, на відміну від якісної, можна об'єктивно змінити. Кількісні дані можна безпосередньо вимірювати, вони мають змістовну інтерпретацію: вартість, розміри, кількість країн, банків, службовців тощо. Над кількісними даними можна виконувати операції як зі звичайними числами: додавати, обчислювати середнє значення тощо. Необхідно зазначити, що до кількісних даних не належать числа, які використовують для кодування або нумерації. Наприклад, різні типи операцій можна закодувати так: 1 – купівля акцій, 2 – продаж акцій, 3 – купівля облігацій, 4 – продаж облігацій. Проте арифметичні дії з такими даними не мають жодного сенсу.

Виділять два основних типи кількісних даних: дискретні та неперервні.

Дискретна змінна/discrete variable може набувати окремих значень, які відрізняються одне від одного тільки на деяке число. Усі можливі значення дискретної змінної можна перерахувати. Такі дані можуть виникати за підрахунку кількості одиниць у сукупності: дітей у сім'ї; автомобілів, що перетнуть перехрестя протягом 5 хв.; дітей, які народилися за добу.

Значення *неперервної змінної/continuous variable* можуть відрізнятися одне від одного на нескінченно малу величину. Вони неперервно заповнюють деякий числовий проміжок. Неперервні дані можна отримати, наприклад під час вимірювання фізіологічних характеристик людини: зріст, вага, артеріальний тиск; час безперебійної роботи приладу; час обслуговування клієнта у банку тощо за припущення, що вимірювання можна робити із довільною точністю. До неперервних також належать змінні, значення яких є дискретними, проте ці значення відрізняються одне від одного на дуже незначну величину. Наприклад, валютний курс євро до долара (пара євро/долар) є неперервною величиною, проте вимірюється із точністю до чотирьох знаків після коми.

Щодо напряму причинно-наслідкового зв'язку, який перевіряють, змінні поділяють на *залежні/dependent* та *незалежні/independent* або фактори/*factors*.

Основною є **класифікація даних за шкалами їх вимірювання**. Для опису або вимірювання даних існують чотири типи шкал: *шкала найменувань/nominal scale*, *порядкова шкала/ordinal scale*, *шкала інтервалів/interval scale* та *шкала відношень/ratio scale*. Вибір типу шкали для характеризації або вимірювання ознаки залежить від природи цієї ознаки. Якщо вона має якісний характер, то вимірювання здійснюють у шкалі найменувань і порядку (якісні дані); якщо вона має кількісний характер, то застосовують шкали інтервалів відношень (кількісні дані).

Номінальні ознаки (ознаки з невпорядкованими станами, класифікаційні ознаки) – це дані, що вимірюють у номінальній шкалі (класифікаційній, шкалі найменувань). **Номінальна шкала** визначає належність об'єкта до певного класу. Між різними варіантами значень номінальної змінної не існує відносин підпорядкованості (напр., показник регіону має варіанти значень: Європа, Азія, Африка тощо – неможна сказати, що одне з них є більш "регіональним" за інше). Цю шкалу використовують для опису якісних даних, що характеризують належність елементів сукупності до деякого класу. Найменування класів можна виразити за допомогою чисел, але числа

варто використовувати лише для відповіді на запитання: чи належать два об'єкти до одного класу. Усім об'єктам одного й того самого класу присвоюють одне й те саме число, а об'єктам різних класів – різні числа. До прикладу, номер спеціальності при розподілі робітників за фахом або за статтю (чоловік = 1, жінка = 2); розподіл товарів у магазині за артикулами. Зміст шкали найменувань означає присвоєння кожному класу певного коду, що необхідно для зберігання та організації пошуку інформації у комп'ютерних системах. Проте проводити будь-які обчислення з такими даними не має жодного сенсу. Прикладами номінальних ознак є назви біологічних видів, навчальних дисциплін, кольори тощо. З погляду автоматизації аналізу даних і застосування стандартних алгоритмів варто обирати такі позначення класів: 0, 1, 2, … Але з цими числами не можна виконувати будь-які дії, крім перевірки їх рівності або нерівності.

Порядкові ознаки (ознаки з упорядкованими станами, ординальні ознаки) – це виміряні у порядкових шкалах дані, які можна порівнювати у певному відношенні: "більше – менше", "легше – важче" тощо. **Порядкова шкала** визначає порядок (ранжує) об'єкти, але не визначає відстані між ними (напр., глибина економічної інтеграції: відсутність регіональної торгівельної угоди, угода с обмеженим предметом дії, зона вільної торгівлі, митний союз, спільний ринок, економічний союз). Якщо використовувати бальні оцінки, то такі якісні показники можна умовно трансформувати у кількісні. Порядкову шкалу використовують для впорядкування об'єктів (ранжування), наприклад розподілу місць серед учасників спортивного змагання або конкурсу. Числа у шкалі (ранги) визначають порядок слідування об'єктів, але не дають можливості визначити, на скільки або у скільки разів один об'єкт переважає інший. Якщо перше місце у конкурсі посів учасник *A*, третє – учасник *B*, п'яте – учасник *C*, сьоме – учасник *D*, то це не означає, що *D* відносно *C* розташований так само близько, як *B* до *A*. У шкалі порядку відсутні поняття масштабу та початку відліку. Прикладами порядкових ознак є військові звання, рейтингові оцінки тощо. Якщо значення порядкової ознаки є числами, то їх

можна застосовувати й для порівняння ступеня вияву класифікаційної ознаки, але відстані між класами при цьому не буде визначено.

Кількісні (**числові**, **варіаційні**) **ознаки** – це ознаки, які вимірюють у кількісних метричних (інтервальних, відносних, циклічних та абсолютних) шкалах вимірювань. **Метрична шкала** визначає величину інтервалу між значеннями показника (напр., експорт). Дії, що можна виконувати з числовими характеристиками даних, залежать від шкали вимірювань.

Шкалу інтервалів використовують для визначення міри відмінності між значеннями ознаки, що притаманна різним елементам сукупності. Класичний приклад інтервальної шкали – вимірювання температури в градусах за Цельсієм. При цьому різниця між 15 °C і 10 °C така сама, що й між 17 °C і 12 °C. У загальному випадку шкала інтервалів може мати довільні точки відліку та масштаб.

Шкала відношень є частковим випадком шкали інтервалів. На відміну від шкали інтервалів, вона має фіксовану точку відліку. Шкала дає можливість відповісти на запитання: у скільки разів значення ознаки, що характеризує одиницю сукупності, перевищує значення цієї ознаки для іншої одиниці. Таке порівняння неможливо здійснити за допомогою шкали інтервалів: неможна стверджувати, що за температури 20 °C буде вдвічі "тепліше", ніж за температури 10 °C, оскільки температура 0 °C не означає відсутність температури загалом. У шкалі відношень вимірюють, наприклад, площу, вагу, довжину, грошові потоки. Нульова точка відліку у цій шкалі означає повну відсутність вимірюваної ознаки.

Дані, що отримані у шкалах вищих рангів, можна привести до шкал нижчих рангів. Наприклад, дані, що виміряні у шкалі відношень, можна привести до інтервальної шкали. Такі перетворення називають зниженням шкали. Необхідність у них зазвичай виникає при обробці даних, що виміряні у шкалах різного типу. Зворотну операцію – перетворення даних, що виміряні у нижчих шкалах до вищих, вважають некоректною. Зниження шкали призводить до втрати частини наявної інформації про досліджувані ознаки.

З метою реєстрації ходу дослідження варто зробити **систематизований опис показників** (змінних) і надати їм скороченої назви, яку буде зручно використовувати для назв аркушів чи у формулах для опису ходу дослідження. Дослідник може сам обрати зручні для нього позначення (напр., $GDPn$ – ВВП у національній валюті, $FRes/GDP$ – валютні резерви щодо ВВП, $GDPpc$ – ВВП на душу населення, $Expgr$ – зростання експорту, $Exprgr$ – зростання реального експорту, $PrivFDebt/FDebt$ – частка приватного боргу у зовнішньому боргу тощо).

Опис може включати:

- повну та скорочену назви;
- джерело (можливо, з кодом показника у джерелі), якщо показник запозичено;
- одиниці виміру;
- формулу розрахунку за потреби;
- примітки – додаткові дані про характер показника (напр., на кінець періоду чи в середньому за рік), про використання лагових значень, спосіб заміни відсутніх даних (якщо відсутні дані замінено на розрахункові дані), максимальні та мінімальні значення (напр., 10 – найвищий рівень економічної свободи, 0 – найнижчий), варіанти значень (напр., 1 – відсутні збройні конфлікти, 2 – обмежений збройний конфлікт, 3 – війна), особливості для окремої групи країн, за який період відсутні дані тощо.

У наукових публікаціях або аналітичних звітах інформацію про показники зазвичай указують у розділі про методологію дослідження або у додатках.

3.2. Формування та види вибірок

Масовим соціальним та економічним явищам відповідають статистичні сукупності, у межах яких ці явища виявляються. Елементами статистичної сукупності можуть бути окремі індивіди або їх групи, а також будь-які об'єкти: країни, галузі та їх підприємства, акції, транспортні засоби, одиниці продукції тощо.

Вихідними поняттями математичної статистики *є генеральна сукупність*/*population* і *вибіркова сукупність* або *вибірка*/*sample*. Якщо до сукупності входять усі можливі елементи,

які відповідають певному явищу, то це генеральна сукупність. Вибіркова сукупність, або вибірка, – це частина (підмножина) генеральної сукупності.

Генеральна сукупність – це множина всіх реально існуючих або можливо уявних однорідних об'єктів, які вивчають під кутом зору їх розподілу за деякою ознакою. Склад генеральної сукупності повністю визначає відповідне явище. Прикладами генеральних сукупностей можуть бути множини окремих людей за віком, множини окремих земель за врожайністю, множини акціонерних банків України за прибутками, множини виробів певного найменування за якістю, множини або групи країн за ступенем розвитку економіки тощо.

Фактично генеральна сукупність – це всі релевантні спостереження (об'єкти), цікаві з практичного погляду, наприклад, усі сучасні країни (зазвичай рекомендації для сучасного періоду та найближчого майбутнього), усі країни з певною ознакою (напр., усі малі країни, якщо рекомендації розроблено за результатами аналізу саме для такого типу країн), усі компанії, усі люди або люди з певною ознакою (дорослі, мешканці міст тощо), усі країно-роки в сучасний історичний період і найближчому майбутньому (якщо аналізують панельні дані).

Оскільки практично будь-яка ознака генеральної сукупності допускає кількісне подання, то замість розподілу одиниць сукупності за цією ознакою говорять про розподіл деякої випадкової величини X, її закон розподілу або числові характеристики. Тоді стохастичний експеримент, з яким пов'язана випадкова величина X, полягає у виборі навмання одного представника цієї сукупності, а значення x, якого набуває випадкова величина X, є значенням ознаки для обраного представника генеральної сукупності. Проте на практиці, замість генеральної сукупності, аналізують **вибірку** (виняток може становити аналіз великих даних) через:

- відсутність у базах даних значень для деяких спостережень;
- значну витратність аналізу великої генеральної сукупності (напр., усього населення землі);
- невизначеність майбутнього.

Тому результати аналізу вибірки узагальнюють (екстраполюють) на всю генеральну сукупність з певною мірою упевненості (достатнім є зазвичай 95 % упевненості). Повтор аналізу на іншій вибірці (в іншому місці, періоді часу, для інших об'єктів/суб'єктів) дозволяє оцінити стійкість одержаних результатів.

Припустимо, що метою статистичного дослідження є виявлення думки громадян деякої країни щодо обрання певного кандидата на посаду президента. У цьому випадку генеральна сукупність включає всіх громадян країни, які мають право голосу. Чисельність, або обсяг такої сукупності, може досягати десятків або сотень мільйонів громадян. Очевидно, що за такої ситуації оперативно організувати опитування всіх потенційних виборців, тобто провести дослідження всіх одиниць сукупності, практично неможливо. Для отримання висновків про властивості генеральної сукупності досліджують деяку її частину або вибірку. У прикладі про вибори президента проводять вибіркове опитування, за результатами якого роблять висновок про рейтинг кандидата.

Якщо генеральна сукупність є достатньо великою, а досліджувана вибірка становить доволі малу її частину, то відмінність між описаними видами відбору є несуттєвою. У граничному випадку, коли сукупність стає нескінченною, а обсяг вибірки залишається сталим, ця різниця загалом зникає.

Випадковість вибірки можна забезпечити наступним чином. Якщо із генеральної сукупності чисельністю N одиниць відбирають випадковим чином n одиниць ($n < N$), то такий відбір називають *простим випадковим відбором*/*simple random sample*. Його реалізують, наприклад, під час розіграшу лотереї. Результатом простого випадкового відбору є проста випадкова вибірка. Схема простого випадкового відбору передбачає реєстрацію елементів сукупності у вигляді деякого реєстру або списку, що дозволяє використовувати таблиці випадкових чисел для формування простої випадкової вибірки. При цьому можна скористатися готовою таблицею випадкових чисел або самостійно згенерувати її за допомогою комп'ютера. До прикладу, аудитору необхідно сформувати випадкову вибірку

50 записів фінансового звіту із загальної кількості у 1000 записів, що занумеровані числами від 1 до 1000. Він може скористатися таблицею випадкових чисел. Для цього необхідно зафіксувати довільну позицію у таблиці випадкових чисел, наприклад, на перетині 5 рядка та 2 стовпчика. Далі складають список із 50 чисел, рухаючись довільним чином таблицею. У кожному обраному п'ятицифровому числі для визначення дробової частини розташовують кому між третім і четвертим знаками, потім округлюють отримане дробове число до найближчого цілого числа. Наприклад, із першого обраного числа 41639 дістають 416,39, звідки після округлення до найближчого цілого отримують число 416. Продовження цього процесу дасть послідовність трицифрових чисел, які будуть порядковими номерами записів фінансового звіту, що утворять випадкову вибірку.

Слід зазначити, що простим випадковим відбором можна отримати як *випадкову вибірку без повернення*, так і *випадкову вибірку із поверненням*.

Крім простого випадкового відбору існують інші методи організації вибірки: систематичний, експертний, районований і багатоступінчастий тощо.

Систематичний відбір/*systematic sample* передбачає формування вибірки, відповідно до деякого плану. Його можна використовувати, коли отримання простої випадкової вибірки вимагає значних витрат. Наприклад, до генеральної сукупності входять 2000 цінних паперів, які зберігаються у спеціальних висувних скринях. Потрібно здійснити випадковий відбір 100 цінних паперів для продажу клієнтам. Теоретично необхідно занумерувати всі цінні папери числами від 0 до 1999, а потім, за допомогою таблиці випадкових чисел, відібрати випадковим чином 100 цінних паперів. Зрозуміло, що така процедура вимагає забагато часу. Швидше й простіше, висуваючи скрині та підраховуючи цінні папери, відбирати кожний 20-й. Систематичний відбір може надати такі самі результати, що й випадковий, якщо елементи генеральної сукупності добре перемішані. Проте, якщо елементи розміщені в

певному порядку, то фактор випадковості вже не буде визначальним при формуванні вибірки.

За **експертного відбору**/*purposive sample* свідомо включають до вибірки ті одиниці, властивості яких найбільшою мірою відповідають меті дослідження. До вибірки включають такі елементи, на підставі яких отримані вибіркові характеристики будуть найкращими оцінками відповідних характеристик генеральної сукупності. Експертний відбір буде ефективним при відборі найбільших вибірок із невеликих генеральних сукупностей. При цьому потрібно добре знати властивості окремих елементів генеральної сукупності. Експертний відбір найчастіше застосовують у міжнародній торгівлі.

Районований або стратифікований відбір/*stratified sample* є різновидністю випадкового відбору. Випадковим чином ділять генеральну сукупність на кілька "районів" і відбирають елементи, які утворюють вибірку, не із всієї генеральної сукупності як цілого, а з кожного "району" окремо. За наявності певних передумов районований відбір може дати точніші результати, ніж простий випадковий відбір. Точність залежить від того, як було проведено "районування". Такий відбір часто використовують за соціологічних опитувань, коли районування проводять за територіальною, соціальною й демографічною ознаками. Наприклад, включають до вибірки чоловіків і жінок пропорційно до їхніх часток у населенні.

Багатоступінчастий відбір/*multistage sampling* передбачає проведення деяких послідовних випадкових відборів, причому вибір одиниць у вибірку відбувається на останній стадії відбору. Припустимо, що необхідно дослідити групу країн. Такий відбір може провести у три кроки: відбір першого ступеня – географічне розташування країни; відбір другого ступеня – рівень розвитку країни; відбір третього ступеня – розмір країни. Цей метод не забезпечує більшу точність оцінки, порівняно з простим випадковим вибором, але його застосування може суттєво скоротити витрати на проведення досліджень.

Багаторівнева вибірка/*multi-level sampling* передбачає формування вибірки з макрооб'єктів, а всередині кожного – формування вибірки з мікрооб'єктів. Наприклад, якщо обирають п'ять міст, а далі у кожному проводять опитування 100 осіб.

Зручна або доступна вибірка/*convenience, accidental, availability sample* передбачає, що до неї включені об'єкти або суб'єкти, до яких найлегше одержати доступ. Наприклад, якщо йдеться про опитування колег ученого, студентів викладачем, однокурсників, друзів. Такий спосіб найменш витратний, хоча й має ризик нерепрезентативної вибірки. Але він все ж є корисним для попереднього дослідження перед проведенням більш масштабного.

Вибірка за методом сніжного кому/*snowball sample* означає, що спочатку обирають кілька суб'єктів для опитування, які рекомендують ученого своїм знайомим; останні після опитування рекомендують уже своїм знайомим і т. д. Таким чином обсяг вибірки збільшується у багато разів. Тут також є ризики нерепрезентативної вибірки, але цей спосіб є ефективним, якщо необхідно опитати суб'єктів, які важкодоступні та не йдуть на контакт без рекомендації людей, яким вони довіряють.

Вибірка за методом модельного екземпляру/*modal instance sampling* включає тільки найбільш типові об'єкти, наприклад, осіб середнього віку із середнім доходом або компанії у формі товариства з обмеженою відповідальністю медіанного розміру.

Гетерогенна вибірка/*heterogeneity sample* включає представників різних груп об'єктів або суб'єктів, але незалежно від пропорцій, в яких вони зустрічаються у генеральній сукупності. У такому випадку цікавою є широта представлених думок, а не середньостатистичні оцінки. Наприклад, якщо вибірка включає 30 малих, 30 середніх і 30 великих компаній, незважаючи на те, що малих компаній набагато більше, ніж інших типів фірм. Або, якщо до вибірки респондентів включають 20 експертів у сфері транспорту та 20 звичайних споживачів транспортних послуг.

Інші способи відбору елементів до вибірки є комбінаціями описаних вище методів.

3.3. Формування таблиці вхідних даних у Microsoft Excel

3.3.1. Підготовка вхідних даних із зовнішніх джерел

Вихідним пунктом статистичного дослідження обробки даних є їх систематизація з метою надання зручної форми та структури для проведення первинного аналізу. Опишемо один із варіантів формування бази даних, з яким часто можна зустрітися при аналізі даних за багатьма об'єктами (напр., країнами) та періодами часу. У нашому прикладі йдеться про панельні дані. На кожному кроці бажано проводити процедури перевірки, аби не припуститися помилки. Помилку набагато легше виправити одразу після її виникнення, ніж пізніше шукати, де вона виникла.

Усі показники або окрема група показників можуть бути представлені в одній книзі *Microsoft Excel*. Кожному показнику виділяють один аркуш. У другому стовпчику зверху до низу йдуть країни в алфавітному порядку (зазвичай англійською мовою). У другому рядку зліва направо перераховують роки (або квартали чи місяці). Бажано робити це, наприклад, у другому стовпчику та другому рядку (а не у першому), аби залишити у таблиці вільне місце для коментарів (назва показника, одиниця виміру, спосіб розрахунку, формули для перевірки відсутності помилок тощо).

На рис 3.1 подано фрагмент аркушу, де розміщено дані за одним показником: короткостроковий зовнішній борг (за даними *World Development Indicators*). Для зручності відображення великих значень числа показано у *науковому стилі*/ *scientific style*.

Важливо одразу визначитися із колом досліджуваних об'єктів і періодів. Наприклад, до дослідження часто не включають невеликі за розміром країни, за якими відсутня переважна частина даних, або країни, які різко відрізняються за характеристиками від України, якщо нас цікавить застосування результатів подальшого аналізу для України. Для більшості показників доступні лише річні дані, але мета аналізу може вимагати використання коротших періодів (місяці, квартали, дні). З одного боку, бажано включати до досліджуваного періоду більше

років; з іншого, – дані за ранні роки не завжди доступні (для України раніше 1995 року їх важко знайти за багатьма показниками, або їх надійність викликає сумнів), так само як і за останній період (хоча іноді можна використовувати більшменш надійні попередні оцінки чи прогнозні дані, напр., у жовтні – прогнозні дані на весь поточний рік). До того ж, система зв'язків між явищами із часом суттєво змінюється. Тому не варто включати до аналізу занадто давні періоди минулого або бажано надавати їм меншої ваги у розрахунках.

Потрібно пересвідчитися, що у різних аркушах (або книгах) за різними показниками значення за однією й тією самою країною в один і той самий період розташовані у комірках із тією самою адресою в межах аркуша. Для цього в усіх аркушах дані щодо України розташовують у рядку з одним і тим самим номером, а дані за 2010 рік – у стовпчику з одним і тим самим номером. Нажаль, часто цього досягають вручну, особливо якщо використано дані із різних джерел. Перевірку на однаковий склад і порядок країн здійснюють копіюванням стовпчика з назвами країн з однієї таблиці (еталонної) до решти таблиць (напр., до першого стовпчика кожної таблиці).

У першому стовпчику прописують формули типу: =IF(A2=B2;1;0). Якщо у результаті у цьому стовпчику в певній комірці з'явиться 0 замість 1, це означатиме: послідовність країн не збігається з еталонною, і це потрібно виправити. На рис. 3.2 подано фрагмент аркушу, де розміщено дані за іншим показником: гранти у формі анулювання боргу, але при цьому у другому стовпчику скопійовано стовпчик з еталонним переліком країн, а в комірці A3 прописано функцію: =IF(B3=C3;1;0) і скопійовано нижче у стовпчику – у першому стовпчику всюди результатом є 1.

На цьому етапі також варто пересвідчитися, що числа записано в стандартному вигляді. Наприклад, якщо в системі *Windows* розділовим значком дробної частини є крапка, а число має вигляд як 3,45 (тобто з комою), то воно сприйматиметься як текст, так само, як і число типу 3 500, оскільки воно містить пробіл. Це можливо виправити командою: *Знайти й замінити*/*Replace*, але потрібно впевнитися, що заміни зроблено правильно.

	A	B	C	D	E	F	G	H	I	J	K	L
1	External debt stocks, short-term (DOD, current US$)											
2				2000	2001	2002	2003	2004	2005	2006	2007	2008
3			Afghanistan							18152000	20997000	16926000
4			Angola	1.32E+09	1.45E+09	1.21E+09	1.07E+09	1.2E+09	2.32E+09	2.13E+09	2.27E+09	2.42E+09
5			Albania	36666000	30619000	29046000	1.49E+08	1227000	2.83E+08	5.91E+08	8.27E+08	7.79E+08
6			Argentina	2.83E+10	2E+10	1.48E+10	2.23E+10	2.65E+10	3.48E+10	3.36E+10	3.81E+10	3.75E+10
7			Armenia	44419000	41972000	2.2E+08	4.04E+08	4.1E+08	2.98E+08	3.07E+08	4.59E+08	4.65E+08
8			Azerbaijan	1.56E+08	1.03E+08	82380000	1.03E+08	1.38E+08	1.86E+08	5.2E+08	1.04E+09	1.17E+09
9			Burundi	65000000	88395000	96346000	47562000	22560000	34059000	37617000	13785000	19303000
10			Benin	65170000	78650000	73610000	33606000	28606000	44045000	44068000	5184000	37606000
11			Burkina Faso	84312000	63472000	12918000	14139000	24096000	22222000	89063000	1.55E+08	1.1E+08
12			Bangladesh	3.34E+08	3.61E+08	5.72E+08	6.17E+08	7.12E+08	6.88E+08	1.18E+09	1.38E+09	1.99E+09
13			Bulgaria	1.45E+09	1.22E+09	1.84E+09	2.66E+09	3.26E+09	4.44E+09	8.04E+09	1.4E+10	1.85E+10
14			Bosnia and Herzegovina	49331000	59730000	3.49E+08	1.13E+08	3.6E+08	8.37E+08	1.17E+09	1.69E+09	9.12E+08
15			Belarus	1.24E+09	1.31E+09	1.63E+09	1.97E+09	2.94E+09	3.5E+09	3.65E+09	6.88E+09	6.96E+09
16			Belize	50000000	50844000	45055000	80000000	319000	5996000	6615000	6358000	6558000
17			Bolivia	4.02E+08	3.8E+08	3.7E+08	3.32E+08	2.72E+08	1.82E+08	2.2E+08	1.77E+08	1.66E+08
18			Brazil	3.1E+10	2.83E+10	2.34E+10	2.46E+10	2.53E+10	2.4E+10	2.03E+10	3.92E+10	3.67E+10

Рис. 3.1

	A	B	C	D	E	F	G	H	I	J	K	L
1	Debt forgiveness grants (current US$)											
2				2000	2001	2002	2003	2004	2005	2006	2007	2008
3	1	Afghanistan	Afghanistan							0	31300000	14900000
4	1	Angola	Angola	1320000	60000	0	0	0	0	0	0	30020000
5	1	Albania	Albania	2040000	0	0	0	0	0	0	0	0
6	1	Argentina	Argentina	0	0	0	9210000	0	0	0	3410000	0
7	1	Armenia	Armenia	0	0	0	0	0	0	0	0	0
8	1	Azerbaijan	Azerbaijan	0	0	0	0	0	0	0	0	0
9	1	Burundi	Burundi	5630000	4490000	3980000	2940000	7080000	14190000	3980000	24440000	44150000
10	1	Benin	Benin	32720000	21360000	24240000	58100000	88760000	22500000	1.01E+09	10840000	9820000
11	1	Burkina Faso	Burkina Faso	1.9E+08	31160000	39090000	48580000	55530000	40510000	1.22E+09	16710000	21250000
12	1	Bangladesh	Bangladesh	1.79E+08	1.56E+08	1.74E+08	93660000	2.71E+08	40950000	2.39E+08	1.29E+08	7.55E+08
13	1	Bulgaria	Bulgaria	1560000	0	0	0	0	0	0	0	0
14	1	Bosnia and He	Bosnia and He	1.25E+08	4250000	3590000	24020000	3590000	3790000	4610000	5990000	6830000
15	1	Belarus	Belarus	0	0	0	0	0	0	0	0	0
16	1	Belize	Belize	0	0	2190000	2380000	2670000	1750000	1770000	1290000	0
17	1	Bolivia	Bolivia	24090000	1.52E+08	4.06E+08	1.17E+08	5.27E+08	37590000	1.82E+09	1.18E+09	2990000
18	1	Brazil	Brazil	0	0	0	0	0	0	0	0	0

Рис. 3.2

3.3.2. Підготовка додаткових розрахованих показників

Припустимо, що сформовано базу даних, які запозичені із зовнішніх джерел показників. Наступним кроком є вставка нових аркушів, де формуються таблички за новими розрахованими показниками. На перетині стовпчиків-років і рядків-країн прописують формули. Наприклад, якщо треба розраховувати показник *Валютні резерви* (частка від ВВП) для комірки C3, то вставляємо туди формулу типу

='FReserveUSD'!C3/'GDPUSD'!C3,

де FReserveUSD та GDP – назви відповідних аркушів, де наявні дані про валютні резерви та ВВП (у дол.). Далі цю формулу копіюють до решти потрібних комірок аркушу, де міститимуться дані про *Валютні резерви* (частка від ВВП). Наприклад, на рис 3.3 подано фрагмент аркушу, де розміщено дані про розрахований подібним чином показник: відношення грантів у формі анулювання боргу до короткострокового зовнішнього боргу.

Залишається лише видалити значення з помилками (комірки з даними щодо Афганістану в 2000-2005 рр. стануть порожніми).

Пошук помилок можна здійснити, зокрема, перевіряючи, чи не вийшли за межі допустимих значень показники. Наприклад, частка сільськогосподарської продукції в експорті не може бути більше 100 % чи одиниці.

Результат розрахунку можна перевірити й вручну у певній комірці. Якщо помилка спричинена неправильною формулою, то потрібно скоригувати її; якщо – неправильним чи незвичним значенням показника, на основі якого проводять розрахунок формули, бажано за можливості виправити це значення; якщо показнику не можна довіряти, то його сприймають як відсутні дані.

Можна у формулах використати додаткові коефіцієнти. Наприклад, якщо потрібні дані у відсотках, а не у частках одиниці, то формула у попередньому прикладі виглядатиме як ='FReserveUSD'!C3*100/'GDPUSD'!C3.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Debt forgiveness grants (current US$)/External debt stocks, short-term (DOD, current US$)											
2				**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
3			Afghanistan	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	0	1.490689	0.880302
4			Angola	0.000997791	4.13E-05	0	0	0	0	0	0	0.012409
5			Albania	0.055637375	0	0	0	0	0	0	0	0
6			Argentina	0	0	0	0.000413	0	0	0	8.96E-05	0
7			Armenia	0	0	0	0	0	0	0	0	0
8			Azerbaijan	0	0	0	0	0	0	0	0	0
9			Burundi	0.086615385	0.050795	0.041309	0.061814	0.31383	0.41663	0.105803	1.772942	2.287209
10			Benin	0.502071505	0.271583	0.329303	1.728858	3.102846	0.510841	22.87465	2.091049	0.261129
11			Burkina Fas	2.252111206	0.490925	3.02601	3.435887	2.304532	1.822968	13.66157	0.107795	0.193106
12			Bangladesh	0.536766467	0.433328	0.304216	0.151799	0.380112	0.059483	0.202784	0.093869	0.380351
13			Bulgaria	0.001075624	0	0	0	0	0	0	0	0
14			Bosnia and I	2.531876508	0.071154	0.010296	0.212566	0.009972	0.004528	0.003937	0.003551	0.007489
15			Belarus	0	0	0	0	0	0	0	0	0
16			Belize	0	0	0.048607	0.02975	8.369906	0.291861	0.267574	0.202894	0
17			Bolivia	0.05986045	0.399289	1.098081	0.35272	1.935499	0.206823	8.268	6.641243	0.018012

Рис. 3.3

Вимірювання у відсотках зазвичай візуально сприймають легше. Але вимірювання у частках одиниці зручніше, якщо розрахований показник надалі використовуватиметься для розрахунку нових показників.

Якщо, наприклад, ВВП вимірюють у доларах, а валютні резерви – у мільйонах доларів, то потрібне корегування на одиниці виміру:

='FReserveUSD'!C3*1000000/'GDPUSD'!C3.

Зокрема, у *World Development Indicators* часто показники вимірюють безпосередньо у доларах чи одиницях національної валюти, у той час як в інших джерелах – у мільйонах чи тисячах доларів.

Якщо показники вимірюють у різних валютах (напр., ВВП у національній валюті), то використовуватиметься формула типу:

='FReserveUSD'! C3*'ExR'C3/'GDPLCU'!C3.

Тут ExR – назва аркушу, де містяться дані про валютний курс в *одиницях національної валюти/local currency unit* за долар. При використанні валютних курсів потрібно пересвідчитися у тому, що застосовано потрібний вид котирування для всіх країн. Не варто допускати, щоб, наприклад, в аркуші з даними про валютний курс у рядку Великобританія курс був у доларах за фунт стерлінгів, а у рядку Україна – у гривнях за долар. Варто подавати всі курси в одиницях національної валюти за долар. Для показників за період необхідно використовувати середній курс за період, для показників на кінець періоду – курс на кінець періоду. Але іноді важко робити вибір, якщо потрібно порівняти один показник за період, а інший – на кінець періоду.

Можна спробувати включити однаковий загальний показник до бази даних у різних формах, зокрема, валютні резерви:

- в абсолютному вимірі (млрд дол.);
- щодо інших змінних (ВВП, зовнішній борг, короткостроковий зовнішній борг, імпорт і сплачені доходи, грошовий агрегат М2);
- приріст у відсотках чи частках одиниці;

- приріст щодо інших змінних (напр., щодо ВВП: валютні резерви/ВВП у поточному періоді мінус валютні резерви/ВВП у попередньому періоді);
- як частку щодо сукупного показника (напр., частка країни у валютних резервах світу, частка валютних резервів у всіх активах у міжнародній інвестиційній позиції країни);
- логарифм;
- щодо середнього рівня у світі;
- трендові значення наступного періоду, наприклад:

2 × валютні резерви поточного року
– валютні резерви попереднього року
= валютні резерви поточного року
+ зміна валютних резервів, порівняно із попереднім роком;

- щодо теоретичного рівня на основі регресії (приклад для іншого показника: співвідношення валютного курсу та коефіцієнта паритету купівельної спроможності – наскільки його фактичне значення відрізняється від теоретичного, що розраховано на основі регресії, де незалежною змінною є відношення ВВП на душу населення країни до середньосвітового рівня ВВП на душу населення);
- стандартизоване значення (у стандартних відхиленнях щодо середнього значення: за всі періоди у певній країні, за всіма країнами за певний період або за всіма країнами за всі періоди), що дозволяє також порівняти значення змінних у різних одиницях виміру;
- категоризоване значення одного із вказаних вище варіантів (напр., 1 – низькі валютні резерви, 2 – середні, 3 – високі), коли змінна у метричній шкалі перетворюється на змінну у порядковій шкалі.

Розраховані показники можуть бути достатньо складними. Наприклад для оцінювання впливу валютних криз (або іншої події) на економічне зростання як залежну змінну розраховано коефіцієнт прискорення економічного зростання ($КПЕЗ_0$)[116] за формулою:

---

[116] Див. : Чугаєв О.А. Валютні кризи на межі XX–XXI століть : моногр. / О.А. Чугаєв. – Київ : "МП Леся", 2007. – 416 с. (https://www.sites.google.com/site/achugaiev/stati/stati-1?authuser=0)

$$КПЕЗ_0 = \frac{(1+y_{t+1})^{\frac{1}{3}} \cdot (1+y_{t+2})^{\frac{1}{3}} \cdot (1+y_{t+3})^{\frac{1}{3}}}{\sqrt{(1+y_{t-1})^{\frac{1}{2}} \cdot (1+y_{t-2})^{\frac{1}{3}} \cdot (1+y_{t-3})^{\frac{1}{6}} \cdot \left(1+y_{t-1}+\frac{2}{3}(y_{t-1}-y_{t-2})+\frac{1}{3}(y_{t-2}-y_{t-3})\right)}},$$

де y_t – приріст ВВП року t (коли відбулася подія).

За допомогою цієї формули порівнюють економічне зростання в наступні три роки (переведене на річну основу) – у чисельнику; з геометричним середнім економічного зростання попередніх трьох років і трендовими його значеннями – у знаменнику. Останнім спостереженням надають більшої ваги, ніж попереднім. Якщо $КПЕЗ_0 > 1$, то за інших рівних умов відповідна подія (напр., валютна криза) у рік t, імовірно, сприяє економічному зростанню, якщо $КПЕЗ_0 < 1$, то – протидіє.

На цьому етапі також потрібно залишити порожні комірки, за якими дані відсутні (оскільки іноді їх позначають текстом, напр., *н.д.*, –, *NA* тощо) або прописано помилку в результаті розрахунків. Якщо комірки з текстом залишаться, то у подальшому аналізі це може не дати результатів або дати не ті результати. До того ж, варто пересвідчитися в тому, що формули не сприймають порожні комірки як такі, що містять значення 0 – ця проблема, зокрема, виникає при відніманні. Наприклад, якщо застосовують формулу =A5-A4, де комірка A5 містить число 30, а A4 – порожня клітинка, то результатом розрахунку буде 30, але насправді це – невизначеність, і результат розрахунку має сам бути замінений на порожню комірку.

Розрахувавши всі потрібні показники у книзі, можна створити нову книгу, до якої скопіювати всі дані із попередньої книги, але вже у формі: *Значення*/*Values*, а не формул. Якщо цього не зробити, то за подальших трансформацій бази даних може статися ситуація, за якої після переміщення формули посилання вказуватимуть не на потрібні комірки, або вони стануть недійсними. Результатом будуть помилки у розрахунках.

Певну проблему можуть створити показники, які мало змінюються з часом. Тут може виникнути ефект завищення кількості спостережень. Припустимо, маємо 10 років і три країни, показник набуває значень: 0 – мир, 1 – війна. Перші дві країни завжди перебувають у стані миру, третя – завжди у стані війни. У результаті є 30 спостережень, які несуть таку

саму інформацію, що й три спостереження за будь-який рік. Якщо аналізувати взаємозв'язок такого проблемного показника зі звичайним, то ефект спотворення буде менший. Але якщо досліджувати взаємозв'язок двох чи більше проблемних показників, то ефект спотворення буде суттєвим. Наприклад, критерії якості регресії можуть указувати на її адекватність, зважаючи на наявність 30 спостережень, хоча насправді вони мали б бути розраховані, виходячи із трьох спостережень. Аналогічну проблему можуть створювати глобальні показники, наприклад, світові ціни на нафту будуть однакові для всіх країн протягом одного періоду.

3.3.3. Формування узагальнюючої таблиці

Перед зведенням усіх даних до єдиної таблиці варто врахувати лаги, наприклад, створити кілька аркушів з даними щодо залежної змінної, де в одному аркуші дані буде зміщено на один стовпчик праворуч (лаг – один період), у другому – на два (лаг – два періоди) і т. д. Розглянемо як залежну змінну зовнішній борг (короткостроковий), решту змінних – як незалежні змінні. Створимо новий аркуш, де міститимуться дані щодо залежної змінної із лагом в один рік. Скопіюємо туди дані з аркушу, де містяться дані щодо залежної змінної (застосуємо опцію *Значення*/*Values* під час вставки), після чого змістимо на один стовпчик ліворуч усі значення (крім найбільш раннього періоду-стовпчика) за допомогою операцій *Вирізати*/*Copy* та *Вставити*/*Paste*. Одержимо таблицю (рис. 3.4).

Так само можна створити додаткові аркуші, де залежна змінна (або змінні) матиме значення, що зміщенні на два роки (два стовпчики), три роки і т. д.

Далі таблиця щодо кожного показника перетворюється на стовпчик (рис. 3.5). Тобто у стовпчику йтимуть дані за всіма країнами почергово за перший період, нижче – за всіма країнами за другий період і т. д., або інший варіант – за першою країною – за всі роки, далі – за другою країною за всі роки і т. д. Це можна робити вручну. Але для автоматизації необхідно прописати макрос, зробивши це вручну лише за однією таблицею під час запису макросу, а для решти таблиць – викликати макрос.

	A	B	C	D	E	F	G	H	I	J	K	L
1	External debt		X	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
2			Y	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	
3			Afghanistan						18152000	20997000	16926000	
4			Angola	1.45E+09	1.21E+09	1.07E+09	1.2E+09	2.32E+09	2.13E+09	2.27E+09	2.42E+09	
5			Albania	30619000	29046000	1.49E+08	1227000	2.83E+08	5.91E+08	8.27E+08	7.79E+08	
6			Argentina	2E+10	1.48E+10	2.23E+10	2.65E+10	3.48E+10	3.36E+10	3.81E+10	3.75E+10	
7			Armenia	41972000	2.2E+08	4.04E+08	4.1E+08	2.98E+08	3.07E+08	4.59E+08	4.65E+08	
8			Azerbaijan	1.03E+08	82380000	1.03E+08	1.38E+08	1.86E+08	5.2E+08	1.04E+09	1.17E+09	
9			Burundi	88395000	96346000	47562000	22560000	34059000	37617000	13785000	19303000	
10			Benin	78650000	73610000	33606000	28606000	44045000	44068000	5184000	37606000	
11			Burkina Faso	63472000	12918000	14139000	24096000	22222000	89063000	1.55E+08	1.1E+08	
12			Bangladesh	3.61E+08	5.72E+08	6.17E+08	7.12E+08	6.88E+08	1.18E+09	1.38E+09	1.99E+09	
13			Bulgaria	1.22E+09	1.84E+09	2.66E+09	3.26E+09	4.44E+09	8.04E+09	1.4E+10	1.85E+10	
14			Bosnia and Herzegovina	59730000	3.49E+08	1.13E+08	3.6E+08	8.37E+08	1.17E+09	1.69E+09	9.12E+08	
15			Belarus	1.31E+09	1.63E+09	1.97E+09	2.94E+09	3.5E+09	3.65E+09	6.88E+09	6.96E+09	
16			Belize	50844000	45055000	80000000	319000	5996000	6615000	6358000	6558000	
17			Bolivia	3.8E+08	3.7E+08	3.32E+08	2.72E+08	1.82E+08	2.2E+08	1.77E+08	1.66E+08	
18			Brazil	2.83E+10	2.34E+10	2.46E+10	2.53E+10	2.4E+10	2.03E+10	3.92E+10	3.67E+10	

Рис. 3.4

	A	B	C	D	E
1			External debt stocks, short-term (DOD, current US$)		
2	Afghanistan	2000			
3		2001			
4		2002			
5		2003			
6		2004			
7		2005			
8		2006	18152000		
9		2007	20997000		
10		2008	16926000		
11	Angola	2000	1322922000		
12		2001	1451099000		
13		2002	1208386000		
14		2003	1074998000		
15		2004	1196102000		
16		2005	2315756000		
17		2006	2132157000		
18		2007	2271190000		
19		2008	2419205000		
20	Albania	2000	36666000		
21		2001	30619000		
22		2002	29046000		

Рис. 3.5

Такі стовпчики із кожної таблиці копіюють до однієї узагальнюючої таблиці, де кожному стовпчику відповідає змінна, а кожному рядку – спостереження (напр., країно-рік), при цьому резервують перший рядок під назви змінних, а перший стовпчик – під назви спостережень. Тепер таблицю можна аналізувати у Microsoft Excel або скопіювати до іншої програми та аналізувати там.

3.4. Відсутні дані

3.4.1. Проблеми відсутності даних і їх діагностика

Відсутність даних може бути результатом кількох причин, зокрема:

- *ненаявність статистичних даних*;
- *можливий вихід розрахованого показника за межі допустимих значень* (напр., якщо потрібно розрахувати логарифм показника, який для певного спостереження набуває від'ємного значення);
- *неприйнятність певного показника для певної групи спостережень, об'єктів або років*. Наприклад, такий показник, як

кумулятивний (сумарний за періоди) дефіцит поточного рахунку платіжного балансу (у % від ВВП) із часу виникнення дефіциту можна виміряти лише для країн, де такий він був щонайменше раз за всю історію визначення статистичних даних; кількість років до наступних виборів можна виміряти лише для країн, де вибори відбуваються.

Іноді частка відсутніх даних настільки велика, що вибірка з наявними даними є занадто малою, і її потрібно збільшити. Відсутні дані можуть становити певну небезпеку, якщо є результатом певної закономірності, наприклад, за аналізу впливу рівня розвитку країни на її міжнародну інвестиційну позицію. У менш розвинених країнах такі статистичні дані зазвичай є менш доступними. Використовуючи регресійний або інший вид аналізу, у результаті можна одержати певну закономірність. Але наскільки вона буде надійною, зважаючи те, що країни з низьким рівнем розвитку та частину країн із середнім рівнем розвитку не брали до уваги? Тому безпечними для результатів аналізу можуть бути лише відсутні дані, які не є результатом важливої для нас закономірності, або частка спостережень з відсутніми даними достатня мала (менше 15-20 %).

Діагностика закономірності у відсутніх даних. Припустимо, здійснюють аналіз зв'язку між припливом інвестицій (у % від ВВП) і податком на прибуток, і за частиною спостережень (напр., країн) дані щодо зміни податку на прибуток відсутні. Діагностику закономірності відсутніх даних можна провести кількома шляхами:

1. Поділити вибірку на дві групи (рис. 3.6): першу – з наявними даними за податком на прибуток, другу – з відсутніми даними за податком на прибуток. Порівняти середні значення припливу інвестицій цих групах. Наявна суттєва різниця (4,33 та 1,5); відсутні дані становлять 25 % усіх спостережень. Отже, закономірність існує: відсутність даних обумовлена певним рівнем одного з показників – припливом інвестицій. Тому потрібно бути дуже акуратним за інтерпретації результатів подальшого аналізу.

	A	B	C	D	E	F	G	H	I
1				Група з відомими даними			Група з відсутніми даними		
2	Інвестиції	Податок		Інвестиції	Податок		Інвестиції	Податок	
3	4	20		4	20				
4	6	15		6	15				
5	1						1		
6	2						2		
7	2	30		2	30				
8	4	35		4	35				
9	3	25		3	25				
10	7	0		7	0				
11									
12			середнє	4.333		середнє	1.5		

Рис. 3.6

2. Створити бінарну змінну, яка набуває значення 1, якщо за податком на прибуток наявні дані; набуває значення 0, якщо відсутні (рис. 3.7). Далі розраховують коефіцієнт кореляції (напр., Пірсона, але бажано – Спірмена) між цією бінарною змінною та припливом інвестиції. Якщо кореляція висока або середня, а частка спостережень з відсутніми даними помітна, то закономірність у відсутніх даних наявна.

	A	B	C
1	Інвестиції	Податок	
2	4	20	1
3	6	15	1
4	1		0
5	2		0
6	2	30	1
7	4	35	1
8	3	25	1
9	7	0	1
10			
11			
12			
13		Кореляція Пірсона	0.634877

Рис. 3.7

3.4.2. Розв'язання проблем відсутності даних

Розв'язати проблему відсутності даних можна кількома способами (обраний спосіб вирішення варто описати в примітках).

1. Використати для аналізу лише спостереження з усіма наявними даними за всіма потрібними змінними (*complete case approach* або *casewise approach*). Наприклад, на рис. 3.8 використано лише позначені рамками три спостереження.

Експорт	Імпорт	ПІІ	Зовнішній борг
12	12	3	10
22	34	4	
33	23	5	30
21	22		40
34		4	50
56	22	2	
43	13	1	20

Рис. 3.8

Цей спосіб можна використовувати, якщо немає закономірності у відсутніх даних. У результаті кількість спостережень може скоротитися до неприйнятно низького рівня, що є недоліком способу. У такому випадку варто додати спостереження, наприклад, за інші роки чи за іншими країнами, хоча це не завжди можливо.

2. Видалити лише ті спостереження та змінні, за якими частка відсутніх спостережень є занадто великою, залишаючи спостереження та змінні без відсутніх даних або із мінімальною їх кількістю. У прикладі на рис. 3.9 використано лише перші три змінні, а останню змінну (зовнішній борг), де частка відсутніх даних найбільша, не використовуватимемо. Не використаємо також п'яте спостереження.

Часто це є оптимальним шляхом і дозволяє суттєво зменшити частку відсутніх даних. За цього способу також видаляють спостереження з відсутніми даними щодо залежної змінної. Замість важливих видалених змінних бажано включити до аналізу змінні, які значно корелюють з видаленими змінними.

3. Попарно не враховувати спостереження при аналізі (*Pairwise approach*), наприклад, при побудові кореляційної матриці. При розрахунку парної кореляції між кожними двома змінними враховують усі спостереження, за якими наявні дані принаймні за цими двома змінними, хоча за іншими змінними у таких спостережень можуть бути відсутні дані. У прикладі на рис. 3.10 для розрахунку кореляції між експортом та імпортом використано всі спостереження, крім п'ятого, а кореляції між прямими іноземними інвестиціями та зовнішнім боргом – перше, третє, п'яте та сьоме.

Експорт	Імпорт	ПІІ	Зовнішній борг
12	12	3	10
22	34	4	
33	23	5	30
21	22		40
		4	50
56	22	2	
43	13	1	20

Рис. 3.9

Експорт	Імпорт	ПІІ	Зовнішній борг
12	12	3	10
22	34	4	
33	23	5	30
21	22		4
34		4	50
56	22	2	
43	13	1	20

Рис. 3.10

4. Застосувати вставку інших даних замість відсутніх значень (*Imputation Methods*):

- застосувати *вставку даних з іншого джерела*. Проте потрібно пересвідчитися, що дані є порівнюваними, оскільки іноді один й той самий показник у різних джерелах може суттєво відрізнятися, наприклад, дані про короткостроковий зовнішній борг можуть відрізнятися у кілька разів, залежно від джерела та способу розрахунку;
- замінити відсутні дані на середні значення змінної (*Mean Substitution*). Недоліком цього способу є зменшення дисперсії змінної, а також кореляції з іншими змінними. Його варто застосовувати, якщо вид аналізу не дозволяє загалом мати відсутні дані або якщо дисперсія змінної не є великою. Наприклад, цей спосіб зазвичай використовують щодо приросту населення. Для нашого прикладу результат буде таким, як на рис. 3.11;
- замінити відсутні дані на значення за попередній період. Це можна робити за впевненості, що воно мало змінилося. Наприклад, спосіб можна використати щодо індексу економічної свободи, але неможна – щодо його зміни, приросту ВВП або інших достатньо волатильних показників;

	Експорт	Імпорт	ПІІ	Зовнішній борг
	12	12	3	10
	22	34	4	30
	33	23	5	30
	21	22	3.166667	40
	34	21	4	50
	56	22	2	30
	43	13	1	20
Середнє	31.57143	21	3.166667	30

Рис. 3.11

▪ замінити відсутні дані на значення, які розраховані на основі тренду (плинної середньої). Наприклад, є часовий ряд з трьома відсутніми даними (рис. 3.12). Показник достатньо рівномірно змінюється із часом. Можна розрахувати одразу середню зміну за період з відсутніми даними. Далі прописують у порожніх комірках відповідні формули (рис. 3.13) та отримують результат (рис. 3.14). Так можна робити, якщо є впевненість, що тренд є достатньо стійким, показник не є за своєю природою волатильним, і під час відповідного періоду не відбувалося екстраординарних подій (напр., війни);

▪ замінити відсутні дані на розраховані за допомогою регресії за відомих значень інших змінних, які суттєво впливають на змінну, щодо якої дані відсутні (*Regression Imputation*). Наприклад, відсутні дані про частку високотехнологічних товарів в експорті за деякими країнами у вибірці, але за всіма країнами з вибірки наявні дані щодо частки витрат на дослідження та розробки у ВВП. За всіма доступними спостереженнями з відомими даними можна побудувати регресію, яка визначає залежність частки високотехнологічних товарів в експорті від частки витрат на дослідження та розробки у ВВП. Далі за допомогою регресії розраховують значення частки високотехнологічних товарів для спостережень із відсутніми даними за цим показником.

Цей метод має низку *недоліків*:

•посилюються наявні залежності між змінними, і результати подальшого аналізу виявляються менш корисними при застосуванні для інших випадків, які перебувають за межами вибірки;

•занижується дисперсія змінної, якщо не включити до її розрахованих значень стохастичну складову;

•ризик ненадійності регресійно-кореляційного зв'язку з іншими змінними;

•розраховані значення можуть опинитися за межами допустимих значень (напр., частка сільськогосподарської продукції в експорті може бути більше 100 %).

- застосувати комбінований спосіб, що є поєднанням попередніх способів для різних змінних чи спостережень.

5. Застосувати процедури на основі моделювання (*Model-Based Procedures*):

- застосувати підхід ЕМ (*EM approach*) – ітеративну двокрокову процедуру. *Е-стадія* передбачає найкращу оцінку відсутніх даних. *М-стадія* розраховує середні, стандартні відхилення та кореляції з поправкою на те, що відсутні дані були заміщені;

- включити відсутні дані до аналізу за умови, що їх виділяють до окремої групи спостережень, яку порівнюють з іншими групами (напр., за частотного аналізу, аналізу середніх чи дисперсійного аналізу);

- застосувати *вставку* замість відсутніх інших даних (див. вище), але при цьому створюють нову бінарну змінну (або псевдозмінну), яка набуває значення 1, якщо спостереження має відсутні дані, та 0, – якщо ні. Це можна застосувати, наприклад, у регресійному аналізі, включивши до числа незалежних змінних таку додаткову бінарну змінну.

	D	E	F	G	H	I	J	K	L	M	N	O
1												
2	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	Середня зміна за період:		
3	23	34	45	40				50	56		2.5	

Рис. 3.12

2000	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
23	34	45	40	=G3+2.5	=H3+2.5	=I3+2.5	50	56

Рис. 3.13

2000	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
23	34	45	40	42.5	45	47.5	50	56

Рис. 3.14

Розділ 4
ПЕРВИННИЙ АНАЛІЗ ДАНИХ

4.1. Описова статистика

При побудові економіко-математичних моделей із застосуванням кількісних методів аналізу МЕВ дослідник зазвичай має справу з великими наборами даних. У практичних задачах часто виникає потреба опису основних особливостей даних однією або кількома числовими характеристиками. Таку техніку називають описовою (дескриптивною) статистикою, а самі числові характеристики – статистиками. При використанні та аналізі таких статистик вивчають дані не на основі попередньо заданої теоретичної моделі, а виходячи зі структури самих даних спостережень.

Описова статистика – це набір основних статистичних показників емпіричної вибірки значень кількісної ознаки. На практиці найчастіше мають справу з *вибірковими характеристиками*, які розраховують за обмеженою кількістю значень досліджуваного показника, що становлять певну вибірку із генеральної сукупності. Вони є оцінками відповідних *генеральних статистичних характеристик* (параметрів розподілу).

Вивчаючи певну ознаку X генеральної сукупності, можна зрозуміти характер закону розподілу випадкової величини X, але параметри цього закону невідомі. Звідси виникає задача: на підставі отриманої вибірки із генеральної сукупності визначити наближені числові значення невідомих параметрів розподілу. Такі наближені числові значення параметрів розподілу називають їх *статистичними оцінками*, або просто оцінками.

Статистичною оцінкою (*статистикою*) невідомого параметра теоретичного розподілу випадкової величини X називають функцію, за допомогою якої знаходять наближене значення параметра.

Точковою оцінкою θ^* невідомого параметра θ розподілу $y = F(\theta)$ генеральної сукупності X називають статистичну оцінку, яку визначає одне число.

Якщо обсяг вибірки досить великий, то точкова оцінка θ^* параметра θ буде близькою до його справжнього значення. Якщо ж обсяг вибірки невеликий, то між точковою оцінкою θ^* і справжнім значенням параметра розподілу θ можуть існувати значні розбіжності. Але точкова оцінка невідомого параметра, що знайдена за вибіркою обсягу n, не вказує на помилку, якої припускаються, набуваючи замість точного значення параметра θ його наближене значення θ^*. Тому виникає питання про надійність точкової оцінки θ^* (можливе її відхилення від істинного значення параметра θ).

Надійністю (*довірчою ймовірністю*) точкової оцінки θ^* параметра θ називають ймовірність γ, з якою виконується нерівність $P\left(\left|\theta-\theta^*\right|<\delta\right)=\gamma$, тобто $\theta\in(\theta^*-\delta;\theta^*+\delta)$. Це означає, що інтервал "покриває" невідоме точне справжнє значення параметра θ із наперед заданою надійністю γ, що близька до одиниці. На практиці надійність оцінки задають наперед, причому число γ обирають близьким до одиниці:

$$\gamma = 0{,}95;\ \gamma = 0{,}99;\ \gamma = 0{,}999.$$

Інтервал $(\theta^*-\delta;\theta^*+\delta)$ називають *довірчим* (*надійним*), а його межі $\theta^*-\delta$ та $\theta^*+\delta$ – *нижньою* та *верхньою довірчими* (*надійними*) *межами*, відповідно.

Точні й теоретично обгрунтовані надійні (довірчі) інтервали для параметрів нормального розподілу величини X подано у таб. 4.1: для параметра $a=M(X)$ – математичного сподівання за вибірковим середнім:

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

і параметра $\sigma^2=D(X)$ – за вибірковим точковими оцінками:

$$\overline{D}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \text{ і } S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2.$$

Таблиця 4.1

Параметри розподілу генеральної сукупності $a = M(X)$ **та** $\sigma = \sqrt{D(X)}$		**Тип розподілу генеральної сукупності**	
$\theta = a$	σ –відомо	*нормальний розподіл* $N(a,\sigma^2)$	*довірчі інтервали* $\bar{x} - c_\gamma \frac{\sigma}{\sqrt{n}} < a < \bar{x} + c_\gamma \frac{\sigma}{\sqrt{n}}.$
$\theta = a$	σ – відомо	*довільний розподіл* $n \geq 30$	$\bar{x} - c_\gamma \frac{\sigma}{\sqrt{n}} < a < \bar{x} + c_\gamma \frac{\sigma}{\sqrt{n}}.$
$\theta = a$	σ – невідомо	*нормальний розподіл* $N(a,\sigma^2)$	$\bar{x} - t_\gamma \frac{S}{\sqrt{n}} < a < \bar{x} + t_\gamma \frac{S}{\sqrt{n}}.$
$\theta = a$	σ – невідомо	*довільний розподіл* $n \geq 30$	$\bar{x} - t_\gamma \frac{S}{\sqrt{n}} < a < \bar{x} + t_\gamma \frac{S}{\sqrt{n}}.$
a – відомо	$\theta = \sigma$	*нормальний розподіл* $N(a,\sigma^2)$	$\frac{\sum_{i=1}^{n}(x_i - a)^2}{x_{n,(1-\gamma)/2}} < \sigma^2 < \frac{\sum_{i=1}^{n}(x_i - a)^2}{x_{n,(1+\gamma)/2}}$
a – невідомо	$\theta = \sigma$		$\frac{(n-1)S^2}{x_{n-1,(1-\gamma)/2}} < \sigma^2 < \frac{(n-1)S^2}{x_{n-1,(1+\gamma)/2}}$

Оскільки значення змінних можуть змінюватись, то потрібно навчитись описувати їх мінливість. Для цього існують *описові* або *дескриптивні статистики/descriptive statistics*: мінімум, максимум, середнє, дисперсія, стандартне відхилення, медіана, квартилі, мода тощо. Ідея цих статистик проста: замість розгляду всіх значень випадкової величини *X* (а їх може бути досить багато – тисячі й мільйони), спочатку варто розглянути описові статистики. Вони дають загальне уявлення про значення, якого набуває випадкова величина *X*. Описові статистики є точковими оцінками параметрів розподілу випадкової величини, яка описує генеральну сукупність.

Нехай $x_1, x_2, ..., x_k$ – вибірка обсягу n із генеральної сукупності X із функцією розподілу $y = F(x)$, дискретний статистичний розподіл абсолютних частот якої задано у табл. 4.2.

Таблиця 4.2

x_k	x_1	x_2	x_3	…	x_k	…
n_k	n_1	n_2	n_3	…	n_k	…

При цьому має місце рівність:

$$\sum_{k=1}^{\infty} n_k = 1.$$

Найпростіший спосіб охарактеризувати вибірку у цілому одним числом – це вказати "середнє положення" або "центр вибірки", навколо якого коливаються значення вибірки. Розглянемо найбільш поширені статистики середнього положення.

Вибірковим початковим емпіричним моментом m-го порядку статистичного розподілу групованої вибірки називають величину:

$$\overline{a_m} = \frac{1}{n}\sum_{i=1}^{k} x_i^m n_i, \qquad (4.1)$$

де x_i – значення i-ої варіанти, n_i – її частота, n – обсяг вибірки, k – кількість груп у вибірці.

Якщо $m = 1$, то величину називають вибірковим середнім і позначають:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i = \sum_{i=1}^{k} x_i \cdot \frac{n_i}{n} = \sum_{i=1}^{k} x_i \cdot w_i$$

або

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i . \qquad (\mathbf{4.2})$$

Вибіркове середнє/mean $\bar{x}$ є однією з основних характеристик статистичного розподілу вибірки. Вибіркове середнє є такою точкою, сума відхилень спостережень ознаки від якої дорівнює нулю. Вибіркове середнє – єдина точка, яка має таку властивість, і це вирізняє її серед усіх інших. Крім того, вибіркове середнє має таку властивість: сума квадратів відстаней між спостережуваними значеннями ознаки та їх середнім арифметичним є мінімальною. Якщо замість середнього арифметичного взяти довільну іншу величину, то сума квадратів відстаней спостережуваних значень і цієї величини буде тільки більше, але жодним чином не менше.

На практиці розраховують середнє за вибіркою, але цікавим є справжнє значення середньої для генеральної сукупності. Для його визначення розраховують *довірчі межі для середнього/confidence limits for mean* із певною мірою ймовірності (зазвичай 95 %). Це означає, що справжня середня перебуває між нижньою та верхньою межами такого інтервалу, і впевненість у цьому становить 95 %. Завжди є невелика ймовірність (5 %), що справжня середня перебуває за межами такого довірчого інтервалу. Широта довірчого інтервалу залежить від величини вибірки (кількості значень) та дисперсії.

Перейдемо до означення основних характеристик розсіювання значень випадкової величини навколо її середнього значення, які розраховують на основі вибірки.

Вибірковим центральним емпіричним моментом m-го порядку статистичного розподілу групованої вибірки називають величину:

$$\bar{\mu}_m = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^m n_i , \qquad (\mathbf{4.3})$$

де x_i – значення i-ої варіанти, n_i – її частота, n – обсяг вибірки, k – кількість груп у вибірці. Зокрема, $\bar{\mu}_0 = 1$, $\bar{\mu}_1 = 0$. Якщо $m = 2$, то величину $\bar{\mu}_2$ називають *вибірковою дисперсією*

/*sample variance*, позначають символом $\overline{D}$[117], обчислюють за формулами:

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad (4.4)$$

або

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\overline{x})^2. \qquad (4.5)$$

Вибіркова дисперсія дорівнює сумі квадратів відхилень спостережень від вибіркового середнього. Але, оскільки вибіркова дисперсія – величина вимірна, яка створює незручності у дослідженнях, то за міру розсіювання значень випадкової величини за результатами вибірки приймають вибіркове середнє квадратичне відхилення.

Вибірковим середнім квадратичним відхиленням статистичного розподілу вибірки називають величину

$$\overline{\sigma} = \sqrt{\overline{D}}. \qquad (4.6)$$

Виявляється, що вибіркове середнє $\overline{x}$ і вибіркова дисперсія $\overline{D}$ є спроможними точковими оцінками для математичного сподівання та дисперсії. При цьому $\overline{x}$ є незміщеною, а $\overline{D}$ – зміщеною оцінкою. Тому часто використовують виправлену дисперсію – S^2, яка відрізняється від звичайної лише нормуючим множником $\frac{(n-1)}{n}$ та є незміщеною точковою оцінкою дисперсії:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2, \qquad (4.7)$$

а за *стандартне відхилення* обирають величину:

$$S = \sqrt{S^2}. \qquad (4.8)$$

Стандартне відхилення/*standard deviation* – величина, яку найчастіше використовують для міри мінливості ознаки. *Дисперсія*/*variance* змінюється від 0 до нескінченності. Крайнє значення 0 означає відсутність мінливості значень ознаки за сталих значень змінної. Що більше дисперсія або стандартне відхилення, то сильніше розкидані значення ознаки щодо середнього.

[117] Термін уперше введено Фішером у 1918 р.

При розв'язуванні практичних задач застосовують інші форми середнього, які можна отримати із середнього степеневого m-го порядку. *Вибірковим середнім степеневим m-го порядку* статистичного розподілу вибірки називають величину:

$$\overline{x_m} = \left(\frac{1}{n} \sum_{i=1}^{k} x_i^m n_i \right)^{\frac{1}{m}}. \qquad \textbf{(4.9)}$$

Якщо $m = 1$, то отримують *вибіркове середнє* (або *середнє арифметичне*) $\overline{x}$ (**4.2**).

Якщо $m = -1$, то отримають вибіркове середнє гармонійне:

$$\overline{x_{-1}} = \left(\frac{1}{n} \sum_{i=1}^{k} x_i^{-1} n_i \right)^{-1} = \frac{n}{\sum\limits_{i=1}^{k} \frac{n_i}{x_i}}. \qquad \textbf{(4.10)}$$

Якщо $m = 0$ (після розкриття невизначеності при обчисленні границі $\lim\limits_{m \to 0} \overline{x_m}$), то отримують *вибіркове середнє геометричне/geometric mean*:

$$\overline{x}_0 = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}} = \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}}, \qquad \textbf{(4.11)}$$

яке застосовують, зокрема, для розрахунку середніх приростів.

Якщо $m = 2$, то отримують *вибіркове середнє квадратичне*:

$$\overline{x_2} = \left(\frac{1}{n} \sum_{i=1}^{k} x_i^2 n_i \right)^{\frac{1}{2}} = \sqrt{\frac{1}{n} \sum_{i=1}^{k} x_i^2 n_i}. \qquad \textbf{(4.12)}$$

Крім зазначених середніх величин, які називають *аналітичними середніми*, у статистичному аналізі застосовують *структурні* або *порядкові середні*. Серед структурних середніх найбільш широко застосовують *моду/mode* та *медіану/median*.

Мода (*mode* x_{mod}) статистичного розподілу вибірки[118] – це варіант із найбільшою частотою. Модою називають значення змінної ознаки, яке найчастіше зустрічають у вибірці (саме модне). Класичний приклад використання моди – ви-

[118] Термін уперше введено Пірсоном у 1894 р.

бір розміру взуття або кольору шпалер. Якщо розподіл має кілька мод, то його називають *мульти-* або *багатомодальним*. Мультимодальність розподілу дає важливу інформацію про природу змінної. Наприклад, у соціологічних опитуваннях, якщо змінна виражає переваги, то мультимодальність може означати існування кількох різних певних думок. Мультимодальність також є індикатором неоднорідності вибірки. Спостереження, можливо, утворені двома або більшою кількістю "накладених" розподілів. Якщо жодне значення не повторюється, то моди немає.

Медіана (*median* x_{med}) статистичного розподілу вибірки[119] – це значення, яке розбиває вибірку на дві рівні частини, тобто ліворуч медіани перебуває стільки саме значень, скільки й праворуч. Медіана має важливу властивість: сума абсолютних відстаней між елементами вибірки та медіаною є мінімальною. Медіана дає уявлення про місце зосередження значення змінної, тобто центру вибірки. У деяких випадках, зокрема, при дослідженні доходів населення, медіана є зручнішою, ніж середнє.

Геометрично медіана – це така точка на осі абсцис, що пряма, яка проходить через цю точку, ділить площу фігури, що обмежена віссю абсцис і графіком функції щільності, навпіл. Медіана є кращою оцінкою середніх величин для змінних у порядковій шкалі або метричних змінних, що не підлягають нормальному розподілу, коли є асиметрія розподілу значень змінної.

Квантиль вибірки[120] представляє число x_p, менше від якого розташована p-та частина (частка) вибірки. Формально p-квантиль неперервного розподілу $y = F(x)$, за якого функція розподілу набуває значення, рівного p:

$$F(x_p) = P(X < x_p) = p, 0 < p < 1. \qquad (\mathbf{4.13})$$

Деякі квантилі дістали особливу назву. Очевидно, медіана випадкової величини – це квантиль рівня 0,5. Квантилі $x_{0,25}$ та $x_{0,75}$ дістали назву *нижня квартиль*/*lower quartile* та *верхня квартиль*/*upper quartile*, відповідно (від слова

[119] Термін уперше введено Гальтоном у 1882 р.
[120] Термін уперше використано Кендаллом у 1940 р.

"кварта" – "чверть"; термін уперше введено Гальтоном у 1882 р.). Із поняттям квантиля тісно пов'язане поняття *відсоткової точки*. Під p % точкою розуміють квантиль x_{1-p}, тобто таке значення випадкової величини X, за якого:

$$P(X \geq x_{1-p}) = p. \qquad (4.14)$$

Три точки: нижня, верхня квартилі та медіана поділяють вибірку на чотири рівні частини за кількістю спостережень.

p%-й – персентиль/percentile – це величина, менше від якої p % усіх спостережень. Наприклад, мінімум є 0 %-й персентиль, нижня квартиль – це 25 %-й персентиль, медіана – це 50 %-й персентиль, верхня квартиль – це 75 %-й персентиль. Можна визначити будь-який персентиль. Наприклад менше 90 %-го персентиля 90 % усіх значень.

Ранг/rank – це місце значення у переліку, в якому всі значення розташовані від найбільшого до найменшого. Ранг 1 має максимальне значення.

Відсотковий ранг/percent rank – це відсоток значень, що менші від нього. Наприклад, максимум має відсотковий ранг 100 %, медіана має 50 %. Операції визначення персентиля та відсоткового рангу зворотні одна одній.

Квартильний розмах/quartiles range змінних – це різниця між верхньою та нижньою квартилями, фактично – інтервал $(x_{0,25}; x_{0,75})$, який містить медіану, і до якого потрапляє 50 % спостережень.

Варіація. Найбільший інтерес являють міри варіації (розсіювання) спостережень навколо середніх величин, зокрема навколо середнього арифметичного (вибіркового середнього). Найпростішим і дуже приблизним показником варіації є *варіаційний розмах вибірки/range* – різниця між *найбільшим /maximum* і *найменшим/minimum* елементами вибірки:

$$R = x_{\max} - x_{\min}. \qquad (4.15)$$

Вибірковим середнім абсолютним відхиленням/mean absolute deviation статистичного розподілу вибірки називають величину, яка є середнім зваженим абсолютних величин відхилень від середнього арифметичного:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{k} |x_i - \bar{x}| n_i . \qquad (4.16)$$

Зауважимо, що "проста" сума:

$$\frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x}) n_i$$

не може характеризувати варіацію, оскільки ця сума дорівнює нулю для будь-якого варіаційного ряду.

Стандартна похибка середньої/standard error of mean – це стандартне відхилення, поділене на корінь квадратний кількості спостережень у вибірці:

$$\delta = \frac{S}{\sqrt{n}}.$$

Розглянемо кілька статистик, які використовують для опису форми розподілу. За міру варіації (розсіювання) зручно брати середнє квадратичне відхилення вибірки $\bar{\sigma}$ або стандартне відхилення *S*, оскільки матимемо характеристику, яка виражена в тих самих одиницях, що й значення ознаки. Крім того, розглядають також і безрозмірну величину – коефіцієнт варіації, який дорівнює відсотковому відношенню середнього квадратичного відхилення до вибіркового середнього.

Коефіцієнтом варіації/coefficient of variation статистичного розподілу вибірки називають відношення стандартного відхилення до середнього та обчислюють за формулою:

$$V = \frac{\bar{\sigma}}{\bar{x}} \cdot 100\ \%, \bar{x} \neq 0. \qquad (4.17)$$

Слід зауважити: якщо коефіцієнт варіації набуває додатних значень і великий (близький до 100 %), то зазвичай це свідчить про неоднорідність значень ознаки.

Для оцінювання відхилення статистичного розподілу вибірки від нормального розподілу використовують вибіркову асиметрію та вибірковий ексцес.

Асиметрія/skewness. *Вибірковим коефіцієнтом асиметрії* статистичного розподілу вибірки[121] називають величину:

[121] Термін уперше використано Пірсоном у 1895 р.

$$\overline{A} = \frac{1}{n \cdot \overline{\sigma}^3} \sum_{i=1}^{k} (x_i - \overline{x})^3. \qquad \textbf{(4.18)}$$

Коефіцієнт асиметрії є мірою несиметричності розподілу. Якщо розподіл симетричний щодо вибіркового середнього $\overline{x}$, то коефіцієнт асиметрії $\overline{A} = 0$, тобто точки рівновіддалені від $\overline{x}$. Якщо $\overline{A} > 0$, то крива розподілу має додатну (правосторонню), якщо $\overline{A} < 0$ – від'ємну (лівосторонню) асиметрію (рис. 4.1).

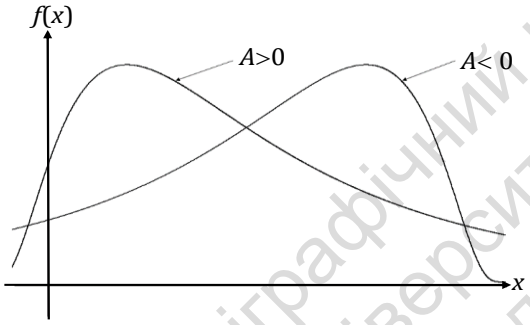


Рис. 4.1

Ексцес. *Вибірковим коефіцієнтом ексцесу* статистичного розподілу вибірки[122] називають величину:

$$\overline{E} = \frac{1}{n \cdot \overline{\sigma}^4} \sum_{i=1}^{k} (x_i - \overline{x})^4 - 3. \qquad \textbf{(4.19)}$$

Ексцес в основному застосовують для неперервних випадкових величин, він характеризує для так звану "крутизну" кривої статистичного розподілу. Якщо розподіл випадкової величини нормальний, то $\overline{E} = 0$. Якщо $\overline{E} > 0$, то крива розподілу випадкової величини буде більш "гостровершинною", ніж крива нормального розподілу; якщо $\overline{E} < 0$, то крива розподілу буде більш "плосковершинною", порівняно з кривою нормального розподілу (рис. 4.2).

Якщо випадкова величина *X*, яка описує генеральну сукупність, має нормальний розподіл, то її асиметрія та ексцес дорівнюють нулю. Тому, що більше відрізняється від нуля асиметрія та ексцес статистичного розподілу вибірки, то менше підстав стверджувати, що вибірка утворена із нормальної генеральної сукупності.

[122] Термін уперше використано Пірсоном у 1905 р.

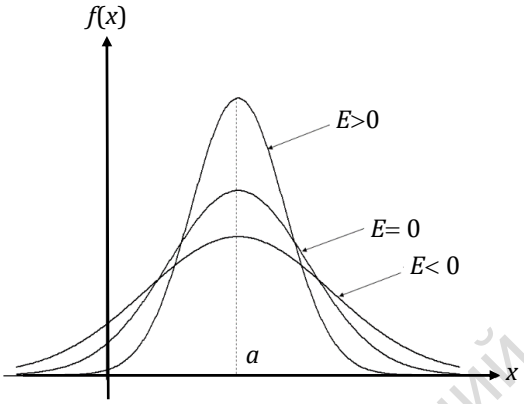


Рис. 4.2

4.2. Розрахунок описової статистики у Microsoft Excel

Описову статистику можна розрахувати за допомогою опції *Descriptive statistics* у надбудові *Data Analysis.* Наприклад, на рис. 4.3 показано вхідні дані за поточним рахунком (у % ВВП) для вибірки країн (за даними *World Development Indicators*), а далі – діалогове вікно (рис. 4.4) та результат аналізу (рис. 4.5). Як додаткові опції вказано, зокрема, друге найменше та друге найбільше спостереження.

	A	B	C
1	Current account balance (% of GDP)		2010
2		Slovak Republic	-3.38
3		Slovenia	-0.81
4		Sweden	6.28
5		Thailand	4.63
6		Tajikistan	-6.79
7		Turkey	-6.49
8		Tanzania	-8.58
9		Uganda	-10.23
10		Ukraine	-2.09
11		Uruguay	-0.40
12		United States	-3.23
13		Venezuela, RB	3.71
14		Vietnam	-4.14
15		South Africa	-2.78
16		Zambia	3.80

Рис. 4.3

Інший варіант – самостійно використати відповідні функції. =AVERAGE (діапазон) повертає середню арифметичну вказаних чисел чи діапазону. Приклади: = AVERAGE (B2:B10) або = AVERAGE (35;121;76;18).

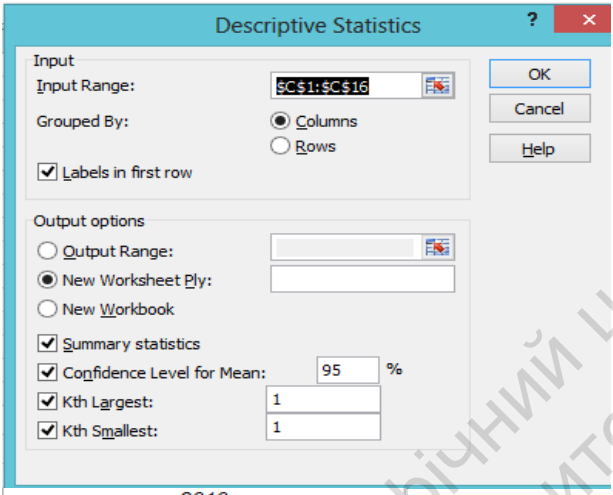


Рис. 4.4

2010	
Mean	-2.03411
Standard Error	1.28219
Median	-2.78174
Mode	#N/A
Standard Deviation	4.965899
Sample Variance	24.66016
Kurtosis	-0.83032
Skewness	0.149812
Range	16.50412
Minimum	-10.2281
Maximum	6.27599
Sum	-30.5116
Count	15
Largest(2)	4.627209
Smallest(2)	-8.57966
Confidence Level(95.0%)	2.750023

Рис. 4.5

=AVERAGEIF (діапазон;умова;діапазон усереднення)

повертає середню арифметичну для комірок у діапазоні, які відповідають указаній умові. Середню розраховують для діапазону усереднення (третій аргумент функції); якщо його не вказано, то для діапазону (перший аргумент функції). Розглянемо приклади в табл. 4.3.

Таблиця 4.3

=AVERAGEIF (B3:B20;"<100")	Розрахунок середньої тих значень із діапазону B3:B20, які менші 100
=AVERAGEIF (B3:B20;">0";C3:C20)	Розрахунок середньої значень із діапазону C3:C20 із тих рядків, де у діапазоні B3:B20 значення більші 0

Закінчення табл. 4.3.

=AVERAGEIF (B3:B20;"=індекс цін*"; C3:C20)	Розрахунок середньої значень із діапазону C3:C20 із тих рядків, де у діапазоні B3:B20 є текст, який починається з *індекс цін*
=AVERAGEIF (B3:B20;"<>Ukraine";C3:C20)	Розрахунок середньої значень із діапазону C3:C20 із тих рядків, де у діапазоні B3:B20 не написано слово Ukraine

=AVERAGEIFS (діапазон усереднення;діапазон умов1;умова1; діапазон умов2;умова2;...)

повертає середню арифметичну для комірок у діапазоні, які відповідають кільком умовам. Середнє розраховують для діапазону усереднення. Діапазони умов і діапазон усереднення мають бути одного розміру. Розглянемо приклад у табл. 4.4.

Таблиця 4.4

	A	B	C	D
1	**Країна експортер**	**Країна імпортер**	**Товар**	**Ціна**
2	Україна	Польща	чавун	500
3	Україна	ФРН	чавун	480
4	Норвегія	Польща	нафта	400
5	Україна	ФРН	зерно	200

=AVERAGEIFS (D2:D7;A2:A7;"Україна";C2:C7;"чавун")

повертатиме середню ціну чавуну (експортований з України).

=CONFIDENCE (альфа;стандартне відхилення;розмір вибірки)

повертає довірчий інтервал для середньої генеральної сукупності з нормальним розподілом. Альфа – рівень значущості для розрахунку рівня надійності. Наприклад, 0,05 – це 95 % рівень надійності. Насправді для одержання максимального (або мінімального) значення довірчого інтервалу потрібно результат розрахунку за цією функцією додати до (або відняти від) середньої за вибіркою (приклад – рис 4.6).

A	B	C	D	E
Mean	0.31			
Standard Deviation	0.14			
Size of the Sample	26			
Confidence interval	=CONFIDENCE(0.05;B2;B3)			
Lower confidence limit	=B1-B4			
Upper confidence limit	=B1+B4			

Рис. 4.6

=GEOMEAN (діапазон) повертає середню геометричну.

=HARMEAN (діапазон) повертає середню гармонійну (обернена величина середньої арифметичної обернених величин значень). Не може бути застосований, якщо одне зі значень дорівнює 0.

Розрахунок середньозваженої за допомогою двох функцій. Розглянемо приклад (табл. 4.5). Формула

=SUMPRODUCT (B2:B5;C2:C5)/SUM(B2:B5)

повертає середньозважену ціну.

Таблиця 4.5

	A	B	C
1	Місяць	Експорт, тис. т	Ціна за тонну
2	Січень	100 000	400
3	Лютий	120 000	430
4	Березень	150 000	380
5	Квітень	110 000	390

=MEDIAN (діапазон) повертає медіану.

=MODE (діапазон) повертає моду.

=STDEV (діапазон) повертає стандартне відхилення для вибірки.

=VAR(діапазон) повертає дисперсію для вибірки.

=AVEDEV(діапазон)повертає середнє абсолютних відхилень значень від середньої.

=MAX (діапазон) повертає найбільше значення. Наприклад, =МАКС(H2:H16;100) повертає найбільше серед чисел зі вказаного діапазону, але якщо найбільше число менше 100, то повертає 100.

=MIN (діапазон) повертає найменше значення.

=SMALL (діапазон;*n*) повертає вказане *n*-е найменше значення з діапазону. Наприклад: =SMALL(B2:B100;2) повертає друге найменше значення з діапазону B2:B100.

=LARGE (діапазон;*n*) повертає вказане *n*-е найбільше значення із діапазону.

Можна комбінувати дві останні функції з іншими (приклад – табл. 4.6).

Для розрахунку рангів і персентилей (квантилей) у надбудові *Data Analysis* існує опція *Rank and Percentile*. Використаємо один із попередніх прикладів. Нижче показано вхідні дані (рис. 4.7), результат (рис. 4.8) і діалогове вікно аналізу (рис. 4.9).

Таблиця 4.6

=SUM(SMALL(B2:B100;{1;2;3;4}))	повертає суму чотирьох найменших значень з діапазону B2:B100
=AVERAGE(LARGE(B2:B100;{1;2}))	повертає середню арифметичну двох найбільших значень з діапазону B2:B100

A	B	C
Current account balance (% of GDP)		2010
	Slovak Republic	-3.38
	Slovenia	-0.81
	Sweden	6.28
	Thailand	4.63
	Tajikistan	-6.79
	Turkey	-6.49
	Tanzania	-8.58
	Uganda	-10.23
	Ukraine	-2.09
	Uruguay	-0.40
	United States	-3.23
	Venezuela, RB	3.71
	Vietnam	-4.14
	South Africa	-2.78
	Zambia	3.80

Рис. 4.7

A	B	C	D
Point	*2010*	*Rank*	*Percent*
3	6.28	1	100.00%
4	4.63	2	92.80%
15	3.80	3	85.70%
12	3.71	4	78.50%
10	-0.40	5	71.40%
2	-0.81	6	64.20%
9	-2.09	7	57.10%
14	-2.78	8	50.00%
11	-3.23	9	42.80%
1	-3.38	10	35.70%
13	-4.14	11	28.50%
6	-6.49	12	21.40%
5	-6.79	13	14.20%
7	-8.58	14	7.10%
8	-10.23	15	0.00%

Рис. 4.8

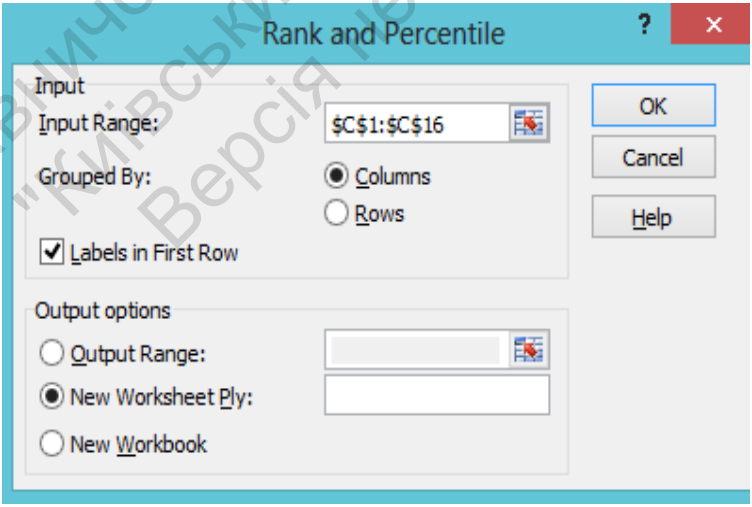


Рис. 4.9

Перший стовпчик у результатах показує номер спостереження у діапазоні вхідних даних; другий – значення досліджуваної змінної; третій – звичайний ранг (найбільшому значенню присвоюється ранг 1); четвертий – відсотковий ранг (найбільшому значенню присвоюється 100 %).

Існують також функції, за допомогою, яких це можна зробити самостійно.

=RANK (число;діапазон;порядок)

повертає ранг числа у переліку чисел (діапазоні), тобто повертає порядковий номер числа, якби всі числа в діапазоні були відсортовані у певному порядку. Якщо порядок=0 або не вказано, то найбільшому числу у діапазоні присвоюється ранг 1. Якщо порядок=1, то ранг 1 присвоюється найменшому числу. Однаковим числам присвоюється однаковий ранг. Наприклад, =RANK (C2;C16:C6;1)

=PERCENTILE (діапазон;*k*)

повертає *k*-ту персентиль для значень із діапазону. *k* набуває значень від 0 (найменше) до 1 (найбільше). У попередньому прикладі

= PERCENTILE (C2:C16;0.857)

повертатиме 3.8 – у даному випадку третє – найбільше значення.

=PERCENTRANK (діапазон;число;розрядність)

повертає відсотковий ранг числа. Розрядність – необов'язковий аргумент, який визначає кількість значущих цифр для значення, що повертається. Якщо він опущений, то повертається значення у тисячних (напр., 0.333, якщо третина значень у діапазоні є нижчими за вказане число). У попередньому прикладі функція

=ПРОЦЕНТРАНГ(C2:C16;3.8;3)

повертатиме 0.857.

=QUARTILE (діапазон;частина)

повертає квартиль діапазону. Якщо частина дорівнює 0, то повертається мінімальне значення; якщо 1 – перша квартиль (25 %-персентиль), якщо 2 – медіана (50 %-персентиль), якщо 3 – третя квартиль (75 %-персентиль), якщо 4 – максимальне значення.

4.3. Викиди

Вхідні дані можуть містити аномальні значення (*крайні точки, грубі похибки* або *викиди/outliers)*. *Викиди* – це спостереження із незвичними характеристиками, які явно відрізняються від решти спостережень (нетипові спостереження). З одного боку, вони також являють сукупність спостережень, з іншого – можуть створювати проблеми щодо надійності результатів аналізу. Аномальні значення можуть суттєво впливати на результати кількісного аналізу, наприклад на побудову моделі, зокрема на вид і коефіцієнти рівнянь регресії. Викиди можна поділи на кілька груп:

- *викиди внаслідок помилки* (вимірювання або при введені даних на комп'ютер). Наприклад, замість частки сільськогосподарських товарів в експорті 30 % можна помилково ввести 130 %. Якщо помилку неможливо виправити, то спостереження потрібно вилучити з аналізу або вважати такими, дані за якими відсутні;
- *викиди як результат неординарної події* (напр., світова війна або розпад СРСР). Потрібно вирішити, чи варто включати такі викиди до аналізу. Якщо є впевненість, що неординарна подія не повторюватиметься, то необхідно такі викиди виключити з аналізу;
- *викиди, які неможна пояснити попередніми причинами*. Якщо викид являє важливу частину генеральної сукупності, то варто провести аналіз двічі – з викидами та без них, порівнюючи результати. Це є одним зі способів перевірки *стійкості результатів/robustness check*;
- *викиди*, які індивідуально за кожною змінною начебто не є викидами, але *в сукупності за всіма змінними мають таку комбінацію значень, що є незвичною*. Наприклад, країна із доволі високою часткою високотехнологічної продукції в експорті та достатньо низьким рівнем розвитку може виглядати незвично. У більшості випадків такі викиди включають до аналізу або проводять аналіз двічі – з викидами та без них.

Варто виключати з аналізу лише дуже невелику частину спостережень (1-3), керуючись тим, що вони можуть бути нерепрезентативними викидами. Проведення аналізу двічі –

з викидами та без них – є страховкою від неправильного рішення щодо викидів. У великій вибірці (кілька сотень спостережень) викиди вже не становлять суттєвої проблеми.

Ідентифікація викидів:

1. *Одномірна/univariate detection*, за якою відбувається аналіз розподілу спостережень щодо кожної змінної окремо. Якщо вибірка для аналізу є невеликою (до 50-80 спостережень), то варто вважати значення змінних викидами (якщо вони відрізняються від середньої більш ніж на 2,5 стандартних відхилення). Якщо вибірка є великою, викидами слід вважати значення, що відхиляються від середньої більш ніж на три-чотири стандартних відхилення. Для цього в Microsoft Excel можна застосувати формулу для розрахунку стандартизованих значень:

=STANDARDIZE(значення;середня;стандартне відхилення);

2. *Двомірна/bivariate detection*, за якої створюється діаграма розсіювання, де за осями *X* та *Y* відкладають дві змінні. Рисують довірчий еліпс, до якого потрапляє переважна більшість спостережень (90-95 %). Спостереження, що перебувають за межами довірчого еліпсу, вважатимуть викидами. Далі створюють нову діаграму розсіювання для іншої пари змінних і т. д.

3. *Багатомірна/multivariate detection*. Уявимо побудову діаграми розсіювання не у двовимірному, а у багатомірному просторі (за багатьма змінними). Кожне спостереження може характеризувати відстань від центру всіх спостережень у багатомірному просторі, яку можна вимірювати як відстань Махаланобіса/*Mahalanobis* D^2. Спостереження з найбільшою відстанню можна вважати викидами. Стане у нагоді також і кластерний аналіз.

Розглянемо детальніше методи відсіювання грубих похибок.

Якщо вибірка малого обсягу ($n \le 25$), то можна скористатися методом обчислення максимального щодо відхилення:

$$\frac{|x_i - \bar{x}|}{S} \le \tau_{1-p}, \qquad (4.20)$$

де x_i – крайній (найбільший або найменший) елемент вибірки; $\bar{x}$, S – відповідно середнє та середнє квадратичне відхилення вибірки; τ_{1-p} – табличне значення статистики τ, що

обчислене за довірчої імовірності $1-p$. Якщо $\tau \le \tau_{1-p}$, то значення x_i не відсіюють; у протилежному випадку його вважають аномальним значенням і рекомендують виключити із даних. Процедуру відсіювання повторюють для тих елементів, які залишились у вибірці обсягу $n-1$ з новими переліченими числовими характеристиками $\bar{x}$ і S.

Можна скористатись іншим методом відсіювання похибок. Для цього обчислюють тест:

$$\tau = \frac{1}{\sqrt{(n-1)/n}} \cdot \frac{|x_i - \bar{x}|}{S}. \qquad (4.21)$$

Подальша процедура відсіювання ідентична попередньому способу. Цей спосіб чутливіший до похибок вимірювання, оскільки має уточнюючий коефіцієнт:

$$k = \frac{1}{\sqrt{(n-1)/n}}.$$

Для відсіювання грубих похибок у вибірках більшого обсягу можна скористатись розподілом Стьюдента (t-розподіл). У цьому методі використовують те, що $\tau_{p,n}$ виражають через критичне значення $t_{p,n-2}$ за формулою:

$$\tau_{p,n} = \frac{t_{p,n-2} \cdot \sqrt{n-1}}{\sqrt{n-2+(t_{p,n-2})^2}}. \qquad (4.22)$$

4.4. Основні розподіли та їх числові характеристики

4.4.1. Нормальний закон розподілу

Нормальний закон розподілу, або закон Гаусса, відіграє дуже важливу роль і займає винятково важливе особливе положення у застосуваннях кількісних методів МЕВ. За певних умов нормальний закон розподілу є граничним законом для багатьох інших законів розподілу.

Випадкова величина X має *нормальний закон розподілу із параметрами* a та σ (використовують запис $X \sim N(a,\sigma)$), якщо щільність розподілу випадкової величини має вигляд:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}, \qquad \textbf{(4.23)}$$

а відповідна *функція розподілу*:

$$F(x)=\frac{1}{\sigma\cdot\sqrt{2\pi}}\int_{\infty}^{x}e^{-\frac{(t-a)^2}{2\sigma^2}}dt. \qquad \textbf{(4.24)}$$

Параметри a та σ мають наступний імовірнісний зміст:

- параметр a дорівнює математичному сподіванню випадкової величини $X: M(X)=a$;
- параметр σ дорівнює середньому квадратичному відхиленню, σ^2 – дисперсії випадкової величини $X: D(X)=\sigma^2$, $\sigma(X)=\sigma$.

Нормальний розподіл випадкової величини з параметрами $a=0, \sigma=1$, тобто $X\sim N(0;1)$, називають *стандартним нормальним розподілом*, щільність якого є *функцією Гаусса*:

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}. \qquad \textbf{(4.25)}$$

Властивості функції Гаусса (рис. 4.10):

1) функція $\varphi(x)$ визначена для всіх $x\in(-\infty;+\infty)$ і $\varphi(x)>0$;

2) $\varphi(x)$ парна функція $\varphi(-x)=\varphi(x)$;

3) $\lim\limits_{x\to\infty}\varphi(x)=0$, тобто вісь Ox є асимптотою графіка функції $y=\varphi(x)$;

4) $\varphi(x)$ має локальний максимум у точці $x=0$ і $\varphi_{max}(x)=\varphi(0)=\frac{1}{\sqrt{2\pi}}$.

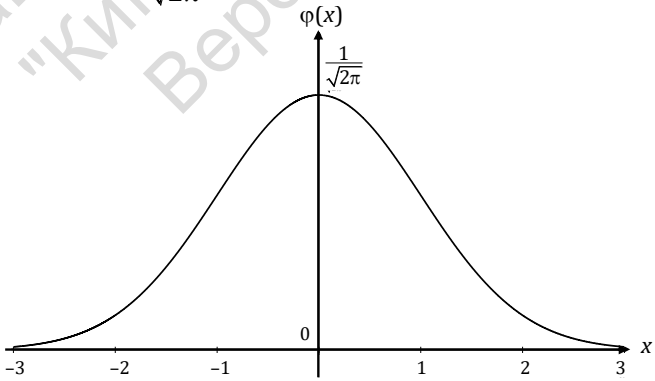


Рис. 4.10

Функція Гаусса табульована (див. додаток В1.) У додатку В1 значення функції $\varphi(x)$ наведено при $0 \le x \le 3{,}99$. Для обчислення значень $y = \varphi(x)$ за від'ємних значень *x* із проміжку $-3{,}99 \le x < 0$ використовують парність функції Гаусса, а при $|x| > 3{,}99$ приймають $\varphi(x) = 0$.

Параметри *a* та σ нормального розподілу мають також простий *геометричний зміст*.

З'ясуємо, як зміна параметрів нормального розподілу впливає на графік функції щільності $y = f(x)$. За зміни параметра *a* крива $y = f(x)$, не змінюючи своєї форми, зміщуватиметься вздовж осі абсцис ліворуч або праворуч (рис. 4.11), залежно від того, зменшується чи збільшується число *a*: $a_1 < a < a_2$.

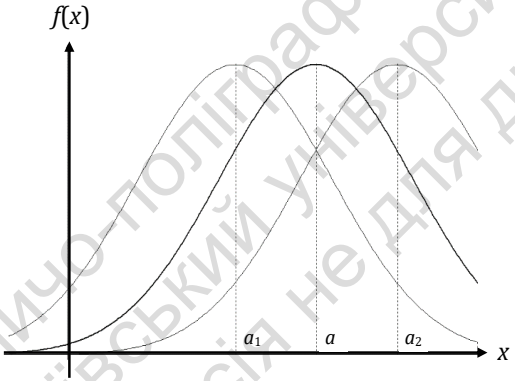


Рис. 4.11

На графіках видно, що параметр *a* виражає положення точки максимуму (піку) щільності розподілу Гаусса, а σ – гостроту піку (чим менше σ, тим пік гостріший, рис. 4.12).

Використовуючи означення центральних моментів для нормального розподілу, можна отримати рекурентне співвідношення:

$$\mu_s = (s-1)\mu_{s-2}\sigma^2, \qquad (4.26)$$

яке дозволяє виражати центральні моменти більш високих порядків через центральні моменти нижчих порядків. Зокрема, оскільки $\mu_1 = 0$, то і $\mu_3 = 0$, звідки випливає, що *кое-*

фіцієнт асиметрії нормально розподіленої випадкової величини X дорівнює нулю, $A=0$. Отже, нормальний розподіл є симетричним щодо свого математичного сподівання.

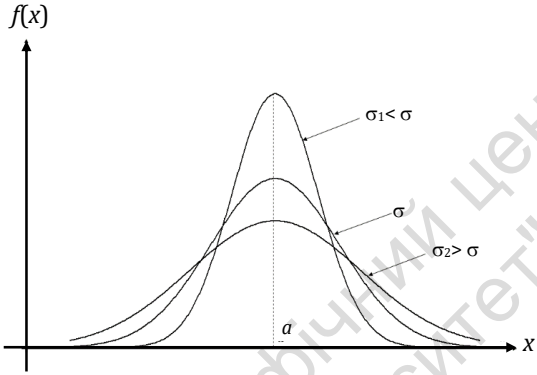


Рис. 4.12

Оскільки $D(X)=\mu_2=\sigma^2$, то, використовуючи рекурентну формулу (**4.26**), знайдемо центральний момент четвертого порядку: $\mu_4=3\sigma^2$, $\mu_2=3\sigma^4$, отже, *ексцес* $E=\dfrac{3\sigma^4}{\sigma^4}-3=0$.

Формула (**4.25**) визначає щільність розподілу випадкової величини $Z=\dfrac{X-a}{\sigma}$, де X – випадкова величина, яка має нормальний закон розподілу із параметрами a та σ. Функцію:

$$\Phi(x)=\int_0^x \varphi(t)\,dt \qquad (\mathbf{4.27})$$

називають *функцією Лапласа.*

Для простішого запису формули (**4.27**) використовують інший запис *функції Лапласа*:

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{z^2}{2}}\,dz. \qquad (\mathbf{4.28})$$

Властивості функції Лапласа:

1) функція $y=\Phi(x)$ визначена для всіх $x\in(-\infty;+\infty)$, $\Phi(x)>0$ при $x>0$ і $\Phi(x)<0$ при $x<0$;

2) $y=\Phi(x)$ – непарна функція, тому $\Phi(-x)=-\Phi(x)$;

3) $\Phi(0)=0$, $\Phi(-\infty)=\lim\limits_{x\to-\infty}\Phi(x)=-0{,}5$; $\Phi(+\infty)=\lim\limits_{x\to+\infty}\Phi(x)=0{,}5$.

Прямі $y=0{,}5$ та $y=-0{,}5$ є горизонтальними асимптотами графіка функції $y=\Phi(x)$ при $x\to+\infty$ та $x\to-\infty$, відповідно;

4) $y=\Phi(x)$ – зростаюча функція, оскільки $\Phi'(x)=\varphi(x)>0$ для всіх x;

5) на проміжку $(0;+\infty)$ графік функції опуклий угору, а на проміжку $(-\infty;0)$ – опуклий донизу, оскільки $\Phi''(x)<0$ при $x>0$ і $\Phi''(x)>0$ при $x<0$.

Функція Лапласа табульована (див. додаток В2). Значення функції $y=\Phi(x)$ у таблиці наведено для $0\le x\le 5$. Для обчислення значень $y=\Phi(x)$ за від'ємних значень із проміжку $-5\le x<0$ використовують непарність функції Лапласа $\Phi(-x)=-\Phi(x)$. Для $x>5$ приймають $\Phi(x)=0{,}5$, а при $x<-5$ приймають $\Phi(x)=-0{,}5$.

Графік функції Лапласа $y=\Phi(x)$ подано на рис. 4.13.

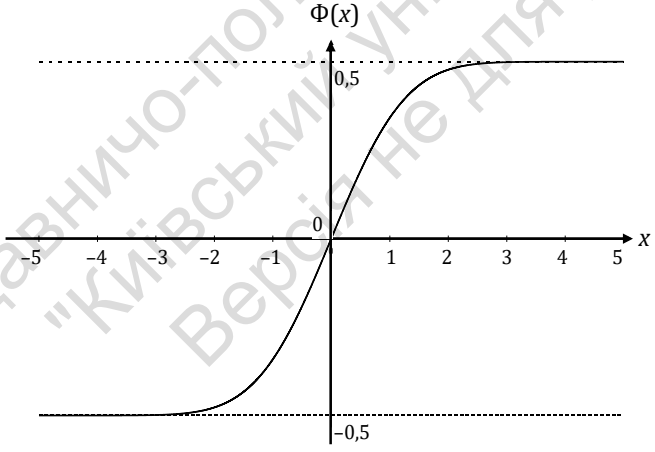


Рис. 4.13

Імовірність попадання нормально розподіленої величини X до проміжку $(\alpha;\beta)$ обчислюють за формулою:

$$P(\alpha<X<\beta)=\Phi\left(\frac{\beta-a}{\sigma}\right)-\Phi\left(\frac{\alpha-a}{\sigma}\right). \qquad \textbf{(4.29)}$$

Розглянемо частинні випадки (**4.29**):

$$P(X<\beta)=P(-\infty<x<\beta)=\Phi\left(\frac{\beta-a}{\sigma}\right)-\Phi(-\infty)=\Phi\left(\frac{\beta-a}{\sigma}\right)+\frac{1}{2},$$

$$P(X>\alpha)=P(\alpha<x<+\infty)=\Phi(+\infty)-\Phi\left(\frac{\alpha-a}{\sigma}\right)=\frac{1}{2}-\Phi\left(\frac{\alpha-a}{\sigma}\right),$$

тому

$$P(X<\beta)=\frac{1}{2}+\Phi\left(\frac{\beta-a}{\sigma}\right), \qquad (4.30)$$

$$P(X>\alpha)=\frac{1}{2}-\Phi\left(\frac{\alpha-a}{\sigma}\right), \qquad (4.31)$$

Імовірність потрапляння нормально розподіленої випадкової величини X на проміжок довжини 2δ, симетричний щодо центра розсіювання, обчислюють за формулою:

$$P(|X-a|<\delta)=2\Phi\left(\frac{\delta}{\sigma}\right). \qquad (4.32)$$

Правило 3σ – "*трьох сігм*". Якщо випадкова величина X має нормальний закон розподілу $X\sim N(a,\sigma^2)$, то ймовірність того, що X відхиляється від свого математичного сподівання – не більш ніж на 3σ, є достатньо близькою до одиниці та не залежить ні від величини математичного сподівання, ні від дисперсії випадкової величини X. Це означає, що подія $\{|X-a|\le 3\sigma\}$ – практично достовірна. Отже, вважатимемо, що можливі значення нормально розподіленої випадкової величини практично не виходять за межі інтервалу $(a-3\sigma;a+3\sigma)$.

На практиці це правило використовують так: якщо закон розподілу випадкової величини X невідомий, але $|X-a|<3\sigma$, то можна припустити, що випадкова величина X має нормальний розподіл.

Нормальний розподіл може бути *одновимірним* (лише для однієї змінної величини) та *багатовимірним* (додатково потрібно, що комбінації кількох змінних також були нормально розподілені). Остання умова є важливою для методів багатомірного статистичного аналізу. Частковою гарантією багатовимірного нормального розподілу багатьох змінних величин є умова, згідно з якою кожна змінна величина, яку

застосовують в аналізі, має одновимірний нормальний розподіл. Якщо існує відхилення від нормального розподілу (змінної чи залишків), то вона часто пов'язана з порушенням інших припущень (напр., гетероскедастичність), що необхідні для того чи іншого виду аналізу. Виправлення інших порушень може привести й до виправлення проблем із відхиленням від нормального розподілу.

Якщо розподіл занадто плоский (немає явного піку), то розв'язанням проблеми є використання в аналізі зворотної величини (напр., $1/Y$ або $1/X$). Якщо розподіл асиметричний, то розв'язанням може бути застосування коренів квадратних змінних або логарифмів чи зворотних величин. Корінь квадратний краще використовувати за від'ємної асиметрії, а логарифм – за додатної.

Зауважимо, що нормальний закон розподілу займає особливе місце в аналізі даних і практиці ймовірносно-статистичних методів – за його допомогою отримано багато важливих розподілів, які розглянемо далі.

4.4.2. Логнормальний розподіл

Неперервна випадкова величина X має *логнормальний* (*логарифмічно-нормальний*) *розподіл*, якщо її натуральний логарифм $\eta = \ln X$ має нормальному розподіл.

Щільність розподілу $f_\eta(x)$ визначають за формулою:

$$f_\eta(x) = \frac{1}{\sigma \cdot \sqrt{2\pi} \cdot x} \cdot e^{-\frac{(\ln x - \ln a)^2}{2\sigma^2}}. \qquad \textbf{(4.33)}$$

Основні числові характеристики логнормального розподілу:

- середнє $M\eta = a \cdot e^{\frac{1}{2}\sigma^2}$;
- дисперсія $D\eta = a^2 e^{\sigma^2}(e^{\sigma^2} - 1)$;
- мода $x_{\text{mod}} = ae^{-\sigma^2}$;
- медіана $x_{\text{med}} = a$;
- асиметрія $A_\eta = (e^{\sigma^2} - 1)^{\frac{1}{2}}(e^{\sigma^2} + 2)$;
- ексцес $E_\eta = (e^{\sigma^2} - 1)(e^{3\sigma^2} + 3e^{2\sigma^2} + 6e^{\sigma^2} + 6)$.

Очевидно, що чим менше значення параметра σ, тим ближче значення моди, медіани та математичного сподівання, а крива розподілу – ближче до симетрії (рис. 4.14.)

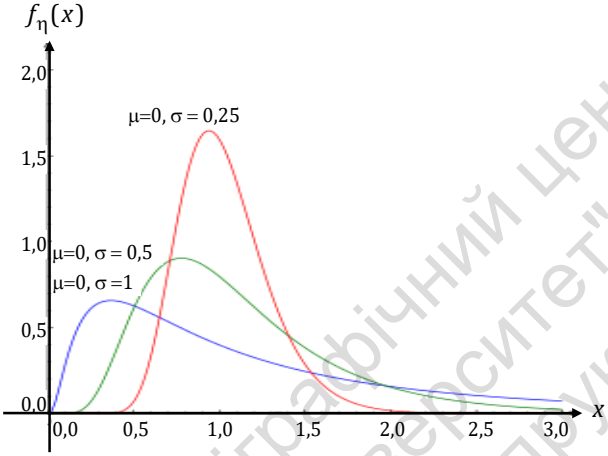


Рис. 4.14

Якщо параметр *a* для нормального розподілу є середнім значенням випадкової величини, то для логнормального – є медіаною. Логнормальний розподіл використовують для опису доходів, банківських внесків, щомісячної заробітної плати тощо.

4.4.3. Розподіли випадкових величин, що є функціями від нормальних величин

Є три види розподілів, які часто застосовують для використання статистичних алгоритмів: χ^2-розподіл, *t*-розподіл Стьюдента, *F*-розподіл Фішера. Опишемо їх властивості.

Розподіл χ^2 (Пірсона)

Сума квадратів n попарно незалежних випадкових величин $X_1, X_2, \ldots, X_n$, які мають нормальний розподіл $X_i \sim N(0,1), i = 1,2,...,n,$ тобто змінна випадкова величина:

$$\chi^2(n) = X_1^2 + X_2^2 + \cdots + X_n^2 = \sum_{k=1}^{n} X_k^2$$

має χ^2 (*хі-квадрат*) – *розподілом з n ступенями вільності.*

Щільність розподілу $f_{\chi^2}(x)$ визначають за формулою:

$$f_{\chi^2(n)}(x) = \begin{cases} \dfrac{x^{\frac{n}{2}-1} \cdot e^{-\frac{x}{2}}}{2^{\frac{n}{2}}\Gamma\left(\dfrac{n}{2}\right)}, & x \geq 0, \\ 0, & x < 0, \end{cases} \qquad \textbf{(4.34)}$$

де $\Gamma(k) = \int_0^{+\infty} u^{k-1} e^{-u} du$ – гама-функція Ейлера.

Основні числові характеристики χ^2-розподілу:

- середнє $M\left(\chi^2(n)\right) = n$;
- дисперсія $D\left(\chi^2(n)\right) = 2n$;
- мода $x_{\text{mod}} = n - 2$;
- асиметрія $A\left(\chi^2(n)\right) = \frac{2^{3/2}}{\sqrt{n}}$;
- ексцес $E\left(\chi^2(n)\right) = \frac{12}{n}$.

Графік щільності $\chi^2(n)$-розподілу для різних значень подано на рис. 4.15.

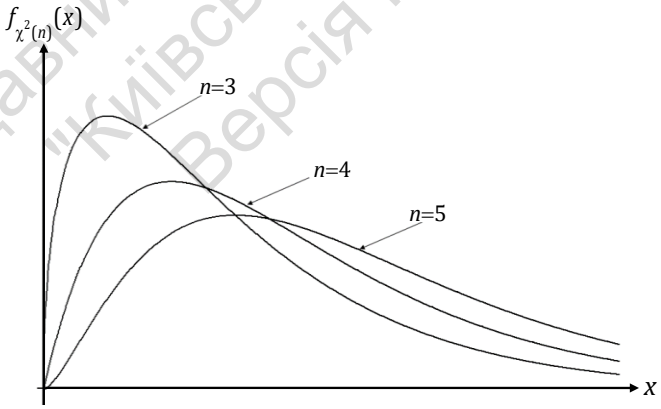


Рис. 4.15

Для розподілу χ^2 складено таблиці виду:

$$P(Y > \chi_\alpha^2) = \int_{\chi_\alpha^2}^{+\infty} f_{\chi^2}(x)dx$$

для кількості ступенів вільності від 1 до 30 (додаток В5). У таблицях для заданих значень імовірностей (здебільшого $\alpha = 0{,}99;\ 0{,}975;\ 0{,}95;\ 0{,}5;\ 0{,}25;\ 0{,}2;\ 0{,}1;\ 0{,}05;\ 0{,}02;\ 0{,}001$) указано значення χ_α^2 для відповідної кількості ступенів вільності. Якщо кількість ступенів вільності $m \geq 30$, то розподіл мало відрізняється від нормального з відповідними математичним сподіванням та дисперсією.

Розподіл Стьюдента (*t*-розподіл)

Розподіл Стьюдента з n ступенями вільності має випадкова величина:

$$t = t(n) = \frac{X_0}{\sqrt{\frac{1}{n}\sum_{k=1}^{n} X_k^2}}, \qquad \textbf{(4.35)}$$

де $X_0, X_1, X_2, \ldots, X_n$ – *n*-попарно незалежних випадкових величин, які мають нормальний розподіл $X_i \sim N(0,1), i = 1,2,...,n.$ Цей розподіл лежить в основі *t*-критерія, який використовують для порівняння середніх двох сукупностей.

Щільність розподілу Стьюдента $f_t(x)$ визначають за формулою:

$$f_t(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}. \qquad \textbf{(4.36)}$$

За достатньо великої кількості ступенів вільності ($n > 30$) *t*-розподіл практично збігається зі стандартним нормальним розподілом.

Основні числові характеристики *t*-розподілу:

- середнє, мода та медіана: $M\big(t(n)\big) = x_{\text{mod}} = x_{\text{med}} = 0;$
- дисперсія $D\big(t(n)\big) = \frac{n}{n-2}$, $(n > 2);$
- асиметрія $A\big(t(n)\big) = 0;$
- ексцес $E\big(t(n)\big) = \frac{6}{n-4}, (n > 4).$

Графік щільності *t*-розподілу Стьюдента за зовнішнім виглядом нагадує графік щільності стандартного нормального розподілу. Проте він значно повільніше прямує до нуля при $|x|\to\infty$, особливо за малих значень n (рис. 4.16). Графік щільності *t*-розподілу деформується при збільшенні числа ступенів вільності таким чином: пік збільшується, хвости крутіше прямують до нуля, і здається, наче графік щільності *t*-розподілу стискається з боків.

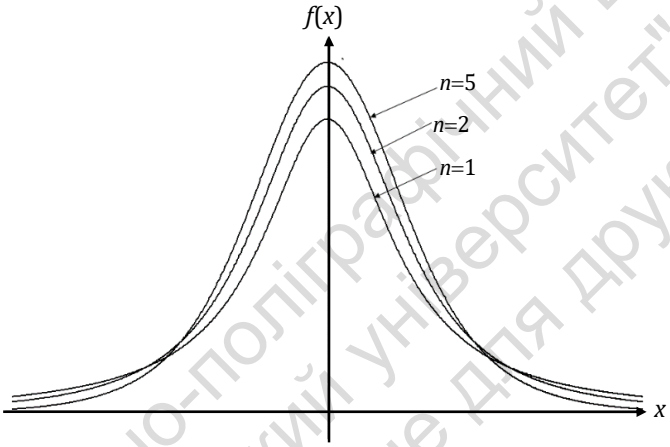


Рис. 4.16

Для розподілу Стьюдента складено таблиці для кількості ступенів вільності від 1 до 20. Якщо кількість ступенів вільності більша, то можна застосовувати нормальний закон розподілу з нульовим математичним сподіванням та одиничною дисперсією (додаток В6).

Розподіл Фішера

Нехай незалежні випадкові величини X_1 та X_2 мають χ^2-розподіл з n_1 та n_2 ступенями вільності, відповідно. Тоді випадкова величина:

$$F(n_1,n_2)=\frac{\frac{1}{n_1}\chi^2(n_1)}{\frac{1}{n_2}\chi^2(n_2)}=\frac{n_2}{n_1}\cdot\frac{X_1}{X_2} \qquad (\mathbf{4.37})$$

має *розподіл Фішера із параметрами* n_1 та n_2.

Щільність розподілу Фішера знаходять за формулою:

$$f_{F(n_1,n_2)}(x)=\begin{cases}\dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)\cdot n_1^{\frac{n_1}{2}}\cdot n_2^{\frac{n_2}{2}}}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)}\cdot\dfrac{x^{\frac{n_1}{2}-1}}{(n_2+n_1x)^{\frac{n_1+n_2}{2}}}, & x\geq 0,\\ 0, & x<0.\end{cases} \quad \textbf{(4.38)}$$

Основні числові характеристики *F*-розподілу:

- середнє $M\left(F(n_1,n_2)\right)=\dfrac{n_2}{n_2-2},(n_2>2);$
- мода $x_{\text{mod}}=\dfrac{(n_1-2)\cdot n_2}{n_1\cdot(n_2+2)},(n_1>2);$
- дисперсія $D\left(F(n_1,n_2)\right)=\dfrac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)},(n_1>1);$
- асиметрія $A\left(F(n_1,n_2)\right)=\dfrac{(2n_1+n_2-2)\cdot\sqrt{8(n_2-4)}}{(n_2-6)\cdot\sqrt{(n_1+n_2-2)n_1}},(n_2>6);$
- ексцес $E\left(F(n_1,n_2)\right)=\dfrac{3(n_2-6)\left(2+0{,}5A^2\left(F(n_1,n_2)\right)\right)}{n_2-8}-3,(n_2>8).$

Графік щільності *F*-розподілу Фішера подано на рис. 4.17.

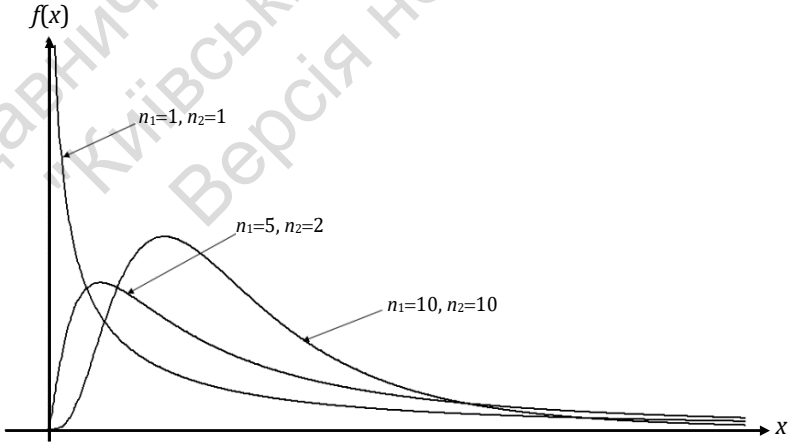


Рис. 4.17

Для розподілу Фішера складено таблиці, в яких для відповідної кількості ступенів вільності для ймовірностей $\alpha = 0{,}05$ та $\alpha = 0{,}01$ наведено значення f_α, яке є розв'язком рівняння $P(F > f_\alpha) = \alpha$ (додаток В7).

4.4.4. Відомі розподіли дискретних випадкових величин

Біноміальний розподіл

Випадкова величина *X*, яка набуває значень $k = 0,1,2,\ldots,n$, має *біноміальний розподіл із параметром* p $(0 < p < 1)$, якщо:

$$P(X = k) = C_n^p p^k (1-p)^{n-k},\ k = 0,1,\ldots,n. \qquad \textbf{(4.39)}$$

Основні числові характеристики біноміального розподілу:

- середнє $M(X) = np$;
- дисперсія $D(X) = npq$;
- середнє квадратичне відхилення $\sigma(X) = \sqrt{npq}$;
- асиметрія $A = \dfrac{q-p}{\sqrt{npq}}$;
- ексцес $E = \dfrac{1-6pq}{npq}$.

Розподіл Пуассона

Випадкова величина *X*, яка набуває значень $0,1,2,\ldots$, має *розподіл Пуассона із параметром* $\lambda > 0$, якщо:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda},\ k = 0,1,2,\ldots. \qquad \textbf{(4.40)}$$

Розподіл Пуассона відіграє важливу роль у теорії масового обслуговування. Його називають ще "розподілом рідкісних подій" (додаток В3).

Основні числові характеристики розподілу Пуассона:

- середнє та дисперсія $M(X) = D(X) = \lambda$;
- середнє квадратичне відхилення $\sigma(X) = \sqrt{\lambda}$;
- асиметрія $A = \dfrac{1}{\sqrt{\lambda}}$.

Геометричний розподіл

Випадкова величина *X*, яка набуває значень $1,2,3,\ldots$, має *геометричний розподіл із параметром* $p (0 < p < 1)$, якщо:

$$P(X = k) = p(1-p)^{k-1},\ k = 1,2,3,\ldots, \qquad \textbf{(4.41)}$$

Основні числові характеристики геометричного розподілу:

- середнє $M(X)=\frac{1}{p}$;
- дисперсія $D(X)=\frac{q}{p^2}$;
- асиметрія $A=\frac{2-p}{\sqrt{q}}$;
- ексцес $E=6+\frac{p^2}{q}$.

4.4.5. Відомі розподіли неперервних випадкових величин

Рівномірний розподіл на відрізку

Випадкова величина має *рівномірний розподіл на відрізку* $[a;b]$, якщо її щільність розподілу має вигляд:

$$f(x)=\begin{cases}\frac{1}{b-a}, & x\in[a;b],\\ 0, & x\notin[a;b].\end{cases} \quad (4.42)$$

Відповідна функція розподілу:

$$F(x)=\begin{cases}0, & x\le a,\\ \frac{x-a}{b-a}, & a<x\le b,\\ 1, & x>b.\end{cases} \quad (4.43)$$

Графік функції $y=f(x)$ подано на рис. 4.18, а функції $y=F(x)$ – на рис. 4.19.

Основні числові характеристики випадкової величини, що має рівномірний розподіл на відрізку $[a;b]$:

$$M(X)=\frac{a+b}{2},\ D(X)=\frac{(b-a)^2}{12},\ \sigma(X)=\frac{(b-a)\sqrt{3}}{6}.$$

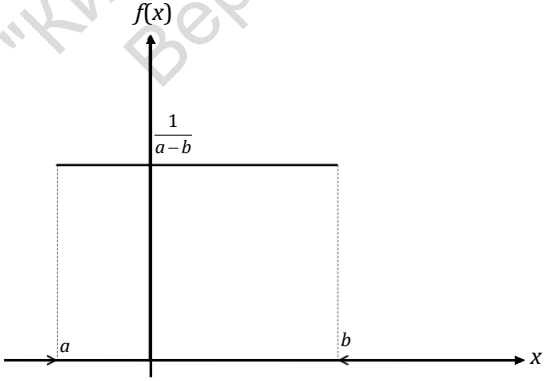


Рис. 4.18

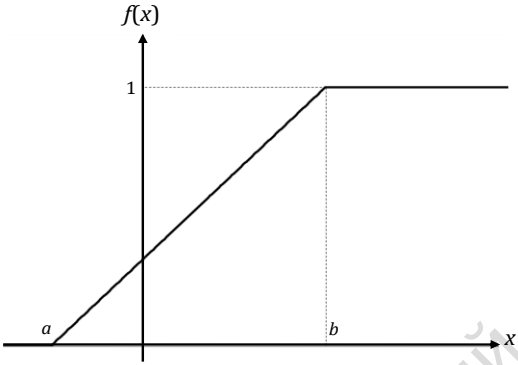


Рис. 4.19

Рівномірний розподіл симетричний щодо свого математичного сподівання, отже всі центральні моменти непарного порядку рівні нулю, тому коефіцієнт асиметрії також дорівнює нулю, тобто $A=0$. Ексцес рівномірно розподіленої величини $E=-1,2$.

Знайдемо ймовірність попадання значень рівномірно розподіленої на відрізку $[a;b]$ випадкової величини X на відрізок $[\alpha;\beta]\subset[a;b]$:

$$P(\alpha<x<\beta)=\int_{\alpha}^{\beta}f(x)dx=\int_{\alpha}^{\beta}\frac{dx}{b-a}=\frac{x}{b-a}\Bigg|_{\alpha}^{\beta}=\frac{\beta-\alpha}{b-a}. \qquad (4.44)$$

Отже, ця ймовірність дорівнює відношенню довжини відрізка $[\alpha;\beta]$ до довжини відрізки $[a;b]$. Числа a та b називають *параметрами* рівномірного розподілу та цілковито його визначають.

Експоненціальний розподіл

Випадкова величина X має *експоненціальний розподіл із параметром* $\lambda>0$, якщо щільність розподілу має вигляд:

$$f(x)=\begin{cases}\lambda e^{-\lambda x}, & x\geq 0,\\ 0, & x<0.\end{cases} \qquad (4.45)$$

Відповідна функція розподілу має вигляд:

$$F(x)=\begin{cases}1-e^{-\lambda x}, & x\geq 0,\\ 0, & x<0.\end{cases} \qquad (4.46)$$

Графік функції щільності $y=f(x)$ подано на рис. 4.20, а функції розподілу $y=F(x)$ – на рис. 4.21.

Числові характеристики випадкової величини X, яка має експоненціальний розподіл, є такими:

$$M(X)=\frac{1}{\lambda},\ D(X)=\frac{1}{\lambda^2},\ \sigma(X)=\frac{1}{\lambda},\ A=2,\ E=6.$$

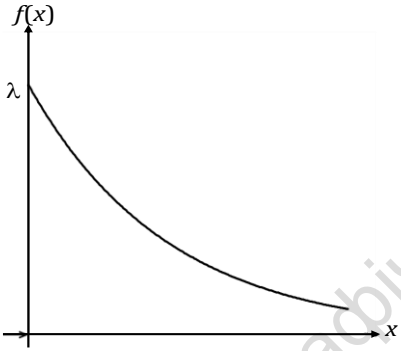


Рис. 4.20

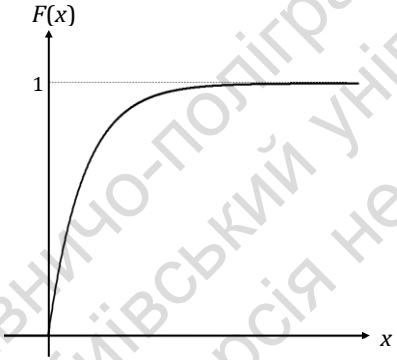


Рис. 4.21

Обчислимо ймовірність попадання значень випадкової величини X, яка має експоненціальний розподіл, до заданого проміжку $(\alpha;\beta)$:

$$P(\alpha<x<\beta)=\lambda\int_{\alpha}^{\beta}e^{-\lambda x}dx=-\int_{\alpha}^{\beta}e^{-\lambda x}d(\lambda x)=-e^{-\lambda x}\Big|_{\alpha}^{\beta}=e^{-\lambda\alpha}-e^{-\lambda\beta}.$$

Отже, $$P(\alpha<x<\beta)=e^{-\lambda\alpha}-e^{-\lambda\beta}. \qquad (\mathbf{4.47})$$

Значення функції $y=e^{-x}$ знаходять із таблиці (додаток В4).

Відомі закони розподілу випадкових змінних величин та їх числові характеристики подано в табл. 4.7.

Таблиця 4.7

Закон розподілу *X*	$M(x)$	$D(x)$	$\sigma(x)$	A	E
Дискретні розподіли випадкових величин					
Біноміальний $P(x=m)=C_n^m p^m q^{n-m}$, $q=1-p$, $m=0,1,2,...,n.$	np	npq	$\sqrt{npq}$	$\frac{q-p}{\sqrt{npq}}$	$\frac{1-6pq}{npq}$
Пуассона $P(X=m)=\frac{\lambda^m}{m!}e^{-\lambda}$, $\lambda>0$, $m=1,2,...$	λ	λ	$\sqrt{\lambda}$	$\frac{1}{\sqrt{\lambda}}$	$\frac{1}{\lambda}$
Геометричний $P(X=m)=pq^{m-1}$, $m=1,2,...$	$\frac{1}{p}$	$\frac{q}{p^2}$	$\frac{\sqrt{q}}{p}$	$\frac{2-p}{\sqrt{q}}$	$6+\frac{p^2}{q}$
Гіпергеометричний $P(X=m)=\frac{C_M^m \cdot C_{N-M}^{n-m}}{C_N^n}$, $m=0,1,...,\min(n,M)$	$\frac{n\cdot M}{N}$	$\frac{nM(N-M)(N-n)}{N^2(N-1)}$			

Закінчення табл. 4.7

Неперервні розподіли випадкових величин					
Рівномірний $f(x)=\begin{cases}\frac{1}{b-a}, & x\in[a;b]\\ 0, & x\notin[a;b]\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{(b-a)\sqrt{3}}{6}$	0	–1,2
Експоненціальний $f(x)=\begin{cases}\lambda e^{-\lambda x}, & x\geq 0\\ 0, & x<0\end{cases}$, $\lambda>0$.	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{1}{\lambda}$	2	6
Нормальний $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$	a	σ^2	σ	0	0

4.5. Визначення моделей розподілу емпіричних даних

Значна частина методів аналізу підходять для аналізу даних, які підлягають нормальному розподілу. Багато показників у міжнародній економіці підлягають нормальному розподілу або близькі до нього (напр., сальдо поточного рахунку щодо ВВП). Але не завжди. Наприклад, динаміка валютного курсу: частими є випадки, за яких національна валюта під час кризи суттєво знецінюється, але нечасто можна спостерігати суттєві подорожчання валют. До того ж, навіть якщо генеральна сукупність підлягає нормальному розподілу, невелика вибірка з неї може суттєво відхилятися від нормального розподілу. Якщо маємо лише інформацію про вибірку, то неможна із упевненістю говорити про розподіл генеральної сукупності. Нормальному розподілу також не підлягають змінні, які вимірюють у неметричній шкалі (номінальній чи порядковій).

За умов відхилення від нормального розподілу результати звичайних (параметричних) методів аналізу (напр., коефіцієнт кореляції Пірсона, дисперсійний аналіз) даватимуть зміщені результати. У такому випадку краще використовувати непараметричні аналоги цих методів. Для змінних, які вимірюють у неметричній шкалі, варто використовувати непараметричні методи.

Але у великих вибірках наявність нормального розподілу вже не є обов'язковою умовою для надання переваги звичайним параметричним методам. До того ж, якщо є впевненість про наявність нормального розподілу, то кращі результати дають звичайні параметричні методи, більш чутливі. Також апріорі за можливості варто використовувати як параметричні, так і непараметричні методи та враховувати результати обох методів.

Про відхилення від нормального розподілу може свідчити, наприклад, значна різниця між середньою арифметичною та медіаною. Про характер розподілу також свідчать такі показники, як асиметрія, ексцес, квартилі.

На практиці часто виникає проблема перевірки відповідності емпіричного розподілу деякому заданому теоретичному. При цьому вирізняють прості та складні гіпотези. Якщо гіпотеза стверджує, що із s параметрів розподілу k мають задані значення, то гіпотезу вважають простою, коли $k = s$, і складною – якщо $k < s$. Різницю $s - k$ називають *кількістю ступенів вільності гіпотези,* а k *– кількістю накладених обмежень.* Особливу роль відіграє перевірка гіпотези: буде розподіл нормальний, чи ні. Прийняття гіпотези про те, що це нормальний розподіл, дає змогу застосовувати більш досліджені параметричні критерії перевірки наступних гіпотез.

Для перевірки відповідності емпіричного розподілу теоретичному застосовують так звані ***критерії згоди***: ω^2, Смирнова, χ^2 тощо.

Критерій ω^2 (Крамера–Мізеса)[123] використовують у випадках, коли необхідно перевірити нульову гіпотезу про відповідність вибірки певному відомому закону розподілу. Розрахункове значення обчислюють за формулою:

$$n\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left(F(x_i) - \frac{2i-1}{2n}\right)^2, \qquad \textbf{(4.48)}$$

де $F(x)$ – теоретична функція розподілу, n – обсяг вибірки.

Критерій Смирнова застосовують у випадку, коли емпіричну функцію розподілу будують за масивом частот. У випадку побудови функції розподілу здійснюють безпосередньо за вихідною вибіркою (при цьому чисельність вихідної вибірки та масиву функції розподілу збігаються), для розрахунку критерію за двобічної гіпотези застосовують формули:

$$D_n = \max_{1\le m\le n}\left\{D_n^{(1)}; D_n^{(2)}\right\},$$

$$D_n^{(1)} = \max_{1\le m\le n}\left\{\frac{m}{n} - F(x_m)\right\}, D_n^{(2)} = \max_{1\le m\le n}\left\{F(x_m) - \frac{m-1}{n}\right\}, \qquad \textbf{(4.49)}$$

а за однобічної – $D_n^{(1)} = D_n$. Функція розподілу D_n є однією й тією самою для всіх неперервних розподілів, а функція розподілу величини $K = D_n\sqrt{n}$ за великих n збігається до *ста-*

[123] Запропонований у 1928-1930 рр. Крамером і фон Мізесом.

тистики Колмогорова–Смирнова, тому в літературі його часто називають **критерієм Колмогорова–Смирнова**. Критерій Смирнова за певних умов можна використовувати також для порівняння емпіричних функцій розподілу двох вибірок (перевірка гіпотези про їх однорідність).

Обмеженнями для застосування цього критерію є:

- вимога щодо неперервності теоретичної функції розподілу;
- необхідність достатньо представницьких вибірок ($n > 200$);
- необхідність незалежних, тобто отриманих не за самою вибіркою, оцінок параметрів розподілу (проста гіпотеза) для порівняння емпіричної й теоретичної функцій розподілу.

Якщо обсяги вибірок $n > 35$, то критичне значення статистики Колмогорова–Смирнова, що відповідає рівню значущості α, можна розрахувати за формулою:

$$K_\alpha \approx \sqrt{\frac{1}{2}\ln\frac{\alpha}{2}}. \qquad \textbf{(4.50)}$$

При застосуванні критерію Смирнова для перевірки нульової гіпотези про однорідність двох вибірок достатньо великого обсягу (n_1, $n_2 > 40$) можна використовувати такі критичні значення:

$$D_c = 1{,}36 \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}},\ \alpha = 0{,}05;$$
$$D_c = 1{,}22 \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}},\ \alpha = 0{,}10. \qquad \textbf{(4.51)}$$

Для складних гіпотез вводять модифіковані статистики Колмогорова, але їх функції розподілу є різними для різних типів неперервних розподілів. На відміну від більшості інших критеріїв, критерій Смірнова за достатнього обсягу досліджуваної вибірки дає змогу встановити різницю емпіричної й теоретичної функцій розподілу, незалежно від параметрів, що її зумовлюють. Зокрема він є чутливим до різниці як вибіркових середніх, так і стандартних відхилень вибірок, коефіцієнтів їх асиметрії та ексцесу. Але такої універсальності досягають за рахунок зменшення потужності критерію.

Критерій χ^2 як критерій згоди застосовують для порівняння емпіричної і теоретичної функцій розподілу. Він оперує не первинними даними, а їх розподілом за класами рівної ширини. Тому необхідно враховувати вимогу щодо мінімальних обсягів ряду спостережень і кількості класів. За різними оцінками, мінімальна допустима кількість класів перебуває у межах 4–7, а кількість елементів у ряді спостережень – у межах 20–200.

Значення критерію розраховують за формулою:

$$\chi^2 = \sum_{i=1}^{k} \frac{(\nu_i - np_i)}{np_i}, \qquad (4.52)$$

де ν_i – абсолютні частоти для k класів; p_i – теоретичні ймовірності обраного розподілу (параметри теоретичного розподілу розраховують за емпіричною вибіркою або задають); n – загальна кількість спостережень (для неперервного розподілу цю величину треба помножити на довжину класового інтервалу d). Кількість ступенів вільності беруть рівною $k–s–1$, де s – кількість параметрів теоретичного розподілу. Зокрема, при обчисленні параметрів теоретичного розподілу за інтервальним варіаційним рядом кількість ступенів вільності беруть рівною $k–2$ для біноміального та $k–3$ –для нормального розподілу.

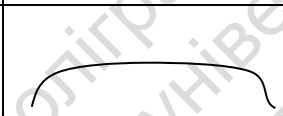
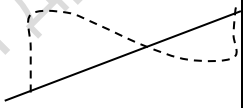
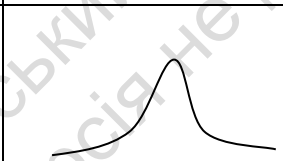
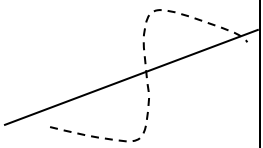
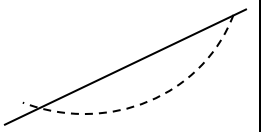
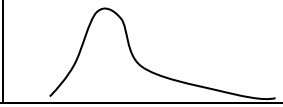
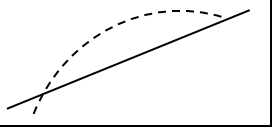
Загальна схема застосування критерію χ^2: спочатку будують емпіричну функцію розподілу; потім на основі її аналізу визначають s параметрів, які необхідні для побудови теоретичної функції розподілу; далі визначають межі класових інтервалів, абсолютні частоти ν_i й теоретичні ймовірності p_i. Після цього розраховують значення критерію χ^2 і порівнюють його з критичним.

4.6. Розподіл даних і методи графічного аналізу

Існують спеціальні статистичні тести, які вказують, з якою впевненістю можливо стверджувати, що розподіл змінної у вибірці відрізняється від нормального: *Тест Шапіро–Уілкса/Shapiro–Wilks test* або *Тест Колмогорова–Смирнова/Kolmogorov–Smirnov test*. Проте ці тести варто ви-

користовувати лише для відносно невеликих вибірок. Якщо вибірка велика, то спеціальні тести сигналізуватимуть про відхилення від нормального розподілу, навіть якщо воно несуттєве. Тому універсальним способом визначення того, чи буде розподіл змінної величини нормальним розподілом, є графічний: гістограма частот, а ще краще – спеціальна *діаграма нормального розподілу/normal probability plot*. В останній ідеальний нормальний розподіл представляє пряма. Якщо спостереження розташовуватимуться вздовж лінії, то розподіл є нормальним. У табл. 4.8 подано варіанти відхилень від нормального розподілу та відповідні графіки нормального розподілу зі спостереженнями.

Таблиця 4.8

Відхилення	**Гістограма**	**Графік нормального розподілу**
Низький пік, розподіл близький до рівномірного		
Гострий пік		
Асиметрія, пік зміщений праворуч		
Асиметрія, пік зміщений ліворуч		

Для графічного зображення варіаційних рядів використовують полігон частот і гістограму. Дискретний варіаційний ряд вибірки табл. 4.2 ілюструють полігоном частот. Для його побудови у прямокутній системі координат наносять точки з координатами (x_i, n_i), де x_i, $i = 1,2,\ldots, k$ – варіанта, а n_i – її частота, і сполучають їх послідовно відрізками прямих. Отриману ламану називають *полігоном частот вибірки*. Якщо побудувати точки (x_i, w_i), де $w_i = \frac{n_i}{n}$ – відносна частота варіанти x_i та сполучити їх аналогічним чином, то буде утворено ламану, яку називають *полігоном відносних частот*.

Для графічного представлення інтервальних варіаційних рядів будують діаграми, які називають гістограмами. Гістограма є найбільш популярним способом графічного розподілу числових даних. Розрізняють гістограми абсолютних і відносних частот.

Наприклад, спостерігаємо значення випадкової змінної ознаки *X* у *n* об'єктів. Згрупуємо їх до вибірки $x_0, x_1, \ldots, x_n$. Виберемо деякий інтервал, $[a;b]$, на якому розташовані всі спостережувані значення. Розіб'ємо цей інтервал на *k* інтервалів $I_i = [a_{i-1}, a_i)$, $i = 1,2,\ldots,k$ однакової ширини $h = \frac{b-a}{k}$, n_i – кількість спостережуваних значень, які попали на інтервал $I_i, i = 1,2,\ldots,k$. Величина n_i – *абсолютна частота/absolute frequency*, а $w_i = \frac{n_i}{n}$ – *відносна частота/relative frequency* інтервалу $I_i, i = 1,2,\ldots,k$. Отримаємо інтервальний варіаційний ряд (табл. 4.9)

Таблиця 4.9

$I_i = [a_{i-1}, a_i)$	$[x_0, x_1)$	$[x_1, x_2)$	$[x_2, x_3)$	…	$[x_{k-1}, x_k)$
n_i	n_1	n_2	n_3	…	n_k

де $x_i - x_{i-1} = h, i = 1,2,\ldots,k$.

Гістограму абсолютних частот будують таким чином: на горизонтальній осі відкладають інтервали I_i і над кожним інтервалом будуть стовпчик висотою $n_i, i = 1,2,\ldots,k$. Для побудови гістограми у прямокутній системі координат на осі

XOY наносять точки $x_0, x_1, \ldots, x_k$. Далі будують прямокутники, основою яких є інтервали $[x_{i-1}, x_i)$, а висота дорівнює $y_i = \frac{n_i}{h}$, де n_i – абсолютна частота *i*-го інтервалу. Побудовану фігуру, що складена із прямокутників, називають гістограмою абсолютних частот.

У гістограмі відносних частот висоту стовпчика визначають так:

$$f_i = \frac{w_i}{h} = \frac{n_i}{hn}.$$

Таким чином, гістограма відносних частот відрізняється від гістограми абсолютних частот лише масштабом за вертикаллю. Нормуючий множник $\frac{1}{hn}$ обирають так, щоб гістограму можна було використовувати як оцінку щільності розподілу вибірки.

Гістограма відносних частот – це східчаста фігура, що утворена із прямокутників з основами $[x_{i-1}, x_i)$, висоти яких $h_i = \frac{w_i}{h}$, $i = 1, \ldots, k$. Для побудови гістограми відносних частот на осі абсцис потрібно відкласти частинні інтервали $[x_{i-1}, x_i)$, $i = 1, 2, \ldots, k$ і на них, як на основах, побудувати прямокутники з висотами h_i. Площа кожного прямокутника дорівнює w_i.

Гістограму відносних частот можна розглядати як графік функції $y = f(x)$, яку ще називають гістограмою – оцінкою щільності розподілу. У той самий час має певні переваги й гістограма абсолютних частот: за висотою її стовпчиків можна побачити, скільки спостережень потрапило до того чи іншого інтервалу.

4.7. Визначення типу розподілу даних у Microsoft Excel

В Microsoft Excel аналіз розподілу можна здійснювати за допомогою опції *Гістограма/Histogram* надбудови *Data Analysis*. Таблиця гістограми складається з меж інтервалів

значень показника і частот значень у межах кожного інтервалу. Використаємо це до нашого прикладу з поточним рахунком платіжного балансу (у % ВВП, рис. 4.22). Додатково потрібно прописати межі бажаних інтервалів для поля *Інтервал карманів/Bin range* у стовпчику D та додаткову опцію – у діалоговому вікні (рис. 4.23): *Виведення графіка/Chart output*. Результати вказано на рис. 4.24.

	A	B	C	D
1	Current account balance (% of GDP)		2010	Інтервали
2		Slovak Republic	-3.38	-10
3		Slovenia	-0.81	-7
4		Sweden	6.28	-4
5		Thailand	4.63	-1
6		Tajikistan	-6.79	2
7		Turkey	-6.49	5
8		Tanzania	-8.58	
9		Uganda	-10.23	
10		Ukraine	-2.09	
11		Uruguay	-0.40	
12		United States	-3.23	
13		Venezuela, RB	3.71	
14		Vietnam	-4.14	
15		South Africa	-2.78	
16		Zambia	3.80	

Рис. 4.22

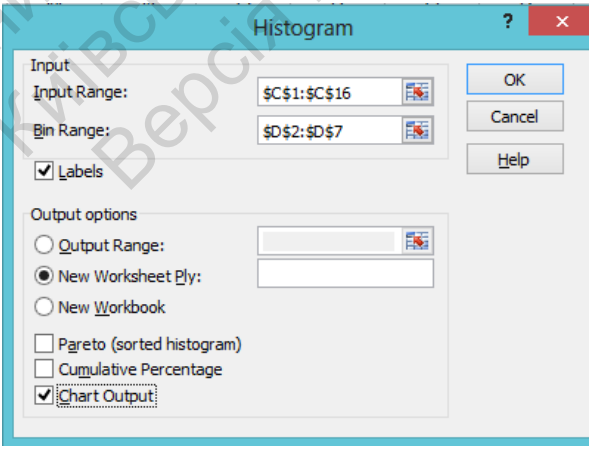
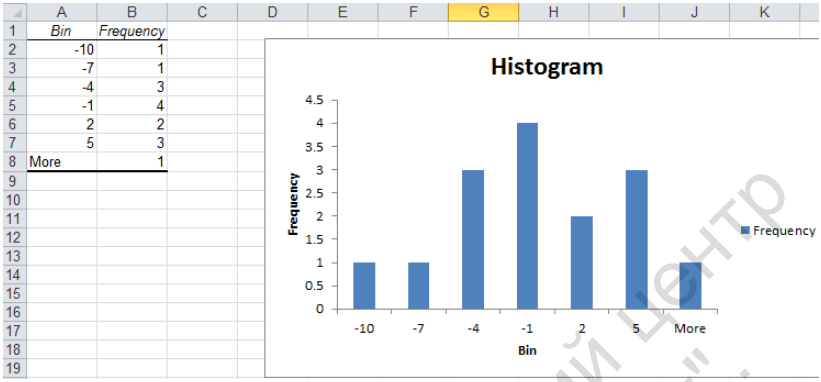


Рис. 4.23



	A	B	C	D	E	F	G	H	I	J	K
1	*Bin*	*Frequency*									
2	-10	1									
3	-7	1									
4	-4	3									
5	-1	4									
6	2	2									
7	5	3									
8	More	1									
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											

Рис. 4.24

Існують також окремі функції, що пов'язані з характером розподілу даних. Наприклад,

=STANDARDIZE (число;середня;стандартне відхилення)

повертає нормалізоване (стандартизоване) значення для розподілу з відомими середньою та стандартним відхиленням.

=SKEW(діапазон)

повертає асиметрію розподілу щодо середньої.

=KURT (діапазон)

повертає ексцес розподілу (міра гостроти піку).

4.8. Генерація випадкових чисел у Microsoft Excel

У Microsoft Excel у надбудові *Пакет аналіза/Data analysis* існує опція *Генерація випадкових чисел/Random number generation*. Вона дозволяє заповнити діапазон випадковими числами з урахуванням потрібного розподілу значень. У прикладі (рис. 4.25) результатом у новому аркуші є стовпчик із 20 комірками, що міститимуть випадкові числа. Ці числа підпорядкува тимуться нормальному розподілу із середньою 50 і стандартним відхиленням 20. Тобто близько 95 % чисел перебуватимуть у межах від 10 до 90 (+/–2 стандартних відхилення).

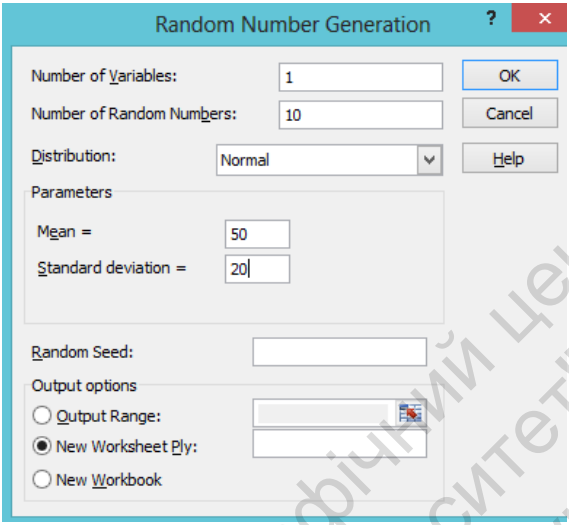


Рис. 4.25

Також існують спеціальні функції:

=RAND( )

повертає випадкове число у діапазоні від 0 до 1. Після кожного натиснення клавіші Enter для вводу функції значення, яке вона повертає, змінюється.

=RANDBETWEEN(нижня межа; верхня межа)

повертає випадкове ціле число до діапазону чисел із вказаними межами. Наприклад,

= RANDBETWEEN(0;1000)

повертає випадкове число між 0 та 1000. Після кожного натиснення клавіші Enter для вводу функції значення, яке вона повертає, змінюється.

4.9. Метод Монте-Карло

Генерація випадкових чисел може бути корисною для прогнозування в умовах ризику. Наприклад, до моделей можна багаторазово включати на вході різні значення екзогенних незалежних змінних (припускають, що вони мають відомий розподіл) і продивлятися весь спектр значень, які моделі дають на виході (визначаючи, напр., середнє очікува-

не значення, коефіцієнт варіації, 95 % довірчий інтервал для прогнозних значень тощо).

Наведемо ілюстративний приклад. Нехай умови задачі такі. Приріст індексу валютного курсу (у %) визначають за формулою:

$$EXR = 0,1 + 0,4GDP - 0,5PR - 30CC, \quad (4.53)$$

де *PR*– інфляція (у %); *GDP* – приріст ВВП (у %); *CC* – бінарна змінна валютна криза (1 – "криза є"; 0 – "кризи немає").

Прогнозований приріст ВВП наступного року перебуває на рівні +3 % (середня або математичне сподівання), приріст ВВП – нормальний розподіл, 95 % довірчий інтервал для прогнозного значення становить (–3 %; 9 %). Тобто стандартне відхилення приймають як 3 %, оскільки розподіл нормальний (ділять різницю між верхньою межею 95 % довірчого інтервалу та середньою приблизно навпіл).

Інфляцію прогнозують на рівні 10 %, і вона має нормальний розподіл. 95 % довірчий інтервал прогнозного значення становить (0 %; 20 %). Тобто стандартне відхилення становить 5 %.

У Microsoft Excel за допомогою опції *Генерація випадкових чисел*/*Random number generation* подібно тому, як і раніше, створюють випадкові значення для двох змінних: приріст ВВП та інфляція в стовпчиках A та B.

Оцінена ймовірність валютної кризи становить 0,1 (тобто із імовірністю 0,9 вона не відбудеться). Далі записують імовірності в окремому місці таблиці та створюють випадкові значення змінної "валютна криза", але обирають дискретний розподіл (рис. 4.26–4.27).

Зверніть увагу на те, що вказано в полі *Вхідний інтервал значень та ймовірностей*/*Value and probability input range*. Замість виводу результатів у новий робочий лист його виводять одразу у потрібний інтервал.

F	G
Валютна криза	
1	0.1
0	0.9

Рис. 4.26

Четверту змінну (валютний курс) розраховують за допомогою формул. У результаті одержують таблицю (рис. 4.28, стовпчик D у двох видах із результатами та формулами).

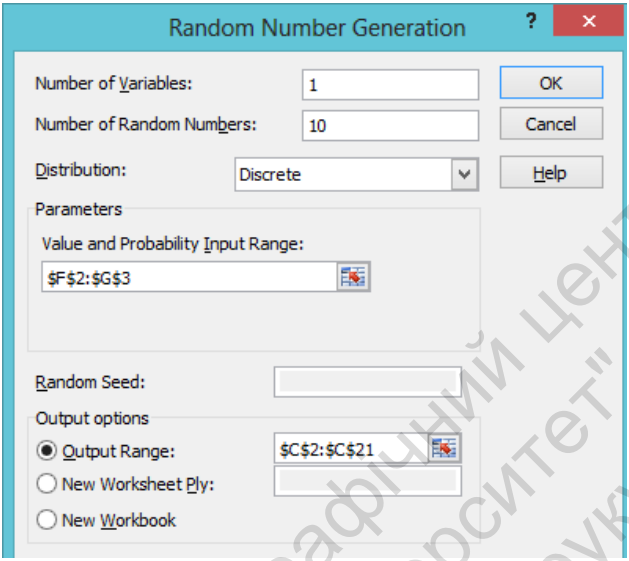


Рис. 4.27

	A	B	C	D	D
1	Приріст ВВП	Інфляція	Валютна криза	Приріст валютного курсу	Приріст валютного курсу
2	5.51	2.99	0	0.81	=0.1+0.4*A2-0.5*B2-30*C2
3	2.13	-5.08	0	3.49	=0.1+0.4*A3-0.5*B3-30*C3
4	6.62	12.06	0	-3.28	=0.1+0.4*A4-0.5*B4-30*C4
5	5.78	11.95	0	-3.56	=0.1+0.4*A5-0.5*B5-30*C5
6	3.24	10.96	0	-4.08	=0.1+0.4*A6-0.5*B6-30*C6
7	6.28	5.63	0	-0.20	=0.1+0.4*A7-0.5*B7-30*C7
8	7.74	8.98	0	-1.29	=0.1+0.4*A8-0.5*B8-30*C8
9	4.09	7.44	0	-1.98	=0.1+0.4*A9-0.5*B9-30*C9
10	5.21	3.11	0	0.63	=0.1+0.4*A10-0.5*B10-30*C10
11	-0.84	10.22	1	-35.34	=0.1+0.4*A11-0.5*B11-30*C11
12	1.52	5.01	0	-1.80	=0.1+0.4*A12-0.5*B12-30*C12
13	3.16	4.25	0	-0.76	=0.1+0.4*A13-0.5*B13-30*C13
14	2.93	16.51	0	-6.98	=0.1+0.4*A14-0.5*B14-30*C14
15	7.07	21.35	0	-7.75	=0.1+0.4*A15-0.5*B15-30*C15
16	5.28	11.71	0	-3.65	=0.1+0.4*A16-0.5*B16-30*C16
17	7.49	-0.59	1	-26.61	=0.1+0.4*A17-0.5*B17-30*C17
18	1.51	8.01	0	-3.30	=0.1+0.4*A18-0.5*B18-30*C18
19	6.57	4.03	0	0.72	=0.1+0.4*A19-0.5*B19-30*C19
20	5.74	9.63	0	-2.42	=0.1+0.4*A20-0.5*B20-30*C20
21	0.78	3.85	0	-1.52	=0.1+0.4*A21-0.5*B21-30*C21

Рис. 4.28

Отже, середнє приросту валютного курсу становить майже –5,6 %, тобто курс зменшиться на 5,6 % (рис. 4.29). Але медіана відрізняється та показує, що курс зменшиться лише на 3,7 %.

Оскільки медіана та середня відрізняються, то можна побачити, що розподіл прогнозного приросту валютного курсу відрізняється від нормального. Максимальне зниження курсу становить 37 %, максимальне зростання 5,5 %. Із імовірністю 70 % динаміка курсу перебуватиме в межах від –11,6 % до –0,04 %.

Це простий приклад. На практиці моделі часто мають складнішу структуру. Кількість випадкових значень, які потрібно створювати, також є набагато більшою – кілька сотень або навіть тисяч, залежно від складності моделей. Так роблять для охоплення максимально більшої кількості можливих комбінацій значень екзогенних змінних.

Exchange Rate growth	
Mean	-5.61501
Standard Error	1.848079
Median	-3.65688
Mode	#N/A
Standard Deviation	8.264862
Sample Variance	68.30794
Kurtosis	11.53014
Skewness	-3.00324
Range	42.36006
Minimum	-36.8638
Maximum	5.496232
Sum	-112.3
Count	20
Largest(2)	-0.04265
Smallest(2)	-11.6004
Confidence Level(95.0%)	3.868074

Рис. 4.29

Розділ 5
ДОСЛІДЖЕННЯ ВАЛЮТНИХ КРИЗ МЕТОДАМИ ЧАСТОТНОГО АНАЛІЗУ

5.1. Метод частотного аналізу впливу політичної стабільності та валютних резервів

Таблиці частот (їх ще називають одновхідними) є засобом аналізу зв'язку між категоріальними змінних. Їх можна успішно використовувати й для дослідження кількісних змінних, хоча можливі труднощі з інтерпретацією результатів. Таблиці показують, як часто виявляється певна ознака/подія у різних групах спостережень. Цей вид аналізу можна ще назвати аналізом часток або аналізом пропорцій.

Наприклад, потрібно дізнатися, яким чином урядові кризи впливають на валютні кризи. Урядова криза (бінарна змінна) буде незалежною змінною, вона ж є групуючою змінною у таблиці частот. Валютна криза (також бінарна змінна) – залежна змінна, частоти якої розраховують емпіричним шляхом за наявними спостереженнями (бажано брати із часовим лагом, напр., валютна криза протягом наступних двох років). Кожне спостереження – країно-рік. Припустимо, спостереження систематизовано у формі таблиці (табл. 5.1). Такий процес називають *крос-табуляцією або групуванням даних*.

Таблиця 5.1

	Наявність валютної кризи (1)	**Відсутність валютної кризи (0)**	**Сума**
Урядова криза наявна (1)	30	120	150
Урядова кризи відсутня (0)	50	600	650
Сума	80	720	800

Це найпростіший варіант крос-табуляції, за якого створюється матриця розміру 2 × 2 (по два значення кожної із двох змінних). Кожна комірка таблиці являє комбінацію значень двох змінних. У комірках містяться дані про частоту

у досліджуваній вибірці відповідної комбінації значень змінних. Частоти по краях таблиці – маргінальні частоти, які показують, як розподіляються спостереження при групуванні лише за однією змінною, без урахування іншої.

Варіанти таблиці частот у відсотковому вигляді (де показано вже не абсолютні частоти, а відносні) подано у табл. 5.2-5.3.

Таблиця 5.2

	Наявність валютної кризи (1)	**Відсутність валютної кризи (0)**	**Сума**
Урядова криза наявна (1)	3.75	15	18.75
Урядова криза відсутня (0)	6.25	75	81.25
Сума	10	90	100

Таблиця 5.3

	Наявність урядової кризи (1)	**Відсутність урядової кризи (0)**	**Середня**
Частка спостережень, за наявності валютної кризи (1), %	**20**=30/150	**7.7**=50/650	**10**=80/800
Кількість спостережень	150	650	

Частку спостережень за валютної кризи (за наявними спостереженнями) використовують як оцінку ймовірності її виникнення у майбутньому, якщо країна опиняється в аналогічних умовах. Ясно, що країна з урядовою кризою має більше шансів опинитися у стані валютної кризи, ніж країна без урядової кризи. Але наскільки надійний цей висновок?

Для перевірки існує χ^2-*тест* (*хі-квадрат критерій*), який показує, наскільки різниця у частотах статистично є значущою. Завдання тесту – перевірити, наскільки відрізняються фактичні частоти за комірками від очікуваних, якщо припустити, що зв'язку між досліджуваними змінними немає. Очікувані частоти для цього прикладу розраховують так, як у табл. 5.4 (фактичні частоти множать на частки маргінальних частот у всіх спостереженнях).

Отже, фактичні частоти 30, 120, 50, 600 відрізняються від очікуваних 15, 135, 65, 535.

Таблиця 5.4

	Наявність валютної кризи (1)	**Відсутність валютної кризи (0)**	**Сума**
1	$15 = 800 \cdot (150/800) \cdot (80/800)$	$135 = 800 \cdot (150/800) \cdot (720/800)$	150
0	$65 = 800 \cdot (650/800) \cdot (80/800)$	$535 = 800 \cdot (650/800) \cdot (720/800)$	650
Сума	80	720	800

Якщо тест показує рівень значущості *p* менше 0.05, то різницю вважають значущою. Результат буде тим більш надійним і значущим, чим більшою є розрахована різниця у частотах на основі вибірки та чим більшою є кількість спостережень у порівнюваних комірках таблиці частот (групах спостережень при групуванні за групуючою змінною). На практиці достатньо надійними вважають результати, коли у кожній порівнюваній групі спостережень (при групуванні за незалежною змінною) більше 50-80 спостережень. Якщо їх 30-50, то результати будуть відносно надійними. За меншої кількості варто бути уважними при інтерпретації результатів, оскільки часто χ^2-тест не може достатньо гарантувати надійність результатів.

Для з'ясування наявності зв'язку між змінними можна застосовувати й інші тести, зокрема:

- *поправка Йетса/Yates* – заниження різниці між очікуваними та фактичними частотами при проведенні χ^2-тесту. Вона є корисною, якщо частоти невеликі (менше 10);
- *критерій МакНемара/McNemar* – 2×2 використовують, якщо спостереження у вибірці є залежними, наприклад, одні й ті самі країни перед і після певної події (зокрема, кризи, глобальні шоки значного зростання світових цін на нафту);
- *коефіцієнт Фі/Phi* – 2×2 *tables* – набуває значень від 0 (немає зв'язку між змінними у таблиці частот) до 1 (повна залежність).

Незалежна змінна не обов'язково має бути категоріальною. У такому випадку її потрібно перетворити на категорі-

альну. Наприклад, потрібно перевірити, наскільки впливає на валютні кризи величина валютних резервів (у % ВВП). Це континуальний показник, який вимірюють у метричній шкалі. Необхідно поділити всі значення на два діапазони значень, наприклад, більше 15 % і менше 15 %. Таким чином створюють категоріальну змінну, яка набуває значень 1 (великі валютні резерви) або 0 (малі валютні резерви). Далі проводять аналіз подібно до аналізу впливу урядових криз.

При виборі порогового рівня, який ділитиме всі значення на два діапазоні, варто користуватися такими міркуваннями для того, щоб:

- кількість спостережень у двох групах не значно відрізнялася;
- це значення добре психологічно сприймалося, наприклад, 15 % сприйматиметься краще, ніж 14.7241 %, навіть якщо останнє поділить спостереження рівно навпіл;
- це значення ділило спостереження на групи, які дозволяють максимізувати різницю у частках досліджуваної ознаки у формі залежної змінної (це дозволяє сигнальний метод).

Можна провести групування й за більшою кількістю груп, наприклад, поділити валютні резерви на три діапазони значень (до 10 %; 10-20 %, більше 20 %). Таблиця частот може мати такий вигляд (табл. 5.5):

Таблиця 5.5

	Валютні резерви		
	малі	**середні**	**великі**
Частка спостережень за наявності валютної кризи (1), %	22	14	10
Кількість спостережень	370	298	314

5.2. Таблиці частот для аналізу взаємодії факторів: відсоткової ставки і валютних резервів

Складнішим варіантом є побудова таблиці частот для кількох змінних одночасно (напр., табл. 5.6) або в іншій формі представлення (табл. 5.7, де для наочності показано й розрахунок часток). За допомогою подібної таблиці можна побачити ефект взаємодії двох факторів (дані умовні). Наприклад, найменша ймовірність валютної кризи за низької (номінальної) відсоткової ставки та високих валютних резервів; найбільша – за високої відсоткової ставки та малих валютних резервів. При цьому для країн із низькою відсотковою ставкою валютні резерви більшою мірою впливають на ймовірність валютної кризи (15–5=10), ніж для країн із високою відсотковою ставкою (20–16,7=3,3). Іншими словами, нарощування валютних резервів часто не є ефективним рятувальним заходом, якщо існує проблема високих номінальних відсоткових ставок. З іншого боку, відсоткова ставка відіграє невелику роль, якщо валютних резервів недостатньо.

Таблицю можна ускладнити, якщо додати групу спостережень з відсутніми даними (табл. 5.8). Із цього прикладу видно, що, незважаючи на доволі велику частку спостережень з невідомою відсотковою ставкою, імовірності валютної кризи є близькими до середніх за спостереженнями з відомою відсотковою ставкою (21 до 18.2, а 8 до 8.9).

Щодо відсутніх даних за валютними резервами можливі два випадки:

1. Суттєва відмінність у групі низьких відсоткових ставок (15 і 9.4), але кількість відсутніх даних щодо відомих даних тут мізерна (20/520). Тому занепокоєння щодо дії ефекту спотворення відсутніх даних немає.

2. Відмінність у групі високих відсоткових ставок (24 і 18.3) є досить великою при тому, що кількість відсутніх даних є теж великою (150/520). Це викликає занепокоєння, оскільки розміщення відсутніх даних може залежати від частоти валютних криз.

Таблиця 5.6

	Валютні резерви						**Сума**
	малі (0)			**великі (1)**			
	наявність валютної кризи (1)	відсутність валютної кризи (0)	сума	наявність валютної кризи (1)	відсутність валютної кризи (0)	сума	
Низька відсоткова ставка (0)	30	170	200	15	285	300	500
Висока відсоткова ставка (1)	70	280	350	25	125	150	500
Сума	100	450	550	40	410	450	1000

Таблиця 5.7

	Валютні резерви		**Середня**
	малі (0)	**малі (0)**	
Низька відсоткова ставка (0)	30/200=**15** (**200**)	15/300=**5** (**300**)	(30+15)/500=9 (500)
Висока відсоткова ставка (1)	70/350=**20** (**350**)	25/150=**16.7** (**150**)	(70+25)/500=19 (500)
Середня	(30+70)/550=18.2 (550)	(15+25)/450=8.9 (450)	(100+40)/1000=14 (1000)

Таблиця 5.8

	Валютні резерви		**Середня**	**Відсутні дані щодо валютних резервів**
	малі (0)	**малі (0)**		
Низька відсоткова ставка (0)	15 (200)	5 (300)	**9.4** (**500**)	**15** (**20**)
Висока відсоткова ставка (1)	20 (350)	16.7 (150)	**18.3** (**500**)	**24** (**150**)
Середня	**18.2** (**550**)	**8.9** (**450**)	140 (1000)	
Відсутні дані щодо відсоткової ставки	**21** (**120**)	**8** (**130**)		

Примітка: у табл. 5.6-5.8 *частка спостережень з валютною кризою, у* % (*у дужках кількість спостережень*)

У результаті висновок про те, що великі валютні резерви лише трохи зменшують імовірність валютної кризи за великих відсоткових ставок, може бути ненадійним. Насправді за великих відсоткових ставок можливі різні варіанти:

- великі валютні резерви лише *трохи зменшують* імовірність валютної кризи (як і здається);
- великі валютні резерви *суттєво зменшують* імовірність валютної кризи (вплив недооцінено);
- великі валютні резерви *не впливають* на ймовірність валютної кризи (вплив переоцінено);
- великі валютні резерви *збільшують* імовірність валютної кризи (характер впливу визначено невірно, хоча цей випадок малоймовірний).

5.3. Частотний аналіз впливу зовнішнього боргу у Microsoft Excel

Для початку потрібно самостійно побудувати таблицю частот, після чого використати функцію

=ХИ2ТЕСТ(діапазон фактичних частот; діапазон очікуваних частот).

Ця функція повертає значущість різниці між очікуваними та фактичними частотами.

Наведемо приклад за умовними даними. Припустимо, потрібно дізнатися, як різні рівні зовнішнього боргу (у % ВВП) впливають на ймовірність валютної кризи. На одному аркуші будують таблиці з даними про фактичні частоти, очікувані частоти та частки спостережень із валютною кризою. На рис. 5.1 подано вхідні дані та результати, на рис. 5.2 – формули. Видно, що різниця є значущою (0.03816 менше 0.05), отже наявні докази на користь того, що величина зовнішнього боргу впливає на ймовірність валютної кризи. Якщо зовнішній борг невеликий (менше 50 % ВВП), то імовірність валютної кризи становить лише 10,7 %, за середнього боргу (50-80 %) – 21 %, за великого (більше 80 %) – 40 %.

	A	B	C	D	E	F
1	Фактичні частоти		зовнішній борг > 80%	50 - 80%	< 50%	
2		валютна криза є	10	8	3	21
3		немає	15	30	25	70
4			25	38	28	91
5						
6	Очікувані частоти					
7			зовнішній борг > 80%	50 - 80%	< 50%	
8		валютна криза є	5.77	8.77	6.46	0.23
9		немає	19.23	29.23	21.54	0.77
10			0.27	0.42	0.31	1.00
11						
12	Частки та кількість спостережень					
13			зовнішній борг > 80%	50 - 80%	< 50%	
14		валютна криза є	0.400	0.211	0.107	0.231
15		кількість	25	38	28	91
16						
17						
18	Значущість χ2 критерію	0.038163205				

Рис. 5.1

	A	B	C	D	E	F
1	Фактичні частоти		зовнішній борг > 80%	50 - 80%	< 50%	
2		валютна криза є	10	8	3	=SUM(C2:E2)
3		немає	15	30	25	=SUM(C3:E3)
4			=SUM(C2:C3)	=SUM(D2:D3)	=SUM(E2:E3)	=SUM(F2:F3)
5						
6	Очікувані частоти					
7			зовнішній борг > 80%	50 - 80%	< 50%	
8		валютна криза є	=F4*F8*C10	=F4*F8*D10	=F4*F8*E10	=F2/F4
9		немає	=F4*F9*C10	=F4*F9*D10	=F4*F9*E10	=F3/F4
10			=C4/F4	=D4/F4	=E4/F4	=F4/F4
11						
12	Частки та					
13	кількість спостережень		зовнішній борг > 80%	50 - 80%	< 50%	
14		валютна криза є	=C2/C4	=D2/D4	=E2/E4	=F2/F4
15		кількість	=C4	=D4	=E4	=F4
16						
17						
18	Значущість χ2 критерію	=CHITEST(C2:E3;C8:E9)				

Рис. 5.2

5.4. Алгоритм дослідження взаємодії факторів валютних криз методами частотного аналізу

Опишемо методологічний алгоритм дослідження взаємодії факторів валютних криз, що запропонований О.А. Чугаєвим[124]. Результати аналізу можна використати для уточнення характеру впливу факторів за різних обставин і визначення заходів з метою попередження валютних криз. Застосування результатів із прогностичною метою можливе, але має суттєві обмеження.

Як залежну змінну використовують логічну змінну наявності валютної кризи у відповідному році у певній країні. Якщо принаймні в один із місяців відповідного року в країні є валютна криза, то залежній змінній присвоюють значення 1. Якщо в жодному місяці цього року не було валютної кризи, то присвоюють значення 0. Наявність валютної кризи у відповідному місяці визначають за індексом I_{ch2}. Перевищення індексом I_{ch2} нуля свідчить про наявність валютної кризи. Цей індекс розраховують за формулою:

$$I_{ch2} = E + R + I - 3 + 0{,}5 \cdot D, \qquad (5.1)$$

де E – компонент зменшення валютного курсу:

$$E = \frac{e_t - \overline{e}}{0{,}06 \cdot \overline{e} + \sigma_e}; \qquad (5.2)$$

R – компонент валютних інтервенцій:

$$R = \frac{\overline{r} - r_t + L_t - \overline{L}}{0{,}12 \cdot \overline{r} + \sigma_r}; \qquad (5.3)$$

I – компонент відсоткової ставки:

$$I = \frac{m_t - \overline{m}}{0{,}03 \cdot (100 + \overline{m}) + \sigma_m}, \qquad (5.4)$$

або якщо немає даних про m, то:

$$I = \frac{c_t - \overline{c}}{0{,}017 \cdot (100 + \overline{c}) + \sigma_c}; \qquad (5.5)$$

[124] Див. : Чугаєв О.А. Валютні кризи на межі XX–XXI століть : моногр. – Київ : "МП Леся", 2007. – 416 с.
https://www.sites.google.com/site/achugaiev/stati/stati-1?authuser=0

D – компонент прискорення зменшення валютного курсу:

$$D = \frac{d_t - \bar{d}}{0{,}037 \cdot \bar{d} + \sigma_d}, \qquad (5.6)$$

де e_t – курс національної валюти до СДР, r_t – золотовалютні резерви (у СДР, включаючи золоті резерви за ринковими цінами), L_t – зовнішні зобов'язання центрального банку, m_t – відсоткова ставка грошового ринку, c_t – облікова ставка центрального банку, d_t – темпи зниження валютного курсу в t-й місяць[125]; $\bar{e}$, $\bar{r}$, $\bar{L}$, $\bar{m}$, $\bar{c}$, $\bar{d}$ – їх середні значення протягом попередніх 12 місяців; σ_e, σ_r, σ_m, σ_c, σ_d – їх стандартні відхилення (крім для зобов'язань монетарної влади) протягом попередніх 12 місяців.

Темпи зниження валютного курсу розраховують так:

$$\frac{e_t}{e_{t-1}} - 1.$$

Абсолютні значення порогових рівнів (0,06; 0,12; 0,03; 0,017; 0,037) розраховують для кожного із показників як середнє плинних коефіцієнтів варіації відповідного показника для всіх країн протягом всього дослідженого періоду (1989–2003 рр.).

Кожне спостереження являє країно-рік, тобто аналізуються панельні дані. За факторами використовують річні дані з лагом в 1 рік, тобто перевіряють чи впливає певне значення фактору на ймовірність валютної кризи наступного року.

Із розгляду варто виключати спостереження, щодо яких відсутні дані за залежною змінною або вже дорівнюють 1 в рік замірювання значення фактору. Таким чином, результати аналізу можна використовувати для прогнозування наявності валютної кризи наступного року тільки у випадку, якщо у поточному році її немає.

Наступним кроком є зменшення мультиколінеарності незалежних змінних. Для цього з аналізу вилучають частину незалежних змінних. Стандартний аналіз основних компо-

[125]Джерело даних для розрахунку – *International Financial Statistics* або інше джерело, напр., статистика центрального банку.

нент в умовах великої кількості змінних може ускладнити економічну інтерпретацію нових штучно побудованих змінних, що не корелюють між собою. Тому для зменшення мультиколінеарності обирають інший метод: виділяють кілька груп змінних, що сильно корелюють між собою (парний коефіцієнт кореляції Пірсона або Спірмана більше 65 %). Із кожної групи факторів, що сильно корелюють між собою, залишають лише одну (сурогатну змінну), яку у подальшому розглядають у кожній групі як фактор, що значною мірою відображає всі інші змінні у групі. Бажано, щоб сурогатна змінна мала:

- значну кількість спостережень із наявними даними щодо цієї змінної;
- більший коефіцієнт кореляції із залежною змінною наступного року;
- більший коефіцієнт кореляції квадрату відхилення змінної від середньої із залежною змінною наступного року (для врахування можливих нелінійних зв'язків);
- більшу кількість сильно корелюючи з нею змінних, які потім виключають з аналізу.

Альтернативним способом, замість використання сурогатної змінної, може бути побудова штучної змінної у формі лінійної комбінації з усіх змінних, що сильно корелюють між собою. Цей варіант дещо точніший, однак ускладнює інтерпретацію результатів.

Після зменшення кількості факторів будують нову таблицю, де з кожного фактора подають таку інформацію:

1. Назва фактора, сурогатна змінна та змінні, що сильно з нею корелюють. Назва фактора може відрізнятися від точної назви сурогатної змінної. Наприклад, фактор із назвою "зростання грошей та інфляція" базується на сурогатній змінній "зростання агрегату М2", яка сильно корелює з кількома показниками грошової маси, обсягів кредитування, відсоткових ставок, інфляції. Іншими словами, у більшості випадків інфляція пов'язана саме з пропозицією грошей.

2. Межі діапазонів значень у рамках зазвичай трьох груп спостережень (найменші значення – група спостережень I, середні значення – II група, найбільші значення – III група). Наприклад, для фактора "Зростання грошей та інфляція" (у %): –21,4 (7) 12 мінімальне; 12,01 (17) 22 середнє; 22.1 (34) 142 значне. У дужках – медіани з кожної групі значень. Бажано при цьому вказувати не тільки межі діапазонів значень сурогатної змінної (що розглядають як фактор), а й відповідні їм межі діапазонів значень сильно корелюючих з нею інших змінних (напр., інфляції за індексом споживчих цін).

Наступним кроком є проведення аналізу частот (часток, пропорцій). За допомогою кросс-табуляції отримують дані за кожною групою значень незалежних змінних: кількість спостережень і частка спостережень, коли залежна змінна наступного року має значення 1. Наприклад, у табл. 5.9 бачимо результати крос-табуляції за двома факторами: за стовпчиками – частка інвестицій до основного капіталу у ВВП r3Invfix (5.9 (15,6) 18 низька; 18.01 (21) 23 середня; 23.01 (27,4) 65.1 значна), за рядками – відсотковий спред між кредитною та депозитною ставками m2Spread (–42 (3) 5 мінімальний; 5.01 (7) 10 середній; 10.01 (13) 164 великий).

Спостереження за кожною змінною поділено на три діапазони/групи (I, II і III або 1, 2 і 3), належність до регіону світу – на п'ять груп. Цифра 6 – позначає групу спостережень з відсутніми даними за відповідною незалежною змінною. На перетині комірок показано кількість спостережень (зверху) і частку спостережень (знизу), що передують виникненню валютної кризи наступного року. Кростабуляцію можна представити й в іншій формі (табл. 5.10): зверху – частоти спостережень, що передують спокійним періодам, а знизу – частоти спостережень, що передують валютним кризам.

Спочатку розглядають вплив, який здійснює групуюча змінна за стовпчиками у рядку з усіма спостереженнями (далі – фактор впливу), у цьому прикладі – це частка інвестицій.

Таблиця 5.9

Рядки:	m2Spread	Стовпчики:	r3Invfix		
	1	2	3	6	Усі групи
1	69	46	115	22	252
	0.2319	0.0435	**0.0522**	0	0.0952
2	61	87	104	8	260
	0.2131	0.1839	0.1442	0	0.1692
3	92	68	50	9	219
	0.1739	0.1618	0.18	0.5556	0.1872
6	46	70	36	106	258
	0.2609	0.1429	0.1111	0.0755	0.1318
Усі групи	268	271	305	145	989
	0.2127	0.1439	**0.1115**	0.0897	0.1446

Таблиця 5.10

Рядки:	m2Spread	Стовпчики:	r3Invfix		
	1	2	3	6	Усі групи
1	53	44	109	22	228
	16	2	6	0	24
2	48	71	89	8	216
	13	16	15	0	44
3	76	57	41	4	178
	16	11	9	5	41
6	34	60	32	98	224
	12	10	4	8	34
Усі групи	211	232	271	132	846
	57	39	34	13	143

Якщо частки в I та III групах значуще відрізняються, то вплив відповідного фактора постійно позитивний або негативний. Якщо частка у II групі значуще відрізняється, то, імовірно, вплив фактора нелінійний та існують найбільш або найменш оптимальні значення незалежної змінної всередині діапазону значень у II групі.

У цьому прикладі як частка 0.21 у групі I, так і частка 0.11 у групі III значуще відрізняються (навіть з рівнем значущості менше 0.01). Отже, що більшою є частка інвестицій до ВВП, то більш захищеною є країна від валютних криз.

Для визначення значущості різниці між частками у групах використовують звичайний χ^2-тест, статистику якого розраховують за формулою:

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}, \quad (5.7)$$

де f_0 – фактичні частоти (за групування спостережень за фактором впливу та за залежною змінною) – кожна з них дорівнює частці спостережень із залежною змінною = 1 або 0 у відповідній групі за фактором впливу, що помножена на кількість спостережень у відповідній групі за фактором впливу; f_e – очікувані частоти, якщо припустити, що вплив фактора відсутній; кожна з них дорівнює частці спостережень із залежною змінною 1 або 0 у всіх групах за фактором впливу, що помножена на кількість спостережень у відповідній групі за фактором впливу.

Різницю вважають значущою (а вплив фактору наявним), якщо тест показує рівень значущості менше 0.05. Додаткові застереження роблять у випадку невеликої кількості спостережень у порівнюваних групах (менше 20).

Далі порівнюють частки у групах з наявними даними із урахуванням дії решти незалежних змінних, тобто після групування за рядками кожного разу за іншою незалежною змінною (фактором умов). У нашому прикладі фактором умови виступає відсотковий спред. Це дозволяє побачити, як залежить імовірність виникнення валютної кризи від розглянутого фактору впливу, залежно від того, якого значення набуває фактор умови.

Значущість відмінності характеру впливу в окремих групах за факторами умов від впливу без урахування фактора умов перевіряють за допомогою запропонованого модифікованого χ^2-тесту, статистику якого розраховують за (**5.7**), але в модифікованому χ^2-тесті f_0 – фактичні частоти (за групування спостережень за фактором впливу та за залежною змінною). Але при цьому для їх розрахунку частки спостережень із залежною змінною 1 у групах за фактором впливу замінюють на ці самі частки спостережень із залежною змінною 1 у групах за фактором впливу, що помножені на

частку спостережень із залежною змінною 1 в усіх спостереженнях і поділені на частку спостережень із залежною змінною 1 у відповідній групі за фактором умов; частку спостережень із залежною змінною 0 у кожній групі за фактором впливу обчислюють як одиниця мінус частка спостережень із залежною змінною 1 у відповідній групі за фактором впливу; f_e – очікувані частоти, якщо припустити, що вплив фактора не відрізняється; кожна з них дорівнює частці спостережень із залежною змінною 1 або 0 у відповідних групах за фактором впливу (але за всіма групами за фактором умов), що помноженій на кількість спостережень у відповідній групі за фактором впливу.

З формулами в Microsoft Office Excel це виглядатиме як на рис. 5.3. Так само відмінність характеру впливу фактора вважають значущою в специфічних умовах, якщо тест показує рівень значущості менше 0.05.

У нашому прикладі у табл. 5.9 в групі I за фактором умов (коли відсотковий спред мінімальний) вплив фактора впливу (частки інвестицій) значуще відрізняється від його впливу за інших рівних умов – цей вплив є сильнішим. У цій групі ймовірність валютної кризи ще зменшується зі зростанням частки інвестицій. Країни виграють більше, якщо поєднують високу частку інвестицій з низьким відсотковим спредом. В табл. 5.11 показано ще кілька прикладів модифікації впливу частки інвестицій у ВВП.

У табл. 5.11 використано позначки, які описують змінні та свідчать про рівень надійності результатів:

1) за стовпчиками – дані за групами спостережень за фактором впливу, якому присвячена таблиця;

2) за рядками – дані за всіма спостереженнями або групами спостережень за факторами умов, де вплив фактора впливу значуще відрізняється від впливу за всіма спостереженнями;

3) жирним шрифтом позначено частки, що значуще відрізняються від інших груп у рядку (за рівня значущості менше 0.05), якщо кількість спостережень не менше 20 (додаткові критерії використовують, якщо кількість спостережень становить 15-19);

	A	B	C	D	E	F	G	H	I	J	K
1	Rows:	m2Spread		Columns:	r3Invfix						
2											
3		1	2	3	6	All					
4											
5	1	69	46	115	22	252	230	69	46	115	22
6		0.232	0.044	0.052	0.000	0.095	0.104	0.342	0.064	0.077	0.000
7											
8	2	61	87	104	8	260	252	61	87	104	8
9		0.213	0.184	0.144	0.000	0.169	0.175	0.188	0.162	0.127	0.000
10											
11	3	92	68	50	9	219	210	92	68	50	9
12		0.174	0.162	0.180	0.556	0.187	0.171	0.156	0.145	0.162	0.499
13											
14	6	46	70	36	106	258	152	46	70	36	106
15		0.261	0.143	0.111	0.076	0.132	0.171	0.235	0.129	0.100	0.068
16											
17	All	268	271	305	145	989	844	268	271	305	145
18		0.213	0.144	0.112	0.090	0.145	0.154	0.213	0.144	0.112	0.090

	L	M	N	O	P	Q	R	S
1								
2								
3								
4								
5	TRUE	0.00479	23.63	2.95	8.86	14.68	6.62	12.82
6	TRUE		45.37	43.05	106.14	54.32	39.38	102.18
7								
8	FALSE	0.69776	11.47	14.12	13.23	12.97	12.52	11.60
9	FALSE		49.53	72.88	90.77	48.03	74.48	92.40
10								
11	FALSE	0.22034	14.38	9.89	8.09	19.57	9.79	5.58
12	FALSE		77.62	58.11	41.91	72.43	58.21	44.43

	G	H	I	J	K
1					
2					
3					
4					
5	=F5-E5	=B5	=C5	=D5	=E5
6	=(F6*F5-E6*E5)/(F5-E5)	=B6*$G$18/$G6	=C6*$G$18/$G6	=D6*$G$18/$G6	=E6*$G$18/$G6
7					
8	=F8-E8	=B8	=C8	=D8	=E8
9	=(F9*F8-E9*E8)/(F8-E8)	=B9*$G$18/$G9	=C9*$G$18/$G9	=D9*$G$18/$G9	=E9*$G$18/$G9
10					
11	=F11-E11	=B11	=C11	=D11	=E11
12	=(F12*F11-E12*E11)/(F11-E11)	=B12*$G$18/$G12	=C12*$G$18/$G12	=D12*$G$18/$G12	=E12*$G$18/$G12
13					
14	=F14-E14	=B14	=C14	=D14	=E14
15	=(F15*F14-E15*E14)/(F14-E14)	=B15*$G$18/$G15	=C15*$G$18/$G15	=D15*$G$18/$G15	=E15*$G$18/$G15
16					
17	=F17-E17	=B17	=C17	=D17	=E17
18	=(F18*F17-E18*E17)/(F17-E17)	=B18	=C18	=D18	=E18

	L	M	N	O	P	Q	R	S
1								
2								
3								
4								
5	=M5<0.05	=CHITEST(N5:P6;Q5:S6)	=H5*H6	=I5*I6	=J5*J6	=H5*B18	=I5*C18	=J5*D18
6	=M5<0.01		=H5-N5	=I5-O5	=J5-P5	=H5-Q5	=I5-R5	=J5-S5
7								
8	=M8<0.05	=CHITEST(N8:P9;Q8:S9)	=H8*H9	=I8*I9	=J8*J9	=H8*B18	=I8*C18	=J8*D18
9	=M8<0.01		=H8-N8	=I8-O8	=J8-P8	=H8-Q8	=I8-R8	=J8-S8
10								
11	=M11<0.05	=CHITEST(N11:P12;Q11:S12)	=H11*H12	=I11*I12	=J11*J12	=H11*B18	=I11*C18	=J11*D18
12	=M11<0.01		=H11-N11	=I11-O11	=J11-P11	=H11-Q11	=I11-R11	=J11-S11

Рис. 5.3

4) зірочкою * позначені частки, що особливо значуще відрізняються (за рівня значущості менше 0,01) від інших груп у рядку, якщо кількість спостережень у групах не менше 30;

5) двома зірочками ** позначено рядки, де вплив фактора впливу особливо значуще відрізняється (за рівня значущості менше 0.01) від впливу за усіма спостереженнями, якщо кількість спостережень у кожній із груп у рядку не менше 30;

6) знаком ^ позначено випадки, за яких надійність результатів під питанням через помітну кількість відсутніх даних;

7) знаком + позначено частки в групах спостережень, за умови, що кількість цих спостережень не менше 80;

8) знаком – позначено частки у групах спостережень за умови, що кількість цих спостережень від 10 до 40 (частки не показано за кількості спостережень менше 10).

Бачимо, що за від'ємних інших інвестицій до країни у банки щодо ВВП імовірність валютної кризи була більшою за середньої частки інвестицій. З одного боку, цей результат є статистично достатньо значущим; з іншого – складно знайти теоретичне підґрунтя, яке пояснювало б таку закономірність.

За низької частки депозитів до запитання у всіх депозитах (а отже за високої довіри до банківської системи), як і за значних золотовалютних резервів висока частка інвестицій вже не є оптимальним варіантом.

У табл. 5.12 аналогічно показано характер впливу зростання курсу національної валюти щодо СДР на ймовірність валютної кризи наступного року. Групи значень за цим фактором розподілені так: –45 % (–12 %) –7 %, тобто помірно значне здешевлення; –7 % (–3 %) 0 % – невелике здешевлення; 0, 01 % (3 %) 44 % – подорожчання.

Загалом найбільший ризик валютної кризи (18 %) спостерігається, коли вже відбувається відчутна девальвація (група 1), порівняно з 13 % ризиком за невеликого здешевлення або подорожчання. Але подорожчання національної валюти стає найгіршим варіантом за таких умов: зростання відсоткових ставок за кредитами (імовірне пояснення: разом із ревальвацією національної валюти подорожчання

кредитів завдає подвійної шкоди експортерам, які є основним джерелом надходження іноземної валюти), значні надходження від туризму щодо ВВП (зниження конкурентоспроможності експорту туристичних послуг) або швидке реальне зростання кінцевого споживання (економічний бум загрожує інфляційними тенденціями, що також підвищує вартість вітчизняних комплектуючих і зарплати для експортерів, що знижує їх конкурентоспроможність). А за високого прогнозованого значення валютних резервів оптимальним для забезпечення валютної стабільності є невелике здешевлення національної валюти, не більше, ніж на 7 %, що знижує ймовірність валютної кризи до 4 %. В останньому рядку для прикладу показано ймовірності для конкретних умов України в 2011 р. У той період помірна девальвація гривні також була би оптимальним варіантом.

5.5. Переваги та недоліки методу частотного аналізу

Ця методика аналізу має такі *переваги*:

- охоплює *значну кількість* потенційних факторів;
- ураховує *ефект взаємодії* факторів;
- дає *можливість з'ясувати нелінійні зв'язки*;
- надає можливість *обрати різні способи* вимірювання схожих факторів (напр., абсолютне значення показника, його зміна, приріст, трендові значення, відхилення від теоретичних або середніх значень, співвідношення з іншими показниками, поточні та кумулятивні показники, застосування даних із різних джерел).

У той самий час існують й елементи *недосконалості*:

- частина закономірностей не мають теоретичного пояснення;
- частина виявлених зв'язків є результатом дії випадковості, а не справжньої закономірності;
- межі діапазонів значень при проведенні крос-табуляції можуть бути обрані не оптимально;

Таблиця 5.11

Групи	Імовірність валютної кризи, %		
	I	II	III
Усі	**21***+	14+	**11***+
c6Othlbank/GDP III** (від'ємні інші інвестиції до країни в банки щодо ВВП)	**8**	20	**8**+
m2Spread I** (мінімальний відсотковий спред)	**23***	4	**5**+
b2DDep/BDep I** (депозити до запитання/усі депозити)	6	12+	12+
a1FRestr/GDP III (значні золотовалютні резерви)	8	10+	10+

Таблиця 5.12

Групи	Імовірність валютної кризи, %		
	I	II	III
Усі	**18**+	13+	13+
m2LRatech III** (зростання відсоткових ставок за кредитами)	15	14+	**28***+
t5Tour/GDP III** (значні надходження від туризму щодо ВВП)	7	**3**+	**12**+
r2FConsgr III** (швидке реальне зростання кінцевого споживання)	12	8+	**21**+
a1FRestr/GDP III** (значне трендове значення валютних резервів наступного року, тобто величина цього співвідношення плюс його зміна, порівняно із попереднім роком)	13	**4**+	13+
Середні з урахуванням умов України в 2011 р.	14.3	8.3	15.7

- кореляція між незалежними змінними не сильна, але існує;
- змінні з більшою кількістю відсутніх даних мають менше шансів дати значущі результати;
- в окремих випадках виявлені закономірності могли би бути іншими, як би була інформація про відсутні дані;
- тест про різницю між частками може дати зміщений результат за умови малої кількості спостережень і близької до нуля частки;
- іноді важко відрізнити, що є причиною, а що – наслідком: валютна криза чи певна величина незалежної змінної;
- можуть траплятися випадки, за яких заходи держави, що виражені в певних значеннях показника, насправді мають антикризовий вплив, але внаслідок їхньої великої політичної чи економічної вартості їх застосовують лише за умови безпосередньої загрози кризи. Тому проведений аналіз помилково характеризує ці заходи як такі, що сприяють виникненню валютної кризи;
- фактори можуть мати інший лаг впливу (напр., негативний вплив середньої, а не великої кількості валютних криз в інших країнах, що спричиняє ефект зараження кризою, можна пояснити тим, за великої кількості валютних криз у країнах регіону вони поширюються на слабкіші країни дуже швидко, і ці країни випадають з аналізу цього самого року як такі, що вже мають валютну кризу).

Розділ 6
ДОСЛІДЖЕННЯ ІНОЗЕМНИХ ІНВЕСТИЦІЙ МЕТОДАМИ АНАЛІЗУ СЕРЕДНІХ

6.1. Методи порівняння середніх

Аналіз даних починають з групування та обчислення описових статистик в групах, наприклад, середніх і стандартних відхилень. Якщо є дві групи даних, то природньо порівнювати середні в цих групах – середній дохід людей у двох групах: тих, які мають вищу освіту, і тих, які вищої освіти не мають. У практичних економічних дослідженнях часто трапляються випадки, за яких середній результат деякої ознаки однієї серії експериментів відрізняється від середнього результату іншої серії. Оскільки середнє – це результати вимірювань, то зазвичай вони завжди відрізняються. Питання тільки в тому, чи можна пояснити виявлену розбіжність середніх неминучими випадковими помилками експерименту або вона викликана певними причинами.

Аналіз середніх нагадує аналіз частот, але замість категоріальної залежної змінної використовують залежну змінну, яку вимірюють у метричній шкалі. Можна також застосувати аналіз середніх і для залежної змінної у порядковій шкалі. Незалежна змінна має бути категоріальною або штучно перетворена на категоріальну, подібно до того, як роблять у рамках частотного аналізу.

Розглянемо методи, що застосовують при порівнянні двох вибірок. За більшої кількості вибірок використовують методи дисперсійного аналізу.

Можливі два варіанти організації даних: з незалежними групами спостережень і залежними групами спостережень. Якщо випадковим чином вибірку розбито на дві частини, і відбувається порівняння показників у першій і другій групах, то, швидше за все, ідеться про незалежні групи.

Критерії й тести, що застосовують для порівняння вибірок, поділяють на дві групи: параметричні й непараметричні. Особливістю параметричних критеріїв є припущення

про те, що розподіл ознаки у генеральній сукупності підпорядкований певному відомому закону. Переважну більшість параметричних тестів розроблено для нормально розподілених даних. Але для деяких типів гіпотез існують параметричні тести, що призначені для вибірок, які підпорядковані іншим законам розподілу.

Зазвичай параметричні критерії є потужнішими за непараметричні. Застосування непараметричних критеріїв у випадках, за яких можна використовувати параметричні, призводить до збільшення ймовірності прийняття помилкової нульової гіпотези, тобто помилки другого роду.

Якщо йдеться про порівняння двох середніх, то зазичай використовують *z-тест* (*z-критерій*), *t-тест* (*t-критерій*) *Стьюдента*, які показують, наскільки значуще відрізняються середні залежної змінної у різних групах спостережень при групуванні за незалежною змінною. Досягнутий рівень значущості *p* *t*-критерію Стьюдента дорівнює ймовірності помилково відхилити гіпотезу про рівність середніх двох вибірок, якщо насправді ця гіпотеза вірна. Наприклад, розглянемо, як відрізняється середній приріст імпорту в країнах із фіксованим валютним курсом від середнього приросту імпорту у країнах з плаваючим валютним курсом (табл. 6.1).

Таблиця 6.1

Група	**Валютний курс**	
	фіксований (I група)	**плаваючий (II група)**
Середній приріст імпорту, %	11	7

При порівнянні середніх, як і при аналізі даних, дуже корисними є візуальні методи. Наприклад, у табл. 6.1 видно суттєву різницю між середніми приростами імпорту. Для візуального представлення різниці у середніх можна також використати *діаграми-короби*/*Box&Wisker plots* (рис. 6.1), на яких точки показують середні значення, межі прямокутники-короби – стандартні відхилення, а "вусики" – відрізки прямих ліній – стандартні помилки, що обчислені окремо (або два стандартні відхилення, залежно від опцій, що обра-

ні у програмному забезпеченні). Тут помітна різниця дисперсії у групах – висота прямокутника I групи *фіксований валютний курс* більше за висоту прямокутника II групи *плаваючий валютний курс.*

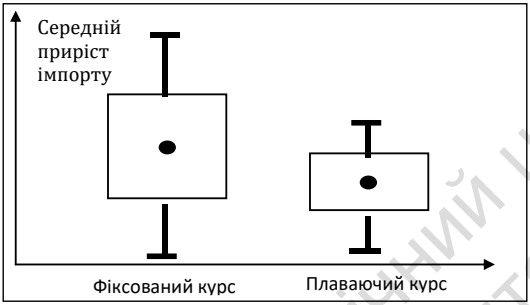


Рис. 6.1

Якщо різниця є значущою (зазвичай ідеться про досягнутий рівень значущості p менше 0,05), це означає, що режим валютного курсу впливає на приріст імпорту (хоча можливий і зворотній вплив приросту імпорту на вибір режиму валютного курсу). Іншими словами, досягнутий рівень значущості p різниці між середніми – це імовірність того, що дві вибірки (два діапазони, масиви значень) взято із генеральних сукупностей, які мають однакові середні. *t*-критерій показуватиме тим більш значущу різницю, чим:

- більша кількість спостережень у досліджуваній вибірці;
- більша різниця у середніх;
- менша дисперсія залежної змінної у кожній групі.

При проведенні аналізу середніх за допомогою звичайного *t*-критерію Стьюдента потрібно виконання *умов*:

1) достатня кількість спостережень у кожній групі (бажано більше 30, щоб результати були надійнішими);

2) залежна змінна повинна мати нормальний розподіл у кожній групі. Якщо ця умова не виконується, то кращим варіантом є використання непараметричних аналогів *t*-критерію Стьюдента. Припущення про нормальний розподіл можна перевірити, досліджуючи розподіл (напр., візуально за допомогою гістограми) або застосовуючи інші критерії;

3) дисперсія залежної змінної має бути приблизно однакова у різних групах. Рівність дисперсії у двох групах можна

перевірити за допомогою F-критерію або використати стійкіший *критерій Левена/Levene's test*. Якщо ця умова не виконується, то використовують спеціальний варіант t-критерію Стьюдента для груп із різною дисперсією;

4) вибірки (спостереження) мають бути незалежними. Якщо ця умова не виконується, то використовують спеціальний варіант t-критерію Стьюдента для залежних вибірок.

6.2. Критерії перевірки гіпотез про рівність середніх

Розглянемо один із варіантів перевірки гіпотези про рівність середніх за припущення, що вибірки незалежні, ознаки мають нормальний розподіл, порівнюють тільки дві сукупності.

Нехай є дві (генеральні) сукупності даних X та Y, які характеризують середні $\bar{x}_0$ та $\bar{y}_0$ і відомі дисперсії σ_x^2 і σ_y^2. Із цих сукупностей беруть дві незалежні вибірки обсягами n_1 і n_2, за якими знайдено вибіркові середні арифметичні $\bar{x}$ та $\bar{y}$, а також виправлені вибіркові дисперсії S_x^2 і S_y^2. Необхідно перевірити гіпотезу H_0 про рівність генеральних середніх $H_0 : \bar{x}_0 = \bar{y}_0$. Іншими словами, перевірити гіпотезу про те, що вибірки належать одній і тій самій генеральній сукупності.

Відомо, що за достатньо великих n_1 і n_2 середні арифметичні $\bar{x}$ та $\bar{y}$ мають наближено нормальний розподіл, тобто:

$$\bar{x} \approx N(\bar{x}_0; \sigma_x / n_1) \text{ і } \bar{y} \approx N(\bar{y}_0; \sigma_y / n_2).$$

У випадку справедливості гіпотези H_0 *Z-критерій*:

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}}}. \qquad \textbf{(6.1)}$$

***Z*-критерій** є випадковою величиною, що підпорядковується стандартному нормальному розподілу. Z-критерій можна застосовувати також для порівняння середніх значень довільно розподілених незалежних вибірок великого обсягу ($n_{1,2} \geq 30$), зважаючи на те, що в цьому разі вибіркові середні мають

приблизно нормальний розподіл, а вибіркові дисперсії є достатньо точними оцінками генеральних дисперсій.

Коли приймають конкуруючу гіпотезу $H_1 : \bar{x}_0 \neq \bar{y}_0$, то критичну область критерія будують таким чином:

- якщо виконується нерівність $|t| > t_{kp}$, то гіпотезу H_0 відхиляють;
- якщо виконується нерівність $|t| \leq t_{kp}$, то гіпотезу H_0 приймають.

При цьому t_{kp} визначають із рівняння:

$$\Phi(t_{kp}) = \Phi(t_{1-p}) = 1 - p,$$

де – рівень значущості критерія, а $y = \Phi(x)$ – функція Лапласа.

Якщо σ_x^2 та σ_y^2 – невідомі, але рівні, тобто $\sigma_x^2 = \sigma_y^2 = \sigma$, то використовують інший *t-критерій*:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{n_1 S_x^2 + n_2 S_y^2}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}, \qquad \textbf{(6.2)}$$

де S_x^2 і S_y^2 – вибіркові дисперсії (оцінки σ_x^2 і σ_y^2).

Відомо, що *t*-критерій (**6.2**) має розподіл Стьюдента з $k = n_1 + n_2 - 2$ ступенями вільності. При цьому:

- якщо виконується нерівність $|t| > t_{k,p}$, то гіпотезу H_0 відхиляють;
- якщо виконується нерівність $|t| \leq t_{k,p}$, то гіпотезу H_0 приймають.

Критерії (**6.1**) і (**6.2**) застосовують для перевірки гіпотези про рівність середніх, якщо обсяги обох вибірок малі та дисперсії рівні; у протилежному випадку застосовують іншу формулу.

За великих обсягів умова рівності дисперсій утрачає актуальність. Обсяги вибірок вважають малими, якщо $n_1 < 30$ і $n_2 < 30$. Якщо обсяги вибірок рівні, тобто $n_1 = n_2$, то формула для обчислення *t-критерія* значно спрощується та набуває вигляду:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S_x^2 + S_y^2}} \cdot \sqrt{\frac{n-1}{n}} \quad . \tag{6.3}$$

Розглянемо процедуру порівняння дисперсій у двох сукупностях *X* та *Y*, які мають нормальний розподіл із дисперсіями σ_x^2 та σ_y^2. Перевіряють гіпотезу про рівність дисперсій, тобто гіпотезу $H_0 : \sigma_x^2 = \sigma_y^2 = \sigma$. Для перевірки із сукупностей беруть дві незалежні вибірки обсягами n_1 і n_2. За вибірками розраховують виправлені вибіркові дисперсії S_x^2 і S_y^2.

Розглядають *F-критерій*, тобто величину:

$$F = \frac{S_x^2}{S_y^2} \tag{6.4}$$

(у чисельнику ставлять більшу дисперсію). Відомо, що величина *F* має розподіл Фішера з $k_1 = n_1 - 1$, $k_2 = n_2 - 1$ ступенями вільності. При цьому:

- якщо виконується нерівність $F > F_{p,k_1,k_2}$, то гіпотезу H_0 відхиляють;
- якщо виконується протилежна нерівність $F \le F_{p,k_1,k_2}$, то гіпотезу H_0 приймають.

6.3. Аналіз чинників припливу інвестицій методом аналізу середніх у Microsoft Excel

Для аналізу середніх у надбудові *Пакет анализу*/*Data analysis* обирають один із трьох можливих варіантів *t*-тесту:

- *Двовибірковий t-тест з однаковими дисперсіями*/*T-test*: *Two samples assuming equal variances*;
- *Двовибірковий t-тест із різними дисперсіями*/*T-test*: *Two samples assuming unequal variances* для незалежних вибірок;
- *Парний двовибірковий t-тест для середніх*/*T-test: Paired two samples for means* для залежних вибірок.

За парного тесту розміри вибірок обов'язково мають бути однаковими, за інших варіантів вибірки – можуть мати різний розмір.

Розглянемо приклад за умовними даними для незалежних вибірок.

Позначимо як *FDIlowincome* – ПІІ у країнах із низьким рівнем доходу (у % ВВП), *FDIhighincome* – у країнах із високим рівнем доходу. Фактично тут один рядок містить не одне спостереження. Кожна комірка – це окрема країна (окреме спостереження). Припустимо, наявні вхідні дані (як на рис. 6.2). Спочатку застосовують *F-тест для різниці дисперсій/F-Test Two-Sample for Variance*. Заповнюють діалогове вікно аналізу (рис. 6.3) та одержують результати (як на рис. 6.4).

C	D
FDIlowincome	FDIhighincome
3	5
5	9
2	4
3	2
2	3
5	11
1	6
1	3
4	4
4	5
3	2
4	2
0	1
1	1
2	2
1	4
2	3
2	2
1	0
0	1

Рис. 6.2

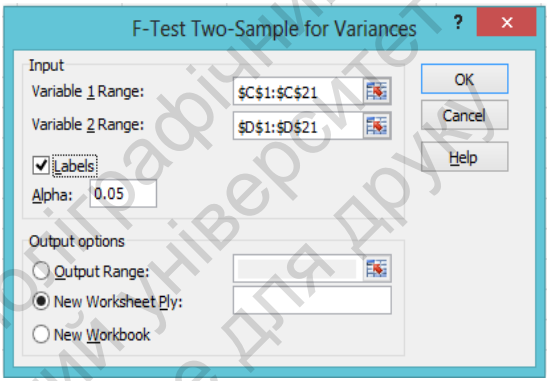


Рис. 6.3

F-Test Two-Sample for Variances

	FDIlowincome	*FDIhighincome*
Mean	2.3	3.5
Variance	2.326315789	7.421052632
Observations	20	20
df	19	19
F	0.313475177	
P(F<=f) one-tail	0.007529238	
F Critical one-tail	0.461201089	
P(F<=f) two-tail	0.015058475	

Рис. 6.4

Тому далі обирають *двовибірковий t-тест* із різними дисперсіями/*T-test: Two samples assuming unequal variances* і заповнюють діалогове вікно аналізу (рис. 6.5). Із результатів на рис. 6.6 видно, що рівень значущості для *t*-тесту із різною

дисперсією в групах становить 0,095989. Із результатів на рис. 6.7 видно, що рівень значущості для *t*-тесту із різною дисперсією в групах становить 0,095989. У країнах із високим доходом на душу населення показники ПІІ на 1,2 % ВВП більші, ніж у країнах із низьким доходом (точніше, на 1,2 пункти ВВП). Але, оскільки 0,5 <*p* <0,1, то йдеться про граничну значущість різниці у середніх. Тобто цілком імовірно, що у генеральній сукупності різниця у середніх існує (отже рівень доходів впливає на ПІІ). Але для такого ствердження все ж не достатньо доказів. Можливість збільшення розміру вибірки дасть шанс точніше з'ясувати наявність чи відсутність вказаного ефекту.

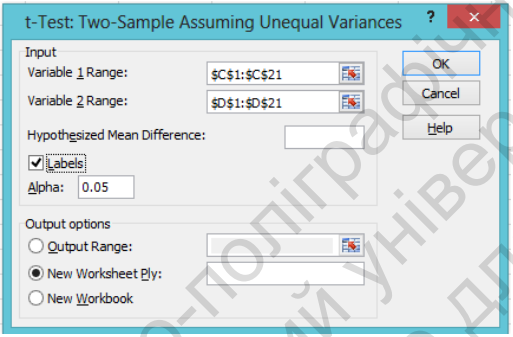


Рис. 6.5

t-Test: Two-Sample Assuming Unequal Variances

	FDIlowincome	*FDIhighincome*
Mean	2.3	3.5
Variance	2.326315789	7.421052632
Observations	20	20
Hypothesized Mean Differer	0	
df	30	
t Stat	-1.718907685	
P(T<=t) one-tail	0.047967861	
t Critical one-tail	1.697260887	
P(T<=t) two-tail	0.095935722	
t Critical two-tail	2.042272456	

Рис. 6.6

Для залежних спостережень використовують *Парний двовибірковий t-тест для середніх/T-test: Paired two samples for means* (рис. 6.7). Позначимо як *FDIbeforeLib* – ПІІ перед лібералізацією руху капіталу, *FDIafterLib* – ПІІ до тих самих країн після лібералізації руху капіталу.

Кожний рядок – це одна країна (діалогове вікно див. на рис. 6.8). Із таблиці результатів (рис. 6.9) видно, що різниця між середніми ПІІ за двома вибірками дуже значуща ($p = 0{,}001576 < 0{,}01$). Самі середні дорівнюють 3,75 і 5,25. Тобто лібералізація руху капіталу збільшує ПІІ на 1,5 % ВВП (точніше 1,5 пункти ВВП) або в 1,4 рази чи на 40 %. Тоді з'являється впевненість у наявності такого ефекту більш ніж на 99 %.

A	B
FDIbeforeLib	FDIafterLib
6	
5	
4	
2	
4	
11	
5	
5	
7	
8	
2	
0	
1	
1	
2	
4	
3	
2	
2	
1	

Рис. 6.7

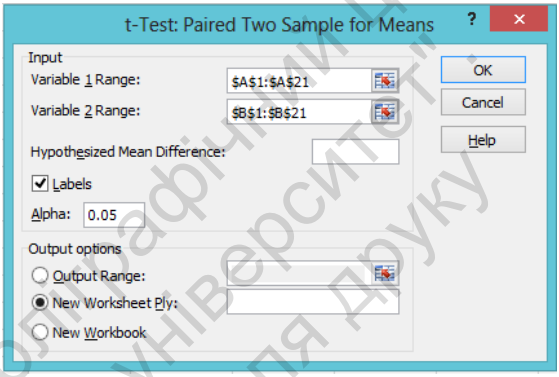


Рис. 6.8

t-Test: Paired Two Sample for Means		
	FDIbeforeLib	*FDIafterLib*
Mean	3.75	5.25
Variance	7.565789474	10.61842105
Observations	20	20
Pearson Correlation	0.829426216	
Hypothesized Mean Difference	0	
df	19	
t Stat	-3.683941988	
P(T<=t) one-tail	0.000788198	
t Critical one-tail	1.729132812	
P(T<=t) two-tail	0.001576397	
t Critical two-tail	2.093024054	

Рис. 6.9

Аналіз середніх також можна проводити без надбудови *Пакет анализа* за допомогою функцій:

=TTEST (діапазон1;діапазон2;хвости;тип)

повертає рівень значущості різниці між середніми. Аргументами діапазон1 і діапазон2 є два набори значень, середні за

якими порівнюють. Хвости: якщо 1 – функція повертає результати одностороннього *t*-тесту, якщо 2 – двостороннього. Тип може набувати значень: 1 – парний тест для залежних вибірок; 2 – тест для незалежними вибірок з однаковими дисперсіями; 3 – тест для незалежними вибірок з різними дисперсіями. Наприклад, =TTEST (H2:H23;I2:I21;2;2).

=FTEST (діапазон1;діапазон2)

повертає рівень значущості різниці між дисперсіями. Аргументами діапазон 1 і діапазон 2 є два набори значень, дисперсії за якими порівнюють.

6.4. Непараметричні критерії

Непараметричні критерії (тести)використовують, якщо розподіл залежної змінної відрізняється від нормального розподілу або невідомий. За цієї умови для аналізу даних некоректно застосовувати параметричні тести. Серед непараметричних тестів важливе місце займають так звані *робастні методи*, що виявляють слабку чутливість до відхилень від стандартних умов, які можна використовувати в широкому діапазоні реальних умов.

Для вказаних непараметричних методів залежні змінні потрібно вимірювати у порядковій чи метричній шкалі. Якщо їх вимірюють у номінальній шкалі, то використовують методи частотного аналізу.

Для незалежних вибірок можна застосовувати **критерій рандомізації компонент**[126]. При порівнянні спряжених вибірок основою методу є перебирання можливих результатів, що побудовані з різницевих оцінок. Нульова гіпотеза відповідає рівності вибіркових середніх. Нехай дані вибірки $x_i, y_i, i = 1,2,...,n,$ де n – кількість пар експериментальних значень. Значення різницевих оцінок визначають за формулою:

$$s_j = \sum_{i=1}^{n} a_{ji} |x_i - y_i|, \; j = 1,2,...,2^n, \qquad (6.5)$$

[126] Розроблено Фішером у 1920 р. для аналізу вибірок малого обсягу.

де $a_{ji}(=1,2,...,2^n; i=1,2,...,n)$ – елементи матриці можливих результатів, що розраховані, згідно з методикою побудови повного ортогонального плану експерименту. Вона є матрицею з 2^n рядків і n стовпців. При цьому i-й стовпець містить величини +1 та –1, що чергуються із кроком 2^{j-1}.

Сума масиву різницевих оцінок:

$$S = \sum_{j=1}^{2^n} S_j.$$

Кількість сприятливих результатів:

$$N = \sum_{j=1}^{2^n} n_j, \quad n_j = \begin{cases} 0, & s_j < S; \\ 1, & s_j \geq S. \end{cases} \qquad (6.6)$$

Однобічне p-значення розраховують за формулою $p = N/2^n$ і порівнюють із заданим значенням довірчого рівня p^*. Якщо $p > p^*$, то нульову про рівність середніх гіпотезу приймають на рівні значущості $1 - p^*$. Для застосування однобічного критерію обсяг вибірки має бути не менше 5 або 6 – за рівнів значущості 0,01 та 0,05, відповідно. Для двобічних критеріїв і тих самих рівнів значущості мінімальні обсяги вибірки дорівнюють 7 та 8, відповідно. За великих обсягів вибірок час обчислень швидко збільшується, тому доцільно використовувати інші критерії.

При порівнянні незалежних вибірок нульова гіпотеза – це належність двох досліджуваних вибірок до генеральних сукупностей з однаковими середніми. Нехай є дві вибірки: $x_1, x_2, ..., x_{n_x}, i = 1,2,...,n_x$ та $y_1, y_2, ..., y_{n_y}, j = 1,2,...,n_y$, де n_x, n_y – кількість елементів у них. Методика тесту основана на перебиранні всіх комбінацій даних. Обчислюють величину:

$$S = \min\left\{\sum_{i=1}^{n_x} x_i; \sum_{j=1}^{n_y} y_j\right\}. \qquad (6.7)$$

Кількість сприятливих результатів визначають за формулою:

$$N = 2\sum_{j=1}^{C_n^m} n_j, \quad n_j = \begin{cases} 0, & s_j < S, \\ 1, & s_j \geq S, \end{cases} \qquad (6.8)$$

де n_j – оцінка j-го результату; C_n^m – загальна кількість результатів; $n = n_x + n_y$ – чисельність об'єднаної вибірки; m – чисельність вибірки, що відповідає мінімальному значенню:

$$S_j = \sum_{i=1}^{m} z_{ij}, j = 1,2,...,C_n^m,$$

де z_{ji} – масив сполучень з об'єднаної вибірки, який будують подібно до розглянутої раніше процедури побудови матриці можливих результатів для спряжених вибірок. Однобічне значення p розраховують так:

$$p = \frac{N}{C_n^m}.$$

Його порівнюють із заданим рівнем значущості α. Нульову гіпотезу відхиляють, якщо $p < \alpha$ або $p > 1-\alpha$. Як і у попередньому випадку, критерій застосовують для відносно малих вибірок. Їх мінімальний допустимий обсяг є таким самим, як і для спряжених вибірок.

***W*-критерій Вілкоксона/Wilcoxon[127] – критерій суми рангів**. Його застосовують для перевірки та порівняння двох вибірок за їх центральною тенденцією, тобто за центрами емпіричних функцій розподілу. Сукупності можуть мати як однакові, так і різні чисельності. Критерій оперує не числовими значеннями даних, а їх рангами – місцями у впорядкованих за згасанням або зростанням рядах даних. При його застосуванні передбачають, що розподіли вибірок є неперервними, а нульова гіпотеза відповідає збігу функцій розподілу вибірок одна з одною.

Критерій застосовують у випадках, за яких ознаки виміряно принаймні у порядковій шкалі. Доцільно використовувати цей критерій, коли величина самих зсувів варіює в певному діапазоні (10-15 % від їх величини), адже розкид значень зсувів має бути таким, щоб з'являлася можливість їх ранжування. Якщо зсуви незначно розрізняються та набувають якихось кінцевих значень (напр., +1, –1 і 0), то формальних перешкод до застосування критерію немає, але,

[127] Запропоново американським хіміком і статистиком Вілкоксоном у 1945 р.

зважаючи на велику кількість однакових рангів, ранжування втрачає сенс, і ті самі результати простіше отримати за допомогою критерію знаків.

Мінімальне значення тесту:

$$W_1 = \frac{n(n+1)}{2},$$

де n – обсяг другої вибірки.

Максимальне значення тесту:

$$W_2 = nm + \frac{n(n+1)}{2},$$

де n – обсяг другої вибірки, m –обсяг першої вибірки.

Процедура обчислення значення критерію є близькою до обчислення критерію рандомізації компонент. Різниця лише в тому, що замість вихідних даних використовують їх ранги. Ранжирування порівнюваних вибірок здійснюють сумісно. Вихідні дані об'єднують до однієї вибірки, упорядковують, визначають ранги елементів об'єднаної вибірки.

Далі формують дві нові вибірки, елементами яких є ранги відповідних елементів вихідних вибірок. Якщо деякі значення збігаються, то відповідним спостереженням призначають середній ранг. Обчислення статистики критерію здійснюють за формулою:

$$S = \min\left\{\sum_{i=1}^{n_1} R_i; \sum_{i=1}^{n_2} S_i\right\}, \qquad (6.9)$$

де R_i – ранги вибірки, що має найменшу, а S_i – найбільшу суму рангів. Для вибірок малого обсягу (до 25 елементів) суму рангів W' вибірки, що має меншу кількість елементів, порівнюють із критичним значенням, яке визначають за спеціальними таблицями. При застосуванні W-критерію Вілкоксона слід зважати на те, що він належить до так званих критеріїв зсуву. Тобто найбільш потужним він буде при виявленні різниці, яка спричинена тим, що одну з вибірок отримано додаванням одного й того самого числа до всіх елементів іншої вибірки. Він є нечутливим до різниці дисперсій порівнюваних вибірок, коефіцієнтів їх асиметрії та ексцесу. Зокрема, якщо дві вибірки мають симетричні функції розподілу з однаковими середніми значеннями, але різними стан-

дартними відхиленнями, то в об'єднаній послідовності елементи однієї вибірки матимуть підвищену кількість елементів із високими та низькими рангами. Елементи іншої вибірки матимуть підвищену кількість елементів із середніми значеннями рангів. Але суми рангів усіх елементів для цих вибірок можуть бути приблизно однаковими.

***U*-критерій Манна–Уітні/Mann-Whitney U test**[128] призначено для перевірки нульової гіпотези про однаковість розподілу досліджуваних сукупностей або для перевірки рівності окремих параметрів цих розподілів, наприклад, середніх значень. Спостереження мають бути непарними. Цей критерій є найпотужнішим непараметричним аналогом *t*-критерію Стьюдента для незалежних вибірок. У деяких випадках його потужність може бути навіть більшою, ніж у *t*-критерію.

Обчислення здійснюють за формулами:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1,$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2, \qquad (6.10)$$

$$U = \max\{U_1; U_2\},$$

де R_1, R_2 – суми рангів вибірок; n_1, n_2 – кількість елементів у них. Якщо $n_{1,2}$, $n > 20$, то розподіл вибірки для *U*-статистики наближається до нормального. Правильність обчислення величин U_1 і U_2 можна перевірити за формулою $n_1 n_2 = U_1 + U_2$.

Модифікована статистика:

$$\frac{U - \mu_U}{\sigma_U}, \qquad (6.11)$$

$\mu_U = \frac{n_1 n_2}{2}$ – математичне сподівання, $\sigma_U^2 = \frac{n_1 n_2 (N+1)}{12}$ – дисперсія, $N = n_1 + n_2$ має стандартний нормальний розподіл. Результати обчислення за цим критерієм збігаються з даними, що отримують за *W*-критерієм Вілкоксона. На цьому критерії базується багатовимірний тест Джонкхієра–Терпстра.

[128] Запропоновано американськими математиками Манном та Уітні в 1947 р.

Критерій серій Вальда–Волфовиця/Wald-Wolfowitz runs test[129] використовують для перевірки нульової гіпотези, згідно з якою дві незалежні випадкові вибірки обсягами n_1 та n_2 не відрізняються одна від одної за досліджуваною ознакою.

Результати спостережень записують як варіаційний ряд об'єднаної вибірки, а їх належність до вихідних вибірок помічають за допомогою додаткової змінної, яка може набувати двох значень, наприклад 0 та 1. Послідовність її значень називають послідовністю кодів. Серією послідовності кодів називають будь-яку послідовність її однакових значень. Наприклад, у послідовності 00101111011 є такі серії: 00,1, 0, 1111, 0, 11. Очевидно, що за умови справедливості нульової гіпотези кількість серій N має бути великою, а за умови її помилковості – відносно малою. Якщо обсяги вибірок є достатньо великими (n_1, $n_2 > 20$), то для перевірки нульової гіпотези можна використовувати статистику:

$$Z = \frac{\left| N - \left(\frac{2n_1 n_2}{n_1 + n_2} \right) + 1 \right| - \frac{1}{2}}{\sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}}, \quad (\mathbf{6.12})$$

яка має стандартний нормальний розподіл.

Ці методи можуть давати інформацію про відмінність не тільки середніх, а й параметрів розподілу в цілому у двох вибірках.

Аналогом t-тесту для різниці середніх у залежних вибірках є **Критерій знаків/Sign test**, **Парний тест Вілкоксона/ Wilcoxon matched pairs test**.

Для візуального представлення за непараметричних тестів також можна використати діаграми-короби. Але в цьому випадку центральні точки вказують медіани, межі коробки – нижню та верхню квартилі, а "вусики" вказують 1 % і 99 % – персентилі.

[129] Розроблено американськими математиками Вальдом і Волфовицем у 1940 р.

Розділ 7
ДИСПЕРСІЙНИЙ АНАЛІЗ МІЖНАРОДНОЇ ТОРГІВЛІ ТА ІНВЕСТИЦІЙ

7.1. Теоретичні основи методу дисперсійного аналізу

Дисперсійний аналіз можна визначити як *параметричний метод*[130], що призначений для оцінювання впливу різних факторів на результат експерименту, а також для наступного планування експериментів. Тому за допомогою дисперсійного аналізу можна досліджувати залежності кількісної ознаки від одного чи багатьох ознак-факторів. Дисперсійний аналіз є сукупністю статистичних методів, що призначені для:

- перевірки гіпотез про зв'язок між певною ознакою та досліджуваними факторами, які не мають кількісного опису;
- установлення ступеня впливу факторів і їх взаємодії.

Мета дисперсійного аналізу – визначення значущості різниці середніх залежної змінної у різних групах спостережень при групуванні за незалежною змінною (групуючою змінною або фактором). Дисперсійний аналіз має більше можливостей, оскільки дозволяє:

- виділяти як дві групи значень за фактором (незалежною змінною), так і більше;
- досліджувати взаємодію факторів;
- досліджувати вплив факторів на кілька залежних змінних одночасно.

Відгуком називають значення вимірюваної ознаки. **Факторами** називають контрольовані чинники, що впливають на кінцевий результат. **Рівнем фактора**, або **способом обробки**, називають значення, що характеризують конкретне виявлення цього фактора. Ці значення подають у номінальній або порядковій шкалі вимірювань. Часто вихідні значення факторів вимірюють у кількісних або порядкових шкалах. Тоді постає проблема групування вихідних даних у ряди спостережень, що відповідають приблизно однаковим значен-

[130] Уперше цей метод розроблено Фішером у 1925 р.

ням фактора. Якщо кількість груп дуже велика, то кількість спостережень у них може виявитися недостатньою для отримання надійних результатів; якщо мала, – то це може призвести до втрати суттєвих особливостей впливу досліджуваного фактора на систему.

Кількість і розміри інтервалів за однофакторного аналізу найчастіше визначають за принципом рівних інтервалів або за рівних частот.

За багатофакторного аналізу застосовують *три типи групування*:

- із рівною кількістю спостережень;
- із різною кількістю спостережень;
- ті, в яких кількості спостережень відповідають певній пропорції.

При цьому існують певні особливості обробки даних, залежно від типу групування.

7.2. Однофакторний дисперсійний аналіз (*ANOVA*)[131]

Основною метою однофакторного аналізу є оцінювання величини впливу конкретного фактора на досліджуваний відгук. Попереднім етапом є перевірка нульової гіпотези про відсутність будь-якого впливу досліджуваного фактора (факторів), тобто гіпотези про те, що зміни значень ознаки у порівнюваних вибірках є випадковими, і всі дані належать до однієї генеральної сукупності.

Якщо нульову гіпотезу відкидають, то наступним етапом є кількісне оцінювання впливу досліджуваного фактора та побудова довірчих інтервалів для отриманих характеристик.

Якщо нульову гіпотезу неможна відкинути, то її приймають, і роблять висновок про відсутність впливу.

Якщо є підстави вважати, що такий вплив має бути наявний (напр., це може випливати з теоретичних уявлень про об'єкт дослідження), то необхідно перевірити наявність інших факторів, що можуть його маскувати.

---

[131] У спеціальній літературі дисперсійний аналіз (або аналіз варіації) часто називають *ANOVA – Analysis of Variance.*

За однофакторного дисперсійного аналізу вихідні дані подають у вигляді таблиць, де кількість стовпчиків дорівнює кількості рівнів фактора, а кількість значень у кожному стовпчику – кількості спостережень за відповідного рівня фактора. У табл. 7.1 подано загальний вигляд вихідної таблиці спостережень при проведенні однофакторного дисперсійного аналізу).

Таблиця 7.1

Результати вимірювань	**Рівні фактора**			
	1	2	...	k
1	x_{11}	x_{12}	...	x_{1k}
2	x_{21}	x_{22}	...	x_{2k}
...	...	...	...	...
n_i	x_{n1}	x_{n2}	...	x_{nk}

Для різних рівнів фактора кількість спостережень може бути різною. При цьому виходять із припущення про те, що результати спостережень для різних рівнів є вибірками із нормально розподілених сукупностей, середні значення та дисперсії яких є однаковими та не залежать від рівнів.

Завданням аналізу є перевірка нульової гіпотези про рівність середніх значень розглядуваних сукупностей.

Метод базується на основній тотожності дисперсійного аналізу. В основі дисперсійного аналізу лежить поділ загальної дисперсії залежної змінної на дві складові: внутрішню групову дисперсію та міжгрупову дисперсію.

Міжгрупова дисперсія, яка є відображенням різниці у середніх за групами, є наслідком наявності ймовірного впливу фактора.

Внутрішньогрупова дисперсія є наслідком дії інших факторів, які не враховано.

Сума квадратів відхилень спостережень від загального середнього (*загальна варіація*) становить:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}-\overline{x}\right)^2=\sum_{j=1}^{k}\left(x_{ij}-\frac{1}{n_j}\sum_{i=1}^{n_j}x_{ij}\right), \qquad \textbf{(7.1)}$$

$$\overline{x}=\frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j}x_{ij},\ N=\sum_{j=1}^{k}n_j,\ \left\langle x_j\right\rangle=\frac{1}{n_j}\sum_{i=1}^{n_j}x_{ij},\ j=1,2,\ldots k,$$

де $\bar{x}$ – загальне середнє; N – загальна кількість; k – кількість вибірок; n_j – кількість елементів у j-й вибірці; $\langle x_j \rangle$ – середнє значення j-ї вибірки.

Перший доданок (*факторна*, або *міжгрупова дисперсія*) є зваженою сумою квадратів відхилень групових середніх від загального середнього. Він характеризує коливання значень, що зумовлені фактором, на основі якого здійснено групування даних.

Другий доданок (*залишкова*, або *внутрішньогрупова варіація*) є сумою квадратів відхилень спостережень від відповідних групових середніх. Він характеризує коливання значень досліджуваної ознаки, що зумовлені неврахованими факторами або випадковими чинниками.

Сутність методу полягає в тому, що за умови правильності нульової гіпотези величини:

$$\sigma_1^2 = \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \langle x_j \rangle\right)^2 \text{ та } \sigma_2^2 = \frac{1}{k-1}\sum_{j=1}^{k} n_j\left(\langle x_j \rangle - \bar{x}\right)^2$$

є незміщеними оцінками дисперсії похибок спостережень σ^2 і мають бути приблизно рівними.

Що більшою є міжгрупова дисперсія, порівняно із внутрішньою груповою, то більш значущою буде різниця між середніми залежної змінної у групах, отже більшим буде й вплив фактора. Цю значущість перевіряють за допомогою *F*-тесту. Перша з них є мірою варіації усередині вибірок і не пов'язана з припущенням про рівність середніх значень, тому $\sigma^2 \approx \sigma_1^2$, незалежно від справедливості нульової гіпотези. Друга оцінка характеризує варіацію між вибірками. За справедливості нульової гіпотези $\sigma_2^2 \approx \sigma^2$, а за її порушенні величина σ_2^2 є тим більшою, чим більшим є відхилення від неї.

Значення критерія розраховують за формулою:

$$F = \frac{(N-k)\sum_{j=1}^{k} n_j\left(\langle x_j \rangle - \bar{x}\right)^2}{(k-1)\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \langle x_j \rangle\right)^2}. \qquad (7.2)$$

Ця величина має *F*-розподіл Фішера з параметрами $k-1$ та $N-k$.

Нульову гіпотезу відхиляють, якщо ймовірність $P(F \geq F^*)$ є достатньо малою, де F^* – значення, що розраховане за емпіричними даними (σ_2^2).

Однофакторний одномірний дисперсійний аналіз/*one-way univariate ANOVA* є найпростішим варіантом дисперсійного аналізу. Якщо за групуючою змінною всі спостереження поділяють лише на дві групи, то дисперсійний аналіз дасть той самий результат, що й *t*-тест для аналізу середніх. Наприклад, при аналізі впливу митних тарифів на відношення імпорту до ВВП поділяють усі спостереження на три групи, залежно від середньозваженої (за всіма товарами) ставки митного тарифу (табл. 7.2).

Таблиця 7.2

Група	**Ставка митного тарифу**		
	низька (I група)	**середня** (II група)	**висока** (III група)
Середня залежної змінної – імпорт/ВВП, %	50	32	18

Дисперсійний аналіз є більш ефективний, ніж порівняння середніх у кожній парі груп по черзі за допомогою *t*-тесту. Якщо *F*-тест показує значущу різницю між середніми, то за допомогою *апостеріорних критеріїв*/*Post Hoc Methods* визначають, саме за рахунок яких з груп виникає різниця. До *апостеріаорних критеріїв* належать:

- *найменша значуща різниця Фішера*/*Fisher's least significant difference*;
- *критерій Шеффе*/*Scheffe test*;
- *критерій Ньюмана–Кеулса*/*Newman–Keuls test & critical ranges*;
- *критерій Дункана*/*Duncan's multiple range test*;
- *критерій Тьюкі*/*Tukey honest significant difference –HSD*;
- *критерій Тьюкі для нерівних N*/*Unequal N HSD*.

Використовують саме апостеріорні критерії, а не серію звичайних *t*-тестів для різниці між середніми, оскільки в останньому випадку за великої кількості порівнянь більшою мірою можливий варіант, за якого невелика частка значущих

різниць між середніми є насправді випадковістю. Але *t*-тест може давати кращі результати, лише якщо для різниці між середніми є теоретичне обґрунтування.

7.3. Метод одномірного багатофакторного дисперсійного аналізу впливу торговельних обмежень на імпорт

Складнішим варіантом дисперсійного аналізу є *багатофакторний одномірний дисперсійний аналіз/n-way univariate ANOVA*. Метою може бути порівняння двох або кількох факторів для визначення різниці їх впливу на відгук, яку часто називають *контрастом факторів*. Наприклад, відбувається аналіз впливу як митних тарифів, так і нетарифних бар'єрів щодо імпорту до ВВП (табл. 7.3 – *Середня залежної змінної за групами – імпорт/ВВП*, %).

Таблиця 7.3

Група	**Ставка митного тарифу**		
	низька (I група)	**середня** (II група)	**висока** (III група)
Низькі нетарифні бар'єри (I група)	60	37	20
Високі нетарифні бар'єри (II група)	38	25	17

Можна досліджувати *три види ефектів*:

- *вплив першого фактора* (митні тарифи) на залежну змінну (високі митні тарифи зменшують імпорт щодо ВВП);
- *вплив другого фактора* (нетарифні бар'єри) на залежну змінну (високі нетарифні бар'єри зменшують імпорт щодо ВВП);
- *взаємодію факторів* (яким чином ті чи інші значення одного фактора здійснюють вплив іншого фактора на залежну змінну – високі нетарифні бар'єри більшою мірою зменшують імпорт/ВВП за нижчих ставок митного тарифу, а вплив митних тарифів не такий сильний за високих нетарифних обмежень).

Перші два ефекти – *головні ефекти*/*main effects*, останній – це *ефект взаємодії*/*interaction effect*. Якщо відбувається групування за трьома факторами, то виявляться три головні ефекти й три парні ефекти взаємодії (кожний фактор із кожним фактором); якщо за чотирма, – то чотири головні ефекти та шість парних ефектів взаємодії. Взаємодію вищих порядків складніше проілюструвати та пояснити.

Досліджуючи ефект взаємодії, потрібно дивитися, як перший фактор змінює вплив другого фактора на залежну змінну, а також як другий фактор змінює вплив першого фактора на залежну змінну.

Ефект взаємодії може бути різним:

- *відсутній*, коли один фактор не змінює вплив іншого фактора на залежну змінну;
- коли певні значення одного фактора зменшують чи збільшують силу впливу іншого фактора на залежну змінну, але *характер впливу не змінюється*;
- коли певні значення одного фактора *змінюють характер впливу іншого фактора* на залежну змінну (позитивний стає негативним або навпаки).

Звісно, є прагнення в одній таблиці відобразити вплив якомога більшої кількості факторів. Але за кожного додаткового групування за додатковим фактором у комірках середні розраховують за все меншою кількістю спостережень, унаслідок чого результати стають менш значущими та надійними. Оскільки більшу кількість факторів одночасно досліджують з урахуванням їхньої взаємодії, то кількість спостережень має бути більшою.

7.4. Умови використання методу дисперсійного аналізу

1. *Кількість спостережень у кожній групі* (чи комірці у таблиці при групуванні за кількома факторами) спостережень *має бути достатньою*. Бажано провести мінімум 30 спостережень, хоча можна проводити аналіз і за меншої кількості з менш надійними результатами. Бажано, щоб у кожній групі (комірці) кількість спостережень не значно відрізнялася.

2. *Спостереження мають бути незалежні* (відсутність серійної кореляції). Якщо беруть значення за різний період часу, навіть для різних об'єктів, то серійна кореляція ймовірна. Інший приклад: якщо певна частина спостережень корелюють одне з одним унаслідок схожих особливостей (напр., якщо дані одержують за допомогою опитування, а одна з груп респондентів одержує заплутані інструкції). Розв'язанням проблеми може бути введення додаткового групування за іншим впливовим фактором або використання суворішого критерію значущості (напр., менше 0.01, а не 0.05).

3. *В усіх групах дисперсія має бути однаковою* (*гомоскедастичність*/*homoscedasticity*). Порушення цієї вимоги означає, що середні та дисперсія за групами корелюють. Тоді *F*-критерій може давати неправильну оцінку значущості різниці між середніми. Наприклад, якщо дисперсія є більшою у групі з більшою середньою залежною змінною, то це може означати, що така середня, імовірно, здається більшою випадково, адже більша дисперсія означає менш надійну оцінку середньої для генеральної сукупності за середньою, що розрахована на основі вибірки. Більша дисперсія, зокрема, може бути наслідком наявності викидів. Порушення цієї вимоги не є критичним, якщо групи (чи комірки) мають приблизно однакову кількість спостережень: найбільша від найменшої відрізняється не більш ніж у 1,5 рази. Якщо гетероскедастичність (різна дисперсія) є проблемою, то потрібно:

- трансформувати залежну змінну;
- використати суворіший критерій значущості, ніж 0,05;
- перевірити на наявність викидів за залежною змінною, провести повторно дисперсійний аналіз без викидів;
- скористатися замість дисперсійного аналізу його непараметричним аналогом.

4. *Залежна змінна має підпорядковуватися нормальному розподілу всередині кожної групи.* Порушення цієї передумови не є критичним, якщо вибірка є великою. Якщо ексцес більше нуля, рівень значущості *p*, який дає *F*-критерій завищений, тобто дисперсійний аналіз указує на відсутність закономірності, коли вона існує; якщо менше нуля, то рівень

значущості буде заниженим. Асиметрія розподілу мало впливає на надійність результатів дисперсійного аналізу.

5. Додатково розглядають лінійність зв'язку між факторами та залежною змінною, а також *відсутність мультиколінеарності*.

7.5. Багатомірний дисперсійний аналіз (*MANOVA*)

У *багатомірному дисперсійному аналізі/multivariate analysis of variance* досліджують вплив фактора або факторів не по черзі, а одночасно на кілька залежних змінних. Наприклад, вплив обмежень руху капіталу одночасно на прямі та портфельні інвестиції до країни. У *MANOVA* замість звичайного *F*-теста використовують такі критерії:

- багатомірний *F*-критерій (*лямбда-критерій Уілкса/Wilks' lambda*;
- *найбільший характеристичний корінь Роя/Roy's greatest characteristic root*;
- *слід Хотеллінга/Hotelling's trace*;
- *критерій Піллаї/Pillai's criterion*, що має переваги, якщо вибірка невелика, кількість спостережень у комірках є достатньо різною або є гетероскедастичність.

Якщо вони показують значущість впливу факторів одночасно на кілька залежних змінних, то далі з'ясовують, яким чином фактори впливають на кожну залежну змінну окремо за допомогою звичайного одномірного *F-критерія з можливими поправками/Bonferroni inequality* або *stepdown analysis*.

Проведення *MANOVA* має перевагу перед проведенням серії одномірних *ANOVA* для кожної залежної змінної, оскільки убезпечує від прийняття випадкової різниці у середніх за закономірність.

Припустимо, аналізують 40 залежних змінних. Якщо серія одномірних *ANOVA* показує, що фактори впливають на 20 із рівнем значущості 0,05, то цілком імовірно, що насправді одна із залежних змінних не залежить від досліджуваних факторів.

MANOVA також може виявити існування певної лінійної комбінації залежних змінних, середні якої суттєво відрізняються за групами, навіть якщо середні індивідуальних залеж-

них змінних мало відрізняються за групами. Іншими словами, *MANOVA* може виявити комбіновану відмінність між групами. Серія одномірних *ANOVA* ігнорує наявність кореляції між залежними змінними.

Мета використання MANOVA:

1. *Дослідження різнорідних залежних змінних* (напр., валютний курс, приріст ВВП, міграція, інфляція тощо), коли потрібно знизити ризик сприйняття випадковостей за закономірності. У разі виявлення впливу на всі змінні одночасно досліджують вплив на кожну залежну змінну окремо.

2. *Дослідження тієї самої залежної змінної у різні періоди часу*, але для тих самих об'єктів, тобто в умовах повторних спостережень. Різні залежні змінні в такому випадку є фактично однією змінною, що виміряна у різні періоди часу, наприклад, якщо перша залежна змінна – це зростання імпорту у 2019 р., друга – зростання імпорту у 2020 р., третя – зростання імпорту у 2021 р. за тими самими країнами. Або перша змінна – зростання експорту перед кризою, друга – під час кризи, третя – після кризи. Фактором може бути, наприклад, дефіцит державного бюджету.

3. *Дослідження однорідних залежних змінних*, коли цікавить *саме одночасний вплив факторів на усі змінні*, а не на кожну окрему. Наприклад, у випадку дослідження валютних криз такими однорідними змінними можуть бути знецінення національної валюти, зменшення чистих валютних резервів, зростання відсоткової ставки. Усі ці змінні є основними індикаторами наявності валютної кризи, і цікавим може бути одночасний вплив факторів на всі залежні змінні у сукупності. Особливо корисним *MANOVA* може бути, коли значення кожної залежної змінної неточні, а використання комбінації залежних змінних нівелює ці неточності. Зокрема, коли індивідуальні залежні змінні є результатом таких суб'єктивних оцінок, як дві змінні: ступінь задоволення іноземних туристів перебуванням у країні та їх бажання приїхати ще раз до країни протягом наступних кількох років.

Більшість *особливостей передумов для MANOVA* збігаються з передумовами для одномірного *ANOVA*, але є й відмінності:

- *MANOVA* потребує більші за обсягом вибірки, ніж одномірний *ANOVA*;
- кореляція для змінних у кожній групі (комірці) має бути однаковою;
- залежні змінні мають підпорядковуватися багатомірному нормальному розподілу, тобто будь-яка лінійна комбінація залежних змінних має підпорядковуватися нормальному розподілу. Оскільки на практиці важко перевірити, чи буде нормальним багатомірний розподіл, то тестують кожну залежну змінну окремо та визначають, чи буде її розподіл нормальним. Якщо залежних змінних тільки дві, то можна побудувати тривимірну гістограму.

7.6. Дисперсійний аналіз впливу рівня економічного розвитку на приплив інвестицій у Microsoft Excel

Припустимо, є дані про ПІІ до країн із різним доходом на душу населення (низьким, високим, але також окремо і з середнім, див. рис.7.1). Не обов'язково, щоб групи мали однакову кількість спостережень.

У надбудові *Пакет аналізу/ Data Analysis* обирають *Однофакторний дисперсіний аналіз/ ANOVA: Single Factor*; далі – *Групування за стовпчиками/Grouped by Columns* (оскільки у вхідних даних – один стовпчик – одна група) та інші опції (рис. 7.2).

FDIlowincome	FDIhighincome	FDImidincome
3	5	6
5	9	11
2	4	3
3	2	6
2	3	1
5	11	10
1	6	8
1	3	4
4	4	5
4	5	3
3	2	6
4	2	8
0	1	6
1	1	4
2	2	3
1	4	4
2	3	1
2	2	4
1	0	5
0	1	3

Рис. 7.1

У робочому аркуші з результатами (рис. 7.3) видно, що найбільші ПІІ – у країнах із середнім доходом на душу населення (5,05 % ВВП), далі – із високим доходом (3,5 % ВВП), і нарешті – із низьким доходом (2,3 % ВВП). Рівень значущості *F*-критерію становить 0,002304, тобто менше 0,05. Це означає, що різниці є значущими.

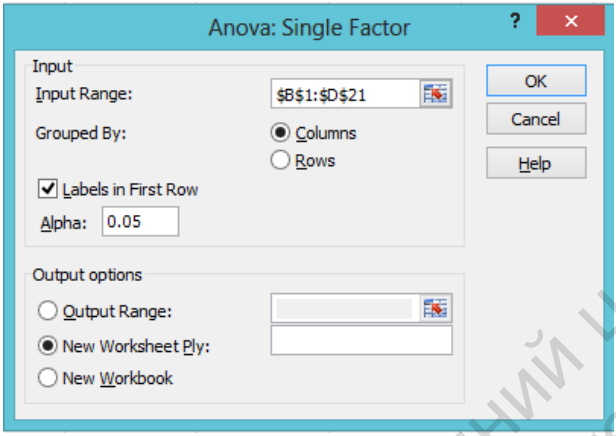


Рис. 7.2

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
FDIlowincome	20	46	2.3	2.326316
FDIhighincome	20	70	3.5	7.421053
FDImidincome	20	101	5.05	7.102632

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	76.03333	2	38.01667	6.768546	0.002304	3.158843
Within Groups	320.15	57	5.616667			
Total	396.1833	59				

Рис. 7.3

У надбудові *Data Analysis* можна використати *F-тест для різниці дисперсій/F-Test Two-Sample for Variance* для попарного порівняння дисперсії. А для візуальної перевірки значень у кожній групі на відповідність нормальному розподілу варто скористуватися опцією *Гістограмма/Histogram*.

Оскільки апостеріорні критерії у Microsoft Excel не доступні, то далі можна попарно порівняти середні у кожній групі за допомогою *t*-тесту, як раніше, або за допомогою дисперсійного аналізу кожний раз для двох груп. Виявиться, що значуще відрізняються середні у 1 і 3 групах (у країнах із низьким і середнім доходом) з рівнем значущості 0,000278.

Значущість відмінності середніх у групах 1 і 2 та 2 і 3 є меншою (0,094 та 0,076).

Microsoft Excel також дозволяє проводити дисперсійний аналіз за допомогою опцій *Двофакторний дисперсійний аналіз з повтореннями/Two-Factor with Replication* та опцій *Двофакторний дисперсійний аналіз без повторень/Two-Factor without Replication*) у надбудові *Data Analysis*.

7.7. Непараметричні методи дисперсійного аналізу

Непараметричним аналогом однофакторного дисперсійного аналізу є *ранговий однофакторний аналіз Краскела–Волліса/Kruskal–Wallis test*[132].

Критерій призначено для перевірки нульової гіпотези про рівність ефектів впливу на досліджувані вибірки з невідомими, але рівними середніми. При цьому кількість вибірок має бути більшою, ніж дві. Сутність нульова гіпотези – в тому, що k вибірок обсягами $n_1, n_2, ..., n_k$, що отримані з однієї і тієї самої генеральної сукупності.

Рангові методи, зокрема, і метод Краскела–Уолліса, не передбачають, що розподіл результатів спостережень є нормальним, їх можна застосовувати як для кількісних даних з невідомим законом розподілу, так і для порядкових ознак. До таблиці, замість спостережень, заносять їх ранги r_{ij}, отримані шляхом упорядкування за зростанням усієї сукупності спостережень x_{ij} (табл. 7.4)

Таблиця 7.4

№ результату	№ вибірки			
	1	2	..	k
1	r_{11}	r_{12}	..	r_{1k}
2	r_{21}	r_{22}	..	r_{2k}
...	...	...	..	...
n_i	r_{n1}	r_{n2}	..	r_{nk}

[132] Розроблено американськими математиком Краскелом та економістом Уоллісом у 1952 р.

Для кожного рівня фактора, тобто для кожного стовпця, розраховують суму рангів:

$$R_j = \sum_{i=1}^{n_j} r_{ij} \qquad (7.3)$$

або відповідні середні ранги:

$$\left\langle R_j \right\rangle = \frac{1}{n_j} \sum_{i=1}^{n_j} r_{ij},\ j = 1,2,...k. \qquad (7.4)$$

Для контролю можна використовувати тотожність:

$$\sum_{j=1}^{k} R_j = \frac{N(N+1)}{2}, N = \sum_{j=1}^{k} n_j,$$

де N – загальна кількість.

Непараметричний **критерій Левене**[133] використовують в умовах, коли немає впевненості у тому, що досліджувані вибірки підпорядковані нормальному розподілу. Розрахункове значення критерія обчислюють за формулою:

$$W = \frac{(N-k)\sum_{j=1}^{k} n_j (Z_j - \overline{Z})^2}{(k-1)\sum_{j=1}^{k}\sum_{i=1}^{n_j} (Z_{ij} - \overline{Z_j})^2}, \qquad (7.5)$$

де $Z_{ij} = \left| x_{ij} - \overline{x_j} \right|$; x_{ij} – значення i-го спостереження в j-ої вибірці; $\overline{x}_j$ – середнє арифметичне спостережень, що потрапили до j-ої вибірки; $\overline{Z}$ – загальне середнє арифметичне значень Z за всіма спостереженнями; $\overline{Z}_j$ – середнє арифметичне Z за спостереженнями, що потрапили до j-ої вибірки.

Розрахункове значення критерію порівнюють із відповідним квантилем F-розподілу з кількостями степенів вільності $(k-1)$ та $(N-k)$.

Більш робастний тест – *критерій Брауна–Форсайта*[134] відрізняється від критерію Левене тим, що значення

[133] Запропоновано американським математиком Левене у 1960 р.

[134] Запропоновано американськими статистиками Брауном та Форсайтом у 1974 р.

$Z_{ij} = \left| x_{ij} - x_j \right|$, де x_j – медіана спостережень, які потрапили до *j*-ої вибірки.

Розглянуті критерії дають змогу встановити різницю дисперсій сукупностей, але не дати кількісну оцінку впливу фактора на досліджувану ознаку, а також встановити, для яких саме сукупностей дисперсії є різними.

Розглянемо приклад виконання рангового однофакторного аналізу.

Нехай є чотири вибірки, що сформовані за допомогою пакету аналізу Microsoft Excel як суміш 100 елементів вибірки, що має нормальний розподіл із параметрами $\bar{x} = 20$, $S = 3$ і рівномірно розподілених вибірок обсягом по 100 елементів кожна, які задано, відповідно, на відрізках:

[17; 22], [18; 22],[18; 22] та [17; 23].

В електронних таблицях Microsoft Excel немає вбудованих засобів для реалізації рангового *однофакторного аналізу Краскела–Уолліса*. Але його неважко здійснити за допомогою наявних функцій. Спочатку необхідно перетворити таблицю вихідних даних на таблицю значень рангів. Для цього використовують функцію RANK ( ). Вікно задання її параметрів показано на рис. 7.4.

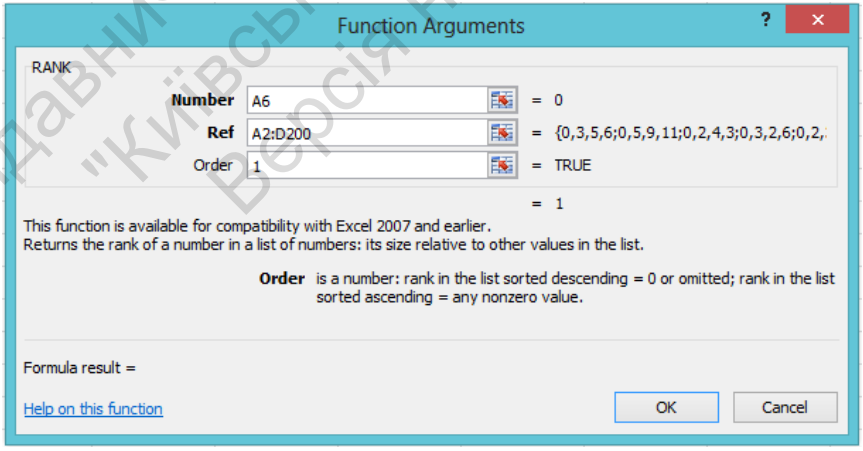


Рис. 7.4

У комірці *Число*/*Number* вказують, для якого саме значення таблиці вихідних даних необхідно обчислити ранг. У комірці *Посилання*/*Ref* дають посилання на весь діапазон, що містить вихідні значення (воно має бути абсолютним). У комірці *Порядок*/*Order* зазначають порядок ранжирування: 0 – за убуванням, інше число – за зростанням. Після цього за формулою

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1), N = \sum_{j=1}^{k} n_j$$

розраховують значення критерію. У нашому випадку воно дорівнює 7,22. Як критичне візьмемо значення оберненої функції розподілу χ^2, яке можна визначити за допомогою функції =CHIINV(0,05;3)=7,82. Її аргументами є рівень значущості (0,05) та кількість степенів вільності (3). Бачимо, що розрахункове значення критерію дещо менше за критичне, тому немає підстав для відхилення нульової гіпотези про однорідність досліджуваних вибірок.

Непараметричними аналогами дисперсійного аналізу для повторних спостережень є:

- ранговий *дисперсійний аналіз Фрідмана*/*Friedman ANOVA test*;
- *Q-критерій Кохрена*/*Cochran's Q test* – для змінних, які вимірюють у номінальній шкалі.

Розділ 8
ДОСЛІДЖЕННЯ ТА АНАЛІЗ МІЖНАРОДНИХ ЕКОНОМІЧНИХ ВІДНОСИН МЕТОДАМИ КОРЕЛЯЦІЙНОГО АНАЛІЗУ

8.1. Кореляційний аналіз кількісних ознак

В економічних дослідженнях та аналізі МЕВ однією з основних задач є аналіз залежностей між змінними ознаками. Але довільна залежність певною мірою є абстракцією, оскільки в оточуючому світі, частиною якої є світова економіка, значення конкретної величини не визначають незмінною формулою її залежності від набору інших величин. Завжди є кілька ознак, які визначають головні тенденції зміни економічного явища, яке розглядають, і в економічній теорії та на практиці обмежуються тим чи іншим колом таких пояснювальних змінних. Закономірності економічних процесів у сучасному світі характеризують ЕММ, за допомогою яких вивчають взаємозв'язки між статистичними показниками. Такі показники перебувають у певних співвідношеннях, виступаючи в ролі незалежних або залежних ознак. Ознаки, які впливають на певний результат, називають *незалежними ознаками* (*факторами*). Ознаки, значення яких сформовані під впливом інших ознак, називають *залежними ознаками* (*показниками*).

Поряд з основними факторами завжди існує взаємодія великої кількості інших, менш важливих, або таких, що важко ідентифікувати, які приводять до відхилень значень показника (залежної змінної) від конкретної формули її зв'язку з факторами (пояснювальними змінними), наскільки б точною ця формула не була. Знаходження, оцінювання та аналіз таких зв'язків, ідентифікація пояснювальних змінних, побудова формул залежності та оцінювання їх параметрів є свого роду мистецтвом, що враховує в кожній конкретній сфері знань (зокрема, у міжнародних економічних відносинах) її внутрішні закони та потреби. Успішність застосування будь-якого кількісного методу аналізу даних залежить від

відповідності аналізованих даних його вихідним припущенням. Методи, що придатні для одного типу даних, можуть призводити до серйозних помилок при їх використанні для даних інших типів.

Першим етапом аналізу будь-яких даних зазвичай є визначення їх типу. Основною є *класифікація даних за шкалами їх вимірювання.*

Залежність між ознаками (змінними) може бути функціональною або стохастичною.

За *функціональної залежності* кожному значенню незалежної змінної (ознаки) відповідає єдине, строго визначене значення залежності змінної (ознаки). Функціональні зв'язки притаманні переважно природничим і технічним системам, на відміну від яких такі зв'язки між показниками у соціально-економічних системах у більшості випадків відсутні. Наприклад, не може існувати строгої функціональної залежності між доходами громадян і їх витратами на споживання, між ціною на певний товар і попитом на нього; між продуктивністю праці та стажем роботи працівників тощо.

Відсутність жорсткої функціональної залежності між змінними (ознаками) у сфері міжнародних економічних відносин пов'язана з низкою причин. Наприклад, при аналізі впливу однієї змінної на іншу може бути не враховано низку факторів, які впливають або на кожну зі змінних окремо, або на всі одночасно. Цей вплив може бути як безпосереднім, так і через цілий ланцюг інших факторів, урахувати які практично неможливо, оскільки вони мають випадкове походження. Тому у дослідженнях у сфері міжнародних економічних відносин зазвичай мають справу не з функціональною, а зі *стохастичною* (*статистичною*, *випадковою*) або *кореляційною* залежністю, вивченням якої і займається кореляційний аналіз.

Кореляцією (кореляційним зв'язком) між ознаками (випадковими величинами) називають наявність статистичного або ймовірнісного зв'язку між ними. При цьому закономірна зміна певних ознак приводить до закономірної зміни середніх значень інших ознак, що пов'язані з ними.

Кореляційним аналізом/correlation analysis називають сукупність методів виявлення кореляційного зв'язку, тому

його можна застосовувати для формалізованого подання моделей зв'язків між окремими компонентами системи або між окремими процесами, що відбуваються в ній. Наявність кореляційного зв'язку не означає існування причинно-наслідкового зв'язку між досліджуваними ознаками. Вона може бути зумовлена тим, що обидві ознаки мають причинно-наслідковий зв'язок із певним іншим фактором. Наприклад, існує кореляція між цінами на нафту й золото. Проте її можна пояснити тим, що обидві ціни виражають у доларах США, і залежать вони від динаміки його індексу.

Кореляційний аналіз здійснюють на початковому етапі вирішення всіх основних задач статистичного аналізу даних. У задачах статистичного аналізу залежностей і побудови ЕММ він дає змогу встановити сам факт існування зв'язку між змінними та оцінити ступінь його виявлення.

У задачах класифікації даних за допомогою кореляційного аналізу отримують вихідну інформацію у вигляді коваріаційних і кореляційних матриць та інших характеристик парних порівнянь. Це дає змогу визначити подібні один до одного або до певних еталонів об'єкти, сформувати класи подібних об'єктів і здійснити класифікацію.

У задачах зменшення розмірності досліджуваного простору ознак за допомогою коваріаційних і кореляційних матриць визначають ознаки, які можна без втрати суттєвої інформації подати через інші наявні дані.

Методика перевірки гіпотези про існування зв'язку між ознаками передбачає такі *етапи*:

І етап. Визначення типу даних.

ІІ етап. Перевірка гіпотези про відсутність зв'язку і, за її відхилення, оцінювання сили зв'язку.

Тип вихідних даних суттєво впливає на вибір методів і критеріїв, які можна застосовувати на наступних етапах аналізу.

Для визначення сили зв'язку використовують різноманітні показники. Їх прагнуть вибрати такими, щоб вони змінювалися від –1 до +1 або від 0 до 1. Значення, що є близькими за модулем до одиниці, свідчать про наявність сильного

зв'язку. Близькі до нуля значення вказують або на відсутність будь-якого зв'язку, або на відсутність зв'язку того типу (найчастіше, лінійного), для якого розроблено відповідний коефіцієнт. Знак коефіцієнта вказує на напрям зв'язку: прямий (для додатних значень) або зворотний (для від'ємних).

Припустимо, що дві змінні пов'язані лінійною залежністю:

$$Y = \beta_0 + \beta_1 X. \quad (\mathbf{8.1})$$

Розглянемо спочатку питання про лінійну залежність двох змінних:

- чи пов'язані між собою лінійно змінні *X* та*Y*?
- яка формула зв'язку змінних *X* та*Y*?

Для відповіді існують спеціальні статистичні методи та відповідно показники, значення яких із певною ймовірністю свідчать про наявність або відсутність лінійного зв'язку між змінними. У першому випадку – це коефіцієнт кореляції величин *X* та *Y*, у другому – коефіцієнти лінійної регресії β_0 і β_1, їхні стандартні похибки й *t*-статистики, за значеннями яких перевіряють гіпотезу про відсутність зв'язку величин *X* та *Y*.

Пояснимо логіку появи такого показника, як коефіцієнт кореляції. Припустимо, що між змінними *X* та *Y* існує лінійна залежність. Наявність такої залежності можна інтерпретувати так: якщо змінна *X* набуває значення, більші за її середнє значення, і зв'язок додатний (коефіцієнт $\beta_1 > 0$), то значення змінної *Y* також має бути більшим від її середнього значення, і співвідношення відхилень *X* та *Y* від їх середніх значень має бути сталим. Якщо змінна *X* набуває значення, меншого за її середнє значення, то значення змінної *Y* також має бути меншим від її середнього значення із тим самим коефіцієнтом пропорційності цих відхилень. Якщо зв'язок між змінними *X* та *Y* від'ємний, то додатне відхилення *X* від її середнього значення має узгоджуватись із від'ємним відхиленням *Y* від її середнього значення, і навпаки. Якщо лінійної залежності між змінними *X* та *Y* немає, то додатні відхилення змінної *X* від її середнього значення має узгоджуватися як з додатними, так і від'ємними відхиленнями *Y* від її середнього значення. Те саме можна сказати й про від'ємні відхилення *X* від її середнього значення. Як міру ступеня

лінійного зв'язку між двома змінними використовують коефіцієнт їхньої кореляції[135].

Розрізняють парні та частинні кореляційні характеристики. Парні характеристики розраховують за результатами вимірювань тільки досліджуваної пари ознак. Тому вони не враховують опосередкованого або спільного впливу інших ознак. Частинні характеристики очищені від впливу інших факторів, але для їх розрахунку необхідно мати вихідну інформацію не тільки про досліджувані ознаки, а й про всі інші, вплив яких необхідно усунути. Для кількісних ознак найчастіше застосовують коефіцієнти кореляції Пірсона та Фехнера. Коефіцієнт кореляції Пірсона[136] (коефіцієнт кореляційного відношення Пірсона, парний коефіцієнт кореляції, вибірковий коефіцієнт кореляції, коефіцієнт Бравайса–Пірсона) вимірює ступінь лінійного кореляційного зв'язку між кількісними скалярними ознаками.

Коефіцієнт кореляції Пірсона/Pearson correlation coefficient обчислюють за формулою:

$$r = \frac{\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2} \cdot \sqrt{\frac{1}{n}\sum_{m=1}^{n}(y_m - \bar{y})^2}}. \qquad (8.2)$$

Тут x_i, y_j – значення *n* змінних *X* та *Y*, відповідно, а $\bar{x}, \bar{y}$ – їх середні арифметичні.

Парний коефіцієнт кореляції Пірсона (**8.2**) є безрозмірною величиною, що не залежить від одиниць вимірювання обох змінних. Він набуває значення $|r| \le 1$. Застосування коефіцієнта Пірсона як міри зв'язку обґрунтований лише за умови, що спільний розподіл пари ознак є нормальним. Тому перед його розрахунком слід перевірити виконання цієї гіпотези. Якщо вона справедлива, то квадрат коефіцієнта кореляції

[135] Методику кількісного оцінювання кореляції між ознаками вперше було запропоновано британським географом, антропологом і психологом Гальтоном у 1888 р.

[136] Запропоновано Пірсоном у 1896 р.

Пірсона дорівнює коефіцієнту детермінації. Значення коефіцієнта кореляції може змінюватися від –1 до +1, що відповідає чіткій лінійній функціональній залежності, яка в першому випадку є спадною, а у другому – зростаючою. Для функціональної залежності $y = \text{const}$ коефіцієнт кореляції, як видно з формули, невизначений, оскільки у цьому випадку знаменник дорівнює нулю. Що ближчим є значення коефіцієнта кореляції до –1 або +1, то більш обґрунтованим є припущення про наявність лінійного зв'язку. Наближення його значення до нуля свідчить про відсутність лінійного зв'язку, але не є доказом відсутності статистичного зв'язку загалом.

Коефіцієнт Пірсона можна виразити також через дисперсії σ_y і $\sigma_{\Delta y}$, друга з яких характеризує розсіювання емпіричних точок щодо рівняння лінійної регресії виду (**8.1**), де β_0, β_1 – коефіцієнти, які визначені за методом найменших квадратів:

$$r = \frac{1}{\sqrt{1 + \left(\frac{\sigma_{\Delta y}}{\sigma_y}\right)^2}}. \qquad (\mathbf{8.3})$$

За умови достатньо великого обсягу спостережень ($n \geq 30$) стандартне відхилення коефіцієнта кореляції Пірсона можна визначити за формулою:

$$\sigma_r = \frac{1 - r^2}{\sqrt{n}}. \qquad (\mathbf{8.4})$$

На рівні значущості 0,01 гіпотезу про наявність кореляційного зв'язку приймають, якщо:

$$\frac{|r|}{\sigma_r} \geq 2{,}6.$$

Прийнято вважати: якщо $|r| \leq 0{,}25$, то кореляція слабка; якщо $0{,}25 < |r| \leq 0{,}75$, то – середня; якщо $|r| > 0{,}75$ – сильна. Якщо $r = 0$, то говорять, що змінні некорельовані. Але при цьому слід ураховувати, що звичайні методи кореляційного аналізу дають змогу перевіряти лише гіпотезу про наявність лінійного зв'язку. Якщо зв'язок є, але він нелінійний, то ви-

сновки, отримані за допомогою кореляційного аналізу, можуть бути помилковими.

Візуально силу кореляції між двома змінними можна представити за допомогою діаграми розсіювання, де за осями відкладено значення двох змінних. Якщо форма сукупності крапок-спостережень наближається до прямої лінії, то кореляція сильна; якщо – до круглої хмари, то – слабка.

На діаграмах розсіювання подано приклади сильної негативної кореляції (рис. 8.1), сильної позитивної кореляції (рис. 8.2), майже відсутньої кореляції (рис. 8.3).

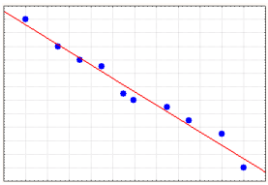


Рис. 8.1

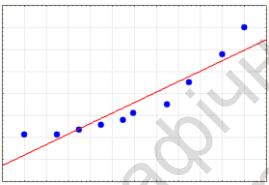


Рис. 8.2

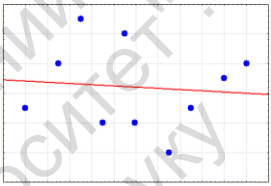


Рис. 8.3

Коефіцієнт кореляції Пірсона часто розглядають як універсальну міру кореляційного зв'язку. У багатьох пакетах загального призначення, зокрема в електронних таблицях MS Excel, не передбачено інших засобів його вимірювання. Але насправді сфера його обґрунтованого застосування є досить вузькою, оскільки лінійність залежності й нормальний розподіл даних навколо неї є скоріше винятком, ніж правилом.

При дослідженні багатовимірних сукупностей випадкових величин із коефіцієнтів кореляції, що обчислені попарно, можна побудувати квадратну симетричну кореляційну матрицю з одиницями на головній діагоналі. Вона є основним елементом при побудові багатьох алгоритмів багатовимірної статистики, наприклад, у факторному аналізі.

Коефіцієнт кореляції Пірсона можна застосовувати для перевірки гіпотези про значущість зв'язку. При аналізі коефіцієнта кореляції оцінюють його значення. Якщо він дорівнює нулю для генеральної сукупності, це зовсім не означає, що він також дорівнює нулю для вибірки. Навпаки він буде обов'язково відхилятися від справжнього значення, але що

більше відхилення, то менш імовірно воно за даного обсягу вибірки. Для вибіркового коефіцієнта кореляції r будують t-статистику, яка має розподіл Стьюдента з $n-2$ ступенями вільності, та обчислюють за формулою:

$$t = r \cdot \frac{\sqrt{n-2}}{\sqrt{1-r^2}}. \quad (\mathbf{8.5})$$

Існує тест для визначення значущості коефіцієнта кореляції (наскільки є впевненість, що він відрізняється від нуля). Перевіряють нульову гіпотезу H_0, згідно з якою істинне значення коефіцієнта кореляції дорівнює нулю: $r=0$. Альтернативною гіпотезою H_1 є гіпотеза про те, що $r \neq 0$. Порівнюючи обчислене за вибіркою значення t-статистики за (**8.5**) із критичними точками, які визначають за таблицями розподілу Стьюдента, можна прийняти або відхилити нульову гіпотезу. Для двосторонньої критичної області заданого рівня значущості α критичну точку t_{kp} знаходять із таблиці як $t_{kp} = t_{\alpha/2,m}$ для кількості ступенів вільності $m = n-2$. Якщо $|t| \le t_{kp}$, то гіпотезу H_0 приймають, якщо $|t| > t_{kp}$ – гіпотезу H_0 відхиляють.

Рівень значущості $\alpha = 1-p$ – це ймовірність припуститися помилки першого роду, тобто відхилити нульову гіпотезу, коли вона є насправді вірною. У цьому випадку існує ймовірність вважати кореляцію ненульовою у той час, коли вона насправді дорівнює нулю.

Число α задають наперед, і найчастіше його обирають рівним 0,1; 0,05; 0,01; 0,001. Значення $\alpha = 0{,}05$ означає, що ймовірність припуститися помилки першого роду є малою: є ризик її припуститися у п'яти випадках зі 100.

Зазвичай допустимим рівнем значущості є 0,05 і менше (менше 0,10 – не дуже суворий критерій, менше 0,01 – більш суворий). Тоді кореляцію вважають статистично значущою. Значущість коефіцієнта кореляції тим більша, чим сильніше його розраховане значення відрізняється від нуля та чим більшою є вибірка спостережень, за якою він розрахований.

Про множинну кореляцію йдеться у випадку, за якого певна ознака може бути пов'язана не з однією, а із сукупністю

кількох інших ознак. У реальних дослідженнях можлива ситуація, за якої на певну ознаку може впливати не одна, а кілька інших. У таких випадках парні показники кореляції даватимуть неправильну інформацію щодо наявності зв'язку між відповідними показниками, оскільки ці їх значення викривлятимуться невраховуваними ознаками.

Якщо йдеться про взаємозв'язок однієї змінної Y з кількома m змінними $X_i, i = 1,2,...,m,$ то використовують коефіцієнт множинної кореляції R як міру щільності взаємозв'язку однієї змінної Y (показника) із сукупністю інших змінних (факторів) $X_i, i = 1,2,...,m.$ Його обчислюють за формулою:

$$R = \frac{\frac{1}{n}\sum_{k=1}^{n}(y_k - \bar{y})(\hat{y}_k - \bar{\hat{y}})}{\sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2} \cdot \sqrt{\frac{1}{n}\sum_{m=1}^{n}(\hat{y}_m - \bar{\hat{y}})^2}}. \qquad \textbf{(8.6)}$$

Квадрат коефіцієнта множинної кореляції R^2 називають *коефіцієнтом детермінації/coefficient of determination*. Він показує, скільки відсотків варіації залежної змінної Y визначає варіація незалежних змінних $X_i, i = 1,2,...,m.$ Коефіцієнт детермінації набуває значення $R^2 \in [0;1].$ Що ближче його значення до одиниці, то суттєвішим є зв'язок між змінними в моделі. При цьому нульове значення коефіцієнта детермінації відповідає відсутності зв'язку, а його рівність одиниці – наявності строго функціонального зв'язку.

Взаємозв'язок між двома змінними, що обчислені за фіксованих значень усіх інших змінних, характеризується коефіцієнтом частинної кореляції. Тоді записують кореляційну матрицю:

$$r = \begin{pmatrix} r_{yy} & r_{yx_1} & ... & r_{yx_m} \\ r_{x_1 y} & r_{x_1 x_1} & ... & r_{x_1 x_m} \\ ... & ... & ... & ... \\ r_{x_m y} & r_{x_m x_1} & ... & r_{x_m x_m} \end{pmatrix}, \qquad \textbf{(8.7)}$$

де r_{yx_i} – парні коефіцієнти кореляції між залежною та незалежними змінними, а $r_{x_k x_i}, k,i = 1,2,...,m$ – парні коефіцієнти кореляції між незалежними змінними.

Якщо досліджувані ознаки задовольняють багатовимірний нормальний розподіл, то частинний коефіцієнт кореляції між двома ознаками *i* та *j* за фіксованих значень інших ознак розраховують за формулою:

$$r_{X_i X_j} = -\frac{R_{ij}}{\sqrt{R_{ii} \cdot R_{jj}}}, \quad (8.8)$$

де R_{ik} – алгебраїчне доповнення для елемента $r_{x_k x_i}$ у кореляційній матриці.

Кореляційний аналіз часто використовують не як самостійний вид аналізу, а як метод первинного аналізу даних, так само як і розрахунок описових статистик, визначення розподілу даних і статистичних викидів. Він дозволяє оцінити структуру даних, попередньо визначити можливі зв'язки між змінними, які варті детальнішого аналізу за допомогою складніших методів дослідження (напр., регресійного аналізу). Показники із високими значеннями коефіцієнта кореляції із залежною змінною варто далі використовувати для детальнішого аналізу.

8.2. Кореляційний аналіз у Microsoft Excel

У Microsoft Excel є можливість скористуватися опцією *Кореляція/Correlation* у надбудові *Пакет аналізу/Data Analysis* для побудови кореляційної матриці для кількох змінних. Для зручності до вхідного діапазону включають комірки з назвами змінних, але при цьому мають бути включені опції *Мітки в першому рядку/Labels in first row* та *Групування за стовпичками/Grouped by Columns*, оскільки змінні розташовані за стовпчиками, а не рядками. Нажаль, неможна автоматично порахувати рівень значущості коефіцієнтів кореляції у цій надбудові. На рис. 8.4-8.6 показано вхідні дані, діалогове вікно аналізу та результат.

Або можна скористатися окремою функцією:

=КОРРЕЛ(діапазон першої змінної;діапазон другої змінної)

Повертає звичайний коефіцієнт кореляції Пірсона між двома змінними, значення яких прописані у двох відповідних діа-

пазонах. Діапазони мають містити однакову кількість комірок (бути одного розміру).

Приклад: =КОРРЕЛ(A2:A25;B2:B25)

	A	B	C	D
1	Exports	Imports	FDI	External Debt
2	12	12	3	30
3	22	34	6	38
4	33	40	5	44
5	21	22	2	46
6	34	25	4	43
7	56	45	2	33
8	43	47	3	35

Рис. 8.4

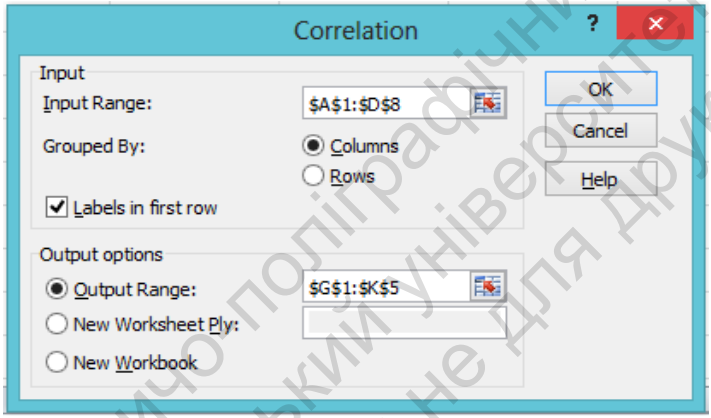


Рис. 8.5

	G	H	I	J	K
		Exports	*Imports*	*FDI*	*External Debt*
Exports		1			
Imports		0.83177	1		
FDI		-0.27694	0.09706	1	
External D		-0.1177	-0.03681	0.222801	1

Рис. 8.6

Для визначення рівня значущості коефіцієнта кореляції потрібно спочатку скористатися (**8.5**) для розрахунку *t*-статистики, а потім – функцією:

=T.DIST.2T(*t*-статистика;ступінь вільності)

Ступінь свободи для парної кореляції розраховують як кількість спостережень мінус 2 (тобто кількість змінних). Функція повертає рівень значущості (*p*-value) коефіцієнта

кореляції. Наприклад, якщо він менше 0,05, то можна принаймні на 95 % бути впевненим, що справжня кореляція у генеральній сукупності – не нульова, отже лінійний (або наближений до нього) зв'язок між змінними існує.

8.3. Подолання проблем у застосуванні кореляційного аналізу світової економіки

Кореляційний аналіз може бути пов'язаний з деякими труднощами.

1. Нелінійний зв'язок. За певних видів нелінійних зв'язків (напр., квадратична залежність), навіть якщо вони сильні, коефіцієнт кореляції може бути достатньо малим або навіть дорівнювати нулю. За інших видів нелінійної залежності (напр., експоненціальна) коефіцієнт кореляції буде високим, хоча не наближатиметься до 1. Розв'язанням цієї проблеми є трансформація змінних: наприклад, розрахунок кореляції між логарифмом однієї змінної та іншою змінною; між однією змінною та кубом іншої змінної; між однією змінною та квадратом відхилення від середньої іншої змінної тощо. Тобто спочатку визначають функцію, яка найкраще відображає залежність, а потім роблять перетворення з однією чи обома змінними. У визначенні такої функції може допомогти діаграма розсіювання.

Наведемо приклад. Є змінні *Y* та *X*, між якими майже відсутній лінійний зв'язок. При користуванні лише звичайною кореляцією Пірсона для попереднього відбору змінних для подальшого аналізу на цьому й зупинилися б. Але з діаграми розсіювання у прикладі на рис. 8.7 бачимо явний нелінійний зв'язок.

Тому додатково до коефіцієнта кореляції Пірсона між *Y* та *X* можна розрахувати кореляцію між *Y* квадратом відхилення *X* від свого середнього значення (рис. 8.8).

Якщо щонайменше один із двох коефіцієнтів кореляції є достатньо великим, то навіть без діаграми розсіювання ясно, що зв'язок між *Y* та *X* існує, і його варто далі аналізувати детальніше.

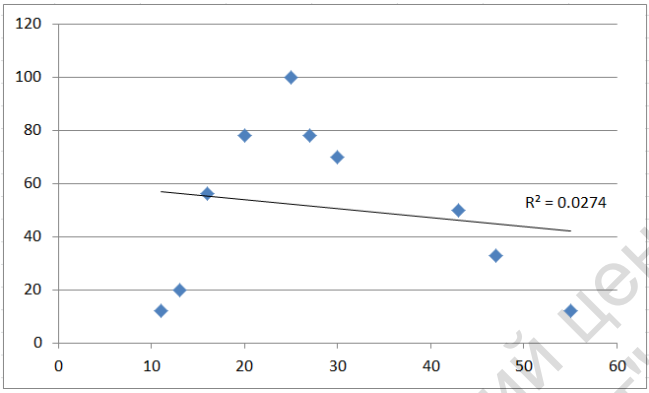


Рис. 8.7

C2 =POWER((B2-AVERAGE($B$2:$B$11));2)

	A	B	C	D	E
1	Y	X	(X-avX)2		
2	12	11	313.29	Correl1	-0.16553655
3	20	13	246.49	Correl2	-0.83593618
4	56	16	161.29		
5	78	20	75.69		
6	100	25	13.69		
7	78	27	2.89		
8	70	30	1.69		
9	50	43	204.49		
10	33	47	334.89		
11	12	55	691.69		

Рис. 8.8

Другий варіант – використання непараметричних аналогів звичайного коефіцієнта кореляції Пірсона (напр., коефіцієнта кореляції Спірмена).

Третій варіант – перетворення однієї зі змінних на категоріальну (розбивають усі значення на групи, напр., великі, середні, малі) та визначення, як відрізняються середні значення другої змінної у цих групах за допомогою методів аналізу середніх чи дисперсійного аналізу.

2. Економічні дані, як і будь-які інші, можуть мати **грубі помилки**/аномальні значення/викиди/впливові спостереження. Вони можуть штучно завищувати чи занижувати значення коефіцієнта кореляції або навіть привести до того, що кореляція матиме не той знак.

На рис. 8.9 наведено приклад: розраховано кореляцію для однієї і тієї самої пари змінних з викидом і без нього. Видно, що кореляція з викидом помірна та негативна (–0,27), а без нього – сильна та позитивна (0,88). За такої ситуації не може йтися про надійні результати кореляційного аналізу.

X	Y		X	Y
10	1		10	1
12	2		12	2
15	3		15	3
34	7		34	7
30	10		30	10
24	8		24	8
17	2		17	2
9	2		9	2
181	***0***	Викид		
Кореляція	-0.27928		Кореляція	0.88872

Рис. 8.9

Для подолання проблеми викидів можна запропонувати такі методи:

- обрати велику вибірку, де викиди вже не створюють великого впливу;
- провести кореляційний аналіз з викидами та без них, порівнюючи результати (напр., ураховувати середній коефіцієнт кореляції як кращу оцінку справжньої кореляції для генеральної сукупності);
- вилучити із аналізу викиди, якщо вони є результатом помилки чи унікальної події, яка у майбутньому навряд чи повторюватиметься.

3. **Кореляції у неоднорідних групах**. До прикладу, коли велике значення коефіцієнта кореляції спричинене даними з цілком неоднорідних груп. Припустимо, є дві групи країн: високорозвинені та слаборозвинені. При цьому між двома показниками (зростання ВВП та інфляція) у високорозвинених країнах існує слабо позитивний зв'язок, а у слаборозвинених – сильний негативний. Якщо розрахувати коефіцієнт кореляції для всієї сукупності країн, то він буде ближче до нейтрального. Але цей результат практично не стане корис-

ним. Для слаборозвинених країн буде недооцінена небезпека інфляції, а для розвинених країн – дефляції.

Проілюструвати цю проблему можна умовним прикладом (рис. 8.10). Для всіх країн кореляція дуже слабка (0,06). Але якщо виділити дві групи країн, то побачимо сильну залежність, але неоднакову за групами країн.

Приріст ВВП	Інфляція	Дохід	
-1	0	Високий	
1	1	Високий	
1.5	2	Високий	
3	2.5	Високий	Кореляція для країн з високим доходом
3	4	Високий	0.92
9	3	Низький	
10	4	Низький	
6	9	Низький	
5	20	Низький	Кореляція для країн з низьким доходом
2	25	Низький	-0.95
			Кореляція для всіх країн
			0.06

Рис. 8.10

Розв'язанням проблеми неоднорідних груп може бути використання діаграм розсіювання для ідентифікації різних груп спостережень за іншою змінною, або розрахунок коефіцієнтів кореляції для різних груп спостережень без побудови діаграми розсіювання. Використання на додаток кластерного аналізу для визначення різних груп спостережень також може бути корисним.

4. Випадкові кореляції за масового аналізу. Ідеться про *хибні кореляції*/*Spurious correlations*. Якщо будують кореляційну матрицю (таблицю, де рядкам і стовпчикам відповідають змінні, на перетині кожного рядка та стовпчика перебуває коефіцієнт кореляції між відповідною парою змінних) для багатьох змінних, то, імовірно, певна частина високих кореляцій буде високою лише випадково, коли насправді (у генеральній сукупності) такої кореляції між відповідними парами змінних не буде. Наприклад, коефіцієнт кореляції, значущий на рівні 0,05, може бути випадковим з імовірністю 5 %. Тому, наприклад, якщо у кореляційній матриці 60 кое-

фіцієнтів кореляцій із рівнем значущості 0,05, то ймовірно, що три серед них відрізняються від нуля лише випадково через збіг обставин.

Саме тому побудова кореляційної матриці – це лише попередній крок при аналізі зв'язку між змінними, а не остаточний доказ наявності зв'язку. Аналіз, який проводять без теоретичного обґрунтування, називають розвідувальним/ *exploratory*. Якщо виявлено високий коефіцієнт кореляції між змінними, але йому немає теоретичного пояснення, то до такого результату варто ставитися обережно. Доцільно перевірити наявність кореляції на іншій вибірці чи на підвибірках цієї самої вибірки.

5. Відсутні дані. Звичайним способом розв'язання проблеми відсутності даних за кореляційного аналізу є попарне *видалення спостережень*/*pairwise method* з відсутніми даними (оскільки повне видалення спостережень, за якими відсутнє значення принаймні щодо *однієї змінної*/*casewise method* може призвести до суттєвого зменшення вибірки). Але може виникнути небезпека систематичного місцезнаходження відсутніх даних (спостереження з відсутніми даними залежать від однієї чи кількох змінних). У результаті середнє значення певної змінної *A*, яке використовують для розрахунку її кореляції зі змінною *B*, може відрізнятися від середнього значення *A*, яке використовують для розрахунку її кореляції зі змінною *C*. Тому коефіцієнти кореляції між різними змінними можуть бути непорівнюваними (зважаючи також на те, що вони обраховані за різною кількістю спостережень). Ще один спосіб (заміна відсутніх даних на середні значення змінних) призводить до заниження кореляції і дисперсії. Таким чином, оптимальним все ж є попарне видалення відсутніх даних. Але можна спробувати всі три способи та порівняти результати.

6. Структурні зміни. Із часом зв'язки між показниками можуть змінюватися. Наприклад, існують періоди, коли курс долара позитивно корелює з цінами на нафту, і періоди, коли – негативно. Це пояснює, яким чином нафтовидобувні країни витрачають свої доларові надходження від нафти: вкладають у доларові активи чи купують імпортні товари,

розраховуючись в інших валютах. Тому розрахований коефіцієнт кореляції в один період часу може не мати практичної корисності в інший період часу.

Варіантом розв'язання цієї проблеми є використання плинних коефіцієнтів кореляції (за аналогією з плинною середньою), наприклад, за 1990-2000, 1991-2001, 1992-2002 рр. тощо, або поділ усього досліджуваного періоду на підперіоди, наприклад, 1971-2010 рр. на 1971-1980, 1981-1990, 1991-2000, 2001-2010, 2011-2020 рр. Це дозволить простежити зміну коефіцієнта кореляції з часом та оцінити його надійність. Наприклад, візьмемо дані за поточним рахунком (у % ВВП) і приростом реального ВВП (у %) за даними *World Economic Outlook* по Україні за 1995-2010 рр. Розраховують плинну кореляцію та кореляції за три підперіоди (рис. 8.11 – вхідні дані та результат і рис. 8.12 – формули). Видно, що кореляція була доволі стійкою в усі роки, крім останніх, коли вона навіть змінила знак.

	A	B	C	D	E
1	Рік	Приріст ВВП	Поточний рахунок / ВВП	Плинні кореляції	Кореляції за 3
2	1995	-12.151	-1.152	(за 5 років)	підперіоди
3	1996	-10.044	-1.184		
4	1997	-2.988	-1.335		
5	1998	-1.949	-1.296		
6	1999	-0.224	1.658	0.51	1995-2000 рр.
7	2000	5.85	1.481	0.71	0.71
8	2001	9.046	1.402	0.73	
9	2002	5.253	3.173	0.58	
10	2003	9.595	2.891	0.21	
11	2004	12.019	6.909	0.65	2001-2005 рр.
12	2005	2.944	2.531	0.54	0.54
13	2006	7.534	-1.617	0.44	
14	2007	7.518	-5.272	0.37	
15	2008	1.945	-12.763	0.65	
16	2009	-14.462	-1.732	-0.11	2006-2010 рр.
17	2010	4.187	-2.884	-0.18	-0.18
18					
19			Кореляція за весь період		0.26

Рис. 8.11

D	E
Плинні кореляції (за 5 років)	Кореляції за 3 підперіоди
=CORREL(B2:B6;C2:C6)	1995-2000 рр.
=CORREL(B3:B7;C3:C7)	=CORREL(B2:B7;C2:C7)
=CORREL(B4:B8;C4:C8)	
=CORREL(B5:B9;C5:C9)	
=CORREL(B6:B10;C6:C10)	
=CORREL(B7:B11;C7:C11)	2001-2005 рр.
=CORREL(B8:B12;C8:C12)	=CORREL(B8:B12;C8:C12)
=CORREL(B9:B13;C9:C13)	
=CORREL(B10:B14;C10:C14)	
=CORREL(B11:B15;C11:C15)	
=CORREL(B12:B16;C12:C16)	2006-2010 рр.
=CORREL(B13:B17;C13:C17)	=CORREL(B13:B17;C13:C17)
	=CORREL(B2:B17;C2:C17)

Рис. 8.12

Іншим способом є використання зважених спостережень. Якщо потрібно дізнатися, як корелює зростання ВВП різних країн, останнім періодам можна надати більшої ваги. Наприклад, 1990 р. надаватимемо ваги 10, 1991 – 11,..., 2020 – 40. Нижче подано приклад (рис. 8.13), коли дані за чотири роки зважені так, що 2007 р. надана вага 2 (тобто перше спостереження враховано двічі), 2008 – вага 3 (тричі), 2009 – вага 4, 2010 р. – вага 5. Видно, що кореляції для незважених і зважених даних трохи відрізняються. Але якби було взято триваліший період часу (кілька десятиліть), то такі кореляції відрізнялися би помітніше.

Зважування дозволяє знайти компроміс між мотивацією включити якомога довший період (забезпечити велику кількість спостережень) і мотивацією включити тільки останні дані (щоб застарілі закономірності не впливали на результат, який буде використано для рекомендацій щодо рішень у майбутньому). З іншого боку, програмне забезпечення автоматично рахуватиме *t*-статистику за завищеною кількістю спостережень (у нашому прикладі за 14, а не справжніми чотирма спостереженнями), отже рівень значущості кореляції буде штучно наближений до 0. Тому необхідно власноруч перерахувати рівень значущості, якщо кореляція порахована на основі зважених даних.

	A	B	C	D	E	F	G
1	Рік	Приріст ВВП	Поточний рахунок / ВВП		Рік	Приріст ВВП	Поточний рахунок / ВВП
2	2007	7.518	-5.272		2007	7.518	-5.272
3	2008	1.945	-12.763		2007	7.518	-5.272
4	2009	-14.462	-1.732		2008	1.945	-12.763
5	2010	4.187	-2.884		2008	1.945	-12.763
6					2008	1.945	-12.763
7					2009	-14.462	-1.732
8					2009	-14.462	-1.732
9					2009	-14.462	-1.732
10					2009	-14.462	-1.732
11					2010	4.187	-2.884
12					2010	4.187	-2.884
13					2010	4.187	-2.884
14					2010	4.187	-2.884
15					2010	4.187	-2.884
16							
17	Кореляція	-0.3856			Кореляція	-0.3860	
18	(незважені дані)				(зважені д		

Рис. 8.13

7. Складність визначення причини та наслідку. Сам по собі коефіцієнт кореляції не відповідає на запитання: коливання якої з двох змінних спричиняють коливання іншої, наприклад, інфляція зміни валютного курсу, чи навпаки. Розв'язанням цієї проблеми є використання лагів. Припустимо, існує кореляція між інфляцією у поточному періоді та зміною валютного курсу у наступному періоді (чи через період, через два періоди тощо). Це, імовірно, означає, що інфляція спричиняє зміни валютного курсу. Проте без теоретичних міркувань неможливо визначити, що є причиною, а що – наслідком у рамках одного й того самого періоду. Спеціальним методом визначення причинних зв'язків є тест Грейнджера (доступний у програмі *E-views*).

На рис. 8.14 (вхідні дані та результати) і рис. 8.15 (формули) показано приклад у *Microsoft Excel* щодо розрахунку кореляції із лагами між приростом ВВП і поточним рахунком платіжного балансу. Кореляція з нульовим лагом дорівнює 0,26 (помірна позитивна кореляція), але напрям такого короткострокового впливу невідомий. Кореляції, де значенням поточного рахунку передують значення приросту ВВП із

лагом один-три роки (0,16; –0,01; –0,12), є дуже слабкими та, імовірно, незначущими, отже середньостроковий вплив приросту ВВП на поточний рахунок не підтверджується. Кореляції, де значенням приросту ВВП передують значення поточного рахунку із лагом один-три роки (0,70; 0,46; 0,33), навпаки, є достатньо суттєвими, отже з високою ймовірністю можна сказати, що поточний рахунок платіжного балансу впливає на ВВП, і найсильніше цей вплив виявляється за рік (кореляція 0,70 найвища).

	A	B	C	D	E	F
1		Y	X			
2	Рік	Приріст ВВП	Поточний рахунок / ВВП			
3	1995	-12.151	-1.152			
4	1996	-10.044	-1.184		Кореляція (Xt,Yt)	0.26
5	1997	-2.988	-1.335		Кореляція (Xt,Yt-1)	0.16
6	1998	-1.949	-1.296		Кореляція (Xt,Yt-2)	-0.01
7	1999	-0.224	1.658		Кореляція (Xt,Yt-3)	-0.12
8	2000	5.85	1.481		Кореляція (Xt-1,Yt)	0.70
9	2001	9.046	1.402		Кореляція (Xt-2,Yt)	0.46
10	2002	5.253	3.173		Кореляція (Xt-3,Yt)	0.33
11	2003	9.595	2.891			
12	2004	12.019	6.909			
13	2005	2.944	2.531			
14	2006	7.534	-1.617			
15	2007	7.518	-5.272			
16	2008	1.945	-12.763			
17	2009	-14.462	-1.732			
18	2010	4.187	-2.884			

Рис. 8.14

E	F
Кореляція (Xt,Yt)	=CORREL(B3:B18;C3:C18)
Кореляція (Xt,Yt-1)	=CORREL(B3:B17;C4:C18)
Кореляція (Xt,Yt-2)	=CORREL(B3:B16;C5:C18)
Кореляція (Xt,Yt-3)	=CORREL(B3:B15;C6:C18)
Кореляція (Xt-1,Yt)	=CORREL(C3:C17;B4:B18)
Кореляція (Xt-2,Yt)	=CORREL(C3:C16;B5:B18)
Кореляція (Xt-3,Yt)	=CORREL(C3:C15;B6:B18)

Рис. 8.15

8. Спільна причина. Несправжня кореляція може бути результатом дії спільної причини, а також ще одним поясненням явища хибної кореляції. Наприклад, можуть корелювати зміни валютних резервів двох країн, які розташовані у різних кінцях світу, не мають суттєвих прямих чи опосередкованих торговельних чи фінансових зв'язків, які б могли пояснити достатній вплив однієї країни на іншу. Насправді ці коливання викликані коливаннями глобальної схильністю інвесторів до ризику. У цьому випадку важливим є теоретичне обґрунтування зв'язків і виявлення ймовірної спільної причини. Далі можна поділити всі спостереження на дві групи: в умовах великої та низької схильності інвесторів до ризику. У межах кожної із груп, можливо, кореляція між змінами валютних резервів обох країн буде майже відсутня. Інший спосіб розв'язання проблеми – використання часткової кореляції (із поправкою на вплив інших змінних) або регресійного аналізу з контрольними змінними.

9. Неаддитивність. Коефіцієнти кореляції не є адитивними. Але коефіцієнти детермінації вже є адитивними. Хоча в цьому випадку може діяти спотворюючий ефект мультиколінеарності. Тому множинний коефіцієнт детермінації між Y та лінійною комбінацією змінних X і W дорівнює сумі коефіцієнтів детермінації між Y та X та між Y та W тільки, якщо кореляція між X і W дорівнює 0.

8.4. Непараметричні методи кореляційного аналізу зв'язку якісних змінних

Коефіцієнт кореляції Пірсона є параметричним критерієм, оскільки припускає, що змінні кількісні, їх вимірюють в інтервальній шкалі, шкалі відношень, або в абсолютній шкалі та мають нормальний розподіл. Непараметричні аналоги коефіцієнта кореляції Пірсона дозволяють аналізувати зв'язки також між змінними, які вимірюють у метричній шкалі, якщо їх розподіл невідомий чи не є нормальним, або у порядковій шкалі. У цьому випадку використовують так

звані рангові коефіцієнти кореляції. Це розповсюджено при обробці даних соціологічних досліджень (анкет), медичних обстежень, рейтингів, експертних оцінок тощо, тобто там, де ознаки вимірюють за допомогою номінальної і порядкової шкали (напр., "стать", "соціально-економічний статус", "діагноз" тощо).

До основи непараметричних методів рангової кореляції покладено принцип нумерації значень статистичного ряду. Кожній одиниці сукупності надають порядковий номер за величиною значення окремої ознаки – *ранг* (натуральне число 1, 2, 3, ...). Ранжування, тобто процедуру упорядкування об'єктів вивчення на основі надання переваг, проводять за кожною ознакою окремо. При ранжуванні значень факторної і результативної ознак слід використовувати один принцип – або від менших значень до більших, або навпаки. Кількість рангів дорівнює обсягу сукупності. Зі збільшенням обсягу ступінь "розпізнаваності" елементів зменшується, тому рангові оцінки щільності зв'язку доцільно використовувати для сукупностей невеликого обсягу. Якщо змінні кількісні, але їх розподіл не відповідає нормальному розподілу, також рекомендують використовувати рангові кореляції. Використовують такі аналоги:

1. *Ранговий коефіцієнт кореляції Спірмена/Spearman rank correlation* (ρ читають "ро") – є прямим аналогом кореляції Пірсона, але розрахунки проводять не за абсолютними значеннями показників, а за їх рангами (порядковими номерами у переліках, що відсортовані за відповідними змінними) за формулою:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}, \qquad (\mathbf{8.9})$$

де $d = R_X - R_Y$ – різниці між рангами змінних, які характеризують об'єкти вибірки, n – обсяг вибірки.

2. *Коефіцієнт кореляції рангів Кендала/Kendall Tau* (τ читають "тау") ґрунтується не на різницях рангів, як попередній коефіцієнт, а на відхиленнях кількості рангів, більшої від порівнюваної s_1, і меншої від порівнювальної s_2, за результативною ознакою *Y*:

$$\tau = \frac{2\sum_{k=1}^{n}(s_1 - s_2)}{n(n-1)}. \qquad (8.10)$$

Коефіцієнти Спірмена та Кендала можуть набувати значень від –1 до +1, тобто характеризують як щільність, так і напрям зв'язку. Якщо ρ і τ дорівнюють +1, то йдеться про повний прямий зв'язок між ознаками, якщо ρ і τ дорівнюють –1, то про повний зворотній зв'язок між ними.

3. *Гамма/gamma-статистика* краща за статистики ρ – Спірмена або τ – Кендала, якщо в даних є багато значень, які збігаються. Гамма-статистика – це ймовірність, точніше, різниця між імовірністю того, що ранговий порядок двох змінних збігається та ймовірністю того, що він не збігається, поділена на вираз: 1 мінус імовірність їх збігів.

Для даних, які вимірюють у номінальній шкалі, непараметричним аналогами можливо вважати χ^2-квадрат, φ-коефіцієнт, точний критерій Фішера, коефіцієнт конкордації Кендалла.

Розглянемо детально *кореляційний аналіз порядкових ознак*. Під *ранговою кореляцією* розуміють статистичний зв'язок між порядковими ознаками. Вихідні дані зазвичай подають у вигляді табл. 8.1, де елемент x_{ik} є рангом *i*-го об'єкта за *k*-ю властивістю.

Таблиця 8.1

Порядковий номер об'єкта	**Порядковий номер досліджуваної ознаки**						
	1	**2**	**...**	***k***	**...**	**...**	***p***
1	x_{11}	x_{12}	...	x_{1k}	...	...	x_{1p}
2	x_{21}	x_{22}	...	x_{2k}	...	...	x_{2p}
...	...	...	...	...	...	...	...
i	x_{i1}	x_{i2}	...	x_{ik}	...	...	x_{ip}
...	...	...	...	...	...	...	...
n	x_{n1}	x_{n2}	...	x_{nk}	...	...	x_{np}

Завданнями аналізу у цьому випадку можуть бути вивчення структури досліджуваних об'єктів; перевірка сукупної узгодженості ознак та умовне ранжирування об'єктів за ступенем щільності зв'язку кожної з них з іншими ознаками; побудова єдиного групового впорядкування об'єктів (задача регресії на порядкових змінних).

У першому випадку кожну послідовність впорядкованих за *k*-ю ознакою *n* об'єктів подають як точку:

$$X^{(k)} = \left(x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)} \right), k = 0,1,...,p$$

у *n*-вимірному просторі ознак.

Найхарактернішими типами структури є:

1) аналізовані точки, що рівномірно розкидані всією областю їх можливих значень. Це означає відсутність будь-якого зв'язку між досліджуваними ознаками;

2) частина точок утворює ядро (кластер) із точок, що розташовані близько одна до одної, а інші – випадково розкидані навколо цього ядра. Це відповідає існуванню підмножини узгоджених ознак;

3) аналізовані точки утворюють кілька кластерів, що розташовані відносно далеко один від одного. Це відповідає наявності кількох таких підмножин, коли існує істотний статистичний зв'язок між ознаками, які належать до однієї і тієї самої підмножини, і не існує значущого зв'язку між ознаками, які належать до різних підмножин.

Прикладом завдання другого типу є визначення узгодженості думок групи експертів із наступним впорядкуванням їх за рівнем компетентності. Для цього розраховують коефіцієнти конкордації для різних сукупностей досліджуваних змінних. Вирішення завдань третього типу зводять до побудови такого впорядкування, яке б у певному значенні було найближчим до кожного з наданих упорядкувань досліджуваних ознак. Для цього часто застосовують середнє арифметичне або медіану наявних базових рангів. Це можна розглядати як задачу найкращого у певному розумінні відновлення невідомого ранжирування за наявними емпіричними даними, що зумовлює можливість її розгляду як задачі регресії.

Коефіцієнт рангової кореляції Спірмена (*показник кореляції рангів Спірмена, коефіцієнт кореляції рангів*)[137], який використовують при дослідженні зв'язку між рядами даних, що виміряні за порядковою шкалою, а також для кількісних даних, але зазвичай це буває недоцільним. У найпростішому випадку досліджувані об'єкти класифікують за двома ознаками. Наприклад, можна спочатку впорядкувати групу студентів за їх здібностями до математики, а потім – до іноземних мов. Місця, які *i*-й студент посяде в обох списках, – його ранги r_i та s_i. Якщо досліджувані ознаки взаємопов'язані, то послідовність рангів $r_1, r_2, ..., r_n$ певною мірою корелює із послідовністю рангів $s_1, s_2,..., s_n$.

Ступінь близькості двох послідовностей відображує величина:

$$S_\rho = \sum_{i=1}^{n} (r_i - s_i)^2. \qquad \textbf{(8.11)}$$

Якщо для нумерації об'єктів попередньо впорядкувати їх за однією з ознак, наприклад за зростанням рангів r_i, то (**8.11**) можна записати так:

$$S_\rho = \sum_{k=1}^{n} \left(k - s_k\right)^2. \qquad \textbf{(8.12)}$$

Величина S_ρ набуде найменшого можливого значення $S_\rho = 0$ тоді й тільки тоді, коли послідовності повністю збігатимуться. Найбільше можливе значення:

$$S_\rho = \frac{1}{3}(n^3 - n)$$

відповідає випадку, за якого послідовності є цілковито протилежними, тобто для будь-яких *i*, *j* з нерівності $r_i > r_j$ випливає $s_i < s_j$, і послідовності рангів першої ознаки $r_i = \{1, 2, ..., n\}$ відповідає послідовність рангів другої $\{n, n-1, ..., 1\}$ $s_i = s_j$. Величину S_ρ незручно застосовувати як міру зв'язку, оскільки на її значення впливає кількість пар варіант досліджуваних рядів *n*.

З огляду на це, як міру зв'язку використовують коефіцієнт рангової кореляції Спірмена, значення якого розраховують за формулою:

[137] Запропоновано британським психологом Спірменом у 1904 р.

$$S_{\rho} = 1 - \frac{6(S_{\rho} + B_x + B_y)}{n^3 - n}, \qquad \textbf{(8.13)}$$

де B_x, B_y – поправки на об'єднання рангів у відповідних рядах, які обчислюють за формулою:

$$B_i = \frac{1}{12}\sum_{i=1}^{m} n_i(n_i^2 - 1), \qquad \textbf{(8.14)}$$

де m – кількість груп об'єднаних рангів у вибірці; n_i – кількість рангів у i-й групі.

Значення коефіцієнта можуть змінюватися в межах від –1 (повна протилежність послідовностей рангів) до +1 (повний збіг послідовностей рангів).

Коефіцієнт рангової кореляції Спірмена можна застосовувати як показник некорельованості вибірок. У цьому випадку розраховують величину:

$$t_p = \sqrt{n-2}\frac{\rho_S}{\sqrt{1-\rho_S^2}}. \qquad \textbf{(8.15)}$$

Для великих за обсягом вибірок ($n > 50$) статистика цього критерію наближається до розподілу Стьюдента з (n – 2) ступенями вільності. Статистика $\sqrt{n-2}\cdot\rho$ для великих вибірок наближається до стандартного нормального розподілу.

Інший підхід використовує як міру подібності двох вибірок мінімальну кількість перестановок сусідніх об'єктів, потрібну для переведення послідовності рангів однієї вибірки до послідовності рангів – іншої. Можна показати, що вона дорівнює кількості інверсій в однієї з цих послідовностей у випадку, коли інша послідовність впорядкована за зростанням.

Наприклад, $n = 4$, послідовність r_i упорядкована за зростанням, а s_i ={4, 3, 1, 2}. Інверсіями є 4 > 3; 4 >1; 4 > 2; 3 >1; 3 > 2. Їх кількість $K = 5$. Найменше можливе значення кількості інверсій $K = 0$ відповідає повному збігу рангових послідовностей, а найбільше $K = \frac{n(n-1)}{2}$ – їх повній протилежності.

Як і в попередньому випадку, кількість інверсій залежить від обсягу вибірки та є незручною для застосування як показника кореляції. Для цього використовують *коефіцієнт ран-*

гової кореляції Кендалла (*коефіцієнт кореляції рангів, ранговий коефіцієнт кореляції*)[138]. Коефіцієнт розраховують за формулою:

$$\tau = 1 - \frac{2K}{\sqrt{\left(\frac{n(n-1)}{2} - B_x\right)\left(\frac{n(n-1)}{2} - B_y\right)}}, \qquad \textbf{(8.16)}$$

де r_j, s_i – масиви рангів аналізованих рядів; n – кількість пар варіант у них. B_x, B_y – поправки на об'єднання рангів у відповідних рядах, які обчислюють за формулою:

$$B_i = \frac{1}{2}\sum_{i=1}^{m} n_i(n_i - 1), \qquad \textbf{(8.17)}$$

де m – кількість груп об'єднаних рангів у вибірці; n_i – кількість рангів у i-й групі.

Для коефіцієнта рангової кореляції Кендалла у випадку великих вибірок статистика:

$$\tau = \sqrt{\frac{9n(n-1)}{(2n+5)}} \qquad \textbf{(8.18)}$$

має розподіл, близький до стандартного нормального закону.

Коефіцієнт рангової кореляції Кендалла призначено для визначення сили кореляційного зв'язку між двома рядами даних за тих самих умов, що й коефіцієнт рангової кореляції Спірмена. Як і для коефіцієнта Спірмена, його значення можуть змінюватися в межах від –1 (повна протилежність послідовностей рангів) до +1 (повний збіг послідовностей рангів).

Слід зазначити, що обчислення коефіцієнта Кендалла є більш трудомістким, але він має низку переваг, порівняно із коефіцієнтом Спірмена. Основними з них є:

- кращий рівень вивченості його статистичних властивостей, зокрема його вибіркового розподілу;
- можливість його застосування для визначення частинної кореляції;
- більша зручність перерахунку при додаванні нових даних.

[138] Запропоновано британським статистиком Кендаллом у 1938 р.

8.5. Кореляційний аналіз номінальних ознак

Типовою ситуацією, за якої необхідна перевірка зв'язку між номінальними ознаками, є обробка результатів соціологічних досліджень, що можуть містити такі комбінації ознак, як освіта, стать, професія, підтримка певної політичної партії, регіон проживання тощо.

При дослідженні зв'язків між *категоризованими ознаками* вихідні дані подають у вигляді таблиці спряженості (табл. 8.2 – спряженості категоризованих ознак). До категоризованих зараховують номінальні ознаки, а також порядкові ознаки, для яких відомий скінченний набір можливих градацій.

Таблиця 8.2

Рівні ознаки 1	**Рівні ознаки 2**				**Разом**
	1	2	…	r	
1	f_{11}	f_{12}	…	f_{1r}	$n1$
2	f_{21}	f_{22}	…	f_{2r}	$n2$
…	…	…	…	…	…
c	f_{c1}	f_{c2}	…	f_{cr}	n_c
Разом	$m1$	$m2$	…	mr	S

Величини f_{ij} показують, скільки разів зустрічалася комбінація ознак, за якої рівень першої має значення i, а рівень другої має значення j; m_j є сумами стовпців, а n_i – сумами рядків. За даними табл. 8.2 можна оцінити значення ймовірностей:

$$p_{ij} = \frac{f_{ij}}{S}, p_i = \sum_{j=1}^{r} p_{ij} = \frac{n_i}{S}, p_j = \sum_{i=1}^{c} p_{ij} = \frac{m_j}{S}. \qquad \textbf{(8.19)}$$

$$f_{ij} \approx \frac{n_i m_j}{S} \qquad \textbf{(8.20)}$$

Величини $\varphi_{ij} = \frac{n_i m_j}{S}$ є очікуваними частотами. Нульову гіпотезу про відсутність зв'язку відхиляють, якщо різницю між ними й частотами, які спостерігають, не можна пояснити випадковими чинниками. Як критерій можна використовувати величину:

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(f_{ij}-\varphi_{ij})^2}{\varphi_{ij}} = S\left(\sum_{i=1}^{c}\sum_{j=1}^{r}\frac{f_{ij}^{\,2}}{n_i m_j}\right), \qquad \textbf{(8.21)}$$

яка за достатньо великого обсягу вибірки наближається до розподілу χ^2 із кількістю ступенів вільності $(r-1)(c-1)$. На практиці для можливості застосування критерія часто вважають достатнім, щоб усі значення f_{ij} були не меншими за п'ять. При збільшенні кількості ступенів вільності мінімальні значення f_{ij} можуть бути дещо меншими.

На практиці частіше використовують *φ-коефіцієнт Пірсона*, або *середньоквадратичну спряженість*:

$$\varphi^2 = \frac{\chi^2}{S},$$

яка може змінюватися від нуля до $\min(c-1;r-1)$.

Існує велика кількість показників ступеня щільності статистичного зв'язку, що призначені для категоризованих змінних, які не є універсальними, а відображають окремі властивості такого зв'язку.

Коефіцієнт спряженості Крамера[139] розраховують за формулою:

$$C = \left(\frac{\sum_{i=1}^{c}\sum_{j=1}^{r}\frac{f_{ij}^{\,2}}{n_i m_j}-1}{\min(c-1;r-1)}\right)^{\frac{1}{2}} = \left(\frac{\varphi^2}{\min(c-1;r-1)}\right)^{\frac{1}{2}}. \qquad \textbf{(8.22)}$$

$C \in [0;1]$. При цьому значення $C = 0$ свідчить про статистичну незалежність аналізованих ознак, а значення $C = 1$ – про можливість однозначного відтворення значень однієї ознаки за відомими значеннями другої. Дисперсію оцінки коефіцієнта Крамера можна отримати із виразу:

$$\sigma_C^2 \approx \frac{1}{n\min(c-1;r-1)}. \qquad \textbf{(8.23)}$$

Її довірчий інтервал:

$$[C - u_{1-\alpha}\sigma_C;\ C + u_{1-\alpha}\sigma_C], \qquad \textbf{(8.24)}$$

де u_q – q-квантиль стандартного нормального розподілу.

[139] Запропоновано Крамером у 1946 р.

Розділ 9
ЛІНІЙНИЙ РЕГРЕСІЙНИЙ АНАЛІЗ

9.1. Основи методу лінійного регресійного аналізу

Побудову будь-якої ЕММ, незалежно від того, на якому рівні та для яких показників її будують, здійснюють як послідовність певних кроків алгоритму.

Алгоритм

Крок 1. Установлення причинно-наслідкового зв'язку між досліджуваними економічними показниками.

Крок 2. Вибір найбільш істотних ознак і встановлення вигляду функції зв'язку.

Крок 3. Знаходження параметрів зв'язку.

Крок 4. Оцінювання достовірності отриманих результатів.

Крок 5. Прогнозування значень результуючої ознаки та їх аналіз.

ЕММ базується на єдності двох аспектів: теоретичного, якісного аналізу взаємозв'язків та емпіричної інформації. Теоретична інформація знаходить своє відображення у специфікації моделі.

Нехай потрібно оцінити зв'язок між змінними (ознаками) *X* та *Y* (напр., зв'язок показників безробіття та інфляції у певній країні за певний проміжок часу). Зокрема, може стояти питання, чи пов'язані між собою ці показники. За позитивної відповіді природно постає задача знаходження формули цього зв'язку. Основою для відповіді на це запитання є статистичні дані про динаміку цих показників (річні, квартальні, місячні тощо). Ці дані утворюють деяку випадкову вибірку з генеральної сукупності, тобто з сукупності всіх можливих показників інфляції та безробіття в заданих умовах. Питання про наявність зв'язку між економічними змінними постає як питання про наявність конкретної формули (специфікації) такої залежності, що є стійкою до кількості спостережень.

Специфікація моделі – аналітична форма ЕММ. На основі досліджуваних чинників її складають певного вигляду функції, які використовують для побудови моделей; вона має

ймовірнісні характеристики, притаманні стохастичним залишкам моделі. Специфікація моделі передбачає відбір чинників для дослідження, оскільки вибір аналітичної форми моделі не можливо розглядати без конкретного переліку незалежних змінних. При цьому у процесі дослідження можна кілька разів повертатись до етапу специфікації моделі, уточнюючи перелік незалежних змінних і вигляд функції, яка застосовується. Адже коли вигляд функції та її складові не відповідають реальним процесам, то йдеться про помилки специфікації:

- ігнорування при побудові моделі істотного фактора;
- введення до моделі незалежної змінної, яка не є істотною для вимірювання зв'язку;
- використання невідповідних математичних форм залежності.

Перша помилка специфікації моделі призводить до зміщення оцінок, яке буде тим більшим, чим більшою є кореляція між введеними та відсутніми в моделі змінними, а напрям зміщення залежатиме від характеру кореляції між введеними та відсутніми величинами. Оцінки параметрів також буде зміщено, тому застосування способів перевірки їх значущості може спричинити хибні висновки щодо значень параметрів генеральної сукупності.

Друга помилка специфікації моделі пов'язана з тим, що до моделі вводять змінну, яка неістотно впливає на залежну змінну, тоді оцінки параметрів будуть незміщені, на відміну від першої помилки. Причому за допомогою звичайних процедур можна дістати також незміщені оцінки дисперсій цих параметрів. Але це не означає, що модель можна беззастережно розширювати за рахунок "неістотних" змінних. Існує ненульова ймовірність того, що в результаті використання вибіркових даних змінна, яка зовсім не стосується моделі, покаже істотний зв'язок із залежною змінною. А це означає, що кількісний зв'язок між змінними буде описано не вірно.

Третя помилка специфікації можлива за припущення, що залежна змінна є лінійною функцією від деякої пояснювальної змінної, тоді як насправді тут краще підійшла б квадра-

тична, кубічна чи степенева залежність. У такому випадку наслідки такі самі, що й у першому випадку, тобто оцінки параметрів моделі матимуть зміщення.

Питання про вибір найкращої форми залежності має базуватись на перевірці узгодженості виду функції із вхідними даними спостереження.

Адекватність побудованої моделі можна встановити, аналізуючи залишки моделі. Їх обчислюють як різницю між фактичними значеннями залежної змінної і тими, що обчислені за моделлю.

Стан міжнародних економічних відносин характеризує велика кількість економічних показників, які, своєю чергою, залежать від різноманітних факторів. Регресійний аналіз дозволяє виявити характер цих взаємозв'язків. За регресійного аналізу моделюють взаємозв'язок однієї залежної економічної змінної (показника, результуючої ознаки або відгуку) від однієї або кількох незалежних економічних змінних (факторів). Вибір або призначення залежної і незалежних змінних є довільним, його здійснюють залежно від поставленої задачі.

У загальному випадку регресійну модель можна записати як:

$$Y = f(x_1, x_2, ..., x_n), \quad (9.1)$$

де Y – залежна змінна (відгук); $x_i, i = 1,2,...,n$ – незалежні змінні (фактори).

За допомогою регресійного аналізу можна вирішувати важливі задачі:

1. Зменшення розмірності простору змінних, які аналізують (факторного простору), за рахунок заміни частини факторів однієї змінної – відгуком. Повніше цю задачу розв'язують за допомогою факторного аналізу.

2. Кількісна оцінка ефекту кожного фактора, тобто множинна регресія, дозволяє задати питання та отримати відповідь на запитання: "що є кращим фактором для…". При цьому стає більш зрозумілими дія окремих факторів на відгук, а також природа економічного явища, яке вивчають.

3. Виявлення прогнозних значень відгуку за певних значень факторів, тобто регресійний аналіз створює базу для

обчислення з метою отримання відповідей на запитання типу: "що буде, якщо…?".

4. У регресійному аналізі у більш явній формі виступає причинно-наслідковий механізм. Прогноз при цьому краще піддається змістовній інтерпретації.

Регресійний аналіз складається з основних етапів:

- *вибір факторів* (множини незалежних змінних), які суттєво впливають на залежну змінну – *відгук*;
- *визначення форми рівняння регресії*;
- *оцінювання параметрів регресійної моделі*.

У загальному випадку класичний регресійний аналіз є параметричним методом. При застосуванні регресійного аналізу припускають, що змінні в моделі кількісні, закон розподілу яких відповідає нормальному закону.

При дослідженні міжнародних економічних відносин часто використовують лінійні регресійні моделі. Існують певні причини широкого застосування лінійних моделей:

- припущення про лінійність є простим, тому його легко сприймати та інтерпретувати;
- метод оцінювання параметрів найменш трудомісткий, можна проводити розрахунки навіть "вручну" за допомогою калькулятора;
- багато математичних методів дослідження пристосовано до лінійних моделей.

Лінійна модель множинної регресії має вид:

$$Y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n + \varepsilon, \qquad (9.2)$$

де Y – залежна змінна (відгук); x_i – i-та незалежна змінна (i-й фактор) $i = 1,2,...,n$; b_i – і-й коефіцієнт регресії (коефіцієнт при i-му факторі), b_0 – вільний член – *константа*/*intercept*; ε – *похибка спостереження* або *залишок*/*residual*, яка є різницею між спостережуваним значенням залежної змінної і тим, що розраховано за регресією (теоретичним або передбаченим). В ідеальному випадку залишки є випадковими величинами, що мають нормальний розподіл з математичним сподіванням, яке дорівнює нулю.

Спочатку за відомими значеннями залежної і незалежних змінних розраховують найкращі коефіцієнти регресії (мето-

дом найменших квадратів або іншими методами), за яких функція регресії стає максимально можливо достовірною. Далі регресію з відомими коефіцієнтами регресії, до якої підставляють нові значення незалежних факторів, використовують для розрахунку невідомих значень залежної змінної.

У результаті регресійного аналізу здійснюють:

- прогнозування;
- пояснення зв'язку між змінними.

9.2. Оцінки параметрів моделі методом найменших квадратів

Розглянемо модель (**9.1**) із двома змінними:

$$Y = f(X) + \varepsilon, \qquad (9.3)$$

де Y – показник (залежна змінна), X – фактор (незалежна змінна), ε – випадкова складова. У цій моделі ідентифіковано змінну X, яка визначає змінну Y. Таку модель називають *простою моделлю* або *парною регресією*. На основі простої моделі розглянемо принципову структуру моделі та основні методи оцінювання її параметрів. Теоретичні знання про взаємозв'язок між економічними показниками мають підказати його конкретну форму. Оскільки одні й ті самі економічні процеси можна описати різними функціями, то потрібно звернутись до статистичного аналізу та за його допомогою зробити вибір серед усіх можливих альтернативних варіантів. Найпростішою вважають лінійну форму зв'язку між двома змінними.

У дослідженнях у сфері міжнародних економічних відносин найбільш широко використовують моделі лінійної регресії, хоча це є спрощеним засобом у моделюванні реальних економічних процесів. Ґрунтовне вивчення та застосування методики побудови лінійних моделей надає необхідну теоретичну базу для створення складніших нелінійних моделей, які більшою мірою відповідають реальним економічним процесам.

Перейдемо до задачі кількісного опису залежності між двома економічними ознаками. Природно очікувати, що значення показника Y не завжди однозначно визначаються зна-

ченням фактора *X*. Крім того, необхідно врахувати всі фактори, які впливають на показник у реальній ситуації. Також для одного значення фактора *X* можна спостерігати різні значення показника. Зазвичай для опису ситуацій із недостатньою інформацією використовують різні ймовірнісні моделі. Нехай на координатній площині задано *n* точок із координатами (x_i, y_i), $(i = 1, 2,..., n)$. Вибір специфікації моделі парної регресії є найбільш очевидним, оскільки його можна виконати візуально, використовуючи графічне зображення емпіричних даних як точок (x_i, y_i) на кореляційному полі у прямокутній системі координат, які утворюють так звану діаграму розсіювання (рис. 9.1). Для точок (рис. 9.1 а) можна припустити, що зв'язок між *Y* та *X* є лінійним $Y = \beta_0 + \beta_1 X$. Для точок (рис. 9.1 б) залежність близька до параболічної:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2.$$

Для точок на рис. 9.1 в явної залежності між *Y* та *X* не спостерігається.

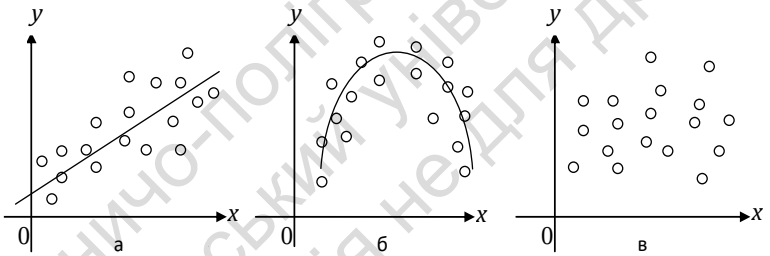


Рис. 9.1

Для множинної регресії виявлення залежності між змінними значно ускладняються. Тут в деяких випадках може спрацювати досвід та інтуїція.

Розглянемо таку задачу. Нехай зв'язок між змінними *y* та *x* є лінійним, тобто графіком залежності є пряма, яку задає рівняння:

$$y = f(x) + \beta_0 + \beta_1 x. \qquad (9.4)$$

Рівняння залежить від двох параметрів – β_0 і β_1. Потрібно знайти "оптимальну" пряму, яка "найменше відхиляється від заданих точок", тобто за даними значеннями x_i та y_i, $i = 1, 2,..., n$ (рис. 9.2) – знайти значення параметрів "оптимальної" прямої.

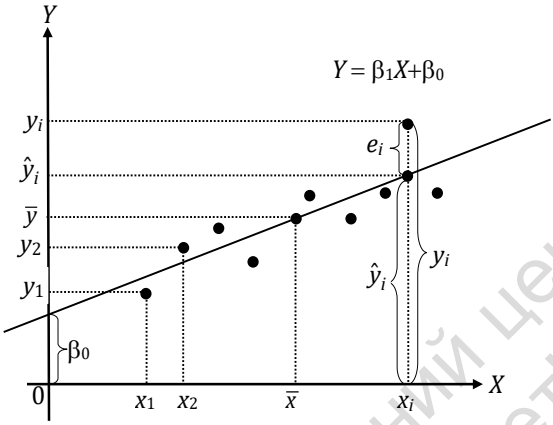


Рис. 9.2

Основні запитання: що розуміти під "найменшим відхиленням прямої від точок", як визначити "міру відхилення прямої від точок"? Виходячи з імовірнісного погляду у випадку нормального розподілу вибіркових даних "найкращі ймовірнісні та статистичні властивості" мають оцінки параметрів прямої, що отримані мінімізацією суми квадратів відхилень. Цей метод оцінювання параметрів оптимальної прямої називають МНК – *методом найменших квадратів/Ordinary Least Squares – OLS*, а отримані оцінки параметрів – *МНК*- або *OLS*-оцінками.

Якщо до рівняння включено лише одну пояснюючу змінну, то потрібно знайти теоретичну модель, яка дістала назву *парної лінійної регресії*:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1,2,...,n, \qquad (9.5)$$

де y_i – *залежна змінна/dependent variable* – випадкова величина, x_i – незалежна змінна або *пояснююча змінна/explanatory variable*), i – номер спостереження; ε_i – випадкові величини, які називають *помилками регресії*, $i = 1, 2,..., n$. Параметр β_1 називають параметром *нахилу лінії регресії/slope*, а β_0 – *параметром зсуву/intercept*.

Теоретичну модель (**9.5**) для парної лінійної регресії можна записати так:

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1, \\ y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ y_n = \beta_0 + \beta_1 x_n + \varepsilon_n. \end{cases} \quad (9.6)$$

У матричній формі співвідношення (**9.6**) матиме вигляд:

$$Y = X\beta + \varepsilon, \quad (9.7)$$

де для змінних введено позначення:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Із умови $M\varepsilon_i = 0, i = 1,2,...,n$ випливає, що:

$$My_i = \beta_0 + \beta_1 x_1, i = 1,2,...,n,$$

тобто середнє значення показника Y за заданого значення X не залежить від помилок регресії. Звідси термін – *несистематичні помилки*.

Для визначення теоретичних коефіцієнтів β_0 і β_1 необхідно використати всі значення змінних Y та X генеральної сукупності, що практично здійснити неможливо. Отже, за міру відхилення прямої від заданих точок (x_i, y_i), $i = 1, 2,..., n$ візьмемо суму квадратів відхилень у кожній точці:

$$S = S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_0 x_1) \right)^2.$$

Тоді параметри прямої, для якої міра відхилення мінімальна, є розв'язком задачі знаходження екстремуму без обмежень:

$$S = S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_0 x_1) \right)^2 \to \min.$$

Відповідно до необхідних умов існування екстремуму параметри оптимальної прямої знаходять як розв'язок системи:

$$\begin{cases} \dfrac{\partial S}{\partial \beta_0} = 2\sum\limits_{i=0}^{n}(y_i - \beta_0 - \beta_1 x_1)(-1) = 0, \\ \dfrac{\partial S}{\partial \beta_1} = \sum\limits_{i=0}^{n}(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0). \end{cases}$$

Після перетворень отримаємо систему лінійних рівнянь:

$$\begin{cases} n\beta_0 + \beta_1 \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i, \\ \beta_0 \sum\limits_{i=1}^{n} x_i + \beta_1 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i, \end{cases} \quad (9.8)$$

яку називають *системою нормальних рівнянь* МНК.

Знайдемо явні формули для розв'язку цієї системи. Для зручності поділимо кожне рівняння системи (**9.8**) на n і дістанемо:

$$\begin{cases} \beta_0 + \beta_1 \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i = \dfrac{1}{n}\sum\limits_{i=1}^{n} y_i, \\ \beta_0 \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i + \beta_1 \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i y_i \end{cases} \quad \text{або} \quad \begin{cases} \beta_0 + \beta_1 \overline{x} = \overline{y}, \\ \beta_0 \overline{x} + \beta_1 \overline{x}^2 = \overline{xy}. \end{cases}$$

Виразимо β_0 із першого рівняння: $\beta_0 = \overline{y} - \beta_1 x_1$ і підставимо до другого рівняння: $(\overline{y} - \beta_1 \overline{x})\overline{x} + \beta_1 \overline{x}^2 = \overline{xy}$.

Після перетворень отримаємо (формально):

$$\widehat{\beta}_1 = \frac{\dfrac{1}{n}\sum\limits_{i=1}^{n} x_i y_i - \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i \dfrac{1}{n}\sum\limits_{i=1}^{n} y_i}{\dfrac{1}{n}\sum\limits_{i=1}^{n} x_i^2 - \left(\dfrac{1}{n}\sum\limits_{i=1}^{n} x_i\right)^2} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{x^2 - (\overline{x})^2} = \frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)} = \operatorname{cor}(x,y)\frac{\sigma_y}{\sigma_x}, \quad (9.9)$$

$$\beta_0 = \overline{y} - \widehat{\beta}_1 \overline{x}, \quad (9.10)$$

де $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,\ \overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2,\ \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i, \quad (9.11)$$

вибіркові (невиправлені) дисперсії:

$$\hat{\sigma}_x^2 = \mathrm{var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \bar{x}^2 - (\bar{x})^2,$$
$$\hat{\sigma}_y^2 = \mathrm{var}(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}^2) = \bar{y}^2 - (\bar{y})^2, \quad (\mathbf{9.12})$$

вибіркові стандартні середньоквадратичні відхилення:

$$\sigma_x = \sqrt{\mathrm{var}(x)}, \ = \sigma_y = \sqrt{\mathrm{var}(y)}, \quad (9.13)$$

вибірковий коефіцієнт коваріації (у MS Excel функція КОВАР(,)):

$$K_{xy} = \mathrm{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{xy} - \bar{x}\cdot\bar{y}, \quad (\mathbf{9.14})$$

вибірковий коефіцієнт кореляції (в MS Excel функція КОРРЕЛ(,)):

$$r_{xy} = \mathrm{cor}(x,y) = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)}\cdot\sqrt{\mathrm{var}(y)}} \quad (\mathbf{9.15})$$

величин *X* та *Y*, відповідно.

Легко показати, що функція $S = S - (\beta_0, \beta_1)$ опукла. Отже, розв'язок системи нормальних рівнянь (**9.8**) буде глобальним мінімумом цієї функції. Таким чином, оптимальну пряму задає рівняння:

$$\hat{y} = \beta_0 + \beta_1 x. \quad (\mathbf{9.16})$$

Зауваження 1. Оптимальна пряма (лінія регресії) проходить через точку з координатами $(\bar{x}, \bar{y})$.

Зауваження 2. Рівняння (**9.16**) можна переписати у вигляді:

$$\hat{y} - \bar{y} = r_{xy} \cdot \frac{\sigma_y}{\sigma_x}(x - \bar{x}). \quad (\mathbf{9.17})$$

Зауваження 3. Легко помітити, що система нормальних рівнянь має єдиний розв'язок тоді й тільки тоді, коли $\sigma_X \neq 0$, тобто коли не всі значення змінної *x* збігаються.

Зауваження 4. Метод найменших квадратів можна застосувати для знаходження параметрів будь-якої функції, що найменше відхиляється від заданих точок.

Природним узагальненням моделі парної регресії є модель множинної регресії, коли розглядають вплив багатьох факторів $X_1, X_2, ..., X_m$ (*регресорів*) на залежну зміну – показ-

ник *Y*. У подібних випадках маємо справу із множинною лінійною моделлю (*регресією*), що описує взаємний зв'язок між залежною змінною *Y* та регресорами $X_1, X_2, ..., X_m$, яку можна подати у вигляді:

$$M\left(\frac{Y}{X_1, X_2, ..., X_m}\right) = \alpha(X_1, X_2, ..., X_m).$$

Цей математичний запис інформує про функціональну залежність умовного математичного сподівання залежної змінної *Y* від *m* регресорів (незалежних, пояснюючих) змінних $X_1, X_2, ..., X_m$.

Розглянемо задачу виявлення статистичного взаємозв'язку між *Y* та *X*. У цій задачі розглядатимемо модель множинної регресії у вигляді:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_j x_{ij} + ... + \beta_m x_{im} + \varepsilon_i, \; i = 1,2,...,n, \quad \textbf{(9.18)}$$

де β_j, $j = 1,2,...,m$ – теоретичні коефіцієнти регресії або параметри теоретичної регресії, які характеризують реакцію залежної змінної y_i, $i = 1,2,...,n$ на зміну кожного регресора X_j, $j = 1,2,...,m$; β_0 – вільний член, який визначає значення y_i, $i = 1,2,...,n$ за умови, коли значення регресорів дорівнюють нулю; x_{ij}, $i = 1,2,...,n$; $j = 1,2,...,m$ – значення X_i-го регресора при *i*-ому спостереженні; ε_i, $i = 1,2,...,n$ – випадкова складова при *i*-ому спостереженні.

Для однозначного визначення параметрів β_j, $j = 1,2,...,m$ необхідно, щоб виконувалась нерівність $n \geq m = 1$, де *n* – число спостережень; *m* – число регресорів у моделі.

У матричній формі теоретичну модель лінійної множинної регресії записують у вигляді:

$$Y = X \cdot \beta + \varepsilon. \quad \textbf{(9.19)}$$

Компоненти β_j, $j = 1,2,...,m$ матриці β є величинами сталими ($\beta_j = \text{const}$), але невідомими. Їх необхідно оцінити шляхом обробки вибірки, тому надалі матимемо справу з емпіричною моделлю:

$$\hat{Y} = X \cdot \hat{\beta} + e, \quad \textbf{(9.20)}$$

де вектор $\hat{\beta}$ є статистичною оцінкою теоретичного вектора β лінійної множинної регресії (**9.19**). Вектор похибок e є статистичною оцінкою випадкового вектора ε цієї самої моделі.

За допомогою емпіричної моделі визначають статистичні оцінки параметрів β_j, $j=1,2,...,m$. При цьому використовують статистичну обробку вибірки.

Компоненти $\hat{\beta}_j$, $j=1,2,...,m$ вектора β є статистичними оцінками компонент β_j, $j=1,2,...,m$ теоретичного вектора β лінійної множинної регресії (**9.18**), а компоненти e_i, $i=1,2,...,n$ вектора похибок e – статистичні оцінки випадкових збурень ε_i, $i=1,2,...,n$ вектора ε.

Якщо теоретичний вектор β є величиною сталою й невідомою, то емпіричний вектор β можна визначити шляхом обробки статистичної інформації вибірки обсягом n. Оскільки вибірка становить лише незначну частину генеральної сукупності ($n \le N$), то інформація, яку одержують за статистичної обробки про регресори X_j моделі буде неповною та для кожної іншої вибірки зазнаватиме певних змін. Отже, компоненти β_j, $j=1,2,...,m$ теоретичного вектора β міститимуть елемент випадковості. Таким чином, $\hat{\beta}_j$, $j=1,2,...,m$, як і сам вектор $\hat{\beta}$, будуть випадковими величинами, які мають певні закони розподілу ймовірностей із відповідними числовими характеристиками.

Оскільки β є статистичною оцінкою для теоретичного вектора β, то постають питання математичної статистики: зміщена чи незміщена ця статистична оцінка; в якому довірчому інтервалі із заданою надійністю γ можуть перебувати теоретичні компоненти (параметри) β_j, $j=1,2,...,m$ і сама функція регресії; як здійснити перевірку на статистичну значущість теоретичних параметрів β_j, $j=1,2,...,m$ за заданим рівнем значущості α. Для вирішення цих питань необхідно визначити числові характеристики для параметрів β_j, $j=1,2,...,m$ і для самої функції регресії, використовуючи при цьому еле-

менти матричної алгебри як інструментарій, застосовуючи який можна без громіздких викладок отримати необхідні результати.

Для застосування методу найменших квадратів щодо помилок регресії припускають, що виконуються умови:

1) $M\varepsilon = 0$, i – (помилки регресії несистематичні);

2) $M(\varepsilon\varepsilon^T) = \sigma^2 I$, де I – одинична матриця порядку m (залишки вектора незалежні один від одного та мають постійну дисперсію);

Отже, $\varepsilon \sim N(0, \sigma^2)$ – випадковий вектор ε помилок регресії має багатовимірний нормальний розподіл із нульовим середнім і коваріаційною матрицею $\sigma^2 I$.

3) $M(X^T\varepsilon) = 0$ (незалежні зміні, не пов'язані із залишками);

4) $\det(X^T X) \neq 0$ (визначник матриці $(X^T X)$, відмінний від нуля. Це означає, що змінні моделі утворюють лінійно незалежну систему векторів, і матриця X має повний ранг).

Оцінимо за методом найменших квадратів параметри моделі (**9.20**). Для цього за аналогією суму квадратів запишемо у матричному вигляді:

$$S = \sum_{i=1}^{n} \varepsilon_1^2 = \varepsilon^T \varepsilon = (Y - X \cdot \beta)^T (Y - X \cdot \beta) = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta.$$

У точці екстремуму для функціонала $\varepsilon^T\varepsilon$ виконується умова:

$$\frac{\partial(\varepsilon^T \varepsilon)}{\partial \varepsilon} \hat{\beta} = -2X^T Y) + 2X^T X\varepsilon = 0,$$

звідки $X^T\beta = X^T$. Тут X^T – матриця, що транспонована до матриці X. Якщо визначник $\det(X^T X) \neq 0$, то система нормальних рівнянь має єдиний розв'язок та *OSL*-оцінка коефіцієнтів моделі:

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \qquad (\mathbf{9.21})$$

Якщо незалежні змінні в матриці X є відхиленнями кожного значення від свого середнього, то матрицю $X^T X$ називають *матрицею моментів*. Тоді числа, що розташовані на головній діагоналі, характеризують дисперсії незалежних змінних, а інші елементи – відповідають коваріаціям.

9.3. Критерії якості регресійної моделі

Вибір найбільш прийнятної регресійної моделі здійснюють через критерії якості регресійної моделі: коефіцієнти детермінації, множинної кореляції, статистики Дарбіна–Уотсона, розподілу залишків тощо. Розраховані значення залежної змінної можуть не збігатися з фактичними її значеннями, але регресія буде тим точнішою, чим меншими будуть залишки.

Введемо позначення y_i – вихідні дані залежної змінної *Y* (відгуку), $\hat{y}_i, i = 1,2,...,n$ – передбачувані значення, тобто значення, що розраховані за регресійною залежністю; *n* – обсяг вибірки, або кількість спостережень.

Коефіцієнт детермінації обчислюють за формулою:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \qquad (9.22)$$

де $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ – середнє значення залежної змінної *Y*.

Коефіцієнт детермінації характеризує відношення розсіювання значень залишків навколо лінії регресії щодо загального розсіювання значень. Коефіцієнт детермінації R^2, який набуває значень від 0 до 1, може свідчити про величину залишків. Чим ближчий він до 1 (за умов виконання умов для регресійного аналізу), тим щільніший лінійний кореляційний зв'язок між залежною змінною та факторами, тим, очевидно, і кращий прогноз.

Коефіцієнт детермінації показує, яку частину дисперсії залежної змінної можна пояснити за допомогою дисперсії включених до регресії факторів. Якщо, наприклад, $R^2 = 0,4$, то 40 % мінливості відгуку можна пояснити рівнянням регресії, 60 % залишкової мінливості залишаються непоясненими. В ідеалі бажано мати пояснення принаймні більшої частини вихідної мінливості. Значення R^2 є індикатором

ступеня відповідності даним. Значення $R^2 \approx 1$ показує, що модель пояснює майже всю мінливість відповідних змінних.

Значущість коефіцієнта детермінації перевіряють за допомогою *F*-тесту. Якщо останній повертає рівень значущості $p < 0{,}05$, то можна принаймні на 95 % бути впевненим, що коефіцієнт детермінації не дорівнює нулю і в генеральній сукупності.

Коефіцієнт множинної кореляції $R = \sqrt{R^2}$ характеризує щільність зв'язку між змінними, а також є оцінкою якості передбачення. Він, як і коефіцієнт парної кореляції, характеризує ступінь лінійного взаємозв'язку. Коефіцієнт детермінації R^2 оцінює адекватність регресійної моделі будь-якого виду. Тому його використовують також як оцінку ступеня нелінійної залежності між змінними.

Для порівняння якості регресійних рівнянь із різною кількістю факторів варто використовувати *скоригований коефіцієнт детермінації*/*adjusted coefficient of determination*. Тоді можна уникнути ситуації, за якої для збільшення коефіцієнта детермінації необхідно включати до регресії якомога більшу кількості факторів, незважаючи на те, що вони можуть здійснювати мінімальний вплив на залежну змінну. За інших рівних умов обирають модель з таким набором факторів, за якої скоригований коефіцієнт детермінації найбільший.

Альтернативним методом вибору оптимальної кількості факторів у моделі є *інформаційні критерії*/*information criteria*. До них належать критерій *Акаіке*/*Akaike criterion*, *критерій Шварца*/*Schwarz criterion* і *критерій Ханнана–Куїна*/*Hannan–Quinn criterion*. За інших рівних умов обирають модель із таким набором факторів, за якої інформаційні критерії найменші.

Знак коефіцієнтів $b_i, i = 1,2,...,n$ указує на характер впливу *i*-го фактора на залежну змінну. Якщо $b_i > 0$ – додатне число, то фактор впливає позитивно (більше значення фактора пов'язане з більшим значенням залежної змінної), якщо від'ємне – негативно (більше значення фактора пов'язане з меншим значенням залежної змінної), якщо дорівнює нулю, то фактор не впливає на залежну змінну. Наскільки значуще

кожен з коефіцієнтів b_1 відрізняється від нуля, показує *t*-тест, який проводять для кожного такого коефіцієнта. Іншими словами, якщо *F*-тест указує на значущість зв'язку між залежною змінною та сукупністю всіх факторів, що включені до регресії, то *t*-тест указує на значущість зв'язку між залежною змінною та окремим фактором у рамках побудованої регресії.

Аналогами звичайних *b*-коефіцієнтів є β-коефіцієнти, якщо до регресії замість абсолютних значень змінних підставляти стандартизовані значення змінних. Тому β-коефіцієнти не залежать від того, в яких одиницях вимірюють змінні. β-коефіцієнти для різних показників у різних одиницях виміру є порівнюваними. Фактор, β-коефіцієнт якого більший, здійснює сильніший вплив.

Обираючи змінні для аналізу, важливо включити до регресійного рівняння всі важливі фактори (уникаючи при цьому проблеми мультиколінеарності) та утримуватися від включення факторів, які мало впливають на залежну змінну. Для попереднього визначення важливості фактора можна використати кореляційний і графічний аналіз за допомогою діаграм розсіювання або *точкових діаграм*/*scatterplots*. Бажано включати незалежні змінні, зв'язок яких із залежною змінною має наявне або потенційне теоретичне пояснення. Якщо регресійний аналіз проводять для прогнозування, то варто включати лише ті змінні, за якими у майбутньому можна знайти дані.

Визначення змінних, які варто включати до регресійної моделі, а які – не варто, можна автоматизувати (таку опцію іноді пропонує програмне забезпечення). Для цього використовують покрокові процедури із включенням або виключенням факторів. Але цілковито покладатися на програмний алгоритм не варто. Обрана за допомогою цих процедур форма регресії може виявитися оптимальною з формально-статистичного погляду, але неоптимальною – з практичного (напр., включені змінні, за якими важко знайти нові дані в майбутньому) або такою, що не підкріплена економічною теорією (що підвищує ризик прийняття випадковості за закономірність).

Покроковий метод включення змінних передбачає, що спочатку будують однофакторну регресію з тим фактором, який має найбільший кореляційний зв'язок із залежною змінною. Далі будують двофакторну регресію: додають із числа інших факторів той, що пояснює найбільшу частину дисперсії, яка не пояснюється дією першого фактора за допомогою *часткового коефіцієнта кореляції*/*partial correlation coefficient*. Потім будують трифакторну регресію: додають з числа решти факторів той, що пояснює найбільшу частину дисперсії, яка не пояснюється дією першого та другого факторів. Аналогічно далі додають четвертий фактор і т. д. Процес зупиняють, коли додаткова користь від включення нової змінної стає мінімальною, тобто частковий *F*-критерій свідчить про незначущість внеску нового фактора. При включенні нових факторів варто перевіряти вже включені до регресії фактори щодо їх залишення чи виключення. На кожному кроці за допомогою *t*-тесту перевіряють значущість *b*-коефіцієнтів; фактори, за яких вони незначущі, виключають із моделі.

Покроковий метод виключення змінних передбачає побудову регресії з усіма можливими факторами. Потім фактори з найменшим внеском виключають.

Альтернативною покроковим процедурам є *побудова всіх можливих видів регресії*. Наприклад, для 10 факторів можна побудувати 1024 варіанти регресії із повним або частковим включенням факторів.

Більша кількість факторів потребує більшої кількості спостережень. При побудові однофакторної регресії бажано мати більше 20-30 спостережень. Умовна рекомендація для багатофакторної регресії передбачає наявність більше 10-20 (мінімум 5) спостережень на кожний фактор, що включений до моделі. Наприклад, для п'ятифакторної моделі бажано мати більше 50-100 спостережень.

Якщо кількість спостережень є дуже великою, наприклад, 1000, то статистичні *F*-тест і *t*-тест показуватимуть, що зв'язок між факторами та залежною змінною буде значущий, навіть якщо він є дуже слабким. Тому за великих вибірок

потрібно забезпечити не лише статистичну, а й практичну значущість результатів. Для визначення практичної значущості варто звернути увагу також і на саму величину коефіцієнта детермінації, парних коефіцієнтів кореляції та β-коефіцієнти. Тому, навіть якщо *F*-тест показує значущість коефіцієнта детермінації, а сам коефіцієнт детермінації невеликий (напр., 0,10 для пояснення впливу факторів або менше 0,30 для прогнозної моделі), то регресія є недостатньо адекватною для прогнозування чи визначення причин змін залежної змінної. Тоді варто змінити специфікацію моделі – трансформувати її: спробувати додати інші фактори або використати нелінійну регресію.

Навіть якщо *t*-тест указує на значущість *b*-коефіцієнта для відповідного фактора, то варто пересвідчитися, чи не є замалим парний коефіцієнт кореляції або β-коефіцієнт для цього фактора.

9.4. Припущення лінійного регресійного аналізу

Як й інші види аналізу лінійний регресійний аналіз виходить із певних припущень. Для впевненості в тому, що регресія адекватно відображає зв'язки між змінними та може бути використана для прогнозу або пояснення впливу, фактично потрібно перевірити на відповідність припущенням не тільки кожну змінну окремо, а й всю лінійну комбінацію факторів (якщо йдеться про багатофакторну регресію).

Тестування на відповідність припущенням відбувається як перед проведенням регресійного аналізу, так і в процесі його проведення з можливим коригуванням специфікації рівняння регресії або трансформацією змінних. Умови для регресійного аналізу:

1. Лінійний характер зв'язку.

2. Гомоскедастичність (незалежність дисперсії від значень змінних).

3. Незалежність залишків (відсутність серійної кореляції).

4. Нормальний розподіл залишків.

5. Відсутність мультиколінеарності.

Відповідність припущенням часто перевіряють, зокрема, за допомогою діаграм розсіювання, де на вертикальній осі відкладають зазвичай залишки, а на горизонтальній – розраховані значення залежної змінної *y*, кожної із залежних змінних $x_i, i = 1,2,...,n$, або часу *t*. Іноді для кращого порівняння використовують не абсолютні значення залишків, а їх стандартизовані значення або трансформовані, відповідно до *t-розподілу Стьюдента/Studentized residuals*. Наведемо приклади відповідних діаграм розсіювання (рис. 9.1).

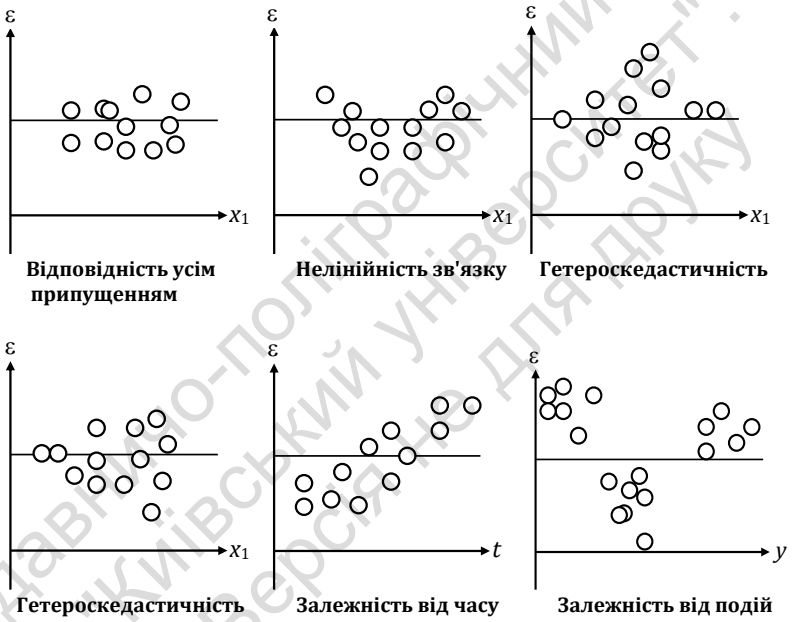


Рис. 9.1

1. Лінійність зв'язку варто перевіряти за допомогою діаграм розсіювання, де за вертикальною віссю відкладають значення залежної змінної *Y*, а за горизонтальною – одного з факторів $x_i (i = 1,2,...,n)$. І таких діаграм розсіювання варто побудувати стільки, скільки факторів наявно у регресії, адже нелінійність зв'язку із залежною змінною може бути у частини факторів, а не у всіх.

Якщо виявлено нелінійний зв'язок залежної змінної із певними факторами, то ці фактори варто трансформувати. Наприклад, замість фактора $x_i (i = 1,2,...,n)$ застосувати функції $\ln(x_i)$, $1/x_i$, $\sqrt{x_i}$ або x_i^2. За необхідності те саме можна зробити й із залежною змінною. Варто також також спробувати заміну абсолютних значень змінних на їх прирости, наприклад, y_t на $(y_t - y_{t-1})/y_t$, а x_{it} – на $(x_{it} - x_{it-1})/x_{it}$, або на різницю логарифмів $\ln(y_t)$ – на $\ln(y_t) - \ln(y_{t-1})$, а $\ln(x_{it})$ – на $\ln(x_{it}) - \ln(x_{it-1})$.

2. Гетероскедастичність передбачає залежність залишків від значення залежної змінної, або залежність дисперсії залежної змінної від її значень. Наприклад, що більші значення залежної змінної, то більшою є її дисперсія, і, відповідно, залишки при побудові регресії. На рис. 9.1 представлені трикутникові та ромбовидна форми гетероскедастичності. Можлива також і метеликова форма, коли дисперсія найменша за середніх значень фактора. Для тестування гетероскедастичності також використовують *тест Левена/Levene's test*.

Для розв'язання проблеми гетероскедастичності можна замість звичайного методу найменших квадратів використати метод *зважених найменших квадратів/weighted least squares* або спробувати трансформувати залежну змінну. Наприклад, $\ln(y_i)$, $1/y_i$, $\sqrt{y_i}$.

3. Регресійні моделі будують за припущення випадковості вибірки (тобто спостереження незалежні). У протилежному випадку коефіцієнти побудованого рівняння регресії можуть бути не стійкими до зміни вихідних даних, що в результаті вплине на достовірність результатів аналізу. Незалежність залишків (відсутність серійної кореляції) передбачає, що розраховані значення залежної змінної не залежать від інших розрахованих значень. Відповідно, і залишки не залежать один від одного. Порушення цієї умови передбачає, таку закономірність: певні залишки є додатними, а решта – від'ємними.

На рис. 9.1 представлено:

- залежність залишків від часу (час часто використовують як змінну, що упорядковує спостереження);
- залежність від подій (напр., коли не враховано фактор сезонності у моделі "зимові та літні місяці"), якщо він є насправді важливим (напр., якщо розглядають зовнішню торгівлю окремими видами одягу), або коли відбуваються спостереження цілковито щодо різнорідних груп об'єктів (напр., країни із низькою часткою неформального сектора економіки та країни з високою його часткою).

Критерій Дарбіна–Уотсона/Durbin–Watson (*DW*) *statistic*[140] використовують для знаходження автокореляції залишків першого порядку регресійної моделі та характеризує ступінь залежності між спостереженнями. Його розраховують за формулою:

$$d = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n}\varepsilon_i^2} \approx 2(1-\rho_1), \qquad (9.23)$$

де ρ_1 – коефіцієнт автокореляції першого порядку.

За відсутності автокореляції помилок статистика Дарбіна–Уотсона дорівнює 2, за позитивної автокореляції прямує до 0, а за негативної – до 4:

- $\rho_1 = 0 \to d = 2$, автокореляція відсутня;
- $\rho_1 = 1 \to d = 0$, додатна автокореляція;
- $\rho_1 = -1 \to d = 4$, від'ємна автокореляція.

На практиці застосування критерію Дарбіна–Уотсона засноване на порівнянні величини d з теоретичними значеннями d_L і d_U для заданого числа спостережень n, числа незалежних змінних k моделі і рівня α значущості:

1) якщо $d < d_L$, то гіпотезу про незалежність випадкових відхилень відкидають (присутня позитивна автокореляція);

2) якщо $d > d_U$, то гіпотезу не відкидають;

[140] Критерій названо на честь Дарбіна та Уотсона.

3) якщо $d_L < d < d_U$, то немає достатніх підстав для ухвалення рішень.

Коли розрахункове значення перевищує 2, то з теоретичними значеннями d_L і d_U і порівнюють не сам коефіцієнт, а вираз $(4-d)$.

Що менший модуль серійної кореляції, то меншими є залежні спостереження у вибірці, отже, і більш адекватною є побудована регресійна модель.

Розв'язанням проблеми залежності залишків можуть бути:

- заміна абсолютних значень на прирости типу y_t на $(y_t - y_{t-1})/y_t$, а x_{it} – на $(x_{it} - x_{it-1})/x_{it}$;
- урахування факторів умов (напр., сезону);
- спеціально сконструйовані регресійні моделі, адаптовані до залежності залишків.

4. Залишки повинні мати нормальний розподіл. Якщо побудовано адекватну регресійну модель, то залишки між вихідними та прогнозними значеннями, що обчислені за рівнянням регресії, мають бути випадковими величинами з математичним сподіванням, яке дорівнює нулю (білий шум). Тому за ступенем відхилення розподілу залишків від нормального закону розподілу також можна судити про якість побудованої моделі.

За недостатньої кількості спостережень відхилення від нормального розподілу змінних можуть призводити до відхилень від нормального розподілу залишків. Чи буде розподіл нормальним перевіряють за допомогою графіка *нормального розподілу/normal probability plot* або гістограми.

Часто, якщо існує порушення однієї умови, то порушуються й інші. І розв'язання проблем з однією умовою часто розв'язує проблеми – з іншими. Наприклад, розв'язання проблем з нелінійністю зв'язку шляхом трансформації змінних може привести до зникнення й гетероскедастичності. Або розв'язання проблеми з відхиленням від нормального розподілу змінних може розв'язати проблему гетероскедастичності.

5. Мультиколінеарність передбачає кореляцію факторів між собою. Включення факторів, які сильно корелюють, до

регресійної моделі недоцільно. Уявімо екстремальний випадок – включення до моделі таких факторів: ціни нафти за тонну та ціни нафти за барель. Кореляція між цими факторами дорівнюватиме 1. Але включати обидва фактори не варто. Ціна нафти за барель не несе жодної додаткової варіації (відповідно, й інформації), порівняно з ціною нафти за тонну. Аналогічно приріст ВВП і приріст ВНД сильно корелюватимуть, і включення обох факторів до моделі недоцільне: другий фактор дає лише мінімум інформації, порівняно з першим. У цьому плані приріст ВВП і приріст ВНД практично є одним фактором, і достатньо включити до моделі лише одну змінну, яка його представляє.

Мультиколінеарність є суттєвою проблемою, оскільки регресійні коефіцієнти при факторах, які сильно корелюють, можуть бути розраховані неправильно, і навіть мати неправильний знак (замість додатного від'ємний, або навпаки, тобто спотворюватимуть висновки про характер впливу фактора, що визначений за коефіцієнтом кореляції його із залежною змінною). Наприклад, якщо парний коефіцієнт кореляції і *b*-коефіцієнт мають різні знаки, то регресія швидше за все не є адекватною: вона неправильно визначає характер впливу цього фактора. Тоді варто перерахувати регресію без цього фактора або інших факторів, які з ним корелюють.

Найпростішим способом визначення мультиколінеарності є вивчення кореляційної матриці факторів. Не можна включати до моделі одночасно фактори, кореляція між якими більше 0,8-0,9. Але відсутність високих парних кореляцій не виключає мультиколінеарності. Наприклад, один із факторів може сильно корелювати з лінійною комбінацією кількох інших факторів, хоча з кожним окремим фактором кореляція буде помірною.

Іншими методами є *толерантність*/*tolerance value* та *фактор зростання дисперсії*/*variance inflation factor – VIF*. Обидва показники є оберненими одна до одної величинами. Їх розраховують для кожного фактора по черзі окремо. Для цього будують регресію, де відповідний фактор розглядають як

залежну змінну від решти факторів. Толерантність є частиною дисперсії такого залежного фактора, яку не пояснюють дисперсією решти факторів. Толерантність нижче 0.1 та *VIF* більше 10 означають наявність сильної кореляції між відповідним залежним фактором і рештою факторів.

До моделі варто включати ті фактори, які сильно корелюють із залежною змінною та слабо корелюють між собою. Ефективним засобом також є спеціальний метод аналізу основних компонент, але іноді його використання ускладнене важкістю інтерпретації некорелюючих штучних незалежних змінних, які він пропонує. У такому випадку звичайний кореляційний аналіз може мати свої переваги. У його рамках усі незалежні змінні поділяють на змінні, що умовно сильно корелюють між собою. Із кожної групи змінних залишають одну (як фактор вона представлятиме не тільки себе, а й решту змінних у групі). Наприклад, якщо обирають зростання індексу споживчих цін, а решта змінних у групі – це зростання індексу цін виробників і зростання грошової маси, то необхідно розуміти, що у подальшому аналізі буде аналізований вплив не просто зростання індексу цін виробників, а в цілому фактора, який можна назвати інфляцією, і зростанням грошової маси. Із групи варто обирати змінну, виходячи з таких міркувань:

- більша кількість спостережень із наявними даними за змінною;
- більший коефіцієнт кореляції змінної із залежною змінною;
- більший коефіцієнт кореляції квадрату відхилення змінної від середньої із залежною змінною.

Як виняток, регресію з мультиколінеарністю можна використовувати, але лише для прогнозування, а не оцінювання величини та характеру впливу факторів на залежну змінну. Розв'язанням проблеми мультиколінеарності також є використання *гребеневої регресії/ridge regression* або *регресії на баз основних компонент/regression on principal components*.

6. Додаткову проблему можуть створювати нестаціонарні часові ряди (коли середні змінних із часом змінюються).

Наприклад, якщо X та Y обидва зростають із часом з невеликою волатильністю, то кореляція між ними буде дуже високою. Але це не означає наявність причинно-наслідкового зв'язку між ними. Надзвичайно багато змінних із часом переважно тільки зростають (або навпаки, тільки зменшуються), але це не означає, що вони всі напряму пов'язані. Тому із нестаціонарними даними потрібно поводитися обережно. Регресію за такими даними варто будувати у тому разі, коли в теорії можна пояснити причинно-наслідковий зв'язок, або його наявність надійно доведена в принципі попередніми дослідженнями у світі (напр., інвестицій і ВВП), а наразі необхідно тільки визначити величину впливу в окремому випадку (інвестицій на ВВП у конкретній країні).

Альтернативою є використання приростів замість абсолютних значень змінних. Для них проблема нестаціонарності не властива, або менш виразна. Проте регресія на основі приростів без часового лагу може оцінити тільки короткостроковий вплив, із лаговими змінними – середньостроковий, і зазвичай тільки регресія на основі абсолютних значень – довгостроковий вплив.

9.5. Стійкість результатів регресійного аналізу

Лінійний регресійний аналіз фактично має ті самі передумови та обмеження, що й кореляційний аналіз. Оскільки ці методи взаємопов'язані, то їх часто розглядають як один: регресійно-кореляційний аналіз. Наприклад, так само, як і кореляційний аналіз, регресійний – не дає відповідь на запитання: що є причиною, а що – наслідком; чи є зміни у залежній змінній наслідком змін у факторах. Відповідь можливо отримати, якщо включити до моделі часові лаги. Припустимо, що як незалежну змінну розглядають приплив інвестицій у поточному році (t), а як залежну – приріст ВВП наступного року $(t+1)$. Тоді, якщо параметри якості такі, як коефіцієнт детермінації, то F-тест, t-тест щодо значущості

регресійного коефіцієнта для приросту ВВП регресії покажуть значущий зв'язок, і можна говорити, що саме приплив інвестицій впливає на приріст ВВП.

З іншого боку, навіть лаги не гарантують установлення напряму зв'язку:

- оскільки зменшення припливу інвестицій може також спричинити очікуваннями зменшення приросту ВВП;
- або внаслідок дії спільної причини (напр., економічний спад за кордоном може призвести як до зменшення припливу інвестицій, так і зменшення приросту ВВП через інший канал, зокрема, зменшення експорту).

Перевірку *стійкості результатів/robustness check* можна провести у кілька способів. При розрахунку значення залежної змінної варто враховувати довірчі інтервали для прогнозного значення. Вони можуть бути достатньо широкими та не давати можливість зробити точний прогноз.

Якісна регресійна модель має надавати достатньо точні розраховані значення залежної змінної на підставі даних щодо факторів за межами сукупності, на основі якої побудовано регресію. Тому варто перевірити, як працює регресійна модель на іншій вибірці. Часто це можна здійснити лише з часом, після одержання даних за новий період. Якщо ж модель спрацьовує дуже добре на наявній вибірці, але дає хибні результати за межами вибірки, на основі якої її було розраховано, то кажуть, що дослідник "підганяє результати" моделі під наявну вибірку. Така ситуація характерна, наприклад, для нелінійної моделі поліноміального тренду, специфікація якої не підкріплена теорією. Тому можна первинну вибірку поділити на дві чи більше частин і розрахувати дві або більше регресії на основі цих підвибірок. Далі перевіряють, наскільки збігаються чи відрізняються коефіцієнти регресій при факторах (величина, знак, рівень значущості) та коефіцієнти детермінації регресій (величина, рівень значущості).

Інший спосіб – використати інші визначення або методи вимірювання змінних (напр., замість приросту експорту щодо попередніх значень різницю у відношеннях експорту до

ВВП у поточному та попередньому роках). Далі перевіряють, наскільки збігаються чи відрізняються коефіцієнти регресій при факторах (знак, рівень значущості) та коефіцієнти детермінації регресій (величина, рівень значущості). Самі регресійні коефіцієнти можуть відрізнятися через інші одиниці вимірювання.

Зазвичай регресійна модель дозволяє достатньо точно розрахувати значення залежної змінної на основі типових для первинної вибірки (на основі якої розраховано коефіцієнти детермінації) значень факторів. Припустимо, що регресійна модель залежності приросту імпорту від приросту світового ВВП і зміни валютного курсу розраховували за вибіркою, де приріст світового ВВП варіювався від 1 до 6 %, а зростання валютного курсу від –10 до 10 %. Якщо потрібно розрахувати приріст імпорту на основі зовсім інших нових значень факторів (скажімо, приріст світового ВВП –3 %, а зростання валютного курсу –40 %), то розрахований приріст імпорту швидше за все буде неточним.

При використанні регресійної моделі потрібно пересвідчитися в тому, що умови не змінилися, порівняно з тими, що були характерні для спостережень, на основі яких розраховували регресію. Наприклад, навряд чи можна використати регресію, що розрахована лише на умови адміністративно-командної економіки, в умовах ринкової економіки; або розраховану лише на умови стабільного економічного зростання – в умовах рецесії; розраховану лише для високорозвинених країн – для країн із низьким рівнем розвитку.

Впливові спостереження (викиди, зокрема) можуть створювати суттєві проблеми в регресійному аналізі. Існує кілька варіантів, як такі спостереження можуть впливати на результати регресійного аналізу (рис. 9.2).

У випадку регресійного аналізу викиди легше за все визначити за величиною залишків (графічно величина залишку може бути представлена як розмір перпендикуляра від спостереження до регресійної прямої), хоча цей метод не дає повної картини про впливові спостереження, які також можна визначати графічно на вказаних вище діаграмах розсіювання.

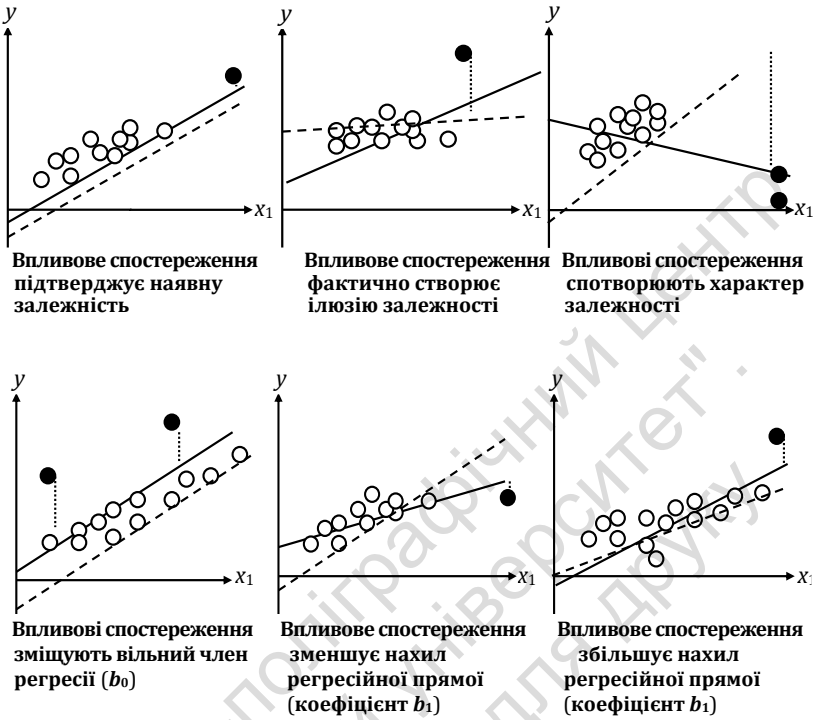


Рис. 9.2

Спостереження, для яких залишки є великими (більше двох-трьох стандартних відхилень) можна вважати викидами. За наявності викидів варто побудувати дві регресії: з викидами та без них і так само порівняти коефіцієнти при факторах і коефіцієнт детермінації.

Припустимо, потрібно розрахувати моделі залежності шестирічного зростання експорту послуг (2011-2017 рр.) від інституційних змінних для 24 країн Центральної і Східної Європи (ЦСЄ). У табл. 9.1 подано альтернативну форму представлення регресійних моделей замість формули для прикладу оцінювання стійкості результатів аналізу. Таблиця також ілюструє, які способи можна використати для оцінювання стійкості результатів аналізу:

Перший спосіб – порахувати моделі за розширеною вибіркою, до якої входять не тільки країни досліджуваного регіону, а й інші, проте країнам ЦСЄ надано більшої ваги (як у моделях 2 і 4). За умови зважування спостережень кількість спостережень штучно завищена, що ускладнює оцінювання статистичної значущості коефіцієнтів. Утім, якщо коефіцієнти регресії мають такий самий знак і мало відрізняються (як коефіцієнти 69,1 і 41,5 при змінній "верховенство права" у першій і другій моделях), це вказує на їх стійкість.

Другий спосіб – застосування альтернативних специфікацій моделей як способу реагування на проблему мультиколінеарності. До прикладу, у моделях 1 і 3 можна включити один й той самий фактор "монетарна свобода" (коефіцієнти при ньому статистично значущі, мають той самий знак і мало відрізняються: 1,47 і 1,81); другим фактором будуть різні фактори по черзі ("верховенство права" або "демократичність"), якщо регресійні коефіцієнти при них стають статистично незначущими у трифакторній моделі. Це дозволяє довести вплив монетарної свободи та спільний вплив щонайменше одного з двох інших факторів (але якого точно – не відомо, адже вплив одного від впливу іншого не можна чітко відрізнити через мультиколінеарність).

Нарешті, можна застосувати інше визначення зростання експорту. Наприклад, зміну відношення експорту до ВВП (замість приросту експорту щодо базового періоду):

$$d = \frac{\exp_{2017}}{GDP_{2017}} \cdot 100\ \% - \frac{\exp_{2011}}{GDP_{2011}} \cdot 100\ \%. \qquad (\mathbf{9.24})$$

Зауважимо, що коефіцієнти регресії у моделях 1-4 і 5-7 непорівнювані, оскільки одиниці вимірювання для різних визначень зростання експорту різняться (відсотки ВВП та експорту у 2011 р.). Головне пересвідчитися, чи залишаються коефіцієнти регресії при відповідних факторах статистично значущими та мають такий самий знак.

Таблиця 9.1

Модель	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Вибірка, визначення *Y*	ЦСЄ	Розширена	ЦСЄ	Розширена	ЦСЄ, альтернативне представлення *Y*		
Y-перетин	20.5*** (5.5)	27.7 *** (3.0)	25.8*** (5.5)	29.5*** (2.9)	3.0*** (0.51)	3.2*** (0.57)	2.4*** (0.56)
Монетарна свобода	1.47** (0.66)	0.82** (0.33)	1.81** (0.69)	0.94*** (0.33)		0.138* (0.071)	0.146** (0.065)
Верховенство права	69.1** (24.8)	41.5*** (12.7)			8.43*** (2.49)		
Демократичність			56.9** (23.3)	38.5*** (11.2)		6.10** (2.40)	
Стримування корупції							6.58*** (1.94)
R^2	0.39	0.06	0.35	0.06	0.34	0.30	0.41
p	0.006	0.000	0.011	0.000	0.003	0.025	0.004
N	24	300	24	300	24	24	24

Джерело: Shnyrkov O., Zablotska R., Chugaiev O. The Impact of Institutions on Services Exports of Central and Eastern European Countries // Baltic Journal of Economic Studies. – 2019. – Vol.5. – No.5. – P. 9-17. http://www.baltijapublishing.lv/index.php/issue/article/view/731

Примітка. Інституційні змінні (*використані як фактори*) *вимірюють як зростання балів.*

Значущість коефіцієнтів: * за $p < 0{,}1$; ** за $p < 0{,}05$; ** за $p < 0{,}01$; згідно з *t*-критерієм.

9.6. Регресійний аналіз з використанням бінарних змінних

У лінійній і більшості нелінійних регресійних моделей залежна змінна є кількісною змінною. Якщо залежна змінна є якісною, то використовують спеціальні види нелінійних регресійних моделей (логіт-регресія або пробіт-регресія).

Зазвичай для лінійної регресії використовують лише кількісні фактори, які вимірюють у метричній шкалі, рідше – у порядковій шкалі. Для включення якісних факторів використовують бінарні змінні або псевдозмінні, *фіктивні змінні /dummy variables*. Бінарна змінна набуває значень 1 або 0 (рідше 1 та –1). Наприклад, 1 – є зона вільної торгівлі, 0 – зона вільної торгівлі відсутня.

Якщо категоріальна змінна, яку вимірюють у номінальній шкалі, може набувати кількох значень k, то для її включення до регресії потрібно $k-1$ бінарних змінних. Наприклад, створимо для категоріальної змінної *Регіон* (яка може набувати чотирьох значень) три бінарні змінні, які можна включити до регресійної моделі (табл. 9.2).

Якщо йдеться про Америку, то всі бінарні змінні, крім Америки, матимуть значення 0. У випадку Африки всі бінарні змінні набуватимуть значення 0. Наведемо приклад інтерпретації результатів. Припустимо, що розрахована за умовними даними регресія має вигляд:

$$i = 5 + 1{,}5\pi - 3D_1 - 2D_2 + 1{,}5D_3, \qquad \textbf{(9.25)}$$

де i – відсоткова ставка (за зовнішніми кредитами) у відсотках; π – інфляція у відсотках.

Таблиця 9.2

Регіон	**Значення бінарної змінної**		
	Європа (D_1)	Азія, Австралія та Океанія (D_2)	Америка (D_3)
Європа	1	0	0
Азія, Австралія та Океанія	0	1	0
Америка	0	0	1
Африка	0	0	0

Аналіз моделі дає можливість стверджувати, що:

- в Африці за нульової інфляції (коли всі змінні набувають значення 0) відсоткова ставка дорівнюватиме 5 %;
- збільшення інфляції на кожний 1 % підвищує відсоткову ставку на 1,5 %;
- належність до регіону Європа (порівняно з Африкою) зменшує відсоткову ставку на 3 %;
- належність до регіону Азія, Австралія та Океанія (порівняно з Африкою) зменшує відсоткову ставку на 2 %;
- належність до регіону Америка (порівняно з Африкою) збільшує відсоткову ставку на 1,5 %.

Якщо, наприклад, потрібно порівняти, наскільки збільшує відсоткову ставку належність до Америки, порівняно з Європою, то дістанемо $4{,}5 = 1{,}5 - (-3)$.

Звичайно, як і у випадку звичайних змінних, бінарні змінні слід включати до регресії, якщо t-тест показує значущість коефіцієнта при відповідній бінарній змінній.

Альтернативний варіант урахування якісних змінних – побудова кількох регресійних моделей. У нашому прикладі потрібно побудувати чотири регресії (за спостереженнями для кожного з чотирьох регіонів). У такому випадку може статися й так, що b-коефіцієнт при змінній π буде різний у різних чотирьох моделях (буде виявлено ефект взаємодії між факторами Регіон та Інфляція), що також може являти інтерес. Доцільніше будувати окремі регресії після побудови регресії з бінарними змінними, пересвідчившись, що відповідна бінарна змінна значуще впливає на залежну змінну.

9.7. Аналіз взаємодії факторів у регресійній моделі

Для урахування *ефекту взаємодії*/*interaction effect* або *moderator effect* між факторами модель можна трансформувати (напр., у випадку двофакторної регресії):

$$Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2. \quad \textbf{(9.26)}$$

Так само потрібно перевіряти за допомогою t-тесту значущість коефіцієнта b_3. Якщо він значущий, то ефект взаємодії існує, тобто вплив одного із факторів залежить від то-

го, якого значення набуває інший фактор. Інший варіант визначення значущості ефекту взаємодії – побудова спочатку регресії без x_1x_2, а потім – із x_1x_2. Якщо коефіцієнти детермінації цих регресій значуще відрізняються, то ефект взаємодії значущий.

Наведемо приклад за умовними даними. Припустимо, що розрахована регресія має вигляд:

$$IMP = 2 + 4GDP + 2E + 0{,}5GDP \cdot E, \quad \textbf{(9.27)}$$

де IMP – приріст імпорту (%); GDP – приріст ВВП (%); E – зростання (реального ефективного) валютного курсу (%).

Сумарний вплив приросту ВВП на приріст імпорту залежить від зростання валютного курсу: $4 + 0{,}5 \cdot E$. Наприклад, якщо валютний курс зростає на 5 %, то приріст ВВП на кожний додатковий відсотковий пункт (в. п.) приведе до зростання імпорту на $6{,}5\,\% = 4 + 0{,}5 \cdot 5$. Якщо валютний курс залишається незмінним, то приріст ВВП на кожний додатковий в. п. приведе до зростання імпорту на $4\,\% = 4 + 0{,}5 \cdot 0$. Якщо валютний курс зменшується на 6 %, то приріст ВВП на кожний додатковий в. п. спричинятиме зростання імпорту лише на $1\,\% = 4 - 0{,}5 \cdot 6$.

Сумарний вплив зростання валютного курсу на приріст імпорту залежить від приросту ВВП: $2 + 0{,}5GDP$.

Трифакторна регресія з урахуванням взаємодії може мати вигляд (остання складова – добуток трьох факторів – може не використовуватися):

$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_1x_2 + b_5x_2x_3 + b_6x_1x_3 + b_7x_1x_2x_3 \quad \textbf{(9.28)}$$

Розгляд ефектів взаємодії може ускладнювати обмежена кількість спостережень. Кожний ефект взаємодії – це ніби додатковий фактор, а що більше факторів у регресії, то більшою має бути кількість спостережень для забезпечення надійності результатів.

Інша проблема – мультиколінеарність. Наприклад, якщо x_1 та x_1x_2 сильно корелюють, то x_1 та x_1x_2 одночасно не можуть перебувати у рівнянні регресії. У цьому випадку варто спробувати спочатку включити до регресійної моделі лише x_1, поряд з іншими факторами, а лише потім – x_1x_2, поряд з іншими факторами. Далі визначають, яка з моделей є більш адекватною.

9.8. Модель лінійної регресії впливу платіжного балансу на економічне зростання у Microsoft Excel

Припустимо, є дані за чотирма найбільшими країнами Центральної та Східної Європи за такими змінними (за даними *World Development Indicators*):

- *GNI_gr_next* – приріст ВНД наступного року з лагом 1 рік щодо вказаного року;
- *PrivGap_ch* – зміна відношення чистого припливу приватного капіталу до ВВП указаного року мінус аналогічне відношення попереднього року;
- *TradeBal* – баланс торгівлі товарами й послугами (сальдо).

Проведемо аналіз за даними (рис. 9.3, фрагмент). У надбудові *Аналіз даних/Data Analysis* оберіть опцію *Регресія/Regression*. Укажіть опції (як на рис. 9.4). Опцію *Константа дорівнює нулю/Constant is Zero* обирають не часто: лише якщо в теорії передбачено, що вільний член регресії (константа) дорівнює нулю.

	A	B	C	D
1		GNI_gr_next	PrivCap_ch	TradeBal
2	Poland 1998	4.4	0.0	-4.8
3	Poland 1999	4.1	-0.1	-5.9
4	Poland 2000	1.1	3.0	-6.4
5	Poland 2001	1.7	-3.8	-3.7
6	Poland 2002	4.5	-0.7	-3.5
7	Poland 2003	7.6	0.2	-2.7
8	Poland 2004	2.6	5.2	-2.4
9	Poland 2005	6.9	-1.9	-0.7
10	Poland 2006	7.8	-4.1	-1.8
11	Poland 2007	3.7	0.5	-2.9
12	Poland 2008	3.1	-1.3	-4.0
13	Poland 2009	3.7	3.9	0.1
14	Ukraine 1998	0.5	-2.5	-2.3
15	Ukraine 1999	6.2	-0.6	5.5
16	Ukraine 2000	7.7	1.1	5.0
17	Ukraine 2001	4.8	-0.4	1.6

Рис. 9.3

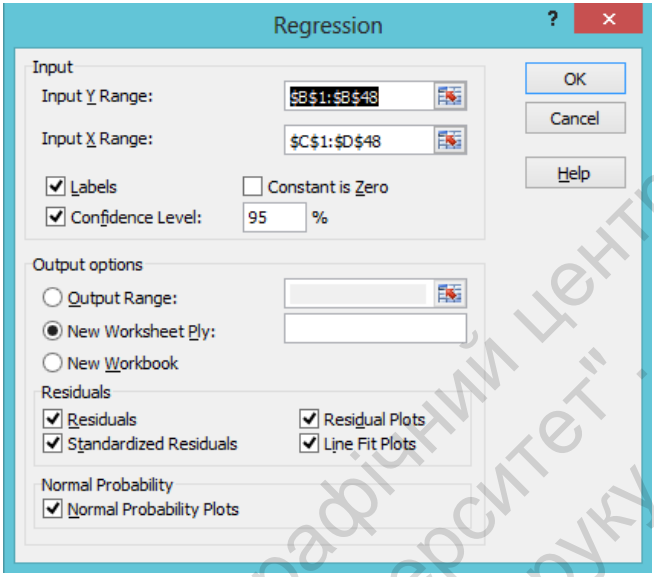


Рис. 9.4

У новому аркуші побачимо результати (рис. 9.5–9.11).

Із першої частини (рис. 9.5) видно, що коефіцієнт детермінації становить лише 0,1998, а скоригований – 0,1635. Використано 47 спостережень. Із другої частини (рис. 9.6) видно рівень значущості коефіцієнта детермінації (0.0074 менше 0.05), тобто він є значущим.

Regression Statistics	
Multiple R	0.447028557
R Square	0.199834531
Adjusted R Square	0.163463373
Standard Error	4.41628501
Observations	47

Рис. 9.5

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	214.3174794	107.1587397	5.494313175	0.007412347
Residual	44	858.1572248	19.50357329		
Total	46	1072.474704			

Рис. 9.6

Із третьої частини (рис. 9.7) можна побудувати рівняння регресії:

$$GNI\_gr\_next = 4{,}094 + 0{,}55 PrivGap\_ch + 0{,}230 TradeBal \quad (9.29)$$

При цьому всі коефіцієнти регресії є значущими (рівні значущості: 0,000, 0,019, 0,020). Указані 95 % довірчі межі для кожного коефіцієнта. Наприклад, коефіцієнт при *PrivGap_ch* перебуває в межах від 0,096 до 1,019.

Із четвертої частини (рис. 9.8, фрагмент), можна побудувати гістограму за допомогою опції *Гістограма/Histogram* надбудови *Пакет аналіза/Data Analysis*, прописавши перед цим інтервал карманів.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.09434466	0.669788508	6.112891767	2.303E-07	2.744474617	5.4442147
PrivCap_ch	0.5572057	0.228911487	2.434153526	0.0190538	0.095864915	1.01854649
TradeBal	0.22995094	0.09537579	2.410999006	0.020152	0.037733661	0.42216821

Рис. 9.7

Observation	*Predicted GNI_gr_next*	*Residuals*	*Standard Residuals*
1	3.007236241	1.42561278	0.330063168
2	2.6888115	1.38080639	0.319689426
3	4.289162695	-3.2052459	-0.742090438
4	1.155673476	0.50900628	0.117847026
5	2.902713836	1.57701076	0.365115391
6	3.575453928	3.98934829	0.923628741
7	6.454055412	-3.877466	-0.897725326
8	2.870198415	4.01886679	0.930462973
9	1.36947335	6.45646193	1.494824058
10	3.707778627	-0.0537875	-0.012453087
11	2.482814542	0.65156174	0.150851996
12	6.278835855	-2.562844	-0.593359158
13	1.177149636	4.92925895	1.141240349
14	5.979617545	2.95766576	0.684769768
15	6.338773945	-2.5504415	-0.590487691
16	9.246171144	-4.0180677	-0.930277965
17	7.121661367	1.3523448	0.313099894
18	5.42999325	0.94921178	0.219765038
19	7.940441869	-1.0754438	-0.248990746
20	6.206313181	2.65700435	0.615159521
21	9.094885383	-1.1744517	-0.271913419
22	5.455081278	0.74241666	0.171887065
23	5.049272691	-13.068339	-3.02563031

Рис. 9.8

Із гістограми (рис. 9.9) видно, що розподіл залишків неідеальний, але близький до нормального. За *стандартизованими викидами/Standardized Residuals* видно, що є два викиди (на рис. 9.8 видно тільки один – спостереження № 23). Тому варто видалити викиди з вибірки та провести аналіз повторно, порівнюючи результати з результатами на основі всіх спостережень. При цьому виявиться, що регресійний коефіцієнт при факторі *PrivGap_ch* статистично незначущий, отже потрібно цей фактор вилучити із моделі та будувати однофакторну модель на базі фактора *TradeBal.*

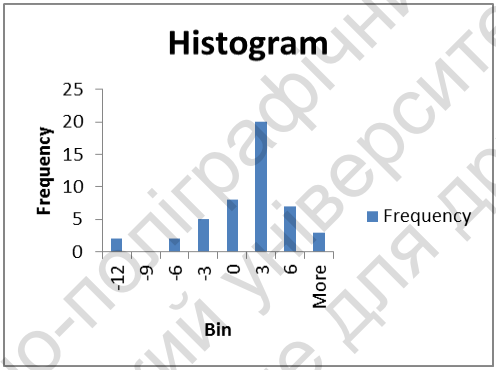


Рис. 9.9

Із графіків залишків і підбору (рис. 9.10–9.11) можна дізнатися про лінійний або нелінійний характер зв'язку за кожним фактором і непрямо – про наявність чи відсутність гетероскедастичності.

У Microsoft Excel можна також самостійно будувати регресію за допомогою функцій LINEST, F.DIST, T.DIST., TREND, FORECAST.

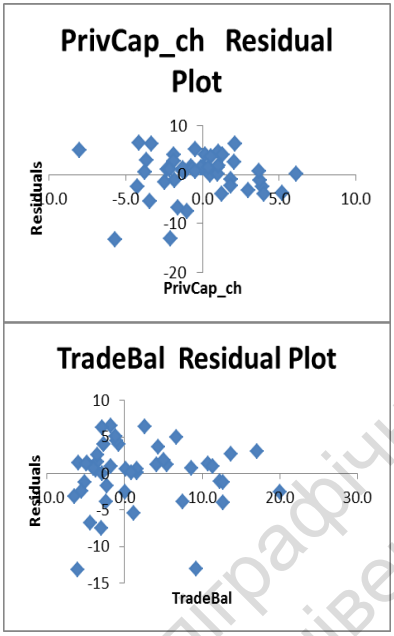


Рис. 9.10

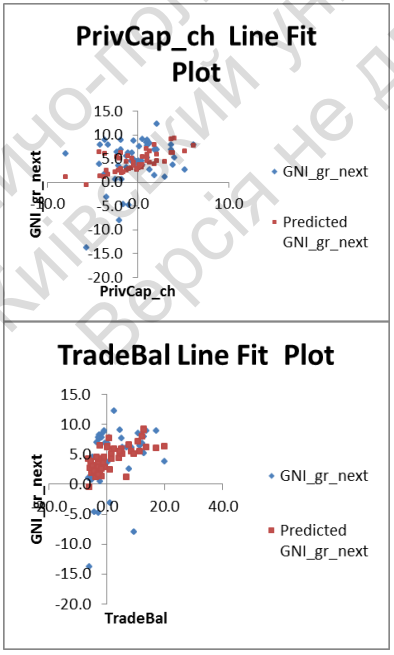


Рис. 9.11

Розділ 10
МЕТОДИ НЕЛІНІЙНОГО РЕГРЕСІЙНОГО АНАЛІЗУ

10.1. Види моделей нелінійної регресії

Серед усіх багатовимірних регресійних моделей найбільш вивчено лінійну, хоча лише деякі з економічних процесів у сфері міжнародних економічних відносин можна моделювати за допомогою такої моделі. Її вибір залежить від типу процесу й тривалості спостереження за ним. Багато процесів у сфері міжнародних економічних відносин за нетривалого спостереження можна із певним наближенням моделювати за допомогою лінійної багатофакторної моделі.

Значну частину нелінійних регресійних моделей можна перетворити на лінійні за допомогою трансформації змінних, а після цього – побудувати їх як звичайні лінійні регресійній моделі за допомогою методу найменших квадратів. Різниця буде лише в інтерпретації моделі.

Розглянемо деякі випадки однофакторної нелінійної моделі, які шляхом перетворень зводяться до лінійної моделі:

$$y = b_0 + b_1 x. \qquad (\mathbf{10.1})$$

Таблиця 10.1 демонструє, яким чином (за допомогою яких замін) можна перетворити змінні та коефіцієнти регресії для ряду розповсюджених нелінійних однофакторних моделей для зведення їх до лінійної моделі (**10.1**).

Наприклад, двофакторну модель:

$$y = b_0 x_1^{b_1} x_2^{b_2} \qquad (\mathbf{10.2})$$

можна перетворити шляхом логарифмування до моделі:

$$\ln y = \ln b_0 + b_1 \ln x_1 + b_2 \ln x_2, \qquad (\mathbf{10.3})$$

яку після заміни $Y = \ln y$, $B_0 = \ln b_0$, $X_1 = \ln x_1$, $X_2 = \ln x_2$ записують у вигляді:

$$Y = B_0 + b_1 X_1 + b_2 X_2. \qquad (\mathbf{10.4})$$

Але деякі нелінійні регресійні моделі неможливо трансформувати у лінійні за допомогою замін. Для них потрібно використовувати інші методи, а не метод найменших квадратів. Наведемо приклади.

Таблиця 10.1

Функція	y	x	b_0	b_1
$y = b_0 + b_1 / x$	y	$1/x$	b_0	b_1
$y = 1/(b_0 + b_1 x)$	$1/y$	x	b_0	b_1
$y = x/(b_0 + b_1 x)$	x/y	x	b_0	b_1
$y = b_0 + b_1^{\,x}$	$\ln(y)$	x	$\ln(b_0)$	$\ln(b_1)$
$y = 1/(b_0 + b_1 e^x)$	$1/y$	e^x	b_0	b_1
$y = b_0 x^{b_1}$	$\ln(y)$	$\ln(x)$	$\ln(b_0)$	b_1
$y = b_0 + b_1 \ln(x+1)$	y	$\ln(x+1)$	b_0	b_1
$y = b_0 x(b_1 + b_2 x)$	$1/y$	$1/x$	b_1/b_0	b_2/b_0
$y = b_0 e^{b_1/x}$	$\ln(y)$	$1/x$	$\ln(b_0)$	b_1
$y = b_0 + b_1 x^n$	y	x^n	b_0	b_1

Експоненціальну регресію/Exponential growth regression визначають формулою (де exp означає число Ейлера в степені вказаної лінійної комбінації факторів):

$$Y = c + \exp(b_0 + b_1 x_1 + b_2 x_2 + ...). \qquad (10.5)$$

Кусково-лінійна регресія/Piecewise linear regression має загальну формулу:

$$Y = (b_{01} + b_{11} x_1 + b_{21} x_2 + ...) \cdot (Y < Y^*) + \\ + (b_{02} + b_{12} x_1 + b_{22} x_2 + ...) \cdot (Y > Y^*), \qquad (10.6)$$

де $Y < Y^*$ та $Y > Y^*$ – логічні вирази, які повертають 1, якщо відповідний вираз відповідає дійсності, або 0, – якщо не відповідає. Тобто за значень залежної змінної, що менша від указаного рівня, треба використати форму регресії $b_{01} + b_{11} x_1 + b_{21} x_2 + ...$, а у разі більшого – $b_{02} + b_{12} x_1 + b_{22} x_2 + ...$.

Точку розриву/Breakpoint Y^* можна вказати серед початкових значень коефіцієнтів або розрахувати за допомогою програмного забезпечення.

Моделі бінарних відгуків (логіт або *пробіт*) використовують, якщо залежна змінна є бінарною (напр., 1 – валютна криза наявна, 0 – валютної кризи немає). Але в регресії можна використати залежну змінну, що є неперервною та може набувати значень від 0 до 1. У логіт-регресії це досягають за допомогою рівняння:

$$p = y = \frac{\exp(b_0 + b_1 x_1 + ... + b_n x_n)}{1 + \exp(b_0 + b_1 x_1 + .. + b_n x_n)}, \quad (\mathbf{10.7})$$

де *y* – неперервна залежна змінна на відрізку [0;1] – аналог імовірності *p* того, що бінарна залежна змінна набуває значення 1. Саме цю формулу використовують для розрахунку прогнозного значення відгуку за відомих значень факторів і регресійних коефіцієнтів.

Для отримання звичайної неперервної залежної змінної на відрізок $(-\infty, +\infty)$ для розрахунку регресійних коефіцієнтів використовують логіт-перетворення:

$$p' = \ln\left(\frac{1}{1-p}\right) = b_0 + b_1 x_1 + ... + b_n x_n. \quad (\mathbf{10.8})$$

Для оцінювання якості логіт-моделі використовують χ^2-критерій і *співвідношення шансів/odds ratio*. Останнє показує, наскільки точно дозволяє передбачити значення залежної змінної розрахована логіт-регресія.

10.2. Функції втрат

Іноді програмне забезпечення дозволяє користувачу обирати *функцію втрат/Loss function*, яка мінімізується при визначенні коефіцієнтів регресії. Відхилення значень спостережень від розрахованих фактично і є втратами точності передбачення за допомогою моделі.

За звичайного методу найменших квадратів мінімізується сума квадратів залишків (відхилень значень спостережень залежної змінної від значень, що розраховані за допомогою регресії). Але можливо, наприклад, замість квадратів залишків використати абсолютні величини залишків. Цей метод зменшить вплив великих залишків на форму регресії.

Метод зважених найменших квадратів передбачає, що залишки можуть бути зважені за допомогою певної іншої змінної, наприклад, з метою здійснення поправки на гетероскедастичність або серійну кореляцію.

До функції втрат може бути включена штрафна функція, яка дорівнює 0, якщо оцінювані параметри перебувають у межах допустимих значень; і дорівнює дуже великій величині, якщо параметри виходять за межі допустимих значень. Наприклад, якщо важливо, щоб вільний член b_0 був невід'ємним, то функція втрат може мати вигляд ($b_0 < 0$ – логічний вираз, який набуває значення 1, якщо оцінений вільний член стає від'ємним):

$$L = (obs - pred)^2 + (b_0 < 0) \cdot 1000000. \quad (\mathbf{10.9})$$

За нелінійного оцінювання часто використовують *метод максимальної правдоподібності/maximum likelihood*. Він дає ті самі оцінки параметрів регресії, що й метод найменших квадратів, якщо виконуються передумови для регресійного аналізу. Перед пошуком мінімуму потрібно вказати початкові значення коефіцієнтів регресії і величину кроків, відштовхуючись від яких програма ітеративним шляхом намагатиметься знайти мінімум функції втрат. Критерій сходження визначає момент, в який ітерації можливо припинити.

На жаль, часто алгоритм знаходить локальний мінімум функції втрат і зупиняється у пошуках, а глобальний (абсолютний) мінімум залишається незнайденим. Відповідно, розраховані коефіцієнти регресії виявляться неоптимальними. Візуально локальні мінімуми є невеликими западинами на графіку функції втрат. Наприклад, на рис. 10.1 показано, як за певного первинного значення коефіцієнта регресії програма зупинить пошук оптимального значення b_1, коли знайде другий локальний мінімум. Тому варто спробувати різні початкові значення.

Розповсюдженим алгоритмом мінімізації функцій втрат є *квазіньютонівський метод/quasi-Newton*. Він передбачає розрахунок першої і другої похідних функції втрат для пошуку її мінімуму.

Крім нього доступні й інші алгоритми:

- *симплекс-метод*/*Simplex*, що є менш чутливим до локальних мінімумів;
- *метод Хука–Дживіса*/*Hooke–Jeeves pattern moves*;
- *метод Розенброка*/*Rosenbrock pattern search* або метод обертання координат;
- комбінація двох перелічених методів.

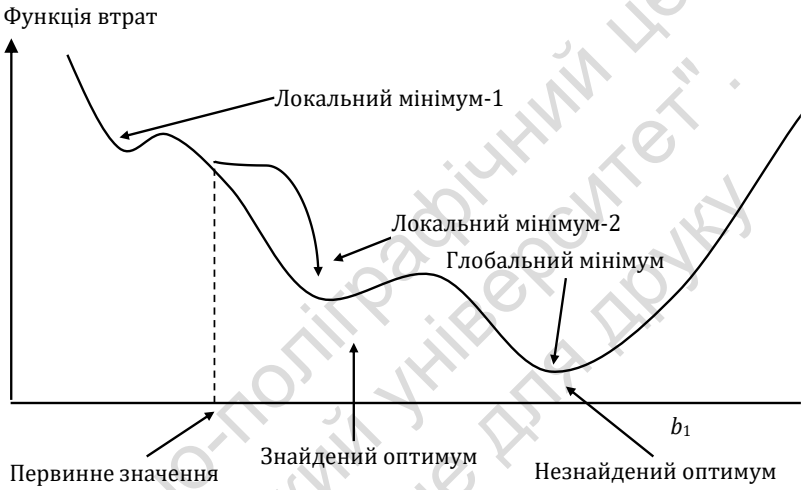


Рис. 10.1

10.3. Модель нелінійної регресії факторів високотехнологічного експорту у Microsoft Excel

У більшості випадків як фактори потрібно використовувати вже трансформовані змінні (напр., квадрат незалежної змінної), щоб звести модель до лінійної, і далі використовувати опцію *Регресія*/*Regression* у надбудові *Пакет Аналізу*/ *Data Analysis*. Але існують також вбудовані функції: LOGEST та GROWTH.

Наведемо приклад використання такої функції масиву:

= LOGEST (відомі значення y; відомі значення x;константа;статистика)

Аргументи функції такі. Відомі значення: *y* – діапазон значень залежної змінної; *x* – діапазон значень незалежних

змінних. Константа – якщо TRUE, то b розраховують; якщо FALSE, то вважають, що $b = 1$. Статистика: якщо TRUE, то розраховують додаткову статистику.

Функцію вводять як функцію масиву за допомогою команди CTRL+SHIFT+ENTER. Рівняння функції має вигляд:

$$y = bm_1^{x_1} m_2^{x_2}.$$

Припустимо, умова задачі – розрахувати залежність частки високотехнологічного експорту в експорті продукції обробної промисловості *HTExp* від змінних: *Lit* – рівень писемності населення не молодше 15 років; *Internet* – частка осіб, що користуються інтернетом. До прикладу, на рис. 10.2 подано значення змінних за даними Світового банку станом на 2009 р.

Вводимо функцію:

= LOGEST (B2:B15;C2:D15;TRUE;TRUE).

Одержуємо результати (рис. 10.3).

Отже, залежність має вигляд:

$$HTExp = 5{,}574(0{,}963127^{Lit}) \cdot (1{,}06635^{Internet}). \quad \textbf{(10.10)}$$

	A	B	C	D
1		HTExp	Lit	Internet
2	Panama	0.02	93.61	27.73
3	Portugal	4.16	94.91	48.61
4	Romania	10.07	97.65	36.25
5	Russian Federation	9.33	99.56	42.09
6	Saudi Arabia	0.26	86.13	36.55
7	Singapore	49.06	94.71	73.35
8	El Salvador	5.03	84.10	14.43
9	Slovenia	6.51	99.68	63.66
10	Trinidad and Tobago	0.26	98.74	36.29
11	Turkey	1.87	90.82	36.76
12	Uruguay	5.37	98.27	55.46
13	Samoa	0.33	98.78	4.93
14	Yemen, Rep.	0.38	62.39	1.80
15	Zambia	1.56	70.88	6.42

Рис.10.2

F	G	H
1.0663453	0.96313	5.574194
0.0296609	0.05695	4.607095
0.3360428	1.80654	#N/A
2.783666	11	#N/A
18.169464	35.8994	#N/A
#N/A	#N/A	#N/A

Рис.10.3

Тепер можна розрахувати залежну змінну, підставляючи відомі значення факторів (рис. 10.4).

HTExp	Lit	Internet
2846.99	5	100

HTExp	Lit	Internet
=H2*POWER(G2;K3)*POWER(F2;L3)	5	100

Рис.10.4

Проте, слід із застереженням використовувати ці результати. Модель потребує трансформації, оскільки коефіцієнт 0,963127 менший за 1, отже писемність негативно впливає на частку високотехнологічного експорту. У той самий час кореляція між писемністю та часткою високотехнологічного експорту слабко позитивна 0,209 (якщо скористатися функцією CORREL). Негативний вплив є дивним з теоретичного погляду. Тому, імовірно, у моделі існує ефект мультиколінеарності, що має спотворюючий характер.

Коефіцієнт детермінації становить 0,336. *F*-статистика 2,78. Кількість ступенів вільності 11. 4,61; 0,057; 0,297 – стандартні похибки для кожного з коефіцієнтів регресії; 1,81 – стандартна похибка для залежної змінної. Але слід враховувати, що додаткову статистику розраховують для лінеарізованої моделі:

$$\ln(HTEep) = \ln(5{,}574) + Lit \cdot \ln(0{,}963127) + \\ + Internet \cdot \ln(1{,}06635). \quad (\mathbf{10.11})$$

Аби дізнатися про рівень значущості *F*-критерію скористаємося функцією: =F.DIST(F; v1; v2), де *F*-значення *F*-статистики, v1 – кількість факторів, v2 – кількість ступені вільності.

У нашому прикладі =FРАСП(F5; 2; G5) повертає 0,1051, тобто коефіцієнт детермінації незначущий (більше 0,05).

10.4. Побудова та аналіз гравітаційної моделі міжнародної торгівлі з використання засобів MS Excel

Гравітаційна модель – це модель макрорівня, яка призначена для дослідження впливу змін економічної динаміки на товарообіг між певними країнами та для пояснення наявних співвідношень між товарообігом для певних країн. Згідно з цією моделлю товарообіг між країнами *i* та *j* буде прямо пропорційний добутку ВВП цих країн та обернено пропор-

ційний – квадрату середньої відстані перевезення товару при здійсненні торгівельних операцій між цими країнами. При виборі гравітаційної моделі встановлюють причинно-наслідковий зв'язок між досліджуваними економічними показниками.

Далі здійснюють специфікацію моделі. Якщо позначити товарообіг як T_{ij}, ВВП країни i – як V_i, ВВП країни j – як V_j, середню відстань перевезень – як R_{ij}, то одержимо рівність:

$$T_{ij} = \gamma \frac{V_i V_j}{R_{ij}^2}, \qquad (\mathbf{10.12})$$

яка, власне, утворює "гравітаційну" модель. Коефіцієнт γ є, строго кажучи, індивідуальним для будь-якої пари країн, оскільки він має враховувати:

- тарифно-митний режим, обмінно-курсову валютну політику (напр., застосування різних офіційних курсів обміну валют при здійсненні різних зовнішньоторговельних операцій), загальноекономічну політику країн-партнерів та інші складові умов торгівлі між ними;
- особливості структури внутрішнього виробництва, що впливають на склад експортного попиту та пропозиції (напр., унаслідок цього товарообіг між близько розташованими нафтовидобувними країнами Середнього Сходу набагато менший за їх товарообіг із розвиненими постіндустріальними країнами);
- специфіку ментальності та історичних традицій партнерів;
- інші чинники.

Якщо зазначені чинники зазнають суттєвих змін у часі, то застосування "гравітаційної" моделі стає проблемним.

Визначимо параметри вибраного рівняння. Параметр γ можна визначити зі статистичних даних. Наприклад, якщо протягом низки років відомі обсяги товарообігу T_{ij}^t (t – номер року, $t = -1, \ldots, n$) і ВВП країн за ті самі роки (позначимо їх як V_i^t, V_j^t), то для визначення γ можна застосувати метод найменших квадратів, згідно з яким γ буде точкою мінімуму функції:

$$F(\gamma)=\sum_{t=1}^{n}\left(T_{ij}^{t}-\gamma\frac{V_i^tV_j^t}{R_{ij}^2}\right)^2,$$

Знайдену похідну за γ цієї функції прирівнюють до нуля, тоді:

$$-2\sum_{t=1}^{n}\left(\left(T_{ij}^{t}-\gamma\frac{V_i^tV_j^t}{R_{ij}^2}\right)\frac{V_i^tV_j^t}{R_{ij}^2}\right)=0.$$

Звідси
$$\hat{\gamma}=\frac{\sum_{t=1}^{n}T_{ij}^{t}\frac{V_i^tV_j^t}{R_{ij}^2}}{\sum_{t=1}^{n}\left(\frac{V_i^tV_j^t}{R_{ij}^2}\right)^2}.$$

Після отримання значення $\hat{\gamma}$ можна обчислювати прогнозні значення товарообігу.

Проведемо аналіз якості моделі. Для оцінювання точності прогнозу за допомогою моделі визначають величину товарообігу для попередніх років і порівнюють її з відомим товарообігом за ті самі роки. Маємо:

$$T_{ij}^{t}=\hat{\gamma}\frac{V_i^tV_j^t}{R_{ij}^2}$$

– прогнозований товарообіг, який можна порівняти з реальним товарообігом T_{ij}^{t} за ті самі роки, обчислюючи середню відносну похибку оцінки:

$$\delta T_{ij}=\frac{1}{n}\sum_{t=1}^{n}\frac{\left|\hat{T}_{ij}^{t}-T_{ij}^{t}\right|}{T_{ij}^{t}}\cdot 100\ \%.$$

У табл. 10.2 (зовнішня торгівля товарами, млн дол. США, T_{ij}^{t}) подано інформацію про обсяги зовнішньої торгівлі товарами між країнами *i* та *j* за 2005-2009 рр., у табл. 10.3 – значення ВВП країн (V_i, V_j, млрд дол. США). Середня відстань перевезень становить $R_{ij}=789$. Використовуючи MS Excel, знайдемо оцінку $\hat{\gamma}$ коефіцієнта γ гравітаційної моделі та обчислимо середню відносну похибку[141].

[141]Див. : Грисенко М.В., Шворак Л.Л, Рижов А.Ю. Економіко-математичне моделювання світогосподарських процесів : навч. посіб. – К. : ВПЦ "Київський університет", 2016. – Ч. ІІІ. Практикум. – 229 с.

Таблиця 10.2

Країна	2005	2006	2007	2008	2009
j	2416,3	3453,7	4005,2	4130,9	4406,5

Таблиця 10.3

Країна	2005	2006	2007	2008	2009
i	100,8	124,9	122,0	140,5	116,2
j	263,7	290,8	324,9	380,1	430,2

Позначимо:

$$X_{ij}^t = \frac{V_i^t V_j^t}{R_{ij}^2},\ t = \overline{1,5}.$$

За допомогою такої заміни гравітаційну модель зводять до вигляду $T_{ij}^t = \gamma X_{ij}^t$, тобто оцінка параметра γ є оцінкою методу найменших квадратів простої лінійної регресії без вільного коефіцієнта. У діапазоні A1:D6 розташуємо умову задачі (рис. 10.5), а у стовпчику E обчислимо значення величин X_{ij}^t.

	A	B	C	D	E
1	t	T_{ij}^t	V_i	V_j	X_{ij}
2	1	2416,3	100,8	263,7	0,0427
3	2	3453,7	124,9	290,8	0,05834
4	3	4005,2	122	324,9	0,06367
5	4	4130,9	140,5	380,1	0,08579
6	5	4406,5	116,2	430,2	0,0803
7					

	A	B	C	D	E
1	t	T_{ij}^t	V_i	V_j	X_{ij}
2	1	2416,3	100,8	263,7	=C2*D2/789^2
3	2	3453,7	124,9	290,8	=C3*D3/789^2
4	3	4005,2	122	324,9	=C4*D4/789^2
5	4	4130,9	140,5	380,1	=C5*D5/789^2
6	5	4406,5	116,2	430,2	=C6*D6/789^2
7					

Рис. 10.5

Для знаходження значення γ скористаємось функцією ЛИНЕЙН() з параметром Конст = 0, тобто покажемо, що модель не містить вільного коефіцієнта (рис. 10.6–10.7).

	A	B	C	D	E
1	t	T_{ij}^t	V_i	V_j	X_{ij}
2	1	2416,3	100,8	263,7	=C2*D2/789^2
3	2	3453,7	124,9	290,8	=C3*D3/789^2
4	3	4005,2	122	324,9	=C4*D4/789^2
5	4	4130,9	140,5	380,1	=C5*D5/789^2
6	5	4406,5	116,2	430,2	=C6*D6/789^2
7					
8	=ЛИНЕЙН(B2:B6;E2:E6;0;1)	=ЛИНЕЙН(B2:B6;E2:E6;0;1)			
9	=ЛИНЕЙН(B2:B6;E2:E6;0;1)	=ЛИНЕЙН(B2:B6;E2:E6;0;1)			
10	=ЛИНЕЙН(B2:B6;E2:E6;0;1)	=ЛИНЕЙН(B2:B6;E2:E6;0;1)			
11	=ЛИНЕЙН(B2:B6;E2:E6;0;1)	=ЛИНЕЙН(B2:B6;E2:E6;0;1)			
12	=ЛИНЕЙН(B2:B6;E2:E6;0;1)	=ЛИНЕЙН(B2:B6;E2:E6;0;1)			

Рис. 10.6

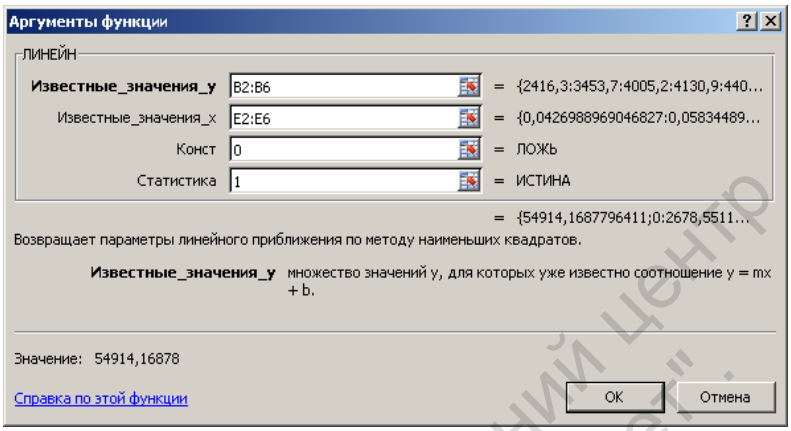


Рис. 10.7

Таким чином, у комірках A8 та A9 знайдено значення коефіцієнта $\hat{\gamma}=54914{,}17$ та його середнього квадратичного відхилення $S=2678{,}55$.

Довірчий інтервал для справжнього значення параметра γ з довірчою ймовірністю 95 % має вигляд:

$$\left[\hat{\gamma}-t_{\frac{\alpha}{2};n-2}\cdot S;\ \hat{\gamma}+t_{\frac{\alpha}{2};n-2}\cdot S\right],$$

де $t_{\frac{\alpha}{2};n-2}$ – критичне значення розподілу Стьюдента, яке визначають за заданим значенням рівня значущості α та числом ступенів свободи $k=n-1$.

Оскільки потрібно знайти 95 %-й довірчий інтервал, то $\alpha=1-0{,}95=0{,}05$, а кількість ступенів свободи $k=n-1=5-1=4$.

Для відшукання значення $t_{\frac{\alpha}{2};n-2}$ скористаємось убудованою функцією СТЬЮДРАСПОБР() (рис. 10.8).

Отримаємо довірчий інтервал:

$$[54914{,}17-2{,}78\cdot 2678{,}55;\ 54914{,}17+2{,}78\cdot 2678{,}55]$$

або

$$[47477{,}32;\ 62351{,}02].$$

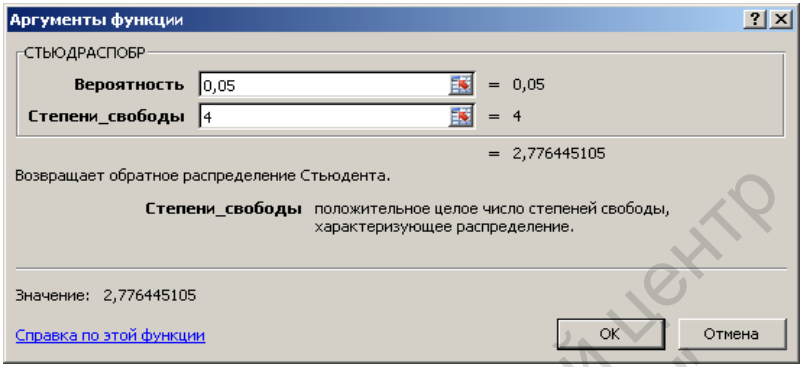


Рис. 10.8

Оскільки число 0 не належить до отриманого довірчого інтервалу, то на рівні $\alpha = 0{,}05$ можна стверджувати, що $\gamma \neq 0$.

Знайдемо у діапазоні F2:F6 значення $\hat{T}_{ij}^{t}$, скориставшись для цього вбудованою функцією ТЕНДЕНЦІЯ() (рис. 10.9).

	A	B	C	D	E	F
1	t	T_{ij}^{t}	V_i	V_j	X_{ij}	$\hat{T}_{ij}^{t}$
2	1	2416,3	100,8	263,7	0,0427	2344,774
3	2	3453,7	124,9	290,8	0,0583	3203,961
4	3	4005,2	122	324,9	0,0637	3496,552
5	4	4130,9	140,5	380,1	0,0858	4710,908
6	5	4406,5	116,2	430,2	0,0803	4409,679

	A	B	C	D	E	F
1	t	T_{ij}^{t}	V_i	V_j	X_{ij}	$\hat{T}_{ij}^{t}$
2	1	2416,3	100,8	263,7	=C2*D2/789^2	=ТЕНДЕНЦИЯ($B$2:$B$6;$E$2:$E$6;E2;0)
3	2	3453,7	124,9	290,8	=C3*D3/789^2	=ТЕНДЕНЦИЯ($B$2:$B$6;$E$2:$E$6;E3;0)
4	3	4005,2	122	324,9	=C4*D4/789^2	=ТЕНДЕНЦИЯ($B$2:$B$6;$E$2:$E$6;E4;0)
5	4	4130,9	140,5	380,1	=C5*D5/789^2	=ТЕНДЕНЦИЯ($B$2:$B$6;$E$2:$E$6;E5;0)
6	5	4406,5	116,2	430,2	=C6*D6/789^2	=ТЕНДЕНЦИЯ($B$2:$B$6;$E$2:$E$6;E6;0)

Рис. 10.9

Приклад обчислення значення $\hat{T}_{ij}^{1}$ для $t = 1$ показано на рис. 10.10. Для обчислення середньої відносної похибки у діапазоні G2:G6 знайдемо відносні похибки для кожного t (рис. 10.11).

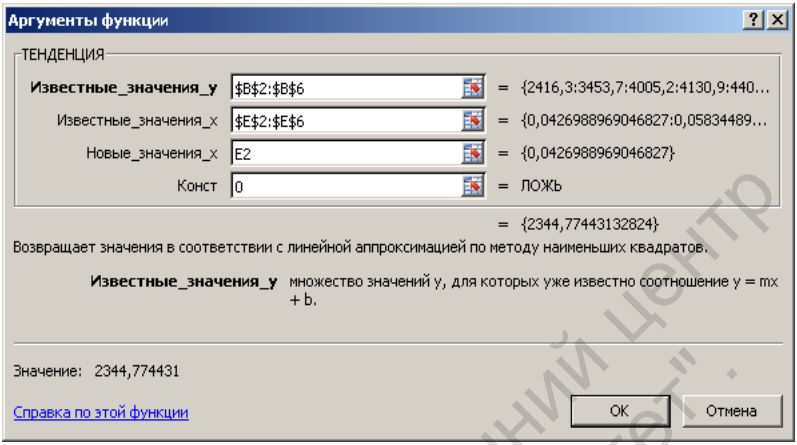


Рис. 10.10

	A	B	C	D	E	F	G
1	t	$T^{\dagger}_{ij}$	V_i	V_j	X_{ij}	$\hat{}T^{\dagger}_{ij}$	delta $T^{\dagger}_{ij}$
2	1	2416,3	101	264	0,0427	2344,774431	0,029601
3	2	3453,7	125	291	0,0583	3203,961201	0,072311
4	3	4005,2	122	325	0,0637	3496,551665	0,126997
5	4	4130,9	141	380	0,0858	4710,907769	0,140407
6	5	4406,5	116	430	0,0803	4409,678649	0,000721

	A	B	C	D	E	F	G
1	t	$T^{\dagger}_{ij}$	V_i	V_j	X_{ij}	$\hat{}T^{\dagger}_{ij}$	delta $T^{\dagger}_{ij}$
2	1	2416,3	100,8	263,7	=C2*D2/789^2	=ТЕНДЕНЦИЯ	=ABS(F2-B2)/B2
3	2	3453,7	124,9	290,8	=C3*D3/789^2	=ТЕНДЕНЦИЯ	=ABS(F3-B3)/B3
4	3	4005,2	122	324,9	=C4*D4/789^2	=ТЕНДЕНЦИЯ	=ABS(F4-B4)/B4
5	4	4130,9	140,5	380,1	=C5*D5/789^2	=ТЕНДЕНЦИЯ	=ABS(F5-B5)/B5
6	5	4406,5	116,2	430,2	=C6*D6/789^2	=ТЕНДЕНЦИЯ	=ABS(F6-B6)/B6
7							

Рис. 10.11

Усереднивши отримані значення, отримаємо:

$$\delta T_{ij} = 0{,}37/5 \cdot 100\,\% = 7{,}4\,\%.$$

10.5. Алгоритм дослідження наслідків утворення зони вільної торгівлі

Розглянемо як приклад методологічний алгоритм для дослідження потенційних наслідків утворення зони вільної торгівлі (ЗВТ) між Україною й Туреччиною[142]. Ці результати можна використати для оцінювання доцільності утворення зони вільної торгівлі, а також наслідків для взаємного торговельного балансу. За основу беруть гравітаційну модель у загальному модифікованому вигляді:

$$TRAde = b_0 \cdot GDP^{b_1} \cdot Dist^{b_2} \cdot e^{b_3 FTA}, \qquad (\mathbf{10.13})$$

де *TRAde* – експорт *exp* (або імпорт *imp*) Туреччини до/з відповідної країни-партнера; *GDP* – ВВП країни-партнера; *Dist* – відстань від столиці країни-партнера до Анкари; *FTA* – одна зі змінних, що характеризують регулювання міжнародної торгівлі.

Якщо прологарифмувати рівняння, то воно матиме вигляд (b_0 і всі змінні, крім *TR*, у формі натурального логарифму):

$$TRAde = b_0 + b_1 GDP + b_2 Dist + b_3 FTA. \qquad (\mathbf{10.14})$$

Змінну *Dist* в окремих моделях не використовують, якщо:

- імпорт Туреччини мало залежить від відстані до країни походження імпорту;
- відстань в окремих випадках корелює зі змінною регулювання торгівлі. Тоді залишають лише змінну регулювання торгівлі, хоча це зменшує надійність оцінки впливу саме зони вільної торгівлі (вплив її важче відрізнити від впливу відстані);

Використаємо почергово кілька варіантів визначення *TR*:

- *FT* – бінарна змінна, яка набуває значення 1 (є ЗВТ або митний союз) або 0 (немає ЗВТ або митного союзу);
- *TF* – зважена середня ставка митного тарифу, який застосовують для всіх продуктів у країнах – контрагентах Туреччини (для країн із ЗВТ вважаємо, що $TF = 0$);

---

[142] Дет. алгоритм див. : Хмара М.П., Чугаєв О.А. Оцінка потенційних наслідків торговельної інтеграції на прикладі формування зони вільної торгівлі між Україною та Туреччиною // Зони вільної торгівлі на початку XXI століття : моногр. – К. : ВПЦ "Київський університет", 2013.

- TN – аналогічний показник імпортного тарифу Туреччини перед утворенням відповідних ЗВТ (для країн без ЗВТ вважаємо, що $TN = 0$);
- $FTAT$ – час перебування у зоні вільної торгівлі (роки).

Перша модель експорту Туреччини до країн-партнерів ($FTA = FT$) виражає вплив зони вільної торгівлі без урахування факторів величини тарифів та часу:

$$Exp = 17{,}500 + 0{,}825GDP - 1{,}072Dist + 0{,}445FTA. \quad (\mathbf{10.15})$$

Тут ураховано параметри якості моделі. Скоригований коефіцієнт детермінації становить 0,50. Значущість F-статистики становить 0,00000. Згідно з t-статистикою всі коефіцієнти є значущими. Суттєві мультиколінеарність і гетероскедастичність не спостерігаються. Розподіл залишків близький до нормального. Серійна кореляція становить 0,14. Впливових спостережень (викидів серед залишків) не виявлено.

Якщо підставити дані по Україні (логарифми ВВП і відстані) і $FTA = 0$, то експорт дорівнюватиме 12,179 (95 %-й довірчий інтервал: 9.18;15.18), якщо $FTA = 1$, то 13,580 (10.59;16.57). Оскільки це логарифми, то відповідні абсолютні значення становлять близько 195000 і 790000 (нас цікавить саме співвідношення цих двох розрахованих величин експорту, навіть якщо вони суттєво відрізняються від фактичних значень). Таким чином, за умов ЗВТ, очікується, що експорт Туреччини до України (він же імпорт України із Туреччини) має бути більший на 305 %, але цей результат не є точним, зважаючи на широкі довірчі інтервали. Це є першою оцінкою потенціалу зростання експорту Туреччини до України за умов утворення ЗВТ.

Аналогічно будуємо дві моделі для імпорту (включаючи статистичні викиди та без них):

$$Imp = 5{,}490 + 1{,}404GDP + 1{,}478FTA, \quad (\mathbf{10.16})$$

$$Imp = 5{,}440 + 1{,}34GDP + 1{,}823FTA. \quad (\mathbf{10.17})$$

Ці дві моделі дещо відрізняються за коефіцієнтами. Аналогічно, за умов ЗВТ, за очікуваннями, імпорт Туреччини з України має бути більшим на 339 % та 519 %, відповідно. Обидві моделі імпорту показують, що потенціал зростання

імпорту Туреччини з України більший за потенціал зростання експорту Туреччини до України.

Розглянемо можливий альтернативний варіант для оцінювання впливу ЗВТ. Пересвідчившись, що вплив ЗВТ значущий і використавши бінарну змінну, можна побудувати по дві окремі пари моделей (для експорту та імпорту) для країн, що входять до ЗВТ, і країн, що не входять до ЗВТ. Цей варіант може дати можливість урахувати взаємодію між належністю до ЗВТ та іншими змінними (коефіцієнти при ВВП і відстані, імовірно, відрізнятимуться).

Проте цифри потенціалу зростання виглядають завеликими, оскільки стартові ставки митного тарифу при утворенні Туреччиною раніше ЗВТ і сучасні ставки митного тарифу країн, які не входять із Туреччиною до ЗВТ, зазвичай більші за поточні ставки у випадку України. А ефект від утворення ЗВТ тим більший, що більші ставки мита, які ліквідують за взаємної вільної торгівлі. Тому наступним кроком є модифікація моделі для експорту таким чином, що $FTA = TF$ (із впливовими спостереженнями та без них):

$$Exp = 17{,}883 + 0{,}858GDP - 0{,}992Dist - 0{,}0867FTA, \quad \textbf{(10.18)}$$

$$Exp = 17{,}934 + 0{,}863GDP - 1{,}108Dist - 0{,}0664FTA. \quad \textbf{(10.19)}$$

За умов ЗВТ експорт Туреччини до України має бути більшим на 27 або 20 %. Бачимо, що потенціал зростання експорту з урахуванням низького рівня тарифних ставок в Україні не такий високий.

Модифікуємо моделі для імпорту так, що $FTA = TN$:

$$Imp = 5{,}812 + 1{,}377GDP + 0{,}220FTA, \quad \textbf{(10.20)}$$

$$Imp = 5{,}949 + 1{,}324GDP + 0{,}241FTA. \quad \textbf{(10.21)}$$

За умов ЗВТ імпорт Туреччини з України має бути більший на 70 або 79 %, але результат не є точним, зважаючи на широкі довірчі інтервали. Потенціал зростання імпорту Туреччини також є менший, ніж за першого методу, але так само більший за потенціал зростання експорту Туреччини.

Трансформуємо тепер моделі так, що $FTA = FTAT$. Для експорту:

$$Exp = 8{,}487 + 0{,}874GDP + 0{,}048FTA. \quad \textbf{(10.22)}$$

Якщо підставити дані по Україні та $FTA = 0$, то експорт дорівнюватиме 12.621; якщо $FTAT = 1$ (рік після утворення ЗВТ), – то дорівнюватиме 12.669, і далі, додаючи по одному року: 12.717, 12.765 ... Тобто за рахунок ЗВТ щорічно експорт Туреччини до України має додатково зростати на 5 %.

Моделі для імпорту Туреччини із країн-партнерів:

$$Imp = 5{,}793 + 1{,}42GDP + 0{,}0657FTA, \quad (\mathbf{10.23})$$

$$Imp = 5{,}923 + 1{,}402GDP + 0{,}0621FTA. \quad (\mathbf{10.24})$$

Аналогічно, щорічно імпорт Туреччини з України має додатково зростати майже на 7 або 6,5 %.

Розраховані моделі експорту та імпорту Туреччини до країн-контрагентів або із країн-контрагентів пояснюють приблизно половину коливань експорту та імпорту Туреччини. Вони відповідають або майже відповідають критеріям якості регресійних моделей. Усі варіанти моделей показали, що, виходячи зі значень незалежних змінних, які характерні для України, утворення ЗВТ з Туреччиною на раніше характерних для Туреччини умовах приведе швидше до покращання сальдо торговельного балансу України.

Статичні моделі у цілому показують потенціал для зростання зовнішньої торгівлі (ураховуючи стартові тарифи: експорт Туреччини до Україну – на 20-27 %, а імпорт Туреччини з України – на 70-79 %). Динамічні моделі показують, що цього потенціалу можна досягнути лише після тривалого періоду (щорічне зростання за рахунок утворення зони вільної торгівлі: експорту Туреччини до України – на 5 %, імпорту Туреччини з України – на 6,5-7 %). Ці цифри є лише середньо очікуваними, фактичні результати можуть суттєво від них відхилятися. В окремих моделях позитивний ефект від угод про ЗВТ може бути наслідком не стільки угоди, скільки географічного розташування країн-партнерів.

Недоліком використаного підходу є те, що він не враховує галузеву структуру взаємної торгівлі, тому для остаточного прийняття рішення варто використовувати цей підхід як додатковий до більш детальних методів.

10.6. Аналіз регресійної моделі з виробничими функціями

Почнемо з прикладу. За даними Держкомстату України за 2000-2011 рр. (табл. 10.4) потрібно знайти оцінки *параметрів виробничої функції Кобба–Дугласа*

$$Y = A \cdot K^{\alpha} L^{\beta} \cdot M^{\gamma},$$

де Y – обсяг ВВП у фактичних цінах (млн грн), K – вартість основних засобів у фактичних цінах (млн грн), L – середня чисельність зайнятого населення (тис. осіб), M – інвестиції до основного капіталу у фактичних цінах (млн грн).

Необхідно перевірити адекватність прийнятої моделі та оцінити середнє значення прогнозу.

Таблиця 10.4

№ спостереження	Рік	Обсяг ВВП, млн дол.	Вартість основних засобів, млн грн.	Кількість зайнятого населення, тис. осіб	Обсяг інвестицій в основний капітал, млн грн
i		*Y*	*K*	*L*	*M*
1	2000	170070	828822	20175,0	23629
2	2001	204190	915477	19971,5	32573
3	2002	225810	964814	20091,2	37178
4	2003	267344	1026163	20163,3	51011
5	2004	345113	1141069	20295,7	75714
6	2005	441452	1276201	20680,0	93096
7	2006	544153	1568890	20730,4	125254
8	2007	720731	2047364	20904,7	188486
9	2008	948056	3149627	20972,3	233081
10	2009	913345	3903714	20191,5	151777
11	2010	1082569	6648861	20266,0	171092
12	2011	1302079	7396952	20324,2	238175

Установлено причинно-наслідковий зв'язок між досліджуваними економічними показниками[143]. Фактично розглядається застосування теорії виробничих функцій для

[143] Див. : Грисенко М.В., Шворак Л.Л, Рижов А.Ю. Економіко-математичне моделювання…

побудови економіко-математичної моделі та її зв'язок з нелінійною множинною регресією.

Специфікація моделі. Ця задача нелінійної множинної регресії перетворенням $y = \ln(x)$ зводиться до лінійної:

$$\ln Y = \ln A + \alpha \ln K + \beta \ln L + \gamma \ln M.$$

Дійсно, введемо нові зміні:

$$Z = \ln Y,\ X_1 = \ln K,\ X_2 = \ln L,\ X_3 = \ln M,$$

нові параметри:

$$\beta_0 = \ln A,\ \beta_1 = \alpha,\ \beta_2 = \beta,\ \beta_3 = \gamma$$

та отримаємо співвідношення для лінійної множинної регресії:

$$Z_1 = \beta_0 + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \beta_3 \cdot X_{i3} + \varepsilon_i.$$

Визначення параметрів вибраного рівняння

Побудову оцінок $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ здійснюють за тим самим алгоритмом. Покажемо, як це можна зробити в MS Excel з використанням функції ЛИНЕЙН(). Для цього розташуємо на новому робочому аркуші в діапазоні A1:E13 вихідні дані (рис. 10.13), а у діапазоні H2:K13 за допомогою функції LN() знайдемо значення *Z*, X_1, X_2, X_3. Виділимо діапазон H16:K20 і застосуємо функції ЛИНЕЙН() з параметрами, що вказані на рис. 10.14. Після натискання комбінації клавіш Ctrl+Shift+Enter отримаємо значення оцінок параметрів регресії, їх середніх квадратичних оцінок і додаткових статистик (рис. 10.15).

H2 | =LN(B2)

	A	B	C	D	E	F	G	H	I	J	K
1	Рік	ВВП, Y	Вартість основних засобів, K	Зайняте населення, L	Обсяг інвестицій у основний капітал, M			Z	X1	X2	X3
2	2000	170070	828822	20 175,00	23629,0			12,044	13,628	9,912	10,070
3	2001	204190	915477	19 971,50	32573,0			12,227	13,727	9,902	10,391
4	2002	225810	964814	20 091,20	37178,0			12,327	13,780	9,908	10,523
5	2003	267344	1026163	20 163,30	51011,0			12,496	13,841	9,912	10,840
6	2004	345113	1141069	20 295,70	75714,0			12,752	13,947	9,918	11,235
7	2005	441452	1276201	20 680,00	93096,0			12,998	14,059	9,937	11,441
8	2006	544153	1568890	20 730,40	125254,0			13,207	14,266	9,939	11,738
9	2007	720731	2047364	20 904,70	188486,0			13,488	14,532	9,948	12,147
10	2008	948056	3149627	20 972,30	233081,0			13,762	14,963	9,951	12,359
11	2009	913345	3903714	20 191,50	151777,0			13,725	15,177	9,913	11,930
12	2010	1082569	6648861	20 266,00	171092,0			13,895	15,710	9,917	12,050
13	2011	1302079	7396952	20 324,20	238175,0			14,079	15,817	9,920	12,381

Рис. 10.13

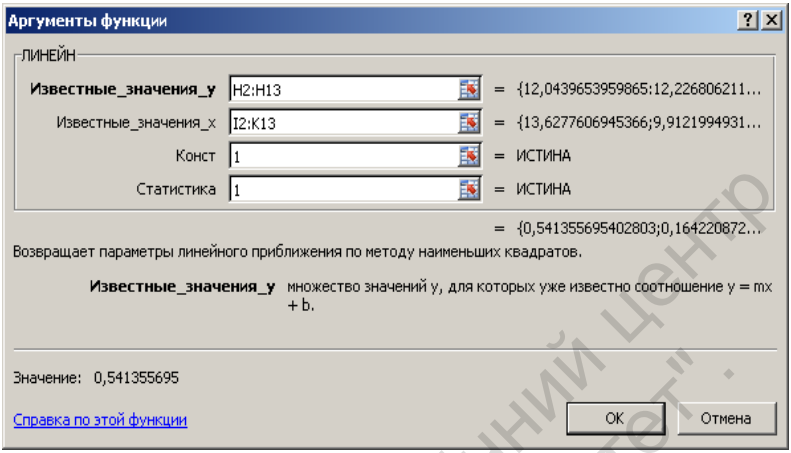
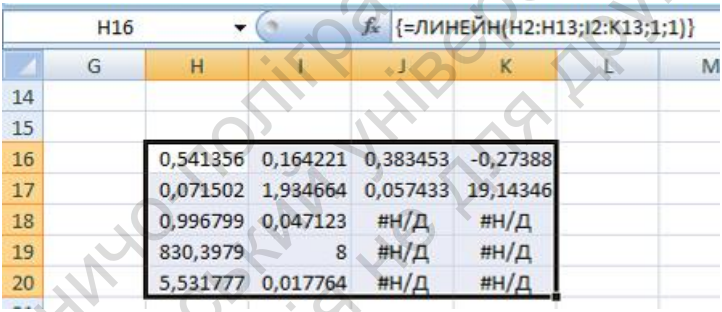


Рис. 10.14



H16 | {=ЛИНЕЙН(H2:H13;I2:K13;1;1)}

	G	H	I	J	K	L	M
14							
15							
16		0,541356	0,164221	0,383453	-0,27388		
17		0,071502	1,934664	0,057433	19,14346		
18		0,996799	0,047123	#Н/Д	#Н/Д		
19		830,3979	8	#Н/Д	#Н/Д		
20		5,531777	0,017764	#Н/Д	#Н/Д		

Рис. 10.15

Значення параметрів регресії розташовані у діапазоні H16:K16 та записані у зворотному порядку:

$$\hat{\beta}_0 = -0{,}27388,\ \ \hat{\beta}_1 = 0{,}38,\ \ \hat{\beta}_2 = 0{,}16,\ \ \hat{\beta}_3 = 0{,}54.$$

Їх середні квадратичні похибки становлять, відповідно:

$$S_{\hat{\beta}_0} = 19{,}14;\ \ S_{\hat{\beta}_1} = 0{,}057;\ \ S_{\hat{\beta}_2} = 1{,}93;\ \ S_{\hat{\beta}_3} = 0{,}07.$$

Для отримання залежності між величинами Y та K, L, M знайдемо оцінку параметра A:

$$\hat{A} = e^{-0{,}274} \approx 0{,}76.$$

Таким чином, $Y = 0{,}76 \cdot K^{0{,}38} \cdot L^{0{,}16} \cdot M^{0{,}54}$.

Здійснимо аналіз якості моделі.

1. *Перевірка загальної якості рівняння регресії*

Для перевірки адекватності моделі порівняємо розраховане значення *F*-статистики $F = 830{,}3979$ (комірка H19) з критичним значенням $F_{\alpha,k_1,k_2} = 4{,}066$ розподілу Фішера з $k_1 = m = 3$ та $k_2 = n - m - 1 = 12 - 3 = 8$ ступенями свободи для рівня значущості $\alpha = 0{,}05$ (рис. 10.16).

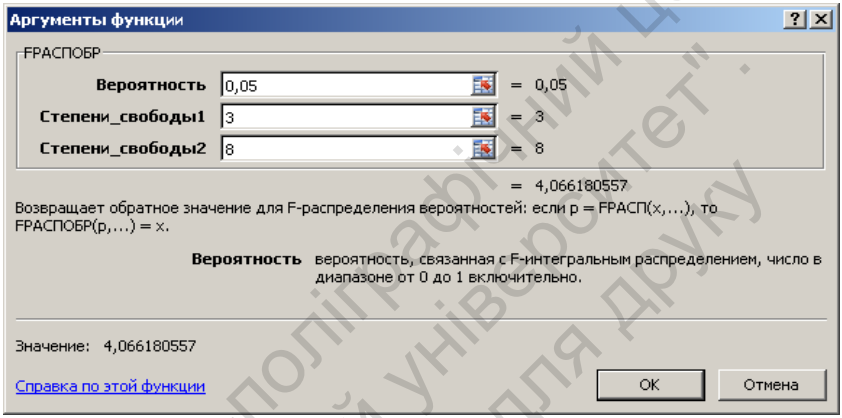


Рис. 10.16

Оскільки $F > F_{\alpha,k_1,k_2}$, то приймаємо гіпотезу про адекватність моделі. Значення коефіцієнта детермінації $R^2 = 0{,}9968$ знайдене у комірці H18. Це свідчить про загальну адекватність моделі.

2. *Перевірка значущості оцінок моделі*

Порівнявши значення оцінок параметрів регресії $\hat{\beta}_j$ з їх середніми квадратичними відхиленнями $S_{\hat{\beta}_j}$, $j_1 = 0,1,2,3$, помітимо, що для вільного коефіцієнта β_0 і параметра $\beta_2 = \beta$ значення середніх квадратичних відхилень значно перевищують значення самих оцінок, що взяті за абсолютною величиною. Зазвичай це свідчить про незначущість відповідних оцінок. Дійсно, знайдемо для кожного *j* відношення:

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{S_{\beta_j}}\beta,$$

порівняємо з критичними значеннями $t''_{кр} = t_{\frac{\alpha}{2},k}$, де $k = n - m - 1$, $t'_{кр} = t''_{кр}$ для заданого рівня значущості α. Якщо $t' < t_{\beta_j} < t''_{кр}$, то на рівні α приймають гіпотезу $H_0 : \beta_j = 0$, інакше – гіпотезу $H_1 : \beta_j \neq 0$. Маємо:

$$t_{\hat{\beta}_0} = \frac{-0{,}27388}{19{,}14} \approx -0{,}0143, \quad t_{\hat{\beta}_1} = \frac{0{,}38}{0{,}057} \approx 6{,}68,$$

$$t_{\hat{\beta}_2} = \frac{-0{,}16}{1{,}93} \approx -0{,}085, \quad t_{\hat{\beta}_3} = \frac{0{,}54}{0{,}07} \approx 7{,}57.$$

Для рівня значущості $\alpha = 0{,}05$ і кількості ступенів вільності $k = n - m - 1 = 12 - 3 - 1 = 8$ критичні значення становлять:

$$t''_{кр}(0{,}025;8) = -t'_{кр}(0{,}025;8) = 2{,}306 \text{ (рис. 10.17)},$$

тому $t_{\hat{\beta}_0}$ і $t_{\hat{\beta}_2}$ належать проміжку $[-2{,}306; 2{,}306]$, у той час як $t_{\hat{\beta}_1}$ і $t_{\hat{\beta}_3}$ – ні. Звідси робимо висновок про значущість на рівні $\alpha = 0{,}05$ оцінок $\hat{\beta}_1$ і $\hat{\beta}_3$ і незначущість оцінок $\hat{\beta}_0$ і $\hat{\beta}_2$. Іншими словами, *залежність ВВП* (Y) *від вартості основних засобів* (K) *та обсягів інвестицій до основного капіталу* (M) *виявилась статистично значущою, у той час як залежність від чисельності зайнятого населення* (L) *– ні.*

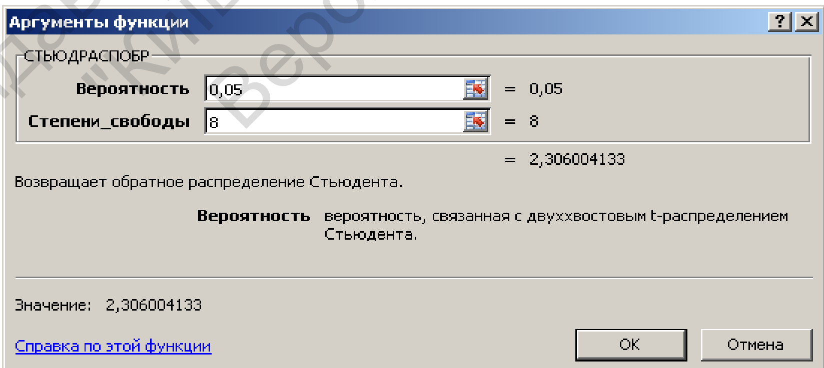


Рис. 10.17

На рис. 10.18 наведено формули MS Excel для отримання довірчих інтервалів для параметрів регресії з довірчою ймовірністю $\gamma = 0{,}95$, звідки (рис. 10.19):

$$-44{,}42 < \beta_0 < 43{,}87;$$
$$0{,}25 < \beta_1 < 0{,}52;$$
$$-4{,}30 < \beta_2 < 4{,}63;$$
$$0{,}38 < \beta_3 < 0{,}71.$$

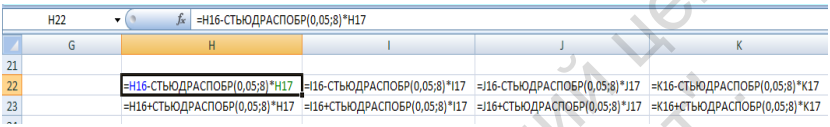


Рис. 10.18

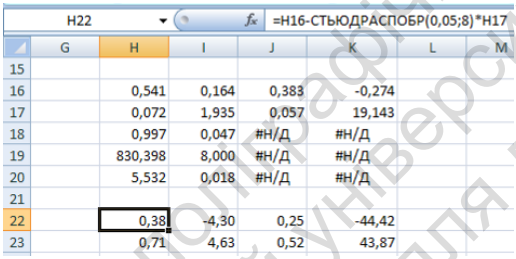


Рис. 10.19

Наслідком статистичної незначущості оцінки параметра $\hat{\beta}_0$ є довірчий інтервал для параметра A, який отримують застосуванням перетворення $y = e^x$ до довірчого інтервалу параметра β_0:

$$e^{-44{,}42} < A < e^{43{,}87} \quad \text{або} \quad 5{,}11 \cdot 10^{-20} < A < 1{,}13 \cdot 10^{19}.$$

Очевидно, що корисність такого довірчого інтервалу є сумнівною.

Прогнозування значень залежної змінної

Оцінку середнього значення прогнозу обчислюють за формулою:

$$\hat{Y}_p = 0{,}76 \cdot K_p^{0{,}38} \cdot L_p^{0{,}16} \cdot M_p^{0{,}54},$$

де $K_p = 9148017{,}0; \cdot L_p = 20354{,}3; \cdot M_p = 26370{,}0.$

Маємо:

$$\hat{Y}_p = 0{,}76 \cdot 9148017^{0{,}38} \cdot 20354{,}3^{0{,}16} \cdot 263730^{0{,}54} \approx 1558647{,}45.$$

Розділ 11
КЛАСТЕРНИЙ АНАЛІЗ ЯК МЕТОД КЛАСИФІКАЦІЇ

11.1. Основи кластерного аналізу

Кластерний аналіз/*cluster analysis* – це метод класифікаційного аналізу. Під час застосування цього методу відбувається розбиття множини досліджуваних об'єктів та ознак на однорідні в деякому розумінні групи, або кластери. Це багатомірний статистичний метод, тому припускають, що вихідні дані можуть мати достатньо великий обсяг, тобто суттєво більшою може бути як кількість досліджень (спостережень), так і ознак, які характеризують ці об'єкти.

Кластерний аналіз призначено для класифікації спостережень, зокрема, країн, фірм, товарів, міст, регіонів, періодів часу або комбінованих просторово-часових спостережень (напр., країно-місяців). При цьому він дає можливість класифікувати спостереження одночасно за кількома ознаками (змінними). Змінні можуть мати будь-який характер. Усі спостереження поділяють на кілька кластерів (груп спостережень) так, що спостереження всередині одного кластера подібні та відрізняються від спостережень з інших кластерів. Кількість кластерів можна обрати апріорі або у процесі кластерного аналізу.

Кластерний аналіз дає можливість зробити розбиття об'єктів не за однією ознакою, а за кількома, що є його перевагою. Крім того, при застосуванні кластерного аналізу, на відміну від більшості математико-статистичних методів, не накладаються жодні обмеження на вид об'єктів, що дозволяє досліджувати множину вихідних даних практично довільної природи.

Кластери – це групи однорідності. Задача кластерного аналізу – на основі ознак об'єктів розбити їх множину на *m* (*m*– ціле) кластерів так, щоб кожен об'єкт належав тільки одній групі розбиття. При цьому об'єкти, що належать одному кластеру, мають бути однорідними (подібними), а об'єкти, які належать різним кластерам, – різнорідними.

Для формалізації задачі класифікації кожний об'єкт зручно інтерпретувати як точку у багатовимірному просторі ознак. Геометрична близькість точок у такому просторі відповідає близькості досліджуваних об'єктів з погляду досліджуваних властивостей. Залежно від мети дослідження задачу класифікації можна сформулювати як розбиття аналізованих об'єктів на певну кількість груп, усередині яких вони розташовані на порівняно малій відстані один від одного, або як виявлення природного розшарування сукупності, яку вивчають, на окремі кластери.

Якщо об'єкти кластеризації подати як точки в n-вимірному просторі ознак (n – кількість ознак, які характеризують об'єкти), то схожість між об'єктами визначають через поняття відстані між точками, оскільки інтуїтивно зрозуміло, що чим менша відстань, тим вони більш схожі.

Для візуалізації результатів будують графіки, що являють проєкції будь-якої пари показників на площину, на яких точки, що належать до одного кластеру, окантовуються. Одним з найбільш важливих і складних питань за кластеризації є вибір оптимальної кількості кластерів. Зазвичай, згідно з вихідною гіпотезою, визначають початкову кількість кластерів, а потім, змінюючи її, емпіричним шляхом вибирають остаточний варіант кластеризації. У результаті кластерного аналізу отримують багаторівневу ієрархічну класифікацію, яка відображає найбільш суттєві особливості взаємини між об'єктами. Таким чином, отримані кластери – це група об'єктів, які мають подібні особливості. Такий аналіз проводять для територіальних об'єктів за низкою показників, для подальшого угруповання районів і виявлення стійких груп.

Наприклад, потрібно класифікувати шість країн за двома ознаками: середня ставка митного тарифу та частка імпорту у ВВП. Класифікацію можна зробити й візуально за допомогою діаграми розсіювання (рис. 11.1). Видно, що найближче розташовані спостереження E та F. А що ближче об'єкти у цьому двомірному просторі, то більш схожими вони є. Можна розглядати їх як один кластер, а решту спостережень – як

окремі кластери. У сумі наявні п'ять кластерів. Якщо потрібні чотири кластери, то спостереження об'єднують як E+F, B+C, D, A. Якщо потрібні три кластери, то спостереження об'єднують як E+F, B+C+D, A. Якщо потрібні два кластери, то спостереження об'єднують як E+F, B+C+D+A. Звичайно, можна всі спостереження розглядати й як один великий кластер.

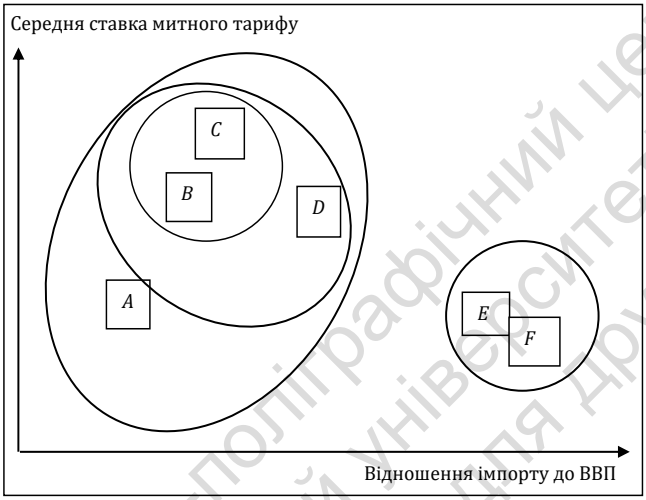


Рис.11.1

Кластерний аналіз дозволяє здійснювати об'єднання за схожістю між об'єктами через поняття відстані між точками, зважаючи на точне місце розташування спостережень один щодо одного у двомірному просторі, де виміри відповідають двом змінним: середня ставка митного тарифу та частка імпорту у ВВП. Але графічний метод практично неможливо використовувати, якщо спостережень дуже багато, зокрема, якщо кількість вимірів (змінних), за якими відбувається класифікація, є великою.

11.2. Визначення відстані між спостереженнями та кластерами

Класичними непараметричними методами класифікації є методи кластерного аналізу (таксономії). За їх допомогою розв'язують проблему такого розбиття (класифікації, клас-

теризації) множини об'єктів, за якого всі об'єкти, що належать до одного класу, були б більш подібними один до одного, ніж до об'єктів інших класів. Із формального погляду, основне завдання методів кластерного аналізу можна сформулювати як визначення класів еквівалентності й рознесення за ними досліджуваних об'єктів. Під класом звичайно розуміють генеральну сукупність, яку описує одномодальна функція щільності ймовірності або, у випадку дискретних ознак, – одномодальний полігон імовірностей. Номери класів не мають змістового навантаження, їх використовують лише для відрізнення один від одного.

Для формування кластерів застосовують міри подібності та відмінності даних, які можна поділити на основні види:

- міри подібності (відмінності) типу "відстань" (об'єкти вважають тим більш подібними один до одного, чим меншою є відстань між ними);
- міри подібності типу "зв'язок" (об'єкти вважають тим більш подібними, чим сильнішим є зв'язок між ними);
- інформаційна статистика.

Як міру відстані (метрику) можна використовувати будь-яку функцію $\rho(X_i,X_j)$, що визначена на множині $\{X_1, X_2,..., X_n\}$ і задовольняє такі вимоги:

- $\rho(X_i, X_j) \geq 0$ для всіх i, j;
- $\rho(X_i, X_j) = 0$ тоді й тільки тоді, коли $X_i = X_j$;
- $\rho(X_i, X_j) = \rho(X_j, X_i)$;
- $\rho(X_i, X_j) \leq \rho(X_i, X_k) + \rho(X_k, X_j)$.

Вибір міри відстані істотно впливає на результат класифікації. Тому для отримання надійних результатів необхідно враховувати мету дослідження, змістову й статистичну природу вектора спостережень та наявні відомості про характер розподілу досліджуваних ознак. Крім того, після закінчення розрахунків слід перевіряти адекватність отриманої класифікаційної моделі.

У кластерному аналізі для кількісної оцінки подібності вводять поняття метрики. Подібність або відмінність між класифікованими об'єктами встановлюється залежно від

метричної відстані між ними. Якщо кожен об'єкт описує k ознак, то він може бути представлений як точка в k-вимірному просторі, і схожість з іншими об'єктами визначатиметься як відповідна відстань. У кластерному аналізі використовують різні формули відстані між об'єктами.

Найчастіше використовують евклідову та мангеттенську відстані, супремум-норму, а також відстань Махаланобіса. Евклідову метрику традиційно застосовують як міру відстані. Мангеттенська відстань є найбільш відомою з класу метрик Мінковського. Відстань Махаланобіса, що не є метрикою, за допомогою дисперсійно-коваріаційної матриці пов'язана з кореляціями змінних. Її широко застосовують у кластерному аналізі та інших методах аналізу даних. Такі міри подібності можна застосувати при реалізації методів ближнього зв'язку, середнього зв'язку Кінга, Уорда, k-середніх Мак-Куїна.

Розглянемо детальніше різні функції відстаней між спостереженнями у багатомірному просторі.

1. Найбільш розповсюдженою є *евклідова відстань/ Euclidean distances – евклідова метрика*, яку розраховують за теоремою Піфагора: відстані між координатами за кожною змінною вводять до квадрату, а потім з їхньої суми визначають корінь квадратний, тобто за формулою:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}. \qquad (\mathbf{11.1})$$

Евклідову відстань доцільно обирати, якщо:

- спостереження належать до генеральних сукупностей, які підпорядковані багатовимірним нормальним законам, а компоненти вектора спостережень є незалежними та мають одну й ту саму дисперсію;
- компоненти вектора спостережень є однорідними з погляду змістової інтерпретації та однаково важливими для класифікації;
- простір ознак має розмірність 1, 2 або 3, і поняття близькості об'єктів у цьому просторі збігається зі звичайною геометричною близькістю.

Евклідова метрика має й недоліки. У випадках, за яких ознаки виміряно у різних одиницях, зміна масштабу одиниць вимірювання може призвести до істотної зміни результатів класифікації. Іншими словами, якщо відстань вимірюють за абсолютними значеннями змінної, а не стандартизованими, то на відстань можуть впливати одиниці виміру. Ця метрика, як і більшість інших, чутлива до одиниць вимірювання осей. Наприклад, якщо сантиметри перевести в міліметри, то зміниться й відстань.

Для запобігання цього використовують різні методи нормування даних, найпоширенішими серед яких є:

$$z_1 = \frac{x - \bar{x}}{\sigma}; \quad z_2 = \frac{x - x_{\min}}{x_{\max} - x_{\min}}; \quad z_3 = \frac{x}{\bar{x}}; \quad z_4 = \frac{x}{x_{\max}}. \quad \textbf{(11.2)}$$

Слід зауважити, що нормування також впливає на результати класифікації. Зокрема, у випадках, коли кластери істотно розділені за деякими ознаками й слабко – за іншими, нормалізація може привести до зменшення дискримінуючих можливостей першої групи ознак через збільшення шумового ефекту інших[144].

2. *Квадрат евклідової відстані* використовують, якщо необхідно надати значної ваги більш віддаленим один від одного об'єктам.

3. Якщо ознаки вимірюють у якісно різних одиницях, то застосування евкідової відстані загалом може виявитися безглуздим. Тому використовують *зважену евклідову відстань*, яку розраховують за формулою:

$$d_{ij}^{*} = \sqrt{\sum_{k=1}^{p} \omega_k (x_{ik} - x_{jk})^2}, \quad \textbf{(11.3)}$$

де ω_k – невід'ємні вагові коефіцієнти, що є пропорційними ступеню важливості критерію з погляду класифікації. Зазвичай приймають $0 \le \omega_k \le 1$. Визначення вагових коефіцієнтів за аналізованою вибіркою зазвичай є недоцільним, оскільки може призвести до істотних помилок. Зокрема,

[144]Див. : Бахрушин В.Є. Методи аналізу даних : навч. посіб. – Запоріжжя : КПУ, 2011.– 268 с.

залежно від певних незначних варіацій змістової і статистичної природи вихідних даних можна обґрунтувати надання їм значень, що пропорційні середньоквадратичній похибці відповідної ознаки або оберненій до цієї похибки величині. Тому варто обирати вагові коефіцієнти за результатами експертних опитувань або інших незалежних попередніх досліджень.

3. *Метрика Мінковського*[145] є узагальненням звичайної евклідової відстані:

$$d_{ij} = \sqrt[r]{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{r}} . \qquad (\mathbf{11.4})$$

У випадку $r = 2$ вона збігається з евклідовою метрикою.

4. У випадку $r = 1$ метрика Мінковського дає *мангеттенську відстань/Manhattan distance* міських кварталів і визначається як сума різниць координат з кожної змінної:

$$d_{ij} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right| \qquad (\mathbf{11.5})$$

5. При $r \to \infty$ метрика Мінковського збігається із супремум-нормою *відстанню Чебишева/Chebyshev distance*, її застосовують, коли бажають визначити два об'єкти, якщо вони відрізняються за якоюсь однією координатою:

$$d_{ij} = \sup \left\{ \left| x_{ik} - x_{jk} \right| \right\}, k = 1, 2, \ldots, p. \qquad (\mathbf{11.6})$$

6. *Геммінгову відстань*[146] використовують як міру відстані об'єктів, які характеризують дихотомічні ознаки. Її розраховують за формулою:

$$d_{ij} = \sum_{s=1}^{p} \left| x_{is} - x_{js} \right|, \qquad (\mathbf{11.7})$$

тобто вона збігається із кількістю значень відповідних ознак, що не збігаються, у i-го та j-го об'єктів (коли ознаки можуть набувати значень 0 або 1).

---

[145] Запропонована видатним німецьким математиком і фізиком Мінковським у 1908 р.

[146] Введено відомим американським математиком Геммінгом у 1950 р.

7. *Відсоток незгоди*/*Percent disagreement* використовують, коли змінні є категоріальними, і коли вихідні дані не мають кількісного вираження.

8. 1 *мінус коефіцієнт кореляції Пірсона*/1-*Pearson r*. За цією опцією до одного кластеру належатимуть не ті спостереження, за якими значення змінних близькі, а ті, які мають щільніший лінійний зв'язок. Наприклад, є три країни, які характеризують такі значення змінних (щодо ВВП): експорт, прямі інвестиції, зовнішній борг і валютні резерви:

A(40;5;60;20), B(45;2;66;15), C(80;9;100;41).

Видно, що країни А та В мають невелику евклідову відстань (значення змінних у країні В мало відрізняються від значень змінних у країні А), але країни А та С мають щільніший лінійний зв'язок (значення змінних у країні С практично вдвічі більші за значення змінних у країні А).

Для порядкових ознак призначені коефіцієнти рангової кореляції Спірмена й Кендалла. Їх можна перетворити до мір подібності типу "відстань" за допомогою формул:

$$d_{ij} = 1 - \rho_s;\ d_{ij} = 1 - \tau. \qquad (\mathbf{11.8})$$

У цьому випадку їх називають *відстанями Спірмена й Кендалла*, відповідно.

Як розрахувати відстань між окремими спостереженнями, указано раніше. Але додатково потрібно знати, як розраховувати відстань між кластерами при прийнятті рішення: які кластери слід об'єднувати на кожному кроці. При конструюванні різноманітних процедур класифікації доцільно використовувати міри близькості кластерів один до одного. Найбільш поширеними є відстані, що вимірюють за принципами найближчого й далекого сусідів, середнього зв'язку та за центрами ваги. Вибір міри близькості кластерів є найбільш суттєвим для агломеративних ієрархічних методів кластерного аналізу.

Розглянемо кілька методів:

- *Правило окремого зв'язку*/*single linkage.* Відстань між кластерами визначають як відстань між їх найближчими елементами. Цей метод ще називають методом "найближчого сусіда", оскільки відстань між двома порівнюваними кла-

стерами визначають як відстань між найближчими об'єктами у різноманітних кластерах. Цей метод має недолік: утворюються занадто продовгуваті кластери, а не колоподібні.

- *Правило повних зав'язків/complete linkage.* Два об'єкти, які належать до однієї групи (кластера), мають коефіцієнт схожості, більший від деякого порогового значення *S*. У термінах евклідової відстані це означає, що відстань між двома точками (об'єктами) кластера не має перевищувати деякого порогового значення *d*. Таким чином, *d* визначає максимально допустимий діаметр підмножини, що утворює кластер. Цей метод ще називають методом "найвіддаленіших сусідів", оскільки за достатньо великого порогового значення *d* відстань між кластерами визначають найбільшою відстанню між довільними об'єктами у різних кластерах.

- *Правило незваженого попарного середнього/unweighted pair-group average.* Відстань між двома порівнюваними кластерами визначають як середню відстань між усіма парами об'єктів з обох кластерів. Метод ефективний, коли об'єкт формують різні групи, але він працює однаково добре й у випадках кластерів ланцюгового типу.

- *Правило зваженого попарного середнього/weighted pair-group average.* Метод ідентичний попередньому, однак при обчисленні відстань між двома порівнюваними кластерами визначають так само, як й у попередньому методі, але з урахуванням зважування на основі величини кластерів.

- *Правило центру маси/centroid* передбачає, що спочатку розраховують центри маси кластерів, а відстань між кластерами визначають як відстань між цими центрами мас.

11.3. Види та практичне призначення кластерного аналізу

Алгоритмів кластерного аналізу достатньо багато. Їх можна поділити на ієрархічні (деревовидні) та неієрархічні.

Ієрархічні процедури – найбільш поширені алгоритми кластерного аналізу за їх реалізації на комп'ютері. Розрізняють агломеративні (від *agglomerate* – збирати) та ітеративні дивизивні (від *division* – розділяти) процедури.

Принцип роботи ієрархічних агломеративних процедур – це послідовне об'єднання груп елементів спочатку самих близьких, потім – усе більш віддалених один від одного. Принципом роботи ієрархічних дивизивних процедур, навпаки, є послідовне розділення груп елементів спочатку найбільш віддалених, потім – більш близьких один від одного. Більшість цих алгоритмів виходить із матриці відстаней. До недоліків цих алгоритмів слід зарахувати громіздкість їх реалізації. На кожному кроці алгоритми потребують обчислення матриці відстаней. Тому реалізація таких алгоритмів за великої кількості спостережень недоцільна.

Загальний принцип роботи агломеративного алгоритму. На першому кроці кожне спостереження розглядають як окремий кластер. У подальшому на кожному кроці роботи алгоритму відбувається об'єднання двох найближчих кластерів і з урахуванням прийнятої відстані за формулою перерахування матриці відстаней, розмірність якої, очевидно зменшується на одиницю. Робота алгоритму закінчується, коли всі спостереження об'єднані до одного класу. Більшість програм, які реалізують алгоритм ієрархічної класифікації, передбачають графічне подання у вигляді дендограми.

Нижче, на рис. 11.2, подано дендрограму, що вказує на ступінь схожості (яка є оберненою величиною відстані) між окремими спостереженнями чи кластерами.

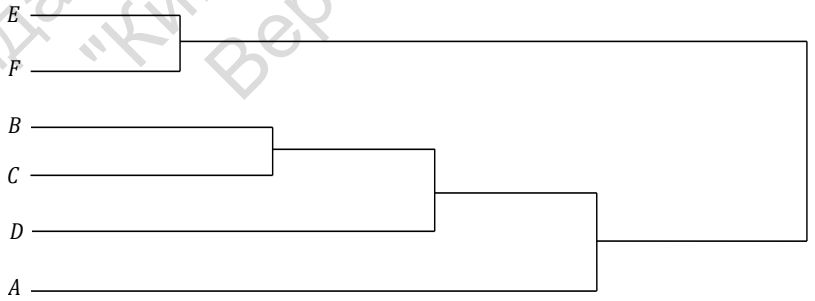


Рис. 11.2

Неірерархічний кластерний аналіз. Припустимо, що є гіпотези відносно числа *m* (*m* – ціле) кластерів. Тоді можна створити рівно *m* кластерів так, щоб вони були настільки різні, наскільки це можливо. Саме для розв'язування задач цього типу призначено *метод k-середніх*/*k-means*. Гіпотеза може базуватися на теоретичних міркуваннях або здогадах. Виконуючи послідовно розбиття на різне число кластерів, можна порівнювати якість отриманих розв'язків.

Неієрархічний метод *k*-середніх передбачає, що спочатку апріорі визначають кількість кластерів, на які треба поділити спостереження. Наприклад, на чотири кластери (як варіант, рішення про кількість кластерів може бути прийняте після ієрархічного аналізу за допомогою побудови дендрограми, коли вже є певне уявлення про відстані між спостереженнями). Програма спочатку випадково розподіляє спостереження за цими кластерами, а далі ітеративним шляхом перерозподіляє спостереження між кластерами. Нарешті всі спостереження розподіляють за чотирма кластерами так, щоб у кожному кластері спостереження були максимально схожими (близькими) та максимально відмінними (далекими) від спостережень щодо решти кластерів.

Після цього можна визначити середні кожної змінної в одному кластері, другому, третьому тощо. За деякими змінними середні можуть суттєво відрізнятися за кластерами, за іншими – майже не відрізнятися. Якщо середні змінних відрізняються достатньою мірою, це означає, що програма успішно кластеризувала спостереження на неоднорідні кластери (групи). Значущість відмінностей у кластерах визначають за допомогою *F*-критерію. Якщо середні змінних за кластерами мало відрізняються (а *F*-критерій при цьому швидше за все сигналізуватиме про незначущість кластеризації), то це означає, що принципових відмінностей між спостереженнями немає. Зазвичай цю проблему розв'язують шляхом приведення значень до схожого масштабу величин. Наприклад, якщо одну змінну вимірюють у середньому шестизначними цифрами, а іншу – однозначними, то можна або поділити шестизначні цифри на коефіцієнт (напр., 1000000), або застосувати стандартизовані значення замість абсолютних.

Кластерний аналіз може бути попереднім етапом для проведення подальшого аналізу. Наприклад, якщо потрібно побудувати не одну регресійну модель для всіх спостережень, а кілька – по одній для кожного кластера. Припустимо, потрібно розрахувати регресійну модель залежності внутрішніх цін на бензин від світових цін на нафту, ступеня монополізації ринку нафтопродуктів і частки імпорту у споживанні бензину. Перед цим можна кластеризувати всі спостереження на основі кількох інших змінних, наприклад, ВВП на душу населення та приріст ВВП (кластеризацію варто проводити на основі змінних, що, як передбачається, змінюють характер або величину впливу факторів на залежну змінну, тобто після перевірки наявності ефекту взаємодії). Припустимо, усі спостереження поділено на чотири кластери:

- **кластер 1** – країни з високим або середнім ВВП на душу населення під час економічного зростання (роки);
- **кластер 2** – країни з високим або середнім ВВП на душу населення під час рецесії або стагнації (роки);
- **кластер 3** – країни з низьким або середнім ВВП на душу населення під час швидкого економічного зростання (роки);
- **кластер 4** – країни з низьким ВВП на душу населення під час рецесії або повільного економічного зростання (роки).

Для кожного кластера будують окрему регресійну модель, або належність до відповідного кластеру вводять до регресійної моделі як бінарну псевдозмінну.

Звичайно, можна безпосередньо ввести до моделі як фактори ВВП на душу населення та приріст ВВП. Але тоді необхідно вводити багато добутків факторів для урахування ефектів взаємодії, що вимагатиме більшої кількості спостережень для надійності результатів в умовах, за яких кількість доступних спостережень часто обмежена. Особливо це може стати проблемою, якщо замість двох змінних (ВВП на душу населення та приріст ВВП) таких змінних багато.

11.4. Використання кластерного аналізу у Tanagra

Припустимо, необхідно класифікувати спостереження-країни за трьома ознаками: відношення валютних резервів до зовнішнього боргу, середньозважена ставка застосованого митного тарифу та відношення сальдо поточного рахунку до ВВП (за даними *World Development Indicators* за 2009 р.) – див. дані у Microsoft Excel на рис. 11.3.

	A	B	C	D
1	**2009**	Total reserves (% of total external debt)	Tariff rate (%)	Current account balance (% of GDP)
2	Romania	37.77	1.51	-4.32
3	Russian Federation	115.21	5.90	3.98
4	El Salvador	27.43	2.47	-1.44
5	Serbia	45.59		-6.88
6	Syrian Arab Republic	349.51	9.75	-2.15
7	Togo	42.88	13.87	-5.60
8	Thailand	235.59	4.92	8.29
9	Tonga	91.58	7.34	-16.63
10	Tunisia	52.03		-2.83
11	Turkey	29.81	2.30	-2.28
12	Tanzania	47.38	10.76	-9.05
13	Uganda	120.28	8.40	-6.73
14	Ukraine	28.45	2.37	-1.48
15	Uruguay	66.11	3.51	0.66
16	Venezuela, RB	62.97	9.44	2.63
17	Yemen, Rep.	109.98	4.24	-9.73
18	South Africa	94.06	3.91	-4.01
19	Zambia	62.05	3.83	4.20

Рис. 11.3

У Tanagra після обрання в меню File опції New у діалоговому вікні (рис. 11.4) обирають у полі Dataset відкритий відповідний файл формату xls. Важливо не забути перед цим закрити цей файл в Excel. Далі натискають ліворуч на Dataset (назва файлу), потім – кольорову іконку зі стрілками (рис. 11.5) і Define Status 1. У новому діалоговому вікні обирають змінні стрілкою (як на рис. 11.6).

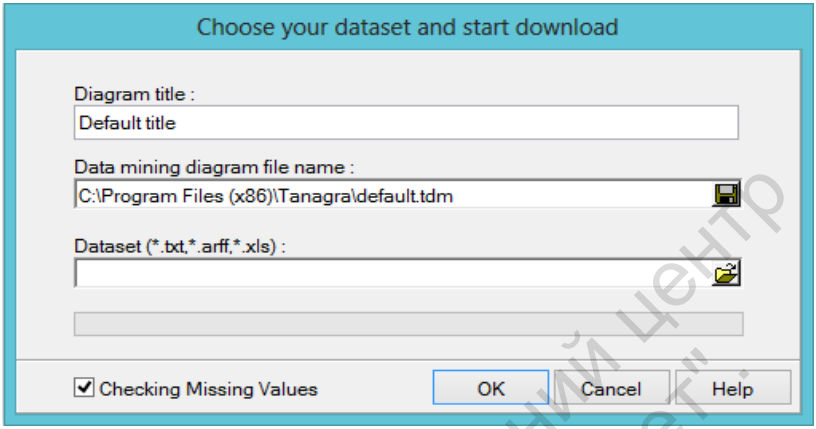


Рис. 11.4

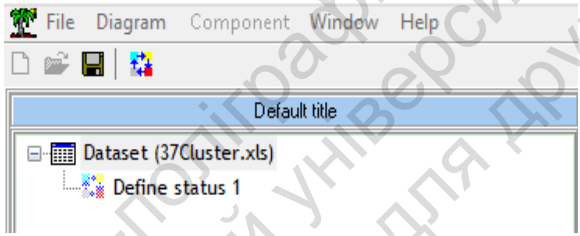


Рис. 11.5

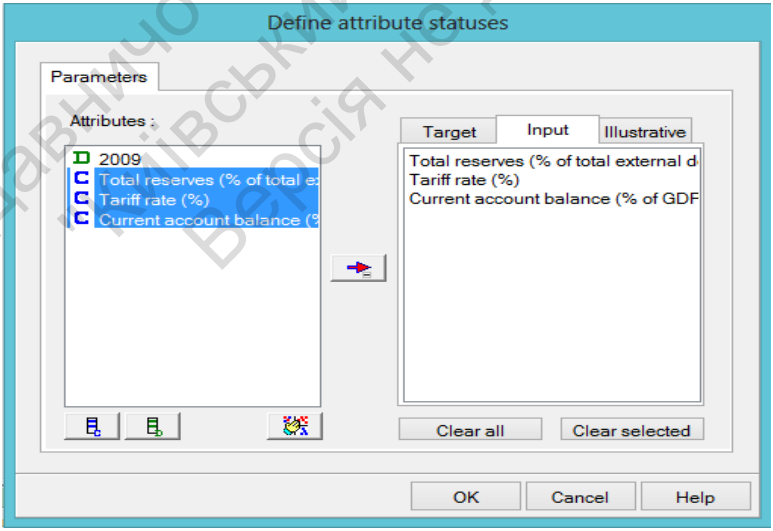


Рис. 11.6

У нижньому меню (рис. 11.7) обирають Clustering – HAC і переміщують відповідну іконку на Define Status 1 для побудови дендрограми (рис. 11.8). Із дендрограми видно, що варто розбити всі спостереження на три кластери.

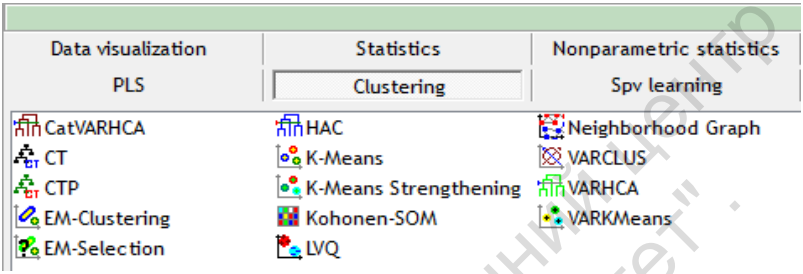


Рис. 11.7

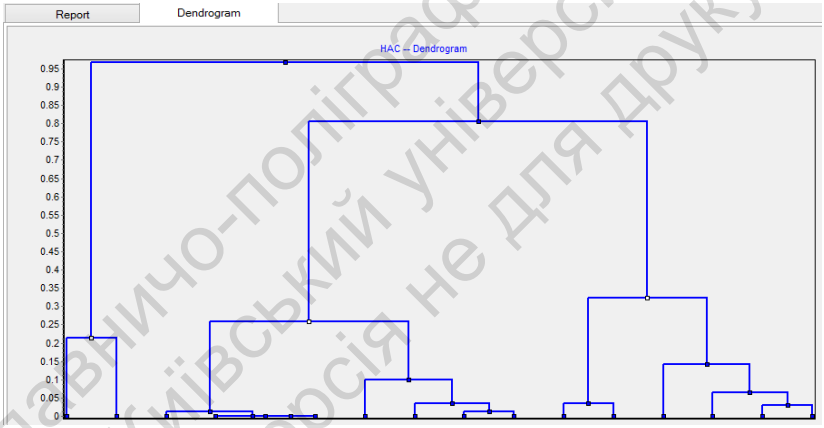


Рис. 11.8

Далі в нижньому меню (рис. 11.7) обирають Clustering – K-Means і переміщують відповідну іконку на Define Status 1; натискають правою кнопкою K-Means 1 та обирають Parameters. У діалоговому вікні, що з'явиться (рис. 11.9), обирають три кластери в Number of clusters і None в Distance normalization, якщо потрібно використати звичайні абсолютні значення, а не стандартизовані. У вкладці Results необхідно помітити Show ANOVA Table. Після натискання OK двічі натискають K-Means 1.

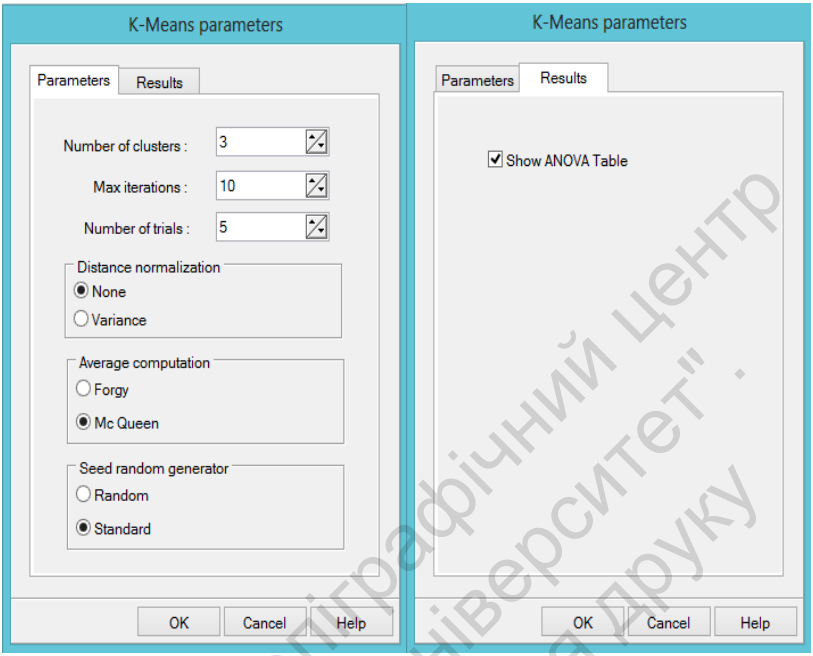


Рис. 11.9

Праворуч з'являться результати кластерного аналізу. Можна прокрутити їх вниз, щоб побачити результати F-тесту (рис. 11.10) та середні змінних за кластерами (рис. 11.11). Праворуч під надписами Proba побачимо рівень значущості F-тесту для кожної змінної окремо. Значення 0,000000 означає дуже значущу відмінність за середніми змінної *Валютні резерви* (щодо *зовнішнього боргу*), отже ця зміна вплинула на розподіл спостережень за кластерами. Значення 0,841465 і 0,136785 означає, що змінні *Ставка імпортного тарифу* та *поточний рахунок платіжного балансу* (*щодо ВВП*) дуже незначно вплинули на розподіл за кластерами. Це можна виправити, якщо замість звичайних значень застосувати стандартизовані значення.

На рис. 11.11. видно, що найбільшими валютні резерви – у третьому кластері (у середньому 292,54 % зовнішнього боргу), середні – у другому кластері (106,22), найменші – у пер-

шому кластері (44,98). Найменші імпортні тарифи – у першому та другому кластерах (5,56 і 5,96%), найбільші – у третьому кластері (7,33). У першому кластері помірний дефіцит поточного рахунку (–1,85 % ВВП), у другому кластері – сильний дефіцит (–6,62), у третьому, – навпаки, профіцит (3,07).

Після цього залишається визначити належність кожного спостереження до кластерів. Для цього обирають K-Means 1 і натискають іконку . У новому діалоговому вікні обирають змінні стрілкою (як на рис. 11.12). Далі в нижньому меню обирають Data visualization – View dataset (рис. 11.13) і перетягують на Define Status 2. Натискають View Dataset 1.

У новому вікні (рис. 11.14) видно, що перший кластер включає Румунію, Сальвадор, Того, Туреччину, Танзанію, Україну, Уругвай, Венесуелу, Замбію; другий кластер – Тонгу, Уганду, Ємен, Південну Африку; третій кластер – Сирію й Тайланд.

В іншому програмному забезпеченні, наприклад Statsoft Statistica, є додаткові можливості, зокрема, вибір методу визначення відстаней, *Матриця відстаней*/*Distance matrix* між спостереженнями, матриця відстаней між кластерами, відстань між спостереженнями та центром кластера (для визначення найбільш і найменш типових спостережень для кожного кластера), графік *профілі спостережень*/*case profiles* для візуалізації різниці між середніми, описова статистика для кожного кластера тощо.

Якщо необхідно ввести нове спостереження до вибірки, то не варто наново переробляти кластерний аналіз. Введення навіть одного чи кількох спостережень може кардинально змінити розподіл спостережень за кластерами. Замість цього необхідно розрахувати відстань між новим спостереженням і центрами наявних кластерів. Нове спостереження можна приєднати до найближчого кластера. Наприклад, копіюють таблицю із середніми значеннями за кластерами, показують нове спостереження та розраховують відстані його від центрів кожного кластера (рис. 11.15). Видно, що найменша відстань у нового спостереження – до першого кластера.

Attribute_Y	Attribute_X	Description				Statistical test		
		Value	Examples	Average	Std-dev	Variance decomposition		
						Source	Sum of square	d.f.
		c_kmeans_1	9	44.9820	15.5523	BSS	101195.5839	2
		c_kmeans_2	5	106.2207	12.7927	WSS	9078.2312	13
		c_kmeans_3	2	292.5473	80.5520	TSS	110273.8152	15
		All	16	95.0647	85.7414	Significance level		
						Statistics	Value	Proba
Total reserves (% of total external debt)	Cluster_KMeans_1					Fisher's F	72.455887	0.000000
		Value	Examples	Average	Std-dev	Variance decomposition		
						Source	Sum of square	d.f.
		c_kmeans_1	9	5.5622	4.5429	BSS	5.1612	2
		c_kmeans_2	5	5.9580	1.9379	WSS	191.7871	13
		c_kmeans_3	2	7.3350	3.4153	TSS	196.9483	15
		All	16	5.9075	3.6235	Significance level		
						Statistics	Value	Proba
Tariff rate (%)	Cluster_KMeans_1					Fisher's F	0.174923	0.841465
		Value	Examples	Average	Std-dev	Variance decomposition		
						Source	Sum of square	d.f.
		c_kmeans_1	9	-1.8529	4.1130	BSS	149.9542	2
		c_kmeans_2	5	-6.6233	7.5654	WSS	418.8036	13
		c_kmeans_3	2	3.0684	7.3843	TSS	568.7578	15
		All	16	-2.7285	6.1577	Significance level		
						Statistics	Value	Proba
Current account balance (% of GDP)	Cluster_KMeans_1					Fisher's F	2.327349	0.136785

Рис. 11.10

Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3
Total reserves (% of total external debt)	44.981969	106.220670	292.547340
Tariff rate (%)	5.562222	5.958000	7.335000
Current account balance (% of GDP)	-1.852877	-6.623261	3.068393

Рис. 11.11

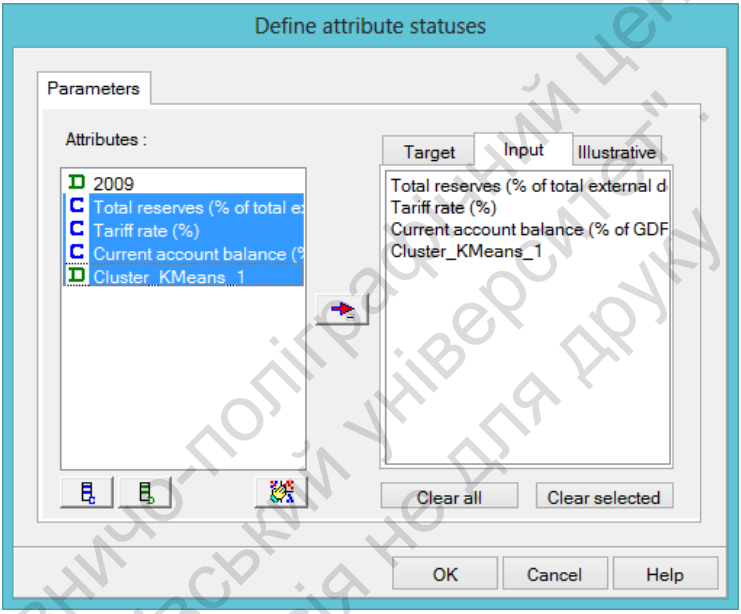


Рис. 11.12

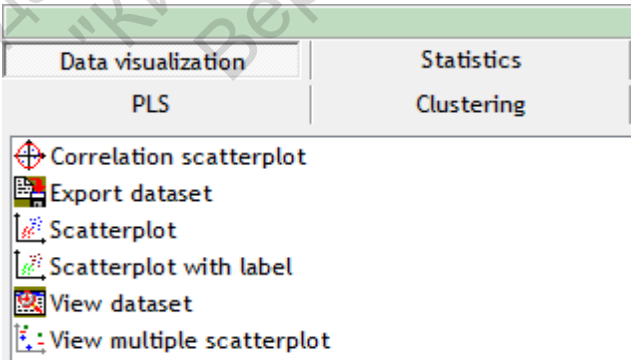


Рис. 11.13

	2009	Total reserves	Tariff rate	Current account	Cluster_KMeans_1
1	Romania	37.7693	1.51	-4.31692	c_kmeans_1
2	Russian Federation	115.21	5.9	3.9775	c_kmeans_2
3	El Salvador	27.4278	2.47	-1.44148	c_kmeans_1
4	Syrian Arab Republic	349.506	9.75	-2.15308	c_kmeans_3
5	Togo	42.8779	13.87	-5.5968	c_kmeans_1
6	Thailand	235.588	4.92	8.28986	c_kmeans_3
7	Tonga	91.5779	7.34	-16.6264	c_kmeans_2
8	Turkey	29.8096	2.3	-2.27661	c_kmeans_1
9	Tanzania	47.3792	10.76	-9.049	c_kmeans_1
10	Uganda	120.275	8.4	-6.73317	c_kmeans_2
11	Ukraine	28.4487	2.37	-1.47747	c_kmeans_1
12	Uruguay	66.1082	3.51	0.66002	c_kmeans_1
13	Venezuela, RB	62.9654	9.44	2.625	c_kmeans_1
14	Yemen, Rep.	109.975	4.24	-9.72827	c_kmeans_2
15	South Africa	94.0648	3.91	-4.00596	c_kmeans_2
16	Zambia	62.0515	3.83	4.19735	c_kmeans_1

Рис. 11.14

	A	B	C	D	E	F	G	H
1		Кластер	Кластер	Кластер	Нове		Відстань між центром кластера	
2		No. 1	No. 2	No. 3	спостереження		і новим спостереженням	
3	Res_ExtDebt	44.98	106.22	292.55	30.0		1	15.38
4	Tarrif	5.56	5.96	7.34	7.0		2	76.24
5	CurAcc_GDP	-1.85	-6.62	3.07	-5.0		3	262.67

G	H
Відстань між центром кластера	
і новим спостереженням	
1	=SQRT((E3-B3)*(E3-B3)+(E4-B4)*(E4-B4)+(E5-B5)*(E5-B5))
2	=SQRT((E3-C3)*(E3-C3)+(E4-C4)*(E4-C4)+(E5-C5)*(E5-C5))
3	=SQRT((E3-D3)*(E3-D3)+(E4-D4)*(E4-D4)+(E5-D5)*(E5-D5))

Рис. 11.15

11.5. Кластеризація країн Європейського Союзу за детермінантами соціалізації їх економічного розвитку

Національні моделі економічної політики країн Європейського союзу об'єднані наразі цілим набором єдиних стратегічних напрямків розвитку, одним з яких є соціалізація економічних процесів. Наприклад, особливу увагу приділяють формуванню та розвитку людського капіталу[147]: розширенню системи соціальної допомоги та забезпечення, активізації діяльності соціальних підприємств третинного сектору економіки, що є загальними рисами розвитку країн ЄС. Класифікація країн ЄС за ключовими детермінантами соціально-економічного розвитку є важливим завданням не лише в аспекті розуміння структури соціально-економічної системи ЄС і визначення особливостей окремих національних моделей, що формують субрегіональні об'єднання та групи в рамках ЄС. Кластерний аналіз країн дозволяє визначити актуальну соціально економічну структуру ЄС, урахування якої для реформування єдиної стратегії та політики розвитку є запорукою сталого характеру розвитку регіону, а також беззаперечною обов'язковою умовою стабільності та непорушності існування самого об'єднання в майбутньому.

Для побудови математичної моделі[148] застосовують багатовимірний математико-статистичний метод класифікаційного аналізу – кластерний аналіз. Результатом кластерного аналізу є об'єднання країн із подібними значеннями показників до одного кластера. На основі кореляційної матриці виділяють головні детермінанти. Основні показники дають

---

147 Див. : Global Wealth Databook 2018: Research Institute. – Credit Suisse Global Wealth Databook, 2018. – 16 p.
https://www.credit-suisse.com/corporate/en/research/research-institute/global-wealthreport.html

148 Див. : Грисенко М.В., Приятельчук О.А. Кластеризація країн Європейського Союзу за детермінантами соціалізації їх економічного розвитку та місце України в даній моделі. Science progress in European countries: new concepts and modern solutions. Hosted by the ORT Publishing the Centre for Scientifically Research "Solution". – 2019. – V. 8. – P. 97-107.

усебічну кількісно-якісну характеристику ступеня соціалізації економіки країн, характеризуючи нагальний рівень соціально-економічного розвитку із врахуванням як загально прийнятих економічних показників, критеріїв суспільного розвитку, так і зважених агрегованих індексів і коефіцієнтів, що враховують як соціальний, так й економічний виміри сталого розвитку.

Основні показники, які розглядають для аналізу, варто формувати за групами:

1. Макроекономічні показники: валовий внутрішній продукт і ВВП на душу населення, що є головним індикатором економічного розвитку та найпоширенішим показником обсягу виробництва товарів і послуг за звітний період часу. Крім того, досліджують окремі кількісні показники: дохід від оподаткування, що характеризує рівень податкового тиску та частку сплачених податків у структурі формування валового продукту; розмір заборгованості, що охоплює загальну заборгованість країни за зовнішніми позиками та неоплаченими за ними відсотками; зовнішньоторговельний баланс; сальдо торгового балансу як кількісний вираз якісної реалізації експортного потенціалу країни; відносні показники прямих іноземних інвестицій у співвідношенні з обсягами ВВП; загальний рівень багатства країни, що включає всі фінансові та нефінансові (зокрема, нерухомість) активи на нетто-основі (тобто за мінусом боргів).

2. Соціально-економічні показники: індекс соціального розвитку, що охоплює такі показники:

- навколишнє середовище та соціальний розвиток, які об'єднані у три напрями соціального прогресу (базові людські потреби, добробут і можливості);
- коефіцієнт Джині, що вимірює ступінь відхилення розподілу доходів між окремими особами або домогосподарствами в межах економіки від абсолютно рівного розподілу;
- індекс людського розвитку як агрегований показник якісного стану людського капіталу й тенденцій його розвитку (зокрема, доступність соціальних послуг, якість їх надання, рівень освіченості населення, якість життя, використання людського ресурсу тощо);

- тривалість життя, що, з одного боку, залежить від стану економічного розвитку, рівня доходів, доступності соціальних послуг, а з іншого – здійснює безпосередній вплив на економіку країни, подовжуючи термін використання людського ресурсу та формуючи потенціал країни;
- рівень мінімальної заробітної плати як індикатор мінімального рівня основних доходів домогосподарств і значення мінімального граничного рівня соціальних гарантій;
- витрати на пенсійне забезпечення, що характеризують середньозважений рівень добробуту громадян пенсійного віку, які перебувають на соціальному забезпеченні, як якісна характеристика виконання взятих державою зобов'язань з гарантованого соціального захисту;
- відрахування з бюджету на охорону здоров'я та освіту, що характеризують рівень соціальної орієнтації економічної політики держав;
- ринок праці як цілковито соціально-орієнтована ринкова категорія, яку характеризують такі загальноприйняті показники, як рівень зайнятості та безробіття серед населення працездатного віку (за методологією Міжнародної організації праці до таких зараховують населення віком від 15 до 70 років, що володіє фізичними та психологічними здібностями до праці).

У результаті математичного моделювання будують дендрограми. Країни ЄС розглядають як однорідну групу з погляду характеру взаємозв'язків між економічними показниками та рівнем соціального розвитку. Кластеризація та вибрані порогові значення дозволяють відсікати за дендрограмою (рис. 11.16 – дендрограма для країн ЄС – метод Варда) відповідні рівні кластерів і надати їм змістовну інтерпретацію.

Аналіз дендрограми дозволяє виділити кластери країн:

1. Австрія, Швеція, Данія, Бельгія, Португалія, Фінляндія.
2. Великобританія, Франція, Греція, Іспанія, Італія.
3. Німеччина.
4. Ірландія, Нідерланди, Мальта, Кіпр, Люксембург.
5. Болгарія, Угорщина, Латвія, Хорватія, Польща, Румунія, Словаччина, Словенія, Чехія, Естонія, Литва.

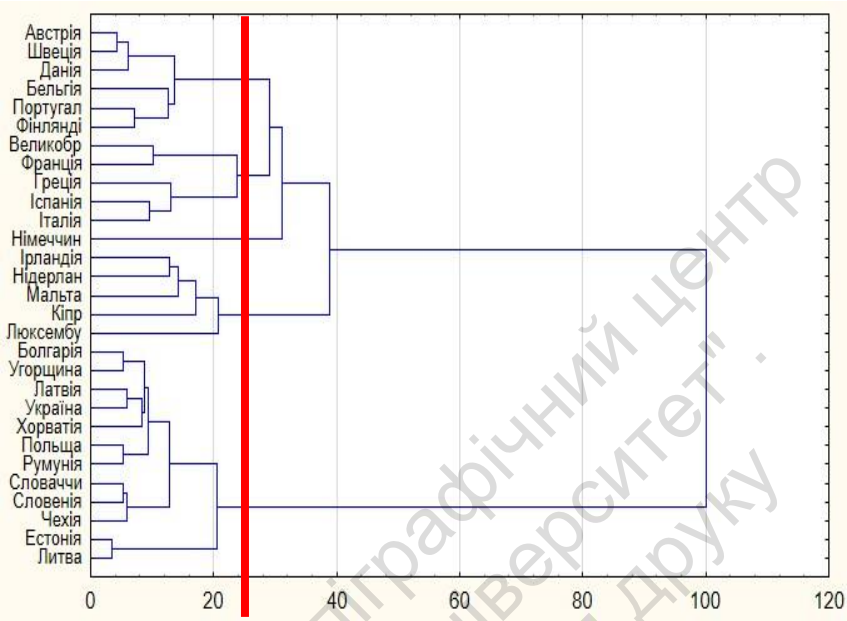


Рис. 11.16

Країни, що об'єднані до окремих груп у результаті кластерного аналізу, крім спільності за сукупністю зазначених показників, мають цілу низку спільних рис, зокрема в напрямах і методах реалізації соціально-економічної політики, особливостях формування національних ринків праці та забезпечення їх інклюзивності, механізмах координації міграційних потоків, методах інтеграції та адаптації мігрантів і біженців, способах формування соціальних гарантів і забезпечення соціального захисту населення, тощо.

Перша група представлена країнами континентальної та скандинавської моделей соціально економічного розвитку. Країни цієї групи характеризують високі обсяги перерозподілу національного багатства через бюджет, реалізація ідей соціальної солідарності та соціального партнерства, активний попереджуючий характер соціальної політики на основі формування високоефективної економіки.

До другої групи ввійшли країни англосаксонської моделі соціально економічної політики та середземноморської моделі, ринок праці яких перебуває на своєрідній стадії трансформації під впливом усталеної європейської континентальної моделі, наявних ідентичностей національних систем та обмеженого їх впливу на загальні світові та, зокрема, європейські тенденції соціально-економічного розвитку. З позиції економічного розвитку країни цієї групи є відмінними, однак спільною рисою є соціальна орієнтація економік країн (високі рівні заробітної плати та відрахувань на соціальну сферу за відсутності чітко задекларованого курсу соціалізації економіки). І якщо у Греції, Італії та Іспанії переважає практика адресної допомоги найбільш уразливим прошаркам населення (однак, ураховуючи сальдо зовнішньоторговельного балансу, незбалансованість ринку праці, низький рівень мінімальної заробітної плати тощо, гарантованих державою соціальних гарантій потребує досить велика частка населення), то Великобританія та Франція – економічно розвинені та стабільні країни нівелюють ці позитивні економічні показники відносно низьким рівнем соціалізації (зокрема, рівень відрахувань з бюджету на сферу соціальних послуг, загальні показники індексів соціального розвитку є відносно низькими).

Окремо розглянемо німецьку модель, яка є основою континентальної та загальної моделі ЄС, для якої характерні високі (більше 50 %) обсяги перерозподілу валового внутрішнього продукту через бюджет, формування страхових фондів переважно за рахунок роботодавців, розвинена система соціального партнерства, політика підтримки повного (або високого) рівня зайнятості[149]. Саме принцип соціального партнерства покладений до основи соціальної економічної моделі. Солідарна участь усіх контрагентів (держави, бізнесу, працівників) у формуванні соціально-орієнтованої ефективної економічної системи стали основоположним принципом її функціонування й запорукою успішного використання різних національних вихідних умов і ресурсної бази.

[149] Див. : Грисенко М.В., Приятельчук О.А. Кластеризація країн Європейського Союзу… – Р. 97-107.

Четверта група країн представлена невеликими за обсягами, політично неактивними, економічно стабільними країнами, що збалансовано поєднують високі макроекономічні показники розвитку зі сталими доходами, відносно високим рівнем добробуту населення країни, що потребує соціального захисту винятково в окремих випадках (тимчасове безробіття, тимчасова втрата працездатності, пенсійне забезпечення тощо). Відсутність різких економічних коливань, політичної нестабільності та соціальних потрясінь визначає загальний соціально економічний курс розвитку цих країн як стабільний, а політику – як виважену та пасивно захисну.

До п'ятої групи належать країни останніх хвиль приєднання до ЄС так званих посткомуністичних країн. Це є, напевно, не причиною, а певним поясненням спільності притаманних їм рис. Рівень соціалізації економіки, так само як і стратегічні напрями та методи реалізації соціально-економічної політики визначають, скоріше, не національні особливості, специфіка внутрішнього ринку праці та матеріально-фінансова база, а наявна загальна концепція соціалізації економіки, що наразі є загально визначеною та діє в рамках ЄС.

Однією з основоположних засад зовнішньої політики України щодо розбудови відносин з Європейським Союзом є забезпечення інтеграції України до європейського політичного, економічного, правового простору з метою набуття членства в ЄС. Зважаючи на пріоритетність євроінтеграційного курсу України доцільно навести спільну модель, отриману за методикою кластерного аналізу країн ЄС та України, що дозволить оцінити наявний рівень інтеграції України до соціально-економічної системи ЄС.

Приєднання України до п'ятої групи, а не явне виокремлення її в результаті кластерного аналізу, свідчить про єдиний вектор її розвитку в рамках регіону ЄС. Тут діє принцип урівноваженості соціальних та економічних вимірів розвитку – за низьких економічних показників обсяги охоплення населення соціальним програмами та відносні показники асигнувань з державного бюджету на сферу соціальних послуг є дуже високими. Однак таке поширення соціальних гарантій і програм соціального захисту за низь-

ких абсолютних економічних показників є свідченням дотаційного характеру розвитку та підтримки, на противагу бажаній активній соціально-економічній політиці, що передбачає створення умов і досконалої матеріально-технічної і ресурсної бази для розвитку та самозабезпечення.

Загалом у результаті математичного моделювання проведення кластерного аналізу можна дійти висновку про існування в рамках європейського регіону на сучасному етапі економічного розвитку чітко виділених груп країн, що об'єднані цілями, напрямками та інструментами соціальної політики, зокрема підходами до формування повної зайнятості (вибір між підтримкою повної зайнятості та стимулювання ефективності й конкурентоздатності виробництва), розвитком соціального сектора (соціальні послуги надаються через державні та/або приватні організації), часткою соціального страхування в бюджетних видатках на соціальні цілі, адресності соціальної політики тощо.

Математичне моделювання та результати економічного аналізу дозволяють дійти висновку про існування принципових відмінностей і характерних рис розвитку національних економік країн ЄС. Разом із тим, стратегія соціально-економічного розвитку залишається спільною.

Європейська стратегія економічного зростання побудована із врахуванням соціальної складової: боротьби з бідністю та соціальною ізоляцією. Головними цілями стратегії є забезпечення економічної, соціальної та територіальної єдності; гарантування поваги до основних прав людей, що зазнають бідності та соціальної ізоляції, і надання їм можливості гідно жити та брати активну участь у житті суспільства; мобілізація інтеграції обездолених до громад, де вони живуть, їх навчання та допомога в пошуку роботи; забезпечення доступу до соціальних пільг.

Реалізація стратегії базується на всебічній соціалізації сфер державного регулювання, зокрема: ринку праці, соціального забезпечення, охорони здоров'я, освіти; реалізації програм соціальної інтеграції, провадження соціальних інновацій; посиленні координації та часткової уніфікації соціально економічної політики держав – членів ЄС.

Розділ 12
СИГНАЛЬНИЙ МЕТОД АНАЛІЗУ ФАКТОРІВ ВАЛЮТНИХ ТА ФІСКАЛЬНИХ КРИЗ

12.1. Основи сигнального методу

Сигнальний метод/*Signalling approach* дає можливість обрати змінні, які найкраще передбачають настання у майбутньому певної події. Наприклад, це може бути криза (валютна, банківська тощо). Наявність кризи розглядають як бінарну змінну, яка набуває значень 0 (кризи немає або спокійний період) або 1 (наявна криза).

Для кожної досліджуваної незалежної змінної (показника) установлюють певний пороговий рівень значення. При перевищенні показником (незалежною змінною) цього *порогового значення*/*threshold* або *cut-off* вважають, що поданий *сигнал*/*signal* про ймовірність кризи (напр., протягом наступних 24 місяців). Сигнал вважають добрим, якщо криза дійсно відбувається, і поганим, – якщо не відбувається. Порогове значення обирають шляхом, що мінімізує відношення поганих (*шуму*/*noise*) і добрих сигналів, тобто прогнозну здатність сигналів. Можна це проілюструвати за допомогою таблиці матриці сигналів (табл. 12.1).

Таблиця 12.1

Сигнал	Криза відбувається	Криза не відбувається
Є	A	B – помилка I роду (type I error)
Немає	C – помилка II роду (type II error)	D

Тут A – кількість випадків, за яких показник давав сигнал, і криза відбувалася; B – кількість випадків, за яких показник давав сигнал, і криза не відбувалася; C – кількість випадків, за яких показник не давав сигнал, і криза відбувалася; D – кількість випадків, за яких показник не давав сигнал, і криза не відбувалася.

Функція *шум/сигнал/Noise-to-signal ratio* (*NSR*), яку необхідно мінімізувати, має вигляд:

$$NSR = \frac{B/(B+D)}{A/(A+C)}. \quad (12.1)$$

Якщо вдається підібрати таке порогове значення показника, що співвідношення шуму до сигналів стає менше 1, то показник можна використовувати для прогнозування. Що менше це співвідношення, то краще показник пояснює ймовірність настання кризи. Іноді застосовують обернений показник ($SNR = 1/NSR$), який максимізується.

Альтернативною функцією, яка може мінімізуватися, є функція *загальної кількості помилок/Total misclassified errors* (*TME*), яка має вигляд:

$$TME = \frac{C}{A+C} + \frac{B}{B+D}. \quad (12.2)$$

Силу сигналу/Signaling power (*SP*) показника розраховують так:

$$SP = 1 - TME. \quad (12.3)$$

Її використовують як зважувальний коефіцієнт при розрахунку ймовірності виникнення кризової події за формулою:

$$CP = \sum w_i I_i, \quad (12.4)$$

де w_i – вага *i*-того показника (незалежної змінної); I_i – розрахована умовна ймовірність кризи, коли значення *i*-того показника перебуває у небезпечній зоні, якщо воно наразі перебуває у небезпечній зоні; або розрахована умовна ймовірність кризи, коли значення *i*-того показника перебуває у безпечній зоні, якщо воно наразі перебуває у безпечній зоні.

Показники не мають корелювати сильно між собою, але навіть за сильної кореляції сигнальний підхід можна використовувати за умови, що показники, які сильно корелюють між собою, мають одержати меншу вагу.

Але слід зважати на те, що розрахована за допомогою *CP* імовірність має недоліки внаслідок ефекту усереднення, оскільки, наприклад, половина показників є проблемними, а інша половина – безпечними. При цьому не враховують ефект взаємодії факторів.

Альтернативним способом розрахунку ймовірності є побудова спочатку певного композитного показника за формулою:

$$CI = \sum w_i d_i , \qquad (12.5)$$

де w_i – вага *i*-того показника (незалежної змінної); d_i – логічна змінна за *i*-тим показником (набуває значення 1, якщо значення і-того показника наразі перебувають у небезпечній зоні; або 0, – якщо значення *i*-того показника – у безпечній зоні).

Після цього будують логіт-регресію, де залежною змінною є логічна змінна наявності кризи, а незалежною – значення *CI*.

12.2. Дослідження валютних криз сигнальним методом у Microsoft Excel

Наведемо приклад застосування сигнального методу для прогнозування валютних криз у Microsoft Office Excel. Вважатимемо, що в країні відбувається валютна криза, якщо протягом року її валюта знецінилися більш ніж на 25 % (хоча на практиці використовують складніші індекси тиску на валютний ринок), згідно із *World Development Indicators*. Як показник, що має давати сигнал, можна вибрати поточний рахунок щодо ВВП (у %) у 2007 р. Сигнал давав би інформацію про те, що в країні мала відбутися валютна криза протягом наступних трьох років (2008-2010).

Далі до стовпчиків ліворуч вводять дані, а праворуч – прописують необхідні формули для одержання результату. Знизу на рисунку подано фрагменти таблиці в двох видах: вхідні дані та результати обрахунку (рис. 12.1) і текст формул (рис. 12.2).

Якщо, наприклад, до комірки E2 (порогове значення) внести значення 5, то в комірці F2 (функція шум/сигнал) з'явиться значення 0.98. Ітераційним шляхом можна змінювати значення у комірці E2 для мінімізації функції шум/сигнал або функції загальної кількості помилок. На жаль, автоматично знайти рішення складно, оскільки може існувати кілька локальних мінімумів чи максимумів функції.

	A	B	C	D	E	F	G
1	Країна	ПР/ВВП2007	Криза2008_10	Сигнал	Порогове значення	Шум/Сигнал	
2	Angola	17.50	0	0	-6	0.529	
3	Albania	-10.75	0	1		Криза	Немає кризи
4	Argentina	2.82	0	0	Сигнал	6	36
5	Armenia	-6.40	0	1	Немає сигналу	3	66
6	Australia	-6.77	0	1			
7	Azerbaijan	27.29	0	0	Ймовірність кризи, якщо сигнал є		
8	Burkina Faso	-15.81	0	1	0.143		
9	Bangladesh	1.25	0	0			
10	Bulgaria	-27.16	0	1	Ймовірність кризи, якщо сигналу немає		
11	Bahrain	15.73	0	0	0.043		
12	Belarus	-6.70	1	1			
13	Bolivia	12.13	0	0	Загальна кількість помилок		
14	Brazil	0.11	0	0	0.686		
15	Brunei Darussalam	39.42	0	0			
16	Botswana	14.51	0	0	Сила сигналу		
17	Canada	0.84	0	0	0.314		

Рис. 12.1

	D	E	F	G
1	Сигнал	Порогове значення	Шум/Сигнал	
2	=IF(B2<$E$2;1;0)	-6	=(G4/(G4+G5))/(F4/(F4+F5))	
3	=IF(B3<$E$2;1;0)		Криза	Немає кризи
4	=IF(B4<$E$2;1;0)	Сигнал	=COUNTIFS(C2:C112;"=1";D2:D112;"=1")	=COUNTIFS(C2:C112;"=0";D2:D112;"=1")
5	=IF(B5<$E$2;1;0)	Немає сигналу	=COUNTIFS(C2:C112;"=1";D2:D112;"=0")	=COUNTIFS(C2:C112;"=0";D2:D112;"=0")
6	=IF(B6<$E$2;1;0)			
7	=IF(B7<$E$2;1;0)	Ймовірність кризи, якщо сигнал є		
8	=IF(B8<$E$2;1;0)	=F4/(G4+F4)		
9	=IF(B9<$E$2;1;0)			
10	=IF(B10<$E$2;1;0)	Ймовірність кризи, якщо сигналу немає		
11	=IF(B11<$E$2;1;0)	=F5/(G5+F5)		
12	=IF(B12<$E$2;1;0)			
13	=IF(B13<$E$2;1;0)	Загальна кількість помилок		
14	=IF(B14<$E$2;1;0)	=F5/(F4+F5)+G4/(G4+G5)		
15	=IF(B15<$E$2;1;0)			
16	=IF(B16<$E$2;1;0)	Сила сигналу		
17	=IF(B17<$E$2;1;0)	=1-E14		

Рис. 12.2

У цьому прикладі після перебирання варіантів одержано значення порогового рівня для поточного рахунку –6 % ВВП. Йому відповідає значення функції шум/сигнал 0,53. Воно менше 1, тобто розглянутий показник можна використовувати для прогнозування валютної кризи. Ясно: якщо сальдо поточного рахунку становить більш ніж –6 % ВВП, то ймовірність валютної кризи протягом наступних трьох років не є великою – трохи більше 4 %. Якщо ж поточний рахунок становить менше –6 % ВВП (великий дефіцит), то ймовірність валютної кризи помітно вища.

Зверніть увагу: формули таблиці прописані так, що нижче значення показника розглядають як більш небезпечне, ніж вище, оскільки небезпечним є дефіцит поточного рахунку. Для показників, у яких небезпечними є вищі значення (напр., відношення зовнішнього боргу до ВВП), формули у рядку D мають бути іншими – типу: =IF(B2>$E$2;1;0)

Тепер на прикладі за умовними даними розглянемо, як можна врахувати дані за кількома показникам одночасно (рис. 12.3–12.4). Припустимо, є три показники, за якими зроблено подібні розрахунки щодо ймовірності валютної кризи (якщо сигнал є або його немає), порогового рівня та напряму небезпечного діапазону (напр., для першого показника значення більше 80 означає більшу ймовірність кризи), сили сигналу. У рядку 9 проставляють поточні значення показників у країні для прогнозування ймовірність кризи. Урешті-решт у клітинці A14 одержують імовірність кризи 0,0515. Значення є невеликим переважно через те, що показник 1 (має найбільшу силу сигналу) не подає сигнал.

	A	B	C	D
1				
2		Показник 1	Показник 2	Показник 3
3				
4				
5	Ймовірність кризи, якщо сигнал є	0.5	0.1	0.2
6	Ймовірність кризи, якщо сигналу немає	0.03	0.07	0.06
7	Напрям небезпечного діапазону	>	<	<
8	Пороговий рівень	80	-3	1
9	Поточне значення показника	60	-5	0
10	Наявність сигналу	0	1	1
11	Сила сигналу	0.55	0.05	0.15
12	Вага показника	0.733	0.067	0.200
13		0.022	0.007	0.040
14	Розрахована імовірність кризи			
15	0.068666667			

Рис. 12.3

	A	B	C	D
1				
2		Показник 1	Показник 2	Показник 3
3				
4				
5	Ймовірність кризи, якщо сигнал є	0.5	0.1	0.2
6	Ймовірність кризи, якщо сигналу немає	0.03	0.07	0.06
7	Напрям небезпечного діапазону		>	<
8	Пороговий рівень	80	-3	1
9	Поточне значення показника	60	-5	0
10	Наявність сигналу	=IF(B9>B8;1;0)	=IF(C9<C8;1;0)	=IF(D9<D8;1;0)
11	Сила сигналу	0.55	0.05	0.15
12	Вага показника	=B11/SUM($B$11:$D$11)	=C11/SUM($B$11:$D$11)	=D11/SUM($B$11:$D$11)
13		=IF(B10=1;B5*B12;B6*B12)	=IF(C10=1;C5*C12;C6*C12)	=IF(D10=1;D5*D12;D6*D12)
14	Розрахована імовірність кризи			
15	=SUM(B13:D13)			

Рис. 12.4

12.3. Застосування сигнального методу для аналізу факторів фіскальних криз

Крім розрахунку порогових рівнів власноруч на практиці можливо використати вже готові розрахунки інших дослідників, підставивши значення показників для нового спостереження. Опишемо алгоритм адаптації наукових висновків[149] за результатами аналізу даних за 29 *розвиненими країнами*/*advanced economies* і 52 *новітніми ринками*/*emerging markets* за 1970-2009 роки) для аналізу фіскальної ситуації в Україні в наступному періоді. Ці результати можна використати для пояснення впливу факторів, визначення заходів для попередження фіскальних криз, прогнозування.

Як залежну змінну використовують логічну змінну, базовану на чотирьох змінних: дефолт за боргом або його реструктуризація, неявний дефолт, надзвичайне фінансування МВФ і значна обмеженість ринкового фінансування. Останню змінну враховують для виявлення епізодів, під час яких формально ситуацію не розглядали як боргову кризу, але ставки на ринках урядових облігацій створювали суттєве напруження. Детальніше визначення компонентів фіскальної кризи подано у табл. 12.2.

Фіскальна криза в країні відбувається, якщо виконується принаймні одна з чотирьох подій. Для розвинених країн – значні довгострокові спреди за *внутрішніми облігаціями* /*Long-term domestic bond spreads* і спреди за 5-річними *свопами кредитного дефолту*/*credit default swap* (*CDS*); для новітніх ринків – індекс облігацій новітніх ринків (*EMBI spreads*) та довгострокові спреди за внутрішніми облігаціями. Спреди розраховують як різницю у доходності облігацій, порівняно з доходністю аналогічних облігацій США.

[149] Див. : Belhocine N., Dobrescu G., Mazraani S., Petrova I. Assessing Fiscal Stress. IMF Working Paper, August 2010. – 34 p.

Фіскальні кризи в одній країні вважають різними, якщо між ними проходить щонайменше два роки (лише перший рік кризи береться до уваги). 47 незалежних змінних розподілені за чотирма групами: фіскальний вплив, фіскальна стабільність, уразливість і демографічні тенденції. У табл. 12.3 подано визначення основних показників; у табл. 12.4 – результати розрахунків щодо кожного показника серед тих, які є найбільш корисними у прогнозі (наведено лише результати для новітніх ринків).

З метою узагальнення прогнозу за всіма показниками одночасно, замість безпосереднього розрахунку ймовірності, будують *індекс фіскального стресу/fiscal stress index*, який обчислюють за формулою:

$$FSI = \sum w_i d_i , \qquad \textbf{(12.6)}$$

де w_i – вага *i*-того показника (незалежної змінної); d_i – логічна змінна за *i*-тим показником (набуває значення 1, якщо значення *i*-того показника перебуває у небезпечній зоні; або 0, якщо значення *i*-того показника – у безпечній зоні). Прогнозні дані індексу фіскального стресу дають можливість провести ранжування країн і розрахувати середній рівень індексу (для розвинених країн 0.62, для новітніх ринків 0.35).

Одержані порогові рівні порівнюють з даними для країни, яка нас цікавить (у нашому прикладі Україна), у подальшому періоді (табл. 12.5). Варто враховувати не всі показники, а лише ті, які дають меншу кількість помилок і за якими можна знайти нові дані.

Оскільки частину показників вилучають, їх вагові коефіцієнти перерозподіляють між тими показниками, що залишаються. Наприклад, FSI для України у 2011 році становив 0,55, а розрахована ймовірність виникнення фіскальної кризи як зважена середня ймовірностей становила 0,11.

Таблиця 12.2

Подія	Критерій	Визначення критерію	Джерело
Дефолт за боргом або його реструктуризація	Неможливість обслуговування боргу або вимушена конвертація боргу	Визначення S&P	Standards and Poor's and IMF's Private Market Financing for Developing Countries
Значне фінансування	Значна програма, підтримувана МВФ	Доступ до більше 100% квоти	IMF's Finance Department database
Неявний дефолт	Висока інфляція	Річна інфляція більше 35 % для розвинених країн або більше 500 % – для новітніх ринків	Standards and Poor's and IMF's Private Market Financing for Developing Countries
Значна обмеженість ринкового фінансування	Тиск на дохідність суверенного боргу	Спред за суверенним боргом більше 1000 базових пунктів або більш ніж на 2 стандартних відхилення середнього рівня для країни	IFS, Bloomberg and Datastream, річні та місячні дані

Таблиця 12.3

Показник українською мовою	Показник англійською мовою	Джерело: примітки
Первинний баланс з поправкою на циклічність	Cyclically adjusted primary balance	Оцінки персоналу: доходи загального уряду – витрати загального уряду (мінус процентні платежі) з поправкою на тренд випуску
Фіскальний імпульс	Fiscal Impulse	Оцінки персоналу, від'ємне значення зміни структурного балансу
Частка податкових надходжень	Public tax revenue share	WEO
Розрив ВВП (% потенційного ВВП)	Output gap (% potential GDP, in abs. value)	Оцінки персоналу
Державний борг (% багатства фінансового сектору)	Public debt (% private sector wealth)	Eurostat (фінансове багатство приватного сектору)
Валовий зовнішній борг (державний і приватний)	Gross external debt (public and private)	WEO (новітні ринки), JEDH (розвинені країни)
Чистий державний борг	Net public sector debt	WEO
Державний борг на душу населення (дол.)	Public debt stock per capita (US dollars)	WBWDI (населення)
Поточний рахунок	Current account	WEO
Короткостроковий державний борг (% від загального державного боргу)	Short-term public debt (% total public debt)	VEE (новітні ринки), BIS (розвинені країни)
Короткостроковий зовнішній борг (% від валових резервів)	Short-term external debt (% gross reserves)	BIS

Закінчення табл. 12.3

Зовнішній державний борг (% загального державного боргу)	External public debt (% total public debt)	BIS
Волатильність частки державних витрат	Volatility of public expenditure share	Оцінки персоналу: стандартне відхилення витрат загального уряду у % до ВВП, поділене на середнє за останні 5 років
Кредитний рейтинг	Credit Ratings	S&P: AAA=10, AA+=9, AA=8, AA-=7, A+=6, A=5, A-=4, BBB+=3, BBB=2, BBB-=1, BB+=0, BB=-1, BB-=-2, B+=-3, B=-4, B-=-5, CCC=-6, CC=-7, C=-8, SD=-9.2
Зміна рейтингу	Rating Actions	Оцінки персоналу: сума балів "+1" у кредитному рейтингу
Витрати на державні пенсії (%ВВП)	Public pension spending (% GDP)	OECD
Витрати на охорону здоров'я (%ВВП)	Healthcare spending (% GDP)	OECD
Державні витрати на охорону здоров'я (% всіх витрат)	Public health spending (% total spending)	OECD
Ефективний вік виходу на пенсію – чоловіки	Effective Retirement Age – Men	OECD
Ефективний вік виходу на пенсію – жінки	Effective Retirement Age – Women	OECD

Примітка: часто показники беруться у відносному вимірі, наприклад у % ВВП.

Таблиця 12.4

Показник	Напрям безпечного діапазону значень	Пороговий рівень	Імовірність фіскальні кризи для безпечного діапазону	Імовірність фіскальні кризи для безпечного діапазону	Сила сигналу (1 - TME)	Ваговий коефіцієнт
Первинний баланс з поправкою на циклічність	>	–2.9	0.152	0.078	0.11	2.3
Фіскальний імпульс	<	2.0	0.136	0.072	0.14	3.0
Частка податкових надходжень	>	11.4	0.173	0.068	0.16	3.4
Розрив ВВП (% потенційного ВВП)	<	1.8	0.141	0.055	0.24	5.0
Державний борг (% багатства фінансового сектору)	<	6.4	0.123	0.000	0.15	3.2
Валовий зовнішній борг (державний і приватний)	<	59.7	0.133	0.075	0.11	2.4
Чистий державний борг	<	62.4	0.145	0.079	0.13	2.8
Державний борг на душу населення (дол.)	<	1715.4	0.141	0.063	0.19	4.0
Поточний рахунок	>	–3.7	0.133	0.050	0.25	5.2

Закінчення табл. 12.4

Короткостроковий державний борг (% від загального державного боргу)	<	35.1	0.215	0.084	0.14	3.0
Короткостроковий зовнішній борг (% від валових резервів)	<	79.0	0.150	0.056	0.25	5.2
Зовнішній державний борг (% від загального державного боргу)	<	16.1	0.140	0.038	0.26	5.5
Волатильність частки державних витрат	<	8.5	0.130	0.073	0.13	2.9
Кредитний рейтинг	>	−1.0	0.119	0.067	0.14	3.1
Зміна рейтингу	>	−1.0	0.225	0.083	0.10	2.1
Витрати на державні пенсії (% від ВВП)	<	5.2	0.088	0.033	0.23	5.0
Витрати на охорону здоров'я (% від ВВП)	<	5.9	0.233	0.099	0.21	4.4
Державні витрати на охорону здоров'я (% всіх витрат)	<	70.6	0.233	0.080	0.25	5.4
Ефективний вік виходу на пенсію – чоловіки	>	66.1	0.173	0.065	0.22	4.7
Ефективний вік виходу на пенсію – жінки	>	65.7	0.148	0.041	0.16	3.5
Решта показників						28.3

Примітка: імовірності фіскальної кризи для безпечного та небезпечного діапазону кожного показника розраховано, виходячи з даних[150] про кількість кризових, спокійних періодів і часток помилок першого та другого типів.

[150] Див. : Belhocine N., Dobrescu G., Mazraani S., Petrova I. Assessing Fiscal Stress. IMF Working Paper, August 2010. – 34 p.

Таблиця 12.5

Показник	Попереднє/оцінене значення, 2011 р.	На основі джерела	Наявність сигналу	Скоригований ваговий коефіцієнт
Частка податкових надходжень	20	Міністерство фінансів України, WEO	0	8.7
Валовий зовнішній борг (державний і приватний)	76	Національний банк України, WEO	1	6.2
Чистий державний борг	38	WEO	0	7.2
Державний борг на душу населення (дол.)	1420	WEO	0	10.3
Поточний рахунок	–3.9	WEO	1	13.4
Короткостроковий зовнішній борг (% від валових резервів)	155	Національний банк України	1	13.4
Зовнішній державний борг (% від загального державного боргу)	56 (64 включаючи гарантований державою)	Міністерство фінансів України	1	14.1
Волатильність частки державних витрат	4	WEO	0	7.5
Кредитний рейтинг	–3	http://chartsbin.com	1	8.0
Витрати на охорону здоров'я (%ВВП)	3.3	Міністерство фінансів України	0	11.3

Але застосована методика не враховує потенційний ефект взаємодії факторів. Кращу оцінку ймовірності може надати побудова логіт-регресії, де залежною змінною є логічна змінна наявності фіскальної кризи, а незалежною – значення FSI. Приклад на рис. 4.5[151] [44] показує, що така залежність може мати нелінійний характер.

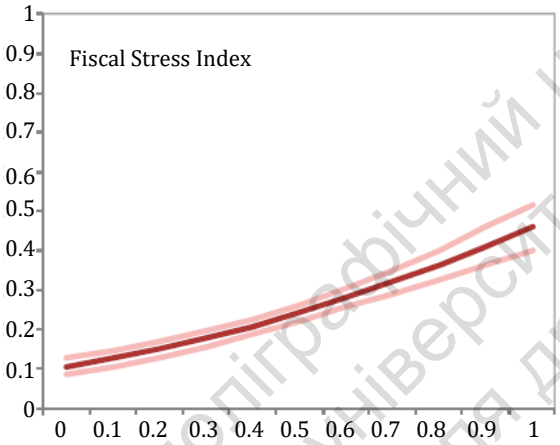


Рис .4.5

Проте для побудови логіт-регресії необхідно мати повні дані за всіма країнами про FSI та періоди фіскальної кризи.

[151] Baldacci E., Petrova I., Belhocine N., Dobrescu G., Mazraani S. Assessing Fiscal Stress. IMF Working Paper WP/11/100, May 2011. – 41 p. http://www.imf.org/external/pubs/ft/wp/2011/ wp11100.pdf.

Додаток А
Основи роботи у програмному забезпеченні Microsoft Office Excel

Розглянемо лише ті аспекти роботи у Microsoft Office Excel, які безпосередньо стосуються кількісного аналізу.

Увага: у виданні застосовано західний стиль позначення розділового знаку для дрібних розрядів чисел, тобто крапка, а не кома, оскільки більшість джерел міжнародної економічної статистики є закордонними і використовують саме цей стиль. Цей параметр можна обрати в Microsoft Windows. Для цього після натиснення кнопки *Пуск* потрібно обрати *Панель керування/Control panel,* далі обрати *Мова й регіональні стандарти/Regional and Language Options*, обрати *Налаштування/Settings/Options*, у полі *Десятковий розділювач/ Decimal separator* обрати крапку (.) або кому (,). Наприклад, якщо в Microsoft Windows як розділовий знак дрібної частини використовують крапку, то Microsoft Excel сприйматиме запис у комірці або формулі типу 3.4, як число, а запис типу 3,4 – як текст. Через це можуть стати неможливими розрахунки з числами типу 3,4. Тому потрібно пересвідчитись, що Microsoft Excel сприймає відповідні записи як число для того, щоб проводити з ними розрахунки, наприклад, позначити всі комірки з числами типу 3,4 і замінити коми на крапки.

Якщо в Microsoft Windows як розділовий знак дрібної частини використовується кома, то виникає протилежна проблема: Microsoft Excel сприйматиме запис у комірці або формулі типу 3,4, як число, а запис типу 3.4 – як текст.

А1. Формули

Формули дозволяють проводити розрахунки та інші операції з даними. Кожна формула починається зі знаку =. Формула повертає результат розрахунку до комірки, в якій вона прописана.

Аргументами формул можуть, наприклад, бути числа, текст, посилання на інші комірки. Посилання на інші комірки можна вводити по-різному:

- прописуючи їх назву у формулі;
- відмічаючи потрібні комірки в процесі написання формули.

Приклади формул показано у табл. А1

Таблиця А1

=3+7*2	Додає 3 до добутку 7 та 2
= ABS(-3)	Обраховує модуль числа -3, тобто повертає 3
= A1+B3	Додає значення комірок А1 та В3
= IF(A3>5)	Перевіряє чи значення у комірці А3 є більшим за 5

Елементи формул. Приклад:

=ABS(-3)*A1^5, де ABS(-3) – функція, яка повертає значення 3; A1 – посилання на комірку А1; -3 та 5 – константи, тобто числа або текст, які вводять у формулу безпосередньо; * та ^ – оператори, тобто символи, які позначають тип розрахунків (* – множення, ^ – введення значення перед цим оператором у ступінь, що вказаний після оператора).

Оператори. Приклади операторів подано у табл. А2. Розрахунки у формулах здійснюють, згідно із правилами арифметики (напр., спочатку – множення, а потім – додавання). Зокрема операції здійснюють в такому порядку:

пробіл , – % ^ * та / + та - & = < > <= >= <>

Якщо є оператори з однаковим пріоритетом, то першою здійснюють операцію, оператор якої є лівішим. За допомогою дужок можна коригувати порядок операцій. Наприклад, у наступній формулі спочатку виконуватиметься додавання, а не множення = 7*(3+5) і результатом формули буде 56.

Таблиця А2

Оператор	Тип операції	Приклад
Арифметичні (результатом є число)		
+	Додавання/addition	1+2
-	Віднімання/subtraction	3-2
*	Добуток/product	5*2
/	Ділення/division	1/2
%	Процент/percent	9%
^	Введення у ступінь/power	2^2

Закінчення табл. А2

Порівняння (результатом є логічне значення Істина/TRUE або Неправда/FALSE)		
=	Дорівнює/equals	A1=A2
>	Більше/more	A1>A2
<	Менше/less	A1<A2
>=	Більше або дорівнює/more or equal	A1>= A2
<=	Менше або дорівнює/less or equal	A1<=A2
<>	Не дорівнює/is not equal	A1<> A2
Текстові		
&	Конкатенація – об'єднання тексту/concatenation – merging text	"Експорт"&"України" дає результат "Експорт України"
Посилання		
:	Діапазон/array/list	A1:A10
,	Об'єднання множин/merging lists	A1:A10,A20:A22
(пробіл/space)	Перетин множин/ intersection of lists (спільні комірки двох діапазонів)	A1:B20 B5:K10

А2. Функції

Функції – це стандартні формули, що вбудовані до програмного забезпечення. Їх можна використати для розрахунку числових значень, одержання поточної дати, пошуку, перетворення тексту, порівняння. Елементи формул. Приклад:

=POWER(3;G6)

Як і формули, функції починаються зі знаку =. Далі йде назва функції. У дужках вказано аргументи, розділені знаком ; . Часто важливим є порядок, у якому прописують аргументи. Аргументами можуть бути:

- числа;
- текст;

- логічні значення (TRUE або FALSE);
- масиви (множини даних у діапазоні комірок);
- помилки.

Аргументи формул також можна класифікувати як:

- константи (постійне значення, яке не розраховують, наприклад, текст або число);
- посилання на комірки;
- формули;
- функції.

Приклад вставки однієї функції до іншої:

=POWER(2+ABS(-3);G6)

У деяких функціях потрібно вказувати лише аргументи певного типу. Деякі функції не потребують аргументів.

При введені тексту функції до комірки можуть з'являтися підказки (за початковими літерами – повна назва функції, після розкриття дужки – характер аргументів, наприклад POWER (*число*;*ступінь*)). Ще один спосіб вставити функцію – натиснути кнопку *Вставити функцію*/Insert function у меню *Формули*/*Fomulas*, обрати з переліку потрібну функцію та у вікні, що з'явиться після натискання кнопки ОК, указати аргументи:

- набрати з клавіатури;
- вставити попередньо скопійовані значення;
- або натиснути мишею на комірку (чи виділити діапазон комірок, натиснувши лівою кнопкою миші та протягнувши її за відповідними комірками), посилання на яку (які) потрібно використати як аргумент.

Після введення функції (як і формули) в комірці відобразиться результат розрахунку або значення помилки. Сам текст функції (або формули) відображатиметься у рядку формул, коли відповідна комірка є обраною.

А3. Перерахунок формул

При зміні значень комірок, на які є посилання в формулі, (впливаючих комірок) Microsoft Office Excel автоматично перераховуватиме результат у комірці, в якій прописана

формула (залежна комірка). Якщо посилання є циклічним (залежна комірка сама впливає на свої впливаючі комірки), то автоматичний перерахунок неможливий.

Вибір способу перерахунку можна змінити. Для цього у вкладці *Файл/File* оберіть *Параметри/Options*, а далі – *Формули/Formulas*. Оберіть необхідний спосіб у розділі *Параметри обчислень/Calculation options*:

- *Автоматично/Automatic*;
- *Автоматично, окрім таблиць даних/Automatic except for data tables;*
- *Вручну/Manual* (в останньому випадку автоматично поряд буде включено режим *Перерахувати книгу перед збереженням/Recalculate workbook before saving*, але його можна відключити).

Зміна способу перерахунку впливатиме на всі відкриті книги. Для перерахунку за ручного способу перерахунку оберіть у вкладці *Формули/Formulas* в групі *Обчислення/Calculation* кнопку *Виконати обчислення в книзі/Calculate now* (для перерахунку всіх відкритих книг, включаючи таблиці даних та оновлення всіх відкритих аркушів діаграм) або *Обчислення в аркуші/Calculate sheet* (для перерахунку активного аркушу, діаграм та аркушів діаграм, що пов'язані з цим аркушем).

А4. Посилання на комірки у формулах

Для того, щоб вказати посилання на комірку, потрібно набрати назву посилання з клавіатури або під час введення формули натиснути на відповідну комірку, на яку здійснюється посилання (в останньому випадку, якщо комірок багато, потрібно виділити діапазон, натиснувши лівою кнопкою миші на крайню комірку діапазону та провести мишею до протилежної крайньої комірки діапазону). Увага: у посиланнях використовують латинські літери, тому якщо замість латинської А буде стояти А кирилицею, то формула повертатиме помилку замість результату розрахунку.

Приклади посилань на комірки вказані у табл. А3. При копіюванні, заповненні за рядками (або стовпчиками) чи перенесенні формули до іншої комірки посилання можуть залишатися незмінними (якщо використані абсолютні посилання) або зміщуватися (якщо використані відносні посилання).

Таблиця А3

B3	Комірка на перетині стовпчика B і рядка 3
A2:A5	Діапазон комірок у стовпчику A, рядків 2-5
B3:K3	Діапазон комірок у рядку 3, стовпчики B-K
A3:C7	Прямокутний діапазон комірок у стовпчиках A-C, рядках 3-7
Експорт!A2:A5	Діапазон A2:A5 в аркуші під назвою *Експорт* у тій самій книзі
'Експорт%'!A2:A5	Діапазон A2:A5 в аркуші під назвою *Експорт%* у тій самій книзі – одинарні лапки використовують, якщо назва аркушу містить не тільки літери
[Україна.xlsx]'Експорт%'!A2:A5	Діапазон A2:A5 в аркуші під назвою *Експорт%* у книзі *Україна.xlsx*
Лист5:Лист7!A1:A8	Діапазон A1:A8 в аркушах *Лист5, Лист6* та *Лист7*

Відносні посилання. Припустимо в комірці A1 є формула =2*A2. A2 – приклад відносного посилання. При копіюванні формули до комірки B3 формула в ній виглядатиме як =2*B4. Тобто у відносному посиланні важливо, де розташована комірка, на яку є посилання, відносно комірки з формулою.

Абсолютні посилання. Припустимо в комірці A1 є формула =2*$A$2. Абсолютні посилання позначає знак $ перед літерою стовпчика та числом рядка. При копіюванні формули до комірки B3 формула в ній виглядатиме як =2*$A$2, тобто зміщення посилання не відбудеться.

Змішані посилання містять абсолютний стовпчик і відносний рядок або відносний стовпчик та абсолютний рядок. Припустимо в комірці A1 є формула =2*$A2. При копіюванні формули до комірки B3 формула в ній виглядатиме як =2*$A4. Приклади змішаних посилань: $A2 або B$6.

А5. Функції масиву

Функція масиву/array function повертає одне або кілька значень, які з'являються у одній чи кількох комірках, відповідно. Аргументами формули масиву є кілька наборів значень (аргументи масиву). Потрібно, щоб аргументи масиву у формулі включали однакову кількість стовпчиків і рядків. Формули масиву вводять не кнопкою *Ввід/Enter*, а одночасним натисканням трьох клавіш *CTRL+SHIFT+Enter*. Приклад розрахунку одного значення показано у табл. А4.

Таблиця А4

	A	B	C
1		Експорт (фізичний обсяг), тонни	Ціна, долари
2	Нафта	100	400
3	Чавун	300	500
4	Загальний експорт (вартісний обсяг)	=SUM(B2:B3*C2:C3)	

Після натискання *CTRL+SHIFT+Enter* формула у рядку формул матиме вигляд {=SUM(B2:B3*C2:C3)}, тобто буде у фігурних дужках.

Приклад розрахунку кількох значень. Припустимо у рядку А – експорт за п'ять років. У рядку В ми хочемо дізнатися розраховані значення експорту в ці роки, згідно з лінійним трендом. Помітимо діапазон B2:B5 (перед вводом формули треба вказати діапазон комірок, який міститиме результати). У рядку формул напишемо формулу =TREND(A1:A5). Натиснемо *CTRL+SHIFT+Enter*. У рядку B з'являться результати розрахунку (табл. А5).

Таблиця А5

	A	B
1	35	39
2	50	49
3	60	59
4	80	69
5	70	79

Після вводу формули масиву, значення у комірках, де містяться результати, неможна змінити. Також неможна вставляти чи видаляти рядки чи стовпчики посеред діапазону з результатами. З одного боку, це не зручно, з іншого, – це убезпечує від внесення випадкових помилкових змін.

Для видалення формули масиву потрібно виділити комірку з результатом розрахунку, виділити формулу у рядку формул, натиснути клавішу *DELETE*, а потім ще натиснути *CTRL+SHIFT+Enter.*

А6. Переміщення та копіювання формул

Для вирізання або копіювання формул (або чисел, тексту) можна використати стандартні кнопки *Вирізати/Cut* та *Скопіювати/Copy* у вкладці *Основне/Home.* Для вставки після виділення діапазону, куди потрібно скопіювати або перемістити, використовується кнопка . Можна обрати спосіб вставки, натиснувши на трикутничок під цією кнопкою. З'являються кнопки, над якими – підказки. Наприклад, можливо вставити:

– формули (відображатимуться значення, але нові комірки міститимуть формули);

– лише значення (відображатимуться значення та нові комірки не міститимуть формули – це зручно, коли нас цікавлять лише значення, які їх потім не потрібно перераховувати, якщо нас цікавлять лише значення, а формула містить відносні посилання, і при переміщенні формули результат може змінитися);

– значення та формат;

– лише формат;

– транспонувати (порівняно з первинним діапазоном комірок у нового діапазону замість стовпчиків будуть рядки, а замість рядків – стовпчики, наприклад, як з табл. А6 до табл. А7).

Таблиця А6

Первинний діапазон	
6	7
2	1

Таблиця А7

Діапазон після копіювання з транспонуванням	
6	2
7	1

Копіювати формули до сусідніх комірок можна також за допомогою маркера заповнення (невеликого чорного квадрату праворуч знизу виділеного діапазону або комірки). Після виділення комірки чи діапазону з потрібною формулою натисніть на маркер заповнення та перемістіть у потрібному напрямі. Формули копіюють до тих комірок, які виділяють при переміщенні цього маркеру заповнення.

Після копіювання або переміщення варто впевнитися, що формули використовують потрібні аргументи, а комірки містять потрібні значення.

А7. Виправлення помилок

Замість значень формули можуть повертати помилки. Приклади помилок:

#DIV/0! – якщо відбувається ділення на нуль або ділення на посилання на комірку з відсутнім значенням;

#VALUE! – якщо формула містить посилання на комірки з різними типами даних; у формулі введено посилання як аргумент, а значення у комірці, на яку вказане посилання, іншого типу, ніж дозволяється, наприклад, якщо H13 містить текст замість числа;

#NAME? – якщо назва формули вказана неправильно; текст як аргумент введений не в лапках; у посиланні на діапазон комірок відсутній знак : ; посилання на інший аркуш без одинарних лапок (') тощо;

#N/A – якщо у формулі масиву аргумент з іншою кількістю стовпчиків і рядків, ніж у діапазоні, якій містить формулу масиву; не вказані обов'язкові аргументи для функції тощо;

#REF! – якщо посилання на комірку не дійсне (напр., посилання є, а комірка була видалена, оскільки було видалено відповідний аркуш, стовпчик тощо; або після копіювання відносні посилання не можуть бути адекватно відображені);

#NUM! – якщо у функції використано не той тип даних в аргументі (напр., у результаті формули =LOG(H13); результатом функції є число занадто велике або мале для того, щоб його можна було представити в програмі;

– позначка комірки, якщо ширина стовпчика недостатня для відображення всіх цифр числа, для її виправлення достатньо розширити стовпець до необхідної ширини;

#NULL! – якщо вказаний перетин двох множин, які насправді не перетинаються, наприклад, =SUM(B2:B3 D2:D3).

А8. Впливові та залежні комірки

Приклад. Якщо в комірці А1 є формула =A2*2, то комірка А1 буде залежною щодо А2, а А2 буде такою, яка здійснює вплив відносно А1. Зв'язки між комірками також можна відобразити графічно стрілками (стрілка йде від комірки, що впливає, до залежної комірки).

Оберіть комірку, для якої потрібно знайти залежності. На вкладці *Формули/Formulas* у групі *Залежності формул/ Formula auditing* натисніть кнопку *Впливаючі комірки/Trace precedents*. З'являться стрілки до цієї комірки з комірок, на які є посилання. Якщо натиснути кнопку *Trace precedents* ще раз, то буде показано комірки, що впливають, другого порядку, які здійснюють вплив на впливові комірки для первинно обраної комірки. Кожного разу при натисканні на *Trace precedents* з'являтимуться залежності все більш глибокого рівня.

Аналогічно натискаючи кнопку *Залежні комірки/Trace Dependents*, можна дізнатися, в яких комірках є посилання на виділену комірку. Так само можна натискати її кілька разів, щоб дізнатися залежності більшого рівня.

За допомогою кнопки *Видалити стрілки*/*Remove arrows* можна повернутися до режиму відображення, коли стрілки між комірками відсутні.

Сині стрілки вказують на комірки без помилок, червоні – на комірки, які спричиняють помилки. Якщо залежна комірка або комірка, що впливає, розташована на іншому аркуші або в іншій книзі, то відображатиметься чорна стрілка зі значком 

А9. Основні логічні функції

Логічні функції можуть бути, зокрема, корисними для перевірки на відсутність помилок або при створенні нових змінних, особливо неметричних. Розглянемо кілька корисних функцій.

=IF(логічний вираз; значення якщо TRUE; значення якщо FALSE) повертає значення (напр., число або текст), вказане як другий аргумент, якщо логічний вираз у першому аргументі вірний, інакше повертає значення, вказане як третій аргумент). Приклад: у комірці D2 напишемо формулу (а потім її скопіюємо в нижчі комірки): =IF(C2="ЗВТ";1;0).

Ця функція корисна, якщо потрібно створити псевдозмінну (для подальшої побудови регресійної моделі), яка набуває значення 1, якщо відповідному спостереженню відповідає режим зони вільної торгівлі, та 0, якщо зони вільної торгівлі немає. У таблиці з вхідними даними наявність зони вільної торгівлі позначена текстом "ЗВТ", як у табл. А8.

Таблиця А8

	A	B	C	D
1	Країни	Зовнішньоторговельний оборот	Регіональна торговельна угода	Наявність ЗВТ
2	Україна-Канада	25000	ЗВТ	1
3	Україна-Бразилія	1000		0

Інші приклади використання формули вказано у табл. А9.

Таблиця А9

=IF(A1=100;A2+A3;"")	Якщо A1=100, розраховують суму A2 та A3, інакше комірка буде пустою
=IF(A1>10;"Високий тариф"; "Низький тариф")	Якщо значення A1 більше 10, то у комірці з формулою буде текст "Високий тариф", інакше – "Низький тариф"
= IF(A1=B1; "ОК"; "Помилка")	Якщо A1=B1 повертається "ОК", інакше "Помилка"
= IF(A1>10;"Високий тариф"; IF(A1>5;"Середній тариф"; "Низький тариф")	Приклад вкладеної функції. Якщо значення A1 більше 10, у комірці з формулою буде текст "Високий тариф", якщо більше 5 і до 10 включно – "Середній тариф", до 5 включно "Низький тариф")

=AND(логічний вираз; логічний вираз; ...)

повертає значення TRUE, якщо кожний з аргументів дорівнює TRUE. Якщо принаймні один з аргументів дорівнює FALSE, то повертає FALSE. Приклади:

Таблиця А10

=AND(1<A2;A2<100)	
=IF(AND(0<A1;A1<100); A1;"Помилка")	Повертатиме значення A1 лише, якщо 0<A1<100, інакше повідомлятиме про помилку

=OR(логічний вираз; логічний вираз; ...)

повертає значення TRUE, якщо принаймні один з аргументів дорівнює TRUE. Якщо всі аргументи дорівнюють FALSE, то повертає FALSE.

=NOT(логічний вираз)

повертає значення TRUE, якщо аргумент дорівнює FALSE. Якщо аргумент дорівнює FALSE, то повертає TRUE.

А10. Основні математичні функції

=ABS(число)

повертає модуль (абсолютну величину) числа (тут і далі замість числа може бути посилання на комірку, в якій міститься число або формула, яка повертає число). Наприклад, =ABS(-10) повертає 10.

=SIGN(число)

повертає 1, якщо аргумент додатне число; 0, – якщо він дорівнює нулю; –1, якщо аргумент – від'ємне число. Наприклад, =SIGN(-10) повертає –1.

=ROUND(число для округлення;кількість десятинних розрядів для округлення)

повертає округлене число.

Приклади:

Таблиця А11

=ROUND(4.2;0)	Повертає 4
=ROUND(A5;0)	Наприклад, якщо в комірці А5 значення дорівнює 5,8, то повертає 6
=ROUND(-4,2;0)	Повертає -4
=ROUND(1.337;2)	Повертає 1,34
=ROUND(53721;-3)	Повертає 54000

Для округлення вверх або вниз використовують схожі функції: ROUNDUP та ROUNDDOWN. Наприклад:

=ROUNDDOWN(-371;-1)

повертатиме –370.

=POWER(число-основа;число-показник ступеню)

повертає результат зведення числа у ступінь (аналог оператора ^). Наприклад, =POWER(10;2) повертає 100.

=EXP(число)

повертає результат зведення числа *e* (2,728182845904) у ступінь, задану аргументом. Наприклад, =EXP(2.5) повертає 12.182494.

=LN(число)

повертає натуральний логарифм додатного числа (логарифм з основою 2,728182845904). Наприклад, =LN(5) повертає 1.6094379.

=LOG10(число)

повертає десятковий логарифм додатного числа (логарифм з основою 10). Наприклад, =LOG10(1000) повертає 3.

=LOG(число;основа)

повертає логарифм додатного числа (логарифм з вказаною основою). Наприклад, =LOG(8; 2) повертає 3.

А11. Функції суми та добутку

=SUM(числа)

повертає суму чисел. Наприклад, =SUM(A1:A10;C5:C9)

Для функції суми також можна використовувати кнопку автосуми Σ у вкладці *Основне*/*Home*, якщо діапазон чисел розташований у суміжних комірках. Перед натисканням цієї кнопки потрібно виділити комірку під стовпчиком чисел або праворуч від рядка чисел, що додаються. У цій комірці буде підрахована сума.

=PRODUCT(числа)

повертає добуток чисел.

=SUMIF(діапазон;критерій;діапазон додавання)

повертає суму чисел, які відповідають певному критерію.

Діапазон – посилання на комірки, які оцінюють за критерієм. Критерій для оцінювання може бути числом, виразом, посиланням на комірку, текстом, функцією.

Критерії у вигляді тексту або з логічними чи математичними знаками мають бути у лапках. В аргументі критерій можна використовувати знак ? (означає будь-який знак) та * (будь-яка послідовність знаків).

Діапазон додавання – необов'язковий аргумент, указує на комірки, які додаються, якщо вони відрізняються від комірок, що вказані у аргументі діапазон. Якщо діапазон додавання не вказано, то додаються значення у комірках, які вказані в аргументі діапазон.

Приклади – табл. А12.

Таблиця А12

=SUMIF(A1:A5;"Експорт";C1:C5)	Повертає суму тих чисел у діапазоні C1:C5, якщо у відповідних комірках в одному й тому рядку з цими числами у діапазоні A1:A5 вказано текст *Експорт*. Наприклад, це може бути корисним, якщо в рядках ідуть різні показники за різними країнами та потрібно підсумувати один показник за кількома країнама
=SUMIF(A1:A100;">0")	Повертає суму тих чисел у діапазоні A1:A100, які перевищують 0. Наприклад, якщо у вказаному діапазоні – баланс поточного рахунку за різними країнами, а потрібна сума лише профіцитів балансу поточного рахунку за кількома країнами (інакше сума дефіцитів і профіцитів наближатиметься до нуля).
=SUMIF(A1:A10;0;B1:B10)	Наприклад, якщо у A1:A10 вказано ставку імпортного мита за конкретними товарами, а в B1:B10 – імпорт цих товарів, то за допомогою цієї функції можна підрахувати величину імпорту товарів, з яких не стягується мито.
=SUMIF(A1:A10;">"&E20;B1:B10)	Повертає суму тих чисел до діапазону B1:B10, якщо у відповідних комірках в одному й тому рядку з цими числами у діапазоні A1:A10 числа перевищують число, що вказано у комірці E20.
=SUMIF(A1:A5;"*порт";C1:C5)	Повертає суму тих чисел у діапазоні C1:C5, якщо у відповідних комірках в одному й тому рядку з цими числами у діапазоні A1:A5 указано текст, який закінчується на *порт*. Наприклад, якщо потрібно підрахувати суму за країнами значень обох показників експорту та імпорту (обидва слова закінчуються літерами *порт).*
=SUMIF(A1:A5;"";C1:C5)	Наприклад, якщо в A1:A5 вказано товари, які імпортуються, а в C1:C5 – величини імпорту цих товарів, то функція повертає суму імпорту тих товарів, назви яких не вказано.

=SUMIFS(діапазон додавання; діапазон умови; критерій; діапазон умови; критерій;…)

повертає суму чисел, які відповідають одразу кільком критеріям.

Приклад – у табл. А13.

Таблиця А13

	A	B	C	D
1	Країна-експортер	Країна-імпортер	Товар	Обсяг
2	Україна	Польща	чавун	100
3	Україна	ФРН	чавун	200
4	Норвегія	Польща	нафта	1500
5	Україна	ФРН	зерно	100
6	США	Мексика	автомобілі	2000

Функція

=SUMIFS(D2:D7;A2:A7;"Україна";C2:C7;"чавун")

повертатиме суму експорту з України чавуна.

=SUMSQ(числа)

повертає суму квадратів чисел.

Наприклад, =SUMSQ(A1:A2). Може бути корисною, якщо самостійно розраховувати деякі статистичні показники. Або, наприклад, якщо з метою попереднього аналізу потрібно дізнатися, наскільки тісно пов'язані два показники (напр., зростання прямих іноземних інвестицій у двох країнах), а в комірках А1 та А2 вказано кореляцію між зростанням прямих іноземних інвестицій у двох країнах: в А1 – за один період (напр., 1991-2010 рр. або 1991-2000), а в А2 – за інший період, напр., за 2001-2010 рр.). Сума квадратів дозволяє уникнути ситуації, за якої від'ємна та додатна кореляція при сумуванні взаємно компенсують одна одну (тоді б ми не побачили наявність суттєвого зв'язку, який міг би змінюватися з часом).

=SUMXMY2(масив1;масив2)

повертає суму квадратів різниці чисел. Як і в інших подібних функціях масиви мають бути однакового розміру за кількістю комірок. Фактично використовується формула:

$$\sum (x - y)^2.$$

Приклади – табл. А14

Таблиця А14

=SUMXMY2(A1:A8;B1:B8)	Повертає суму: $(A1-B1)^2+(A2-B2)^2+...+(A8-B8)^2$
=SUMXMY2({2;3;9};{6;5;11})	Повертає суму квадрату різниць двох масивів констант

=SUMPRODUCT(масив1;масив2;...)

повертає суму добутків пар чисел у відповідних масивах.

Приклад – табл. А15.

Таблиця А.15

	A	B	C	D	E
1	Місяць	Експорт, тонн	Імпорт, тон	Експортна ціна за тонну	Імпортна ціна за тонну
2	Січень	100	30	100	80
3	Лютий	80	40	90	70
4	Березень	60	20	60	80
5	Квітень	70	30	70	75

=SUPRODUCT(B2:C5;D2:E5)

повертає вартісний зовнішньоторговельний оборот за чотири місяці:

100*100+30*80+80*90+40*70+60*60+20*80+70*70+30*75,

аналогом цієї функції є функція =SUM(B2:C5*D2:E5)

=SUMX2MY2(масив1;масив2)

повертає суму різниць квадратів чисел. Фактично використовується формула:

$$\sum(x^2 - y^2).$$

Приклад:

=SUMX2MY2(A1:A2;B1:B2)

повертає суму $(A1^2-B1^2)+(A2^2-B2^2)$

=SUMX2PY2(масив1;масив2)

повертає суму сум квадратів чисел. Фактично використовується формула:

$$\sum(x^2 + y^2).$$

Наприклад,

=SUMX2PY2(A1:A2;B1:B2)

повертає суму $(A1^2+B1^2)+(A2^2+B2^2)$.

A12. Табличні функції

=COUNT(діапазон)

повертає кількість комірок у діапазоні, які містять числа. Приклад – табл. A16.

Таблиця A16

Індекс експортних цін (2000=100)	**A**	**B**	**C**	**D**
1	Країна/рік	1993	1994	1995
2	Україна	50	60	75
3	Білорусь	н. д.	70	65

=COUNT(B2:D3)

повертає кількість спостережень (країно-років) з наявними числовими даними (5).

Якщо потрібно підрахувати також й комірки, що містять й інші дані (логічні значення, текст, значення помилок), то використовують функцію =COUNTA(діапазон).

=COUNTIF(діапазон; критерій)

повертає кількість комірок, які відповідають певному критерію. Приклади – табл. A17.

=COUNTIFS(діапазон умови; критерій; діапазон умови; критерій;…)

повертає кількість комірок, які відповідають одразу кільком критеріям. Приклади – табл. A18.

=FREQUENCY(масив даних; масив інтервалів)

повертає масив чисел, які вказують на частоту появи даних у певних діапазонах значень. Є формулою масиву (отже вводять за допомогою CTRL+SHIFT+ENTER). Приклад – табл. A19.

Дані в діапазоні B8:B11 з'явилися після вставки туди формули масиву =FREQUENCY(B2:B7;C2:C4).

Функція FREQUENCY у поєднанні з функціями SUM та IF дозволяє проводити сумування унікальних значень. Приклад – табл. A20.

Таблиця А17

=COUNTIF(B1:B100;"ЗВТ")	Повертає кількість комірок у діапазоні B1:B100, в яких наявний текст ЗВТ (зона вільної торгівлі)
=COUNTIF(B1:B100;"<>"&E3)	Повертає кількість комірок у діапазоні B1:B100, в яких значення не дорівнює значенню у комірці E3
=COUNTIF(B1:B100;">=10")-COUNTIF (B1:B100;">30")	Повертає кількість комірок у діапазоні B1:B100, в яких значення становить від 10 до 30
=COUNTIF(A1:A5;"*порт")	Повертає кількість комірок у діапазоні A1:A5, в яких указано текст, який закінчується на *порт*.
=COUNTIF(A1:A5;"*")	Повертає кількість комірок у діапазоні A1:A5, в яких указано будь-який текст
=COUNTIF(A1:A5;"???рс")	Повертає кількість комірок у діапазоні A1:A5, в яких указано будь-яке слово із 5 букв, що закінчуються на *рс* (напр. *GDPpc*)
=COUNTIF(A1:A5;"так")/ROWS(A1:A5)	Повертає середню кількість комірок, коли зустрічається слово *так* (напр., якщо в стовпчику A вказано на наявність валютної кризи (*так* або *ні*) у певній країні у певний рік)

Таблиця А18

=COUNTIFS(B1:B5;"=так";C1:C5;">0")	Повертає кількість рядків, в яких у діапазоні B1:B5 комірки мають значення *так*, а в діапазоні C1:C5 комірки мають число більше 0
=COUNTIFS(B1:B5;"=так";C1:C5;"<"&E4)	Повертає кількість рядків, в яких у діапазоні B1:B5 комірки мають значення *так*, а в діапазоні C1:C5 комірки мають число менше, ніж вказане у комірці E4
=COUNTIFS(B1:B5;"=так";C1:C5;">31.12.2000")	Повертає кількість рядків, в яких у діапазоні B1:B5 комірки мають значення *так*, а в діапазоні C1:C5 комірки мають дату після 31 грудня 2000 року

Таблиця А19

	A	B	C
1	Товари	Ставка імпортного тарифу	Інтервали
2	Товар 1	0	5
3	Товар 2	5	10
4	Товар 3	0	20
5	Товар 4	25	
6	Товар 5	12	
7	Товар 6	2	
8		3	
9		1	
10		1	
11		1	

Таблиця А20

	A
1	3
2	8
3	7
4	8
5	1
6	3
7	1

=SUM(IF(FREQUENCY(A1:A7;A1:A7)>0;A1:A7))

у сумі дає 3+8+7+1=15.

Функція FREQUENCY разом з іншими функціями дозволяє проводити розрахунок кількості унікальних значень. Наприклад, табл. А21–А22.

=CONCAT(масив даних; масив інтервалів)

об'єднує кілька стовпчиків з даними. Наприклад, вставимо до D2 формулу =CONCAT(A2;",";B2;"-";C2) і скопіюємо її до D3 (табл. А23).

Таблиця А21

	A	B
1	Україна	3
2	Білорусь	8
3	Болгарія	
4	Польща	8
5	Україна	1
6	Польща	3
7	Україна	1

Таблиця А22

=SUM(IF(FREQUENCY(A1:B7;A2:B7)>0;1))	Повертає кількість унікальних числових значень до діапазону A1:B7 без урахування пустих комірок і тексту (3)
=SUM(IF(FREQUENCY(MATCH(A1:A7; A1:A7;0);MATCH(A1:A7;A1:A7;0))>0;1))	Повертає кількість унікальних текстових або числових значень до діапазону A1:A7, діапазон не має містити пусті комірки (4)
=SUM(IF(FREQUENCY(IF(LEN(A1:A7)>0; MATCH(A1:A7;A1:A7;0);"");IF(LEN(A1:A7)>0; MATCH(A1:A7;A1:A7;0);""))>0;1))	Повертає кількість унікальних числових або текстових значень до діапазону A1:B7 без урахування пустих комірок і тексту (4)

Таблиця А23

	A	B	C	D
1	Показник	Одиниця виміру	Ціни	Повна характеристика показника
2	Експорт	долар	за місяць	Експорт, долар за місяць
3	М2	національна валюта	на кінець місяця	М2, національна валюта на кінець місяця

А13. Надбудови

Надбудовою/*Add-in* в Microsoft Excel є компоненти, які надають додаткові можливості. Стандартними надбудовами є *Пакет аналізу*/*Data Analysis* та *Пошук рішення*/*Solver*, які включені вже до Microsoft Excel (існують й зовнішні надбудови, напр., PhStat). Але у вже інстальованому Microsoft Excel надбудови потрібно додатково інсталювати або активувати.

Для активації стандартних надбудов потрібно відкрити вкладку *Файл*/*File*, *Параметри*/*Options*, обрати *Надбудови*/*Add-ins.* У полі *Керування*/*Manage* обрати *Надбудови Excel*/*Excel add-ins* і натиснути *Перейти*/*Go*. Установіть позначки проти *Data Analysis* та *Solver* та натисніть *ОК*. Після цього у вкладці *Дані*/*Data* в групі *Аналіз*/*Analysis* можна побачити кнопки *Аналіз даних*/*Data Analysis* та *Пошук рішення*/*Solver*. При натисканні кнопки *Аналіз даних*/*Data Analysis* з'являється вікно, в якому можна вибрати конкретний інструмент аналізу (рис. А1).

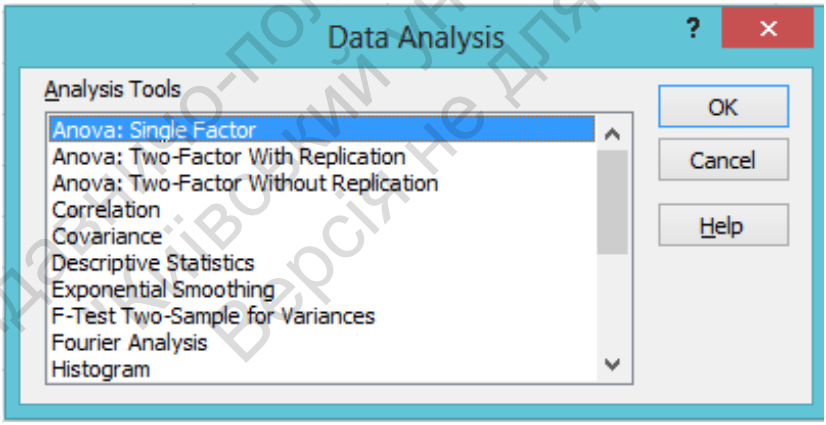


Рис. А1

Наприклад, інструмент *Вибірка*/*Sampling* створює з вихідного діапазону (який розглядають як генеральну сукупність) вибірку. Це може бути корисним, наприклад, коли побудований за всіма даними графік занадто перевантажений графічними елементами. Цей інструмент дозволяє ви-

брати кожне *n*-не значення з діапазону, що вказаний у полі *Вхідний інтервал*/*Imput range* (якщо обрати *Періодичний*/*Periodic* й вказати у полі *Період*/*Period* число *n*, напр., 5) або створити вибірку певного розміру за випадковим принципом (якщо обрати *Випадковий*/*Random* і вказати кількість елементів вибірки у *Число вибірок*/*Number of samples*). Якщо у першому рядку цього діапазону вказано назви стовпчиків (роки, назви показників тощо), то потрібно поставити позначку у полі *Мітки*/*Labels*. Наприклад, нижче вказано вікно інструменту (рис. А2).

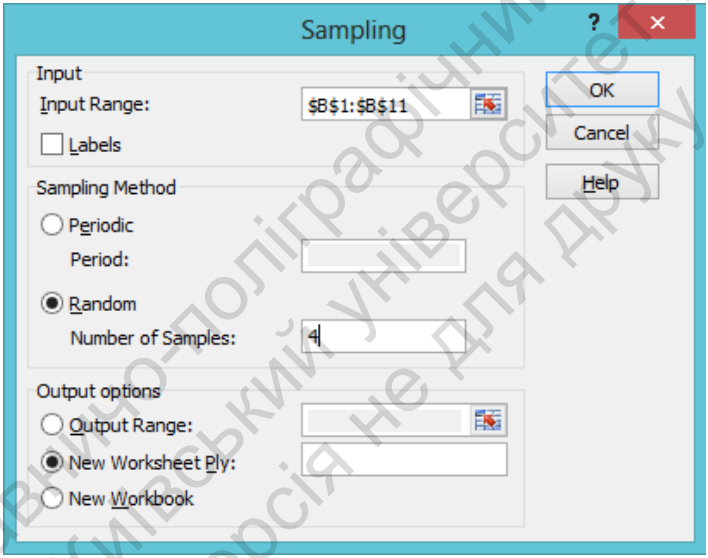


Рис. А2

Інструмент *Плинна* (*ковзна*) *середня*/*Moving average* використовують для розрахунку плинної середньої. Наприклад, зробимо це для даних експорту товарів і послуг України (дані умовні). У полі *Інтервал*/*Interval*, якщо вказати кількість періодів, за які розраховують плинну середню (напр., 3). Додатково можна розрахувати стандартні похибки та показати графік. На рис. А3 вказано вхідні дані та результат у таблиці, на рис. А4 – діалогове вікно аналізу, на рис. А5 – графік.

Подібне призначення має інструмент *Експоненційне згладжування/Exponential smoothing*.

	A	B	C	D
1	Експорт			
2	33		#N/A	#N/A
3	44		#N/A	#N/A
4	23		33.33333	#N/A
5	11		26	#N/A
6	55		29.66667	18.0144
7	33		33	16.99782
8	43		43.66667	14.63127
9	42		39.33333	1.586984
10	12		32.33333	11.84624
11	23		25.66667	11.93966

Рис. А3

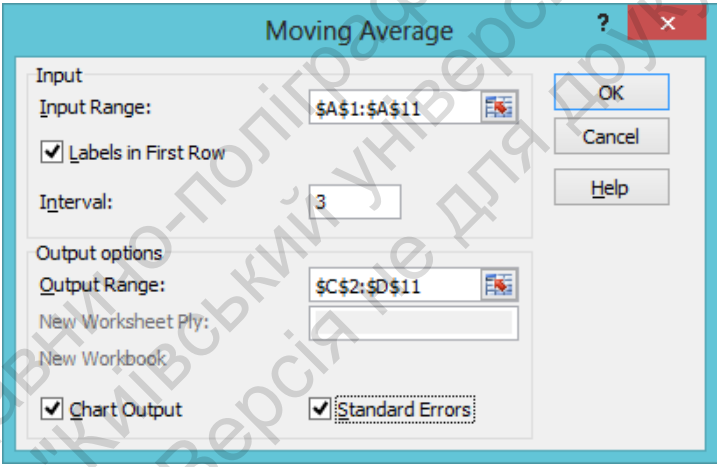


Рис. А4

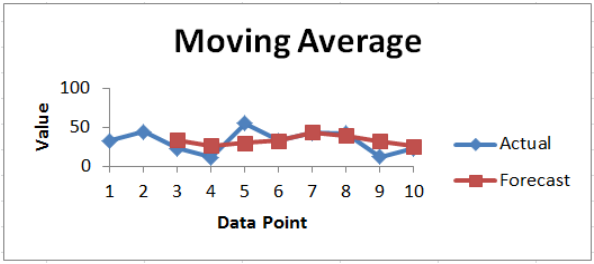


Рис. А5

А14. Діаграми

У програмі доступні такі види діаграм:

- *Стовпчасті*/*Column* діаграми.
- *Лінійні*/*Line графіки* (для визначення динаміки змінних з часом або по впорядкованих категоріях).
- *Діаграми з областями*/*Area* (мають схоже призначення).
- *Точкові* (*крапкові*) *діаграми* або *діаграми розсіювання*/*X Y Scatter*/*Scatterplots* (для визначення залежностей між змінними).
- *Біржові*/*Stock* діаграми.
- *Поверхневі*/*Surface* діаграми (тривимірні).
- *Кругові*/*Pie* діаграми (для відображення часток цілого).
- *Кільцеві діаграми*/*Doughnut* (мають схоже призначення, але дають можливість відобразити декілька рядів даних).
- *Бульбашкові*/*Bubble* діаграми (тривимірні, схожі на точкові діаграми, де третя змінна показує вагу за допомогою розміру бульбашки).
- *Пелюсткові*/*Radar* діаграми (багатовимірні).

Для створення діаграм потрібно:

- ввести вхідні дані (залежно від типу діаграми вони мають бути впорядковані певним чином);
- виділити дані;
- на вкладці *Вставлення*/*Insert* обрати *Діаграми*/*Charts*;
- обрати тип діаграми.

Далі за потреби елементи діаграми редагують.

До діаграм можна додавати лінії тренду та рівняння регресійної залежності.

Додаток В
Таблиці значень функцій

В1. Значення функції Гаусса $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

x	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	0,3989	0,3989	0,3988	0,3986	0,3984	0,3982	0,3980	0,3977	0,3973
0,1	0,3970	0,3965	0,3961	0,3956	0,3951	0,3945	0,3939	0,3932	0,3925	0,3918
0,2	0,3910	0,3902	0,3894	0,3885	0,3876	0,3867	0,3857	0,3847	0,3836	0,3825
0,3	0,3814	0,3802	0,3790	0,3778	0,3765	0,3752	0,3739	0,3725	0,3712	0,3697
0,4	0,3683	0,3668	0,3653	0,3637	0,3621	0,3605	0,3589	0,3572	0,3555	0,3538
0,5	0,3521	0,3503	0,3485	0,3467	0,3448	0,3429	0,3410	0,3391	0,3372	0,3352
0,6	0,3332	0,3312	0,3292	0,3271	0,3251	0,3230	0,3209	0,3187	0,3166	0,3144
0,7	0,3123	0,3101	0,3079	0,3056	0,3034	0,3011	0,2989	0,2966	0,2943	0,2920
0,8	0,2897	0,2874	0,2850	0,2827	0,2803	0,2780	0,2756	0,2732	0,2709	0,2685
0,9	0,2661	0,2637	0,2613	0,2589	0,2565	0,2541	0,2516	0,2492	0,2468	0,2444
1,0	0,2420	0,2396	0,2371	0,2347	0,2323	0,2299	0,2275	0,2251	0,2227	0,2203
1,1	0,2179	0,2155	0,2131	0,2107	0,2083	0,2059	0,2036	0,2012	0,1989	0,1965
1,2	0,1942	0,1919	0,1895	0,1872	0,1849	0,1826	0,1804	0,1781	0,1758	0,1736
1,3	0,1714	0,1691	0,1669	0,1647	0,1626	0,1604	0,1582	0,1561	0,1539	0,1518
1,4	0,1497	0,1476	0,1456	0,1435	0,1415	0,1394	0,1374	0,1354	0,1334	0,1315
1,5	0,1295	0,1276	0,1257	0,1238	0,1219	0,1200	0,1182	0,1163	0,1145	0,1127
1,6	0,1109	0,1092	0,1074	0,1057	0,1040	0,1023	0,1006	0,0989	0,0973	0,0957
1,7	0,0940	0,0925	0,0909	0,0893	0,0878	0,0863	0,0848	0,0833	0,0818	0,0804
1,8	0,0790	0,0775	0,0761	0,0748	0,0734	0,0721	0,0707	0,0694	0,0681	0,0669
1,9	0,0656	0,0644	0,0632	0,0620	0,0608	0,0596	0,0584	0,0573	0,0562	0,0551

Закінчення додатку В1

2,0	0,0540	0,0529	0,0519	0,0508	0,0498	0,0488	0,0478	0,0468	0,0459	0,0449
2,1	0,0440	0,0431	0,0422	0,0413	0,0404	0,0396	0,0387	0,0379	0,0371	0,0363
2,2	0,0355	0,0347	0,0339	0,0332	0,0325	0,0317	0,0310	0,0303	0,0297	0,0290
2,3	0,0283	0,0277	0,0270	0,0264	0,0258	0,0252	0,0246	0,0241	0,0235	0,0229
2,4	0,0224	0,0219	0,0213	0,0208	0,0203	0,0198	0,0194	0,0189	0,0184	0,0180
2,5	0,0175	0,0171	0,0167	0,0163	0,0158	0,0154	0,0151	0,0147	0,0143	0,0139
2,6	0,0136	0,0132	0,0129	0,0126	0,0122	0,0119	0,0116	0,0113	0,0110	0,0107
2,7	0,0104	0,0101	0,0099	0,0096	0,0093	0,0091	0,0088	0,0086	0,0084	0,0081
2,8	0,0079	0,0077	0,0075	0,0073	0,0071	0,0069	0,0067	0,0065	0,0063	0,0061
2,9	0,0060	0,0058	0,0056	0,0055	0,0053	0,0051	0,0050	0,0048	0,0047	0,0046
3,0	0,0044	0,0043	0,0042	0,0040	0,0039	0,0038	0,0037	0,0036	0,0035	0,0034
3,1	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026	0,0025	0,0025
3,2	0,0024	0,0023	0,0022	0,0022	0,0021	0,0020	0,0020	0,0019	0,0018	0,0018
3,3	0,0017	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014	0,0013	0,0013
3,4	0,0012	0,0012	0,0012	0,0011	0,0011	0,0010	0,0010	0,0010	0,0009	0,0009
3,5	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007	0,0007	0,0007	0,0006
3,6	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005	0,0005	0,0005	0,0005	0,0004
3,7	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003	0,0003	0,0003	0,0003
3,8	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002	0,0002	0,0002	0,0002	0,0002
3,9	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0001	0,0001

В2. Значення функції Лапласа $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt$

x	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767

2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

В3. Значення $P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$ (***розподіл Пуассона***)

$k\backslash\lambda$	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
0	0,90484	0,81873	0,74082	0,67032	0,60653	0,54881	0,49659	0,44933	0,40657
1	0,09048	0,16375	0,22225	0,26813	0,30327	0,32929	0,34761	0,35946	0,36591
2	0,00452	0,01637	0,03334	0,05363	0,07582	0,09879	0,12166	0,14379	0,16466
3	0,00015	0,00109	0,00333	0,00715	0,01264	0,01976	0,02839	0,03834	0,04940
4		0,00005	0,00025	0,00072	0,00158	0,00296	0,00497	0,00767	0,01111
5			0,00002	0,00006	0,00016	0,00036	0,00070	0,00123	0,00200
6					0,00001	0,00004	0,00008	0,00016	0,00030
7								0,00002	0,00004

$k\backslash\lambda$	1	2	3	4	5	6	7	8	9	10
0	0,36788	0,13534	0,04979	0,01832	0,00674	0,00248	0,00091	0,00034	0,00012	0,00005
1	0,36788	0,27067	0,14936	0,07326	0,03369	0,01487	0,00638	0,00268	0,00111	0,00045
2	0,18394	0,27067	0,22404	0,14653	0,08422	0,04462	0,02234	0,01073	0,00500	0,00227
3	0,06131	0,18045	0,22404	0,19537	0,14037	0,08924	0,05213	0,02863	0,01499	0,00757
4	0,01533	0,09022	0,16803	0,19537	0,17547	0,13385	0,09123	0,05725	0,03374	0,01892
5	0,00307	0,03609	0,10082	0,15629	0,17547	0,16062	0,12772	0,09160	0,06073	0,03783
6	0,00051	0,01203	0,05041	0,10420	0,14622	0,16062	0,14900	0,12214	0,09109	0,06306
7	0,00007	0,00344	0,02160	0,05954	0,10444	0,13768	0,14900	0,13959	0,11712	0,09008
8		0,00086	0,00810	0,02977	0,06528	0,10326	0,13038	0,13959	0,13176	0,11260
9		0,00019	0,00270	0,01323	0,03627	0,06884	0,10140	0,12408	0,13176	0,12511
10		0,00004	0,00081	0,00529	0,01813	0,04130	0,07098	0,09926	0,11858	0,12511
11			0,00022	0,00192	0,00824	0,02253	0,04517	0,07219	0,09702	0,11374
12			0,00006	0,00064	0,00343	0,01126	0,02635	0,04813	0,07277	0,09478

Закінчення додатку В3

13			0,00001	0,00020	0,00132	0,00520	0,01419	0,02962	0,05038	0,07291
14				0,00006	0,00047	0,00223	0,00709	0,01692	0,03238	0,05208
15				0,00002	0,00016	0,00089	0,00331	0,00903	0,01943	0,03472
16					0,00005	0,00033	0,00145	0,00451	0,01093	0,02170
17					0,00001	0,00012	0,00060	0,00212	0,00579	0,01276
18						0,00004	0,00023	0,00094	0,00289	0,00709
19						0,00001	0,00009	0,00040	0,00137	0,00373
20							0,00003	0,00016	0,00062	0,00187
21								0,00006	0,00026	0,00089
22								0,00002	0,00011	0,00040
23									0,00004	0,00018
24									0,00002	0,00007
25										0,00003
26										0,00001

В4. Значення функції $y = e^{-x}, x \geq 0$

x	0	1	2	3	4	5	6	7	8	9
0,0	1,00000	0,99005	0,98020	0,97045	0,96079	0,95123	0,94176	0,93239	0,92312	0,91393
0,1	0,90484	0,89583	0,88692	0,87810	0,86936	0,86071	0,85214	0,84366	0,83527	0,82696
0,2	0,81873	0,81058	0,80252	0,79453	0,78663	0,77880	0,77105	0,76338	0,75578	0,74826
0,3	0,74082	0,73345	0,72615	0,71892	0,71177	0,70469	0,69768	0,69073	0,68386	0,67706
0,4	0,67032	0,66365	0,65705	0,65051	0,64404	0,63763	0,63128	0,62500	0,61878	0,61263
0,5	0,60653	0,60050	0,59452	0,58860	0,58275	0,57695	0,57121	0,56553	0,55990	0,55433
0,6	0,54881	0,54335	0,53794	0,53259	0,52729	0,52205	0,51685	0,51171	0,50662	0,50158
0,7	0,49659	0,49164	0,48675	0,48191	0,47711	0,47237	0,46767	0,46301	0,45841	0,45384
0,8	0,44933	0,44486	0,44043	0,43605	0,43171	0,42741	0,42316	0,41895	0,41478	0,41066
0,9	0,40657	0,40252	0,39852	0,39455	0,39063	0,38674	0,38289	0,37908	0,37531	0,37158
1,0	0,36788	0,36422	0,36059	0,35701	0,35345	0,34994	0,34646	0,34301	0,33960	0,33622
1,1	0,33287	0,32956	0,32628	0,32303	0,31982	0,31664	0,31349	0,31037	0,30728	0,30422
1,2	0,30119	0,29820	0,29523	0,29229	0,28938	0,28650	0,28365	0,28083	0,27804	0,27527
1,3	0,27253	0,26982	0,26714	0,26448	0,26185	0,25924	0,25666	0,25411	0,25158	0,24908
1,4	0,24660	0,24414	0,24171	0,23931	0,23693	0,23457	0,23224	0,22993	0,22764	0,22537
1,5	0,22313	0,22091	0,21871	0,21654	0,21438	0,21225	0,21014	0,20805	0,20598	0,20393
1,6	0,20190	0,19989	0,19790	0,19593	0,19398	0,19205	0,19014	0,18825	0,18637	0,18452
1,7	0,18268	0,18087	0,17907	0,17728	0,17552	0,17377	0,17204	0,17033	0,16864	0,16696
1,8	0,16530	0,16365	0,16203	0,16041	0,15882	0,15724	0,15567	0,15412	0,15259	0,15107
1,9	0,14957	0,14808	0,14661	0,14515	0,14370	0,14227	0,14086	0,13946	0,13807	0,13670
2,0	0,13534	0,13399	0,13266	0,13134	0,13003	0,12873	0,12745	0,12619	0,12493	0,12369
2,1	0,12246	0,12124	0,12003	0,11884	0,11765	0,11648	0,11533	0,11418	0,11304	0,11192
2,2	0,11080	0,10970	0,10861	0,10753	0,10646	0,10540	0,10435	0,10331	0,10228	0,10127
2,3	0,10026	0,09926	0,09827	0,09730	0,09633	0,09537	0,09442	0,09348	0,09255	0,09163

2,4	0,09072	0,08982	0,08892	0,08804	0,08716	0,08629	0,08543	0,08458	0,08374	0,08291
2,5	0,08208	0,08127	0,08046	0,07966	0,07887	0,07808	0,07730	0,07654	0,07577	0,07502
2,6	0,07427	0,07353	0,07280	0,07208	0,07136	0,07065	0,06995	0,06925	0,06856	0,06788
2,7	0,06721	0,06654	0,06587	0,06522	0,06457	0,06393	0,06329	0,06266	0,06204	0,06142
2,8	0,06081	0,06020	0,05961	0,05901	0,05843	0,05784	0,05727	0,05670	0,05613	0,05558
2,9	0,05502	0,05448	0,05393	0,05340	0,05287	0,05234	0,05182	0,05130	0,05079	0,05029
3,0	0,04979	0,04929	0,04880	0,04832	0,04783	0,04736	0,04689	0,04642	0,04596	0,04550
3,1	0,04505	0,04460	0,04416	0,04372	0,04328	0,04285	0,04243	0,04200	0,04159	0,04117
3,2	0,04076	0,04036	0,03996	0,03956	0,03916	0,03877	0,03839	0,03801	0,03763	0,03725
3,3	0,03688	0,03652	0,03615	0,03579	0,03544	0,03508	0,03474	0,03439	0,03405	0,03371
3,4	0,03337	0,03304	0,03271	0,03239	0,03206	0,03175	0,03143	0,03112	0,03081	0,03050
3,5	0,03020	0,02990	0,02960	0,02930	0,02901	0,02872	0,02844	0,02816	0,02788	0,02760
3,6	0,02732	0,02705	0,02678	0,02652	0,02625	0,02599	0,02573	0,02548	0,02522	0,02497
3,7	0,02472	0,02448	0,02423	0,02399	0,02375	0,02352	0,02328	0,02305	0,02282	0,02260
3,8	0,02237	0,02215	0,02193	0,02171	0,02149	0,02128	0,02107	0,02086	0,02065	0,02045
3,9	0,02024	0,02004	0,01984	0,01964	0,01945	0,01925	0,01906	0,01887	0,01869	0,01850
4,0	0,01832	0,01813	0,01795	0,01777	0,01760	0,01742	0,01725	0,01708	0,01691	0,01674
4,1	0,01657	0,01641	0,01624	0,01608	0,01592	0,01576	0,01561	0,01545	0,01530	0,01515
4,2	0,01500	0,01485	0,01470	0,01455	0,01441	0,01426	0,01412	0,01398	0,01384	0,01370
4,3	0,01357	0,01343	0,01330	0,01317	0,01304	0,01291	0,01278	0,01265	0,01253	0,01240
4,4	0,01228	0,01216	0,01203	0,01191	0,01180	0,01168	0,01156	0,01145	0,01133	0,01122
4,5	0,01111	0,01100	0,01089	0,01078	0,01067	0,01057	0,01046	0,01036	0,01025	0,01015
4,6	0,01005	0,00995	0,00985	0,00975	0,00966	0,00956	0,00947	0,00937	0,00928	0,00919
4,7	0,00910	0,00900	0,00892	0,00883	0,00874	0,00865	0,00857	0,00848	0,00840	0,00831
4,8	0,00823	0,00815	0,00807	0,00799	0,00791	0,00783	0,00775	0,00767	0,00760	0,00752

4,9	0,00745	0,00737	0,00730	0,00723	0,00715	0,00708	0,00701	0,00694	0,00687	0,00681
5,0	0,00674	0,00667	0,00660	0,00654	0,00647	0,00641	0,00635	0,00628	0,00622	0,00616
5,1	0,00610	0,00604	0,00598	0,00592	0,00586	0,00580	0,00574	0,00568	0,00563	0,00557
5,2	0,00552	0,00546	0,00541	0,00535	0,00530	0,00525	0,00520	0,00514	0,00509	0,00504
5,3	0,00499	0,00494	0,00489	0,00484	0,00480	0,00475	0,00470	0,00465	0,00461	0,00456
5,4	0,00452	0,00447	0,00443	0,00438	0,00434	0,00430	0,00425	0,00421	0,00417	0,00413
5,5	0,00409	0,00405	0,00401	0,00397	0,00393	0,00389	0,00385	0,00381	0,00377	0,00374
5,6	0,00370	0,00366	0,00362	0,00359	0,00355	0,00352	0,00348	0,00345	0,00341	0,00338
5,7	0,00335	0,00331	0,00328	0,00325	0,00321	0,00318	0,00315	0,00312	0,00309	0,00306
5,8	0,00303	0,00300	0,00297	0,00294	0,00291	0,00288	0,00285	0,00282	0,00279	0,00277
5,9	0,00274	0,00271	0,00269	0,00266	0,00263	0,00261	0,00258	0,00255	0,00253	0,00250
6,0	0,00248	0,00245	0,00243	0,00241	0,00238	0,00236	0,00233	0,00231	0,00229	0,00227
6,1	0,00224	0,00222	0,00220	0,00218	0,00215	0,00213	0,00211	0,00209	0,00207	0,00205
6,2	0,00203	0,00201	0,00199	0,00197	0,00195	0,00193	0,00191	0,00189	0,00187	0,00185
6,3	0,00184	0,00182	0,00180	0,00178	0,00176	0,00175	0,00173	0,00171	0,00170	0,00168
6,4	0,00166	0,00165	0,00163	0,00161	0,00160	0,00158	0,00156	0,00155	0,00153	0,00152
6,5	0,00150	0,00149	0,00147	0,00146	0,00144	0,00143	0,00142	0,00140	0,00139	0,00137
6,6	0,00136	0,00135	0,00133	0,00132	0,00131	0,00129	0,00128	0,00127	0,00126	0,00124
6,7	0,00123	0,00122	0,00121	0,00119	0,00118	0,00117	0,00116	0,00115	0,00114	0,00112
6,8	0,00111	0,00110	0,00109	0,00108	0,00107	0,00106	0,00105	0,00104	0,00103	0,00102
6,9	0,00101	0,00100	0,00099	0,00098	0,00097	0,00096	0,00095	0,00094	0,00093	0,00092
7,0	0,00091	0,00090	0,00089	0,00088	0,00088	0,00087	0,00086	0,00085	0,00084	0,00083
7,1	0,00083	0,00082	0,00081	0,00080	0,00079	0,00078	0,00078	0,00077	0,00076	0,00075
7,2	0,00075	0,00074	0,00073	0,00072	0,00072	0,00071	0,00070	0,00070	0,00069	0,00068
7,3	0,00068	0,00067	0,00066	0,00066	0,00065	0,00064	0,00064	0,00063	0,00062	0,00062
7,4	0,00061	0,00061	0,00060	0,00059	0,00059	0,00058	0,00058	0,00057	0,00056	0,00056

Закінчення додатку В4

7,5	0,00055	0,00055	0,00054	0,00054	0,00053	0,00053	0,00052	0,00052	0,00051	0,00051
7,6	0,00050	0,00050	0,00049	0,00049	0,00048	0,00048	0,00047	0,00047	0,00046	0,00046
7,7	0,00045	0,00045	0,00044	0,00044	0,00044	0,00043	0,00043	0,00042	0,00042	0,00041
7,8	0,00041	0,00041	0,00040	0,00040	0,00039	0,00039	0,00039	0,00038	0,00038	0,00037
7,9	0,00037	0,00037	0,00036	0,00036	0,00036	0,00035	0,00035	0,00035	0,00034	0,00034
8,0	0,00034	0,00033	0,00033	0,00033	0,00032	0,00032	0,00032	0,00031	0,00031	0,00031
8,1	0,00030	0,00030	0,00030	0,00029	0,00029	0,00029	0,00029	0,00028	0,00028	0,00028
8,2	0,00027	0,00027	0,00027	0,00027	0,00026	0,00026	0,00026	0,00026	0,00025	0,00025
8,3	0,00025	0,00025	0,00024	0,00024	0,00024	0,00024	0,00023	0,00023	0,00023	0,00023
8,4	0,00022	0,00022	0,00022	0,00022	0,00022	0,00021	0,00021	0,00021	0,00021	0,00021
8,5	0,00020	0,00020	0,00020	0,00020	0,00020	0,00019	0,00019	0,00019	0,00019	0,00019
8,6	0,00018	0,00018	0,00018	0,00018	0,00018	0,00018	0,00017	0,00017	0,00017	0,00017
8,7	0,00017	0,00016	0,00016	0,00016	0,00016	0,00016	0,00016	0,00016	0,00015	0,00015
8,8	0,00015	0,00015	0,00015	0,00015	0,00014	0,00014	0,00014	0,00014	0,00014	0,00014
8,9	0,00014	0,00014	0,00013	0,00013	0,00013	0,00013	0,00013	0,00013	0,00013	0,00012
9,0	0,00012	0,00012	0,00012	0,00012	0,00012	0,00012	0,00012	0,00012	0,00011	0,00011
9,1	0,00011	0,00011	0,00011	0,00011	0,00011	0,00011	0,00011	0,00010	0,00010	0,00010
9,2	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00010	0,00009	0,00009	0,00009
9,3	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00008	0,00008
9,4	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008	0,00008
9,5	0,00007	0,00007	0,00007	0,00007	0,00007	0,00007	0,00007	0,00007	0,00007	0,00007
9,6	0,00007	0,00007	0,00007	0,00007	0,00007	0,00006	0,00006	0,00006	0,00006	0,00006
9,7	0,00006	0,00006	0,00006	0,00006	0,00006	0,00006	0,00006	0,00006	0,00006	0,00006
9,8	0,00006	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005
9,9	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005	0,00005

В5. Значення χ^2_α, для яких $P(\chi^2 > \chi^2_\alpha) = \alpha$ залежить від кількості степенів вільності k та ймовірності α (*розподіл* χ^2)

	Ймовірність α									
k	**0,99**	**0,975**	**0,95**	**0,5**	**0,25**	**0,2**	**0,1**	**0,05**	**0,02**	**0,001**
1	0,0002	0,0010	0,0039	0,4549	1,3233	1,6424	2,7055	3,8415	5,4119	10,8276
2	0,0201	0,0506	0,1026	1,3863	2,7726	3,2189	4,6052	5,9915	7,8240	13,8155
3	0,1148	0,2158	0,3518	2,3660	4,1083	4,6416	6,2514	7,8147	9,8374	16,2662
4	0,2971	0,4844	0,7107	3,3567	5,3853	5,9886	7,7794	9,4877	11,6678	18,4668
5	0,5543	0,8312	1,1455	4,3515	6,6257	7,2893	9,2364	11,0705	13,3882	20,5150
6	0,8721	1,2373	1,6354	5,3481	7,8408	8,5581	10,6446	12,5916	15,0332	22,4577
7	1,2390	1,6899	2,1673	6,3458	9,0371	9,8032	12,0170	14,0671	16,6224	24,3219
8	1,6465	2,1797	2,7326	7,3441	10,2189	11,0301	13,3616	15,5073	18,1682	26,1245
9	2,0879	2,7004	3,3251	8,3428	11,3888	12,2421	14,6837	16,9190	19,6790	27,8772
10	2,5582	3,2470	3,9403	9,3418	12,5489	13,4420	15,9872	18,3070	21,1608	29,5883
11	3,0535	3,8157	4,5748	10,3410	13,7007	14,6314	17,2750	19,6751	22,6179	31,2641
12	3,5706	4,4038	5,2260	11,3403	14,8454	15,8120	18,5493	21,0261	24,0540	32,9095
13	4,1069	5,0088	5,8919	12,3398	15,9839	16,9848	19,8119	22,3620	25,4715	34,5282
14	4,6604	5,6287	6,5706	13,3393	17,1169	18,1508	21,0641	23,6848	26,8728	36,1233
15	5,2293	6,2621	7,2609	14,3389	18,2451	19,3107	22,3071	24,9958	28,2595	37,6973
16	5,8122	6,9077	7,9616	15,3385	19,3689	20,4651	23,5418	26,2962	29,6332	39,2524
17	6,4078	7,5642	8,6718	16,3382	20,4887	21,6146	24,7690	27,5871	30,9950	40,7902
18	7,0149	8,2307	9,3905	17,3379	21,6049	22,7595	25,9894	28,8693	32,3462	42,3124
19	7,6327	8,9065	10,1170	18,3377	22,7178	23,9004	27,2036	30,1435	33,6874	43,8202
20	8,2604	9,5908	10,8508	19,3374	23,8277	25,0375	28,4120	31,4104	35,0196	45,3147

Закінчення додатку В5

21	8,8972	10,2829	11,5913	20,3372	24,9348	26,1711	29,6151	32,6706	36,3434	46,7970
22	9,5425	10,9823	12,3380	21,3370	26,0393	27,3015	30,8133	33,9244	37,6595	48,2679
23	10,1957	11,6886	13,0905	22,3369	27,1413	28,4288	32,0069	35,1725	38,9683	49,7282
24	10,8564	12,4012	13,8484	23,3367	28,2412	29,5533	33,1962	36,4150	40,2704	51,1786
25	11,5240	13,1197	14,6114	24,3366	29,3389	30,6752	34,3816	37,6525	41,5661	52,6197
26	12,1981	13,8439	15,3792	25,3365	30,4346	31,7946	35,5632	38,8851	42,8558	54,0520
27	12,8785	14,5734	16,1514	26,3363	31,5284	32,9117	36,7412	40,1133	44,1400	55,4760
28	13,5647	15,3079	16,9279	27,3362	32,6205	34,0266	37,9159	41,3371	45,4188	56,8923
29	14,2565	16,0471	17,7084	28,3361	33,7109	35,1394	39,0875	42,5570	46,6927	58,3012
30	14,9535	16,7908	18,4927	29,3360	34,7997	36,2502	40,2560	43,7730	47,9618	59,7031
31	15,6555	17,5387	19,2806	30,3359	35,8871	37,3591	41,4217	44,9853	49,2264	61,0983
32	16,3622	18,2908	20,0719	31,3359	36,9730	38,4663	42,5847	46,1943	50,4867	62,4872
33	17,0735	19,0467	20,8665	32,3358	38,0575	39,5718	43,7452	47,3999	51,7429	63,8701
34	17,7891	19,8063	21,6643	33,3357	39,1408	40,6756	44,9032	48,6024	52,9952	65,2472
35	18,5089	20,5694	22,4650	34,3356	40,2228	41,7780	46,0588	49,8018	54,2438	66,6188
36	19,2327	21,3359	23,2686	35,3356	41,3036	42,8788	47,2122	50,9985	55,4889	67,9852
37	19,9602	22,1056	24,0749	36,3355	42,3833	43,9782	48,3634	52,1923	56,7305	69,3465
38	20,6914	22,8785	24,8839	37,3355	43,4619	45,0763	49,5126	53,3835	57,9688	70,7029
39	21,4262	23,6543	25,6954	38,3354	44,5395	46,1730	50,6598	54,5722	59,2040	72,0547
40	22,1643	24,4330	26,5093	39,3353	45,6160	47,2685	51,8051	55,7585	60,4361	73,4020

В6. Значення функції розподілу Стьюдента для кількості ступенів вільності від 1 до 20

	Степені вільності, k									
x	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000
0,1	0,5317	0,5353	0,5367	0,5374	0,5379	0,5382	0,5384	0,5386	0,5387	0,5388
0,2	0,5628	0,5700	0,5729	0,5744	0,5753	0,5760	0,5764	0,5768	0,5770	0,5773
0,3	0,5928	0,6038	0,6081	0,6104	0,6119	0,6129	0,6136	0,6141	0,6145	0,6148
0,4	0,6211	0,6361	0,6420	0,6452	0,6472	0,6485	0,6495	0,6502	0,6508	0,6512
0,5	0,6476	0,6667	0,6743	0,6783	0,6809	0,6826	0,6838	0,6847	0,6855	0,6861
0,6	0,6720	0,6953	0,7046	0,7096	0,7127	0,7148	0,7163	0,7174	0,7183	0,7191
0,7	0,6944	0,7218	0,7328	0,7387	0,7424	0,7449	0,7467	0,7481	0,7492	0,7501
0,8	0,7148	0,7462	0,7589	0,7657	0,7700	0,7729	0,7750	0,7766	0,7778	0,7788
0,9	0,7333	0,7684	0,7828	0,7905	0,7953	0,7986	0,8010	0,8028	0,8042	0,8054
1	0,7500	0,7887	0,8045	0,8130	0,8184	0,8220	0,8247	0,8267	0,8283	0,8296
1,1	0,7651	0,8070	0,8242	0,8335	0,8393	0,8433	0,8461	0,8483	0,8501	0,8514
1,2	0,7789	0,8235	0,8419	0,8518	0,8581	0,8623	0,8654	0,8678	0,8696	0,8711
1,3	0,7913	0,8384	0,8578	0,8683	0,8748	0,8793	0,8826	0,8851	0,8870	0,8886
1,4	0,8026	0,8518	0,8720	0,8829	0,8898	0,8945	0,8979	0,9005	0,9025	0,9041
1,5	0,8128	0,8638	0,8847	0,8960	0,9030	0,9079	0,9114	0,9140	0,9161	0,9177
1,6	0,8222	0,8746	0,8960	0,9076	0,9148	0,9196	0,9232	0,9259	0,9280	0,9297
1,7	0,8307	0,8844	0,9062	0,9178	0,9251	0,9300	0,9335	0,9362	0,9383	0,9400
1,8	0,8386	0,8932	0,9152	0,9269	0,9341	0,9390	0,9426	0,9452	0,9473	0,9490
1,9	0,8458	0,9011	0,9232	0,9349	0,9421	0,9469	0,9504	0,9530	0,9551	0,9567

2	0,8524	0,9082	0,9303	0,9419	0,9490	0,9538	0,9572	0,9597	0,9617	0,9633
2,1	0,8585	0,9147	0,9367	0,9482	0,9551	0,9598	0,9631	0,9655	0,9674	0,9690
2,2	0,8642	0,9206	0,9424	0,9537	0,9605	0,9649	0,9681	0,9705	0,9723	0,9738
2,3	0,8695	0,9259	0,9475	0,9585	0,9651	0,9694	0,9725	0,9748	0,9765	0,9779
2,4	0,8743	0,9308	0,9521	0,9628	0,9692	0,9734	0,9763	0,9784	0,9801	0,9813
2,5	0,8789	0,9352	0,9561	0,9666	0,9728	0,9767	0,9795	0,9815	0,9831	0,9843
2,6	0,8831	0,9392	0,9598	0,9700	0,9759	0,9797	0,9823	0,9842	0,9856	0,9868
2,7	0,8871	0,9429	0,9631	0,9730	0,9786	0,9822	0,9847	0,9865	0,9878	0,9888
2,8	0,8908	0,9463	0,9661	0,9756	0,9810	0,9844	0,9867	0,9884	0,9896	0,9906
2,9	0,8943	0,9494	0,9687	0,9779	0,9831	0,9863	0,9885	0,9901	0,9912	0,9921
3	0,8976	0,9523	0,9712	0,9800	0,9850	0,9880	0,9900	0,9915	0,9925	0,9933
3,1	0,9007	0,9549	0,9734	0,9819	0,9866	0,9894	0,9913	0,9927	0,9936	0,9944
3,2	0,9036	0,9573	0,9753	0,9835	0,9880	0,9907	0,9925	0,9937	0,9946	0,9953
3,3	0,9063	0,9596	0,9771	0,9850	0,9893	0,9918	0,9934	0,9946	0,9954	0,9960
3,4	0,9089	0,9617	0,9788	0,9864	0,9904	0,9928	0,9943	0,9953	0,9961	0,9966
3,5	0,9114	0,9636	0,9803	0,9876	0,9914	0,9936	0,9950	0,9960	0,9966	0,9971
3,6	0,9138	0,9654	0,9816	0,9886	0,9922	0,9943	0,9956	0,9965	0,9971	0,9976
3,7	0,9160	0,9670	0,9829	0,9896	0,9930	0,9950	0,9962	0,9970	0,9975	0,9979
3,8	0,9181	0,9686	0,9840	0,9904	0,9937	0,9955	0,9966	0,9974	0,9979	0,9983
3,9	0,9201	0,9701	0,9850	0,9912	0,9943	0,9960	0,9971	0,9977	0,9982	0,9985
4,0	0,9220	0,9714	0,9860	0,9919	0,9948	0,9964	0,9974	0,9980	0,9984	0,9987
4,1	0,9239	0,9727	0,9869	0,9926	0,9953	0,9968	0,9977	0,9983	0,9987	0,9989
4,2	0,9256	0,9739	0,9877	0,9932	0,9958	0,9972	0,9980	0,9985	0,9988	0,9991
4,3	0,9273	0,9750	0,9884	0,9937	0,9961	0,9975	0,9982	0,9987	0,9990	0,9992
4,4	0,9289	0,9760	0,9891	0,9942	0,9965	0,9977	0,9984	0,9989	0,9991	0,9993
4,5	0,9304	0,9770	0,9898	0,9946	0,9968	0,9979	0,9986	0,9990	0,9993	0,9994

Продовження додатку В6

x	11	12	13	14	15	16	17	18	19	20
0	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000
0,1	0,5389	0,5390	0,5391	0,5391	0,5392	0,5392	0,5392	0,5393	0,5393	0,5393
0,2	0,5774	0,5776	0,5777	0,5778	0,5779	0,5780	0,5781	0,5781	0,5782	0,5782
0,3	0,6151	0,6153	0,6155	0,6157	0,6159	0,6160	0,6161	0,6162	0,6163	0,6164
0,4	0,6516	0,6519	0,6522	0,6524	0,6526	0,6528	0,6529	0,6531	0,6532	0,6533
0,5	0,6865	0,6869	0,6873	0,6876	0,6878	0,6881	0,6883	0,6884	0,6886	0,6887
0,6	0,7197	0,7202	0,7206	0,7210	0,7213	0,7215	0,7218	0,7220	0,7222	0,7224
0,7	0,7508	0,7514	0,7519	0,7523	0,7527	0,7530	0,7533	0,7536	0,7538	0,7540
0,8	0,7797	0,7804	0,7810	0,7815	0,7819	0,7823	0,7826	0,7829	0,7832	0,7834
0,9	0,8063	0,8071	0,8078	0,8083	0,8088	0,8093	0,8097	0,8100	0,8103	0,8106
1	0,8306	0,8315	0,8322	0,8329	0,8334	0,8339	0,8343	0,8347	0,8351	0,8354
1,1	0,8526	0,8535	0,8544	0,8551	0,8557	0,8562	0,8567	0,8571	0,8575	0,8578
1,2	0,8723	0,8734	0,8742	0,8750	0,8756	0,8762	0,8767	0,8772	0,8776	0,8779
1,3	0,8899	0,8910	0,8919	0,8927	0,8934	0,8940	0,8945	0,8950	0,8954	0,8958
1,4	0,9055	0,9066	0,9075	0,9084	0,9091	0,9097	0,9103	0,9107	0,9112	0,9116
1,5	0,9191	0,9203	0,9212	0,9221	0,9228	0,9235	0,9240	0,9245	0,9250	0,9254
1,6	0,9310	0,9322	0,9332	0,9340	0,9348	0,9354	0,9360	0,9365	0,9370	0,9374
1,7	0,9414	0,9426	0,9435	0,9444	0,9451	0,9458	0,9463	0,9468	0,9473	0,9477
1,8	0,9503	0,9515	0,9525	0,9533	0,9540	0,9546	0,9552	0,9557	0,9561	0,9565
1,9	0,9580	0,9591	0,9601	0,9609	0,9616	0,9622	0,9627	0,9632	0,9636	0,9640
2	0,9646	0,9657	0,9666	0,9674	0,9680	0,9686	0,9691	0,9696	0,9700	0,9704
2,1	0,9702	0,9712	0,9721	0,9728	0,9735	0,9740	0,9745	0,9750	0,9753	0,9757
2,2	0,9750	0,9759	0,9768	0,9774	0,9781	0,9786	0,9790	0,9794	0,9798	0,9801
2,3	0,9790	0,9799	0,9807	0,9813	0,9819	0,9824	0,9828	0,9832	0,9835	0,9838
2,4	0,9824	0,9832	0,9840	0,9846	0,9851	0,9855	0,9859	0,9863	0,9866	0,9869

Закінчення додатку В6

2,5	0,9852	0,9860	0,9867	0,9873	0,9877	0,9882	0,9885	0,9888	0,9891	0,9894
2,6	0,9877	0,9884	0,9890	0,9895	0,9900	0,9903	0,9907	0,9910	0,9912	0,9914
2,7	0,9897	0,9903	0,9909	0,9914	0,9918	0,9921	0,9924	0,9927	0,9929	0,9931
2,8	0,9914	0,9920	0,9925	0,9929	0,9933	0,9936	0,9938	0,9941	0,9943	0,9945
2,9	0,9928	0,9933	0,9938	0,9942	0,9945	0,9948	0,9950	0,9952	0,9954	0,9956
3	0,9940	0,9945	0,9949	0,9952	0,9955	0,9958	0,9960	0,9962	0,9963	0,9965
3,1	0,9949	0,9954	0,9958	0,9961	0,9963	0,9966	0,9967	0,9969	0,9971	0,9972
3,2	0,9958	0,9962	0,9965	0,9968	0,9970	0,9972	0,9974	0,9975	0,9976	0,9978
3,3	0,9965	0,9968	0,9971	0,9974	0,9976	0,9977	0,9979	0,9980	0,9981	0,9982
3,4	0,9970	0,9974	0,9976	0,9978	0,9980	0,9982	0,9983	0,9984	0,9985	0,9986
3,5	0,9975	0,9978	0,9980	0,9982	0,9984	0,9985	0,9986	0,9987	0,9988	0,9989
3,6	0,9979	0,9982	0,9984	0,9986	0,9987	0,9988	0,9989	0,9990	0,9990	0,9991
3,7	0,9982	0,9985	0,9987	0,9988	0,9989	0,9990	0,9991	0,9992	0,9992	0,9993
3,8	0,9985	0,9987	0,9989	0,9990	0,9991	0,9992	0,9993	0,9993	0,9994	0,9994
3,9	0,9988	0,9989	0,9991	0,9992	0,9993	0,9994	0,9994	0,9995	0,9995	0,9996
4,0	0,9990	0,9991	0,9992	0,9993	0,9994	0,9995	0,9995	0,9996	0,9996	0,9996
4,1	0,9991	0,9993	0,9994	0,9995	0,9995	0,9996	0,9996	0,9997	0,9997	0,9997
4,2	0,9993	0,9994	0,9995	0,9996	0,9996	0,9997	0,9997	0,9997	0,9998	0,9998
4,3	0,9994	0,9995	0,9996	0,9996	0,9997	0,9997	0,9998	0,9998	0,9998	0,9998
4,4	0,9995	0,9996	0,9996	0,9997	0,9997	0,9998	0,9998	0,9998	0,9998	0,9999
4,5	0,9995	0,9996	0,9997	0,9998	0,9998	0,9998	0,9998	0,9999	0,9999	0,9999

В7. Значення f_α, для яких $P(F_{k_1,k_2} > f_\alpha) = \alpha$ залежить від кількості ступенів вільності k_1 і k_2 для ймовірностей $\alpha = 0{,}05$ та $\alpha = 0{,}01$ (*розподіл Фішера*)

0,05	k_2									
k_1	1	2	3	4	5	6	7	8	9	10
1	161,4476	18,5128	10,1280	7,7086	6,6079	5,9874	5,5914	5,3177	5,1174	4,9646
2	199,5000	19,0000	9,5521	6,9443	5,7861	5,1433	4,7374	4,4590	4,2565	4,1028
3	215,7073	19,1643	9,2766	6,5914	5,4095	4,7571	4,3468	4,0662	3,8625	3,7083
4	224,5832	19,2468	9,1172	6,3882	5,1922	4,5337	4,1203	3,8379	3,6331	3,4780
5	230,1619	19,2964	9,0135	6,2561	5,0503	4,3874	3,9715	3,6875	3,4817	3,3258
6	233,9860	19,3295	8,9406	6,1631	4,9503	4,2839	3,8660	3,5806	3,3738	3,2172
7	236,7684	19,3532	8,8867	6,0942	4,8759	4,2067	3,7870	3,5005	3,2927	3,1355
8	238,8827	19,3710	8,8452	6,0410	4,8183	4,1468	3,7257	3,4381	3,2296	3,0717
9	240,5433	19,3848	8,8123	5,9988	4,7725	4,0990	3,6767	3,3881	3,1789	3,0204
10	241,8817	19,3959	8,7855	5,9644	4,7351	4,0600	3,6365	3,3472	3,1373	2,9782
11	242,9835	19,4050	8,7633	5,9358	4,7040	4,0274	3,6030	3,3130	3,1025	2,9430
12	243,9060	19,4125	8,7446	5,9117	4,6777	3,9999	3,5747	3,2839	3,0729	2,9130
13	244,6898	19,4189	8,7287	5,8911	4,6552	3,9764	3,5503	3,2590	3,0475	2,8872
14	245,3640	19,4244	8,7149	5,8733	4,6358	3,9559	3,5292	3,2374	3,0255	2,8647
15	245,9499	19,4291	8,7029	5,8578	4,6188	3,9381	3,5107	3,2184	3,0061	2,8450
16	246,4639	19,4333	8,6923	5,8441	4,6038	3,9223	3,4944	3,2016	2,9890	2,8276
17	246,9184	19,4370	8,6829	5,8320	4,5904	3,9083	3,4799	3,1867	2,9737	2,8120
18	247,3232	19,4402	8,6745	5,8211	4,5785	3,8957	3,4669	3,1733	2,9600	2,7980
19	247,6861	19,4431	8,6670	5,8114	4,5678	3,8844	3,4551	3,1613	2,9477	2,7854
20	248,0131	19,4458	8,6602	5,8025	4,5581	3,8742	3,4445	3,1503	2,9365	2,7740

21	248,3094	19,4481	8,6540	5,7945	4,5493	3,8649	3,4349	3,1404	2,9263	2,7636
22	248,5791	19,4503	8,6484	5,7872	4,5413	3,8564	3,4260	3,1313	2,9169	2,7541
23	248,8256	19,4523	8,6432	5,7805	4,5339	3,8486	3,4179	3,1229	2,9084	2,7453
24	249,0518	19,4541	8,6385	5,7744	4,5272	3,8415	3,4105	3,1152	2,9005	2,7372
25	249,2601	19,4558	8,6341	5,7687	4,5209	3,8348	3,4036	3,1081	2,8932	2,7298
26	249,4525	19,4573	8,6301	5,7635	4,5151	3,8287	3,3972	3,1015	2,8864	2,7229
27	249,6309	19,4587	8,6263	5,7586	4,5097	3,8230	3,3913	3,0954	2,8801	2,7164
28	249,7966	19,4600	8,6229	5,7541	4,5047	3,8177	3,3858	3,0897	2,8743	2,7104
29	249,9510	19,4613	8,6196	5,7498	4,5001	3,8128	3,3806	3,0844	2,8688	2,7048
30	250,0951	19,4624	8,6166	5,7459	4,4957	3,8082	3,3758	3,0794	2,8637	2,6996
31	250,2301	19,4635	8,6137	5,7422	4,4916	3,8038	3,3713	3,0747	2,8588	2,6946
32	250,3567	19,4645	8,6111	5,7387	4,4878	3,7998	3,3670	3,0703	2,8543	2,6900
33	250,4757	19,4654	8,6085	5,7354	4,4842	3,7959	3,3630	3,0662	2,8500	2,6856
34	250,5878	19,4663	8,6062	5,7323	4,4808	3,7923	3,3592	3,0623	2,8460	2,6815
35	250,6934	19,4672	8,6039	5,7294	4,4775	3,7889	3,3557	3,0586	2,8422	2,6776
36	250,7933	19,4680	8,6018	5,7267	4,4745	3,7856	3,3523	3,0551	2,8386	2,6739
37	250,8878	19,4687	8,5998	5,7241	4,4716	3,7826	3,3491	3,0518	2,8352	2,6704
38	250,9774	19,4694	8,5979	5,7216	4,4689	3,7797	3,3461	3,0486	2,8320	2,6670
39	251,0624	19,4701	8,5961	5,7192	4,4663	3,7769	3,3432	3,0456	2,8289	2,6639
40	251,1432	19,4707	8,5944	5,7170	4,4638	3,7743	3,3404	3,0428	2,8259	2,6609

0,01	k_2									
k_1	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	4052,1807	98,5025	34,1162	21,1977	16,2582	13,7450	12,2464	11,2586	10,5614	10,0443
2	4999,5000	99,0000	30,8165	18,0000	13,2739	10,9248	9,5466	8,6491	8,0215	7,5594
3	5403,3520	99,1662	29,4567	16,6944	12,0600	9,7795	8,4513	7,5910	6,9919	6,5523
4	5624,5833	99,2494	28,7099	15,9770	11,3919	9,1483	7,8466	7,0061	6,4221	5,9943

Продовження додатку В7

5	5763,6496	99,2993	28,2371	15,5219	10,9670	8,7459	7,4604	6,6318	6,0569	5,6363
6	5858,9861	99,3326	27,9107	15,2069	10,6723	8,4661	7,1914	6,3707	5,8018	5,3858
7	5928,3557	99,3564	27,6717	14,9758	10,4555	8,2600	6,9928	6,1776	5,6129	5,2001
8	5981,0703	99,3742	27,4892	14,7989	10,2893	8,1017	6,8400	6,0289	5,4671	5,0567
9	6022,4732	99,3881	27,3452	14,6591	10,1578	7,9761	6,7188	5,9106	5,3511	4,9424
10	6055,8467	99,3992	27,2287	14,5459	10,0510	7,8741	6,6201	5,8143	5,2565	4,8491
11	6083,3168	99,4083	27,1326	14,4523	9,9626	7,7896	6,5382	5,7343	5,1779	4,7715
12	6106,3207	99,4159	27,0518	14,3736	9,8883	7,7183	6,4691	5,6667	5,1114	4,7059
13	6125,8647	99,4223	26,9831	14,3065	9,8248	7,6575	6,4100	5,6089	5,0545	4,6496
14	6142,6740	99,4278	26,9238	14,2486	9,7700	7,6049	6,3590	5,5589	5,0052	4,6008
15	6157,2846	99,4325	26,8722	14,1982	9,7222	7,5590	6,3143	5,5151	4,9621	4,5581
16	6170,1012	99,4367	26,8269	14,1539	9,6802	7,5186	6,2750	5,4766	4,9240	4,5204
17	6181,4348	99,4404	26,7867	14,1146	9,6429	7,4827	6,2401	5,4423	4,8902	4,4869
18	6191,5287	99,4436	26,7509	14,0795	9,6096	7,4507	6,2089	5,4116	4,8599	4,4569
19	6200,5756	99,4465	26,7188	14,0480	9,5797	7,4219	6,1808	5,3840	4,8327	4,4299
20	6208,7302	99,4492	26,6898	14,0196	9,5526	7,3958	6,1554	5,3591	4,8080	4,4054
21	6216,1184	99,4516	26,6635	13,9938	9,5281	7,3722	6,1324	5,3364	4,7856	4,3831
22	6222,8433	99,4537	26,6396	13,9703	9,5058	7,3506	6,1113	5,3157	4,7651	4,3628
23	6228,9903	99,4557	26,6176	13,9488	9,4853	7,3309	6,0921	5,2967	4,7463	4,3441
24	6234,6309	99,4575	26,5975	13,9291	9,4665	7,3127	6,0743	5,2793	4,7290	4,3269
25	6239,8251	99,4592	26,5790	13,9109	9,4491	7,2960	6,0580	5,2631	4,7130	4,3111
26	6244,6239	99,4607	26,5618	13,8940	9,4331	7,2805	6,0428	5,2482	4,6982	4,2963
27	6249,0708	99,4621	26,5460	13,8784	9,4182	7,2661	6,0287	5,2344	4,6845	4,2827
28	6253,2031	99,4635	26,5312	13,8639	9,4043	7,2527	6,0157	5,2214	4,6717	4,2700

Закінчення додатку В7

29	6257,0530	99,4647	26,5174	13,8503	9,3914	7,2402	6,0034	5,2094	4,6598	4,2581
30	6260,6486	99,4658	26,5045	13,8377	9,3793	7,2285	5,9920	5,1981	4,6486	4,2469
31	6264,0142	99,4669	26,4925	13,8258	9,3680	7,2176	5,9813	5,1876	4,6381	4,2365
32	6267,1711	99,4679	26,4812	13,8147	9,3574	7,2073	5,9712	5,1776	4,6282	4,2267
33	6270,1383	99,4689	26,4705	13,8042	9,3474	7,1976	5,9618	5,1683	4,6190	4,2174
34	6272,9323	99,4698	26,4605	13,7943	9,3380	7,1885	5,9528	5,1595	4,6102	4,2087
35	6275,5679	99,4706	26,4511	13,7850	9,3291	7,1799	5,9444	5,1512	4,6020	4,2005
36	6278,0581	99,4714	26,4421	13,7762	9,3207	7,1718	5,9364	5,1433	4,5941	4,1927
37	6280,4147	99,4721	26,4337	13,7679	9,3127	7,1641	5,9289	5,1358	4,5867	4,1853
38	6282,6481	99,4728	26,4257	13,7600	9,3052	7,1568	5,9217	5,1287	4,5797	4,1783
39	6284,7677	99,4735	26,4180	13,7525	9,2980	7,1498	5,9149	5,1220	4,5730	4,1716
40	6286,7821	99,4742	26,4108	13,7454	9,2912	7,1432	5,9084	5,1156	4,5666	4,1653

ЛІТЕРАТУРА

1. Бахрушин В.Є. Методи аналізу даних : навч. посіб. / В.Є. Бахрушин. – Запоріжжя : КПУ, 2011.– 268 с.
2. Вахненко Т. Моделювання макроекономічних факторів зовнішніх запозичень та їх впливу на розвиток економіки України / Т. Вахненко // Економіка України. – 2007. – №7. – С. 15-24.
3. Вдовиченко А. Визначення детермінантів заощаджень та споживання населення України на основі емпіричного дослідження / А. Вдовиченко // Економіка України. – 2009. – №9. – С. 40-52.
4. Грисенко М.В. Кластеризація країн Європейського Союзу за детермінантами соціалізації їх економічного розвитку та місце України в даній моделі. Science progress in European countries: new concepts and modern solutions. Hosted by the ORT Publishing the Centre for Scientifically Research "Solution" / М.В. Грисенко, О.А. Приятельчук. – 2019. – V. 8. – P. 97-107.
5. Грисенко М.В. Математична статистика для економістів-міжнародників : навч. посіб. / М.В. Грисенко, А.Ю. Рижов. – К. : ВПЦ "Київський університет". – 2012. – 212 с.
6. Грисенко М.В. Кількісні методи аналізу міжнародних економічних відносин : навч. посіб. / М.В. Грисенко, А.А. Чугаєв. – К. : Ін-т міжнар. відносин Київський національного університету імені Тараса Шевченка, 2012. – 235 с.
7. Грисенко М.В. Економіко-математичне моделювання світогосподарських процесів : навч. посіб. / М.В. Грисенко, Л.О. Шворак. – К. : ВПЦ "Київський університет", 2016. – Ч. І. Теоретичні основи. – 271 с.
8. Грисенко М.В. Економіко-математичне моделювання світогосподарських процесів : навч. посіб. / М.В. Грисенко, Л.О. Шворак. Частина ІІ. Прикладні моделі. – К. : ВПЦ "Київський університет", 2016. – Ч. 2. – 223 с.; Ч
9. Грисенко М.В. Економіко-математичне моделювання світогосподарських процесів : навч. посіб. / М.В. Грисенко, Л.О. Шворак, А.Ю. Рижов. – К. : ВПЦ "Київський університет", 2016. – Ч. ІІІ. Практикум. – 229 с.
10. Економічні санкції у сучасному світовому господарстві : моногр. / за ред. О.І. Шниркова. – К. : ВПЦ "Київський університет", 2018. – 239 с.

11. Клименко О. Моделювання та дослідження динаміки еміграції трудових ресурсів / О. Клименко, Т. Зубка // Банківська справа. – 2009. – №2. – С. 67-70.
12. Лондар С.Л. Економетрія засобами MS EXCEL : навч. посіб. / С.Л. Лондар, Р.В. Юринець. – К. : Вид-во Європейського університету, 2004. – 238 с.
13. Макроекономічне моделювання та короткострокове прогнозування / за ред. І.В. Крючкової. – Харків : Форт, 2000. – 336 с.
14. Майборода Р.Є. Комп'ютерна статистика : підруч. / Р.Є. Майборода. – К. : ВПЦ "Київський університет", 2019. – 589 с.
15. Світова економіка : підруч. / за ред. О.І. Шниркова, В.І. Мазуренко, О.І. Рогача. – К. : ВПЦ "Київський університет", 2018. – 616 с.
16. Скрипниченко М.І. Секторальні та міжкраїнні моделі економічного розвитку / М.І. Скрипниченко. – К. : Фенікс, 2004. – 256 с.
17. Точилін В.О. Прикладна економіко-математична модель "Тімес-Україна" для оптимізації енергетичних потоків та прогнозування енергетичного балансу України / В.О. Точилін, Р.З. Подолець, О.А. Дячук, Ю.А. Олександренко // Наука та інновації. – 2010. – № 2. – С. 48-66.
18. Тронь В.П. Нечітка стратегія чітких рішень / В.П. Тронь. – К. : Національна академія державного управління при Президентові України, Українська академія наук з державного управління, 2007. – 748 с.
19. Хмара М.П. Оцінка потенційних наслідків торговельної інтеграції на прикладі формування зони вільної торгівлі між Україною та Туреччиною / М.П. Хмара, О.А. Чугаєв // Зони вільної торгівлі на початку XXI століття : моногр. / А.С. Філіпенко, В.С. Будкін, О.І. Шнирков. – К. : ВПЦ "Київський університет", 2013.
20. Черняк О.І. Застосування байєсівських мереж в економіці / О.І. Черняк, Л.В. Кучерук // Вісн. Харків. ун-ту імені В. Н. Каразіна. Економічна серія. – 2009. – № 869. – С. 199-209.
21. Чугаєв О.А. Валютні кризи на межі XX–XXI століть : моногр. / О.А. Чугаєв. – К. : "МП Леся", 2007. – 416 с.
https://www.sites.google.com/site/achugaiev/stati/stati-1?authuser=0)
22. Чугаєв О.А. Економічна сила країни у глобальному господарстві. дис. ... д-ра екон. наук : 08.00.02. Київ. ун-т ім. Тараса Шевченка, 2018. – 638 с.
http://scc.univ.kiev.ua/upload/iblock/4a3/dis_Chugaiev O. A.pdf
23. Чугаєв О.А. Глобальні виміри економічної сили країни : моногр. / О.А. Чугаєв. – К. : ВПЦ "Київський університет", 2017. – 495 с.

24. Чугаєв О.А. Синхронність економічних циклів як складова економічної сили / О.Чугаєв // Глобалізаційні виклики розвитку національних економік : матер. міжнар. наук.-практ. конф. 19.10.2016. – Київ : нац. торг.-екон. ун-т, 2016. – Ч.1 – С. 545-555.
25. Belhocine N. Assessing Fiscal Stress / N. Belhocine, G. Dobrescu, S. Mazraani, I. Petrova // IMF Working Paper, August 2010. – 34 p.
26. Baldacci E. Assessing Fiscal Stress / E. Baldacci, I. Petrova, N. Belhocine, G. Dobrescu, S. Mazraani // IMF Working Paper WP/11/100, May 2011. – 41 p.
(http://www.imf.org/external/pubs/ft/wp/2011/ wp11100.pdf).
27. Driver R., Westaway P. Concepts of equilibrium exchange rates / R. Driver, P. Westaway // Working Paper no. 248. – Bank of England, 2004. – 64 p.
28. Grcic B. The Pollak's Macroeconomic Monetary Model / B. Grcic. – Faculty of Economics Split. – 12 p.
29. Hair J., Anderson R., Babin B., Black W. Multivariate Data Analysis / J. Hair, R. Anderson, B. Babin, W. Black // 8th edition. Cengage Learning EMEA, 2018.
30. Hrysenko M. (2020). Modelling the factors influencing migration processes in the European Union / M. Hrysenko, O. Pryiatelchuk // Economic Annals-XXI. – 2020. – 183(5-6), 26-42.
https://doi.org/10.21003/ea.V183-03
31. Hrysenko M. Modeling of state socioeconomic systems in the countries of the European region / M. Hrysenko, O. Pryiatelchuk, L. Shvorak // Problems and Perspectives in Management. – 2019 17 (3), 452-463.
https://dx.doi.org/10.21511/ppm.17(3).2019.36
32. Lune H., Berg B. L. Qualitative Research Methods for the Social Sciences. – Pearson, Ninth edition/global edition, 2017.
33. Kaminsky G. Leading Indicators of Currency Crises / G. Kaminsky, S. Lizondo, C. Reinhart // Working Paper. – 1997. – № 97/79. – Wash. D.C. : IMF. – 43 p.
34. Laxton D. MULTIMOD Mark III The Core Dynamic and Steady-State Models / D. Laxton, P. Isard, H. Faruqee, E. Prasad, B.Turtelboom // Occasional paper. – 1998. – № 164. – Washington, DC : International Monetary Fund. – 73 p.
35. Levine D. Applied Statistics for Engineers and Scientists: Using Microsoft Excel & Minitab / D. Levine, P. Ramsey, R. Smidt // Upper Saddle River. – 2001. – New Jersey: Prentice Hall Inc.

36. Shnyrkov O. Resilience of the EU Exports to Ukraine under the COVID-19 pandemic / O. Shnyrkov, O. Chugaiev // EURINT, 2020, V. 7. – P. 80-100.
(https://eurint.uaic.ro/proceedings/articles/EURINT2020_SHN.pdf)
37. Shnyrkov O. The Impact of Institutions on Services Exports of Central and Eastern European Countries / O. Shnyrkov, R. Zablotska, O. Chugaiev // Baltic Journal of Economic Studies. – 2019. – V. 5. – № 5. – P. 9-17.
http://www.baltijapublishing.lv/index.php/issue/article/view/731
38. The IMF-FSB Early Warning Excersise. Design and methodological Toolkit. – IMF, September 2010. – 41 p.
39. Steven J. Taylor, Robert Bogdan, Marjorie DeVault Introduction to Qualitative Research Methods: A Guidebook and Resource / J. Taylor Steven, Bogdan Robert, DeVault Marjorie. – New Jersey : Wiley & Sons, 2016.
40. TIBCO Software Inc. Data Science Textbook, 2020.
https://docs.tibco.com/data-science/textbook/
41. Trochim, William M.K. The Research Methods Knowledge Base, 2020 https://conjointly.com/kb
42. Roland W. Scholz, Olaf Tietje. Embedded Case Study Methods: Integrating Quantitative and Qualitative Knowledge / Roland W. Scholz, Olaf Tietje. – London : Sage Publications, 2002.
43. Using Excel for Statistical Analysis
http://www.scribd.com/doc/49234740/Stats-Using-Excel1
44. Web Pages that Perform Statistical Calculations
https://statpages.info/

ЗМІСТ