**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

ЖИТОМИРСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ФРАНКА

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА

ПОЛІСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

# ОСНОВИ СТАТИСТИЧНОГО МОДЕЛЮВАННЯ

Житомир – "Рута" – 2022

**УДК 311:330.101.52**
**О-73**

*Рекомендовано вченою радою Житомирського державного університету імені Івана Франка протокол №4 від 25 березня 2022 р.*

**Рецензенти:**

*Денель Г.* - доктор хаб., професор Економічного університету у Познані, завідувачка кафедри статистики Економічного університету у Познані, м. Познань, Польща.

*Огірко І. В.* - д. ф.-м. н., професор, завідувач кафедри інформаційних та мультимедійних технологій Української академії друкарства, м. Львів.

*Горобець С. М.* - к. пед. н., доцент кафедри комп'ютерних наук та інформаційних технологій Житомирського державного університету імені Івана Франка, м. Житомир.

**О-73** Основи статистичного моделювання: навч. посібник / за загальною редакцією С.В. Чугаєвської, Н.В. Ковтун. Житомир: Видавництво ПП "Рута", 2022. 604 с.

**ISBN 978-617-581-536-6**

У навчальному посібнику висвітлені питання обробки та моделювання статистичної інформації стосовно масових явищ суспільного життя. Прикладний напрям посібника проілюстровано великою кількістю задач та прикладів їх практичного застосування для оцінки, аналізу і прогнозування суспільно-економічних явищ. Для студентів, аналітиків даних, викладачів, здобувачів, науковців, фахівців інформаційно-консультаційних послуг, котрі займаються прикладними статистичними дослідженнями.

**УДК 311:330.101.52**

**ISBN 978-617-581-536-6**

# ПЕРЕДМОВА

Сучасні напрями розвитку національної та світової економіки потребують використання аналітичного інструментарію щодо збирання, класифікації інформації про масові суспільні та природні явища, проведення обробки, детальної оцінки та планування ситуацій в економічній, фінансовій, аудиторській, бухгалтерській, маркетинговій та управлінській сферах діяльності. Теорія статистики виступає фундаментальною основою такого інструментарію, а її методи статистичного моделювання забезпечують якісну оцінку тенденцій розвитку суспільних явищ та процесів, кількісного та якісного вимірювання причинно-наслідкових зв'язків, вирішення задач планування та прогнозування відповідно до видів діяльності.

Сучасні знання з основ статистичного моделювання є запорукою формування важливих компетентностей фахівців як сфери прикладної математики, ІТ – технологій, так і економічних досліджень з обробки баз даних, спеціалістів з бізнес-аналітики, здатних: розробити план та програму аналітичної роботи на підприємстві; застосувати сукупність прийомів і способів дослідження; визначити систему синтетичних та аналітичних показників для характеристики об'єкту аналізу; здійснити перевірку достовірності, повноти і вірогідності звітних даних; виявити взаємодіючі фактори і обчислити їх вплив на зміну величини показників; виявити невикористані можливості (перспективні резерви) підвищення ефективності діяльності; провести підсумкову оцінку проведених досліджень і узагальнити результати аналізу, відповідно до його цілей і завдань.

Зростання потреби обробки аналітичної інформації великих масивів зумовлено формуванням у користувачів навичок використання сучасних комп'ютерних статистичних програм, котрі дають змогу оцінити ефективність управління підприємствами, значно розширюють і поглиблюють можливості вирішення завдання аналізу їх діяльності, оцінити тенденції розвитку регіонів у цілому, змоделювати можливі сценарії вирішення викликів суспільства. Відбувається перегляд деяких традиційних підходів статистичного моделювання до його методики з урахуванням сучасних реформацій. Статистичне моделювання набуває властивостей зв'язуючої ланки для застосування його основних методів при вирішенні практичних задач фахівцями різних сфер діяльності: від інженерії до медицини, економіки, військової справи та хіміко-біологічних досліджень, стає більш тісно пов'язаним з функціями системної аналітики, маркетингу, менеджменту.

Видатний український філософ та громадський діяч Михайло Грушевський ще у далекому 19 столітті написав: «Покоління, яких жде Україна тепер, повинні бути людьми діла реального, практичного – спеціалісти-адміністратори, фінансисти, економісти, знавці військового і морського діла, техніки…». Саме методам вирішення таких різноманітних, але, безумовно, важливих практичних задач присвячено даний навчальний посібник з Основ статистичного моделювання.

Запропонований навчальний посібник, підготовлений на сучасному рівні, його прикладний напрям проілюстровано великою кількістю задач та прикладів їх практичного застосування для оцінки, аналізу і прогнозування суспільно-економічних явищ. Він стане корисним для студентів, аналітиків даних, викладачів, здобувачів, науковців, фахівців інформаційно-консультаційних послуг тощо.

Навчальний посібник підготовлено колективом авторів-фахівців з багаторічним досвідом у галузі прикладних статистичних

досліджень трьох університетів: Житомирського державного університету імені Івана Франка, Київського національного університету імені Тараса Шевченка, Поліського національного університету:

- Розділ 1. Методологічні засади статистичного моделювання – к. пед. наук, доц. Королюк О.М.;
- Розділ 2. Моделі статистичних класифікацій – к. екон. н., доц. Трубнік Т.Є.;
- Розділ 3. Моделювання рядів розподілу – к. екон. н., доцент Шибіріна С.О., д. екон. наук, проф. Ковтун Н.В.;
- Розділ 4. Методи і моделі багатовимірного регресійного аналізу – к. фіз.- мат. наук, доц. Щехорський А.Й., к. екон. наук, доц. Чугаєвська С.В.;
- Розділ 5. Моделювання рядів динаміки - к. екон. наук, доц. Чугаєвська С.В., к. фіз.- мат. наук, доц. Щехорський А.Й.;
- Розділ 6. Використання логіт та пробіт регресійних моделей в у статистичному аналізі бінарної класифікації – д. екон. наук, проф. Ковтун Н.В., аспірант Шамайда Д.М.;
- Розділ 7. Нелінійні моделі у статистичному аналізі – к. техн. наук, доц. Бродський Ю.Б.;
- Розділ 8. Моделювання індексних систем – д. екон. наук, проф. Ковтун Н.В., к. екон. наук, проф. Мазуренко О.К., к. екон. наук, доц. Горна М.О;
- Розділ 9. Статистичне моделювання випадкових процесів – д. фіз.-мат. наук, проф. Погоруй А.О.;
- Розділ 10. Задачі оптимізації у статистичному моделюванні – д. екон. наук, проф. Николюк О.М., к. техн. наук, доц. Бродський Ю.Б.;
- Розділ 11. Статистичні методи моделювання ризиків – к. фіз.-мат. наук, доц. Франовський А.Ц., к. фіз.-мат. наук, доц. Щехорський А.Й., к. екон. наук, доц. Чугаєвська С.В.;

- Розділ 12. Метод головних компонент – к. екон. н., доц. Поплюйко Я.В.;
- Розділ 13. Моделі відновлення статистичних даних - д. екон. наук, проф. Ковтун Н.В., аспірант Фаталієва А.Я.

Автори сподіваються, що розроблений навчальний посібник допоможе користувачам оволодіти методикою статистичного моделювання і застосовувати отримані знання у їх практичній діяльності. Авторський колектив висловлює вдячність рецензентам за влучні зауваження та рекомендації.

Відгуки, зауваження та пропозиції щодо поліпшення змісту й структури посібника просимо надсилати на адресу: matem.analiz@gmail.com .

# Розділ 1. МЕТОДОЛОГІЧНІ ЗАСАДИ СТАТИСТИЧНОГО МОДЕЛЮВАННЯ

## 1.1. Методологія та методи наукового дослідження

Сьогодення потребує фахівців, які володіють дослідницькими методами, методиками, що спрямовані на вирішення практичних завдань у різних галузях економіки, техніки й технологій та інших сферах людського життя. Важливою умовою успішного здійснення дослідницької діяльності є володіння основами наукової методології.

Термін *методологія* походить від грецьких *methodos* – спосіб, метод і *logos* – наука, знання. Виходячи з цього, *методологію* логічно визначити як певну сукупність засобів та прийомів пізнання і вчення про них. Проте формулювання науковцями поняття «методологія» дещо відрізняються. Зокрема, *методологія* формулюється як:

– вчення про правила мислення при створенні науки, проведенні наукових досліджень[1];

– вчення про систему наукових принципів і способів дослідницької діяльності[2];

– сукупність принципів, засобів, методів і форм організації та проведення наукового пізнання поставленої проблеми[3];

– теорія методів дослідження, створення концепцій; система знань про теорію науки або про систему методів дослідження[4].

---

[1] Рассоха І. М. Конспект лекцій з навчальної дисципліни «Методологія та організація наукових досліджень». Х. : ХНАМГ, 2011. 76 с.

[2] Краус Н. М. Методологія та організація наукових досліджень : навч.-метод. посіб. Полтава : Оріяна, 2012. 183 с.

[3] Юринець В. Є. Методологія наукових досліджень : навч. посіб. Львів : ЛНУ ім. І. Франка, 2011. 178 с.

[4] Шейко В. М., Кушнаренко Н. М. Організація та методика науково-дослідницької діяльності : підруч. К. : Знання-Прес, 2002. 295 с.

Н. М. Краус підкреслює, що методологія наукового дослідження розглядає найбільш суттєві особливості і ознаки методів дослідження, розкриває їх за спільністю і глибиною аналізу.

Узагальнивши існуючи підходи, Г. С. Цехмістрова виділяє дві складові поняття: 1) сукупність прийомів дослідження, що застосовуються в певній науці; 2) вчення про методи пізнання та перетворення дійсності. Це дозволило сформулювати: *методологія* – це концептуальний виклад мети, змісту, методів дослідження, які забезпечують отримання максимально об'єктивної, точної, систематизованої інформації про процеси та явища[5].

Виділяють основні *функції методології* в науці:

- визначати способи здобуття наукових знань, які відображають динаміку процесів та явищ;

- передбачати особливий шлях, за допомогою якого може бути досягнуто науково-дослідної мети;

- забезпечувати всебічність отримання інформації щодо процесу чи явища, що вивчається;

- допомагати введенню нової інформації;

- забезпечувати уточнення, збагачення, систематизацію термінів і понять у науці;

- створювати систему наукової інформації, яка базується на об'єктивних явищах, і логіко-аналітичний інструмент наукового пізнання[6].

Нагадаємо**,** *принцип* – це керівна ідея, головне правило, найперша вимога, яка випливає із закономірностей, встановлених наукою.

---

[5] Цехмістрова Г.С. Основи наукових досліджень : навч. посіб. К. : Вид. дім «Слово», 2004. 240 с.

[6] Цехмістрова Г.С. Основи наукових досліджень : навч. посіб. К. : Вид. дім «Слово», 2004. С. 78-79.

Методологія наукових досліджень базується на *принципах*: єдності теорії і практики; системності; розвитку; об'єктивності; декомпозиції; абстрагування[7].

*Принцип єдності теорії і практики* реалізується: по-перше, у формі впровадження в практику фундаментальних наукових концепцій, розроблених у ході академічних наукових досліджень; по-друге, на шляху наукової думки від практичних завдань до розроблення спеціальних конкретно-наукових концепцій.

*Принцип системності* передбачає взаємне узгодження всіх напрямків вивчення об'єкту, а також усунення протиріч між ними. У результаті такого підходу створюється система організації дослідження, у межах якої усі його складові частини діють узгоджено і спрямовані на ефективне функціонування всієї системи.

*Принцип розвитку*. В усіх явищах природи та суспільства, у духовному житті людини постійно виникає щось нове. Перебіг багатьох процесів відбувається від простого до складного, за висхідною лінією (прогресивно). Водночас, існують і зворотні процеси, коли події відбуваються за низхідною лінією (регресивні).

*Принцип об'єктивності* стверджує: у питаннях науки жодна думка не є вирішальною. Це стосується й думки авторитетних учених, громадської думки, думки державних інстанцій, керівництва тощо. Також і дослідник не має права нічого додавати від себе ні на етапі спостереження за явищем, ні в процесі формулювання висновків.

*Принцип декомпозиції* базується на використанні структури задачі та її розділенні на простіші (елементарні) частини (підзадачі), які пізніше, у певний спосіб, знову поєднуються в єдине ціле.

Завдяки *принципу абстрагування* формується ідеальний образ реальності. Наукова абстракція підпорядкована певним вимогам: по-

[7] Юринець В. Є. Методологія наукових досліджень : навч. посіб. Львів : ЛНУ ім. І. Франка, 2011. 178 с.

перше, треба знати, від чого ми абстрагуємося; по-друге, визначити, до якої межі можна коректно абстрагуватися; по-третє, потрібно враховувати, що інтервал абстрагування для створення ідеалу залежить лише від об'єктивних умов.

Кожне наукове дослідження ґрунтується на сукупності пізнавальних засобів, методів, прийомів, які вибудовано в чіткій послідовності. Зазвичай, під *методологічною основою* дослідження розуміють певні основні, вихідні, базові положення, з позиції яких надається пояснення основних наукових явищ і розкриваються їх закономірності, які частіше існують поза наукою, в межах якої проводяться вишукування, і не випливають із проведених досліджень. Від чіткого визначення методологічної основи значною мірою залежить, чи досягне дослідник мети своїх наукових пошуків.

Знання методології в науковому дослідженні дозволяє впорядкувати отримані результати, оцінити їх практичну значущість, а також встановити актуальні напрямки подальших досліджень або ж окреслити перспективи пошуку альтернативних шляхів вирішення сформульованої проблеми.

*Філософська (фундаментальна) методологія* – це вищий рівень методології науки, що визначає загальну стратегію принципів пізнання особливостей явищ, процесів, сфер діяльності. Вона виконує два типи функцій. По-перше, виявляє сенс наукової діяльності та її взаємозв'язків з іншими сферами діяльності, тобто розглядає науку відносно практики, суспільства, культури людини. По-друге, філософська методологія вирішує завдання вдосконалення, оптимізації наукової діяльності, видаючись за межі філософії, хоча спирається на розроблені нею світоглядні й загальнометодологічні орієнтири та постулати.

Фундаментальна методологія базується на *принципах діалектики* (відбивають найбільш суттєві властивості об'єктивної дійсності й свідомості з урахуванням досвіду, набутого в процесі пізнавальної діяльності людини), *детермінізму* (об'єктивної

причинної зумовленості явищ) та *ізоморфізму* (взаємовідношень об'єктів, що відбивають тотожність їх побудови тощо) дослідження[8]. Від тлумачення філософських принципів залежить обґрунтування методологічного підходу у дослідженні в конкретній галузі.

Вагому роль у науковому пізнанні відіграють *закони діалектики*, головними з яких є:

• *закон єдності та боротьби протилежностей*, що відображає джерело розвитку – протиріччя як взаємозв'язок і взаємообумовленість протилежностей;

• *закон взаємного переходу кількісних змін у якісні*, що розкриває механізм розвитку як поступове нагромадження кількісних змін, яке в певний момент обумовлює докорінні якісні перетворення, виникнення нової властивості, що відповідно, здійснює зворотний вплив на характер і темпи кількісних змін;

• *закон заперечення заперечень*, який виражає поступальний, послідовний характер розвитку явищ і процесів реального світу, показує, що поступальний розвиток має форму висхідної спіралі; це процес, який начебто повторює пройдене, але на щабель вище.

*Загальнонаукова методологія* використовується у переважній більшості наук, оскільки будь-яке наукове відкриття має не лише предметний, але й методологічний зміст, спричиняє критичний перегляд понятійного апарату, чинників, передумов і підходів до інтерпретації досліджуваного матеріалу.

*Принципи*, що належать до загальнонаукової методології, – це принципи історизму, системності, функціональності, термінологічний, пізнавальний (когнітивний), моделювання, кібернетичний та ін.

[8] Шейко В. М., Кушнаренко Н. М. Організація та методика науково-дослідницької діяльності : підруч. К. : Знання-Прес, 2002. 295 с.

*Принцип історизму* дає можливість вивчати будь-яке явище з погляду того, як воно колись виникло, які визначальні етапи мало у своєму розвитку, чим стало у певний час і чим буде в майбутньому.

*Термінологічний принцип* передбачає вивчення історії термінів і дефініцій, які вони позначають, розробку або уточнення змісту та обсягу понять, встановлення взаємозв'язку і підпорядкування понять, їх місця в категорійному апараті теорії, на якій ґрунтується дослідження.

*Принцип системності* дає змогу визначити стратегію наукового дослідження. Будь-який предмет повинен розглядатися як упорядкована єдність відносно самостійних частин (підсистем, елементів), кожна з яких виконує певні функції в існуванні предмета.

*Принцип функціональності* – елементи структури системи взаємнопов'язані, її діяльність обумовлюється призначенням, отже, створювати та досліджувати систему необхідно після визначення її функцій. У разі визначення нових функцій системи доцільно змінювати її структуру, а не намагатися «прилаштувати» цю функцію до попередньої структури.

*Принцип моделювання* полягає в тому, що вивчення об'єкта замінюється вивченням аналога – моделі. Реальний процес у цьому випадку відображується, частіше, у вигляді математичних і логічних схем.

Завдяки *кібернетичному принципу* формується уявлення про систему як тріаду «вхід-процес-вихід» і відбувається вивчення того, як певний об'єкт обробляє інформацію, реагує на неї та змінюється під її впливом. Особлива увага при цьому приділяється дослідженню зворотного зв'язку між об'єктом і суб'єктом, який аналізується з урахуванням змін оточення, що можуть мати місце після здійснених впливів.

*Пізнавальний,* або *когнітивний, принцип* особливо ефективний у вивченні динаміки науки та її співвідношення з суспільством, в обґрунтуванні провідного значення знання в поведінці індивідуума.

Слід враховувати, що для аналізу формування знання необхідне вивчення практичної і теоретичної діяльності людини у співвідношенні з її соціальним аспектом. У центрі досліджуваних проблем знаходиться людина як член соціуму, комунікант.

*Конкретнонаукова (частковонаукова) методологія* – сукупність ідей або специфічних методів певної науки, які є основою для вирішення конкретної дослідницької проблеми; іншими словами, це наукові концепції, на які спирається дослідник[9].

*Концепція* (від лат. *conceptio* – розуміння) – система поглядів, те або інше розуміння явищ і процесів; єдиний, визначальний задум. *Концепція дослідження* – система вихідних теоретичних положень, яка є основою дослідницького пошуку. У процесі дослідження прийняті початкові положення перевіряють, розвивають, коригують, а за необхідності – відкидають (відбувається зміна або модернізація концепції).

Важливо розрізняти значення понять «методологія», «метод», «методика».

*Метод* – це спосіб досягнення мети, розв'язання конкретної задачі; сукупність прийомів (операцій) практичного впливу чи теоретичного освоєння об'єктивної реальності з метою її пізнання[10]. По-іншому, *метод* – це інструмент, за допомогою якого вирішують ті чи інші проблеми, встановлюють закономірні зв'язки певних явищ і процесів.

*Функція методу* полягає в отриманні нової інформації про навколишню дійсність, встановленні сутності явищ і процесів, розкритті законів і закономірностей розвитку, формування і функціонування об'єктів, які досліджуються.

---

[9] Шейко В. М., Кушнаренко Н. М. Організація та методика науково-дослідницької діяльності : підруч. К. : Знання-Прес, 2002. 295 с.

[10] Рассоха І. М. Конспект лекцій з навчальної дисципліни «Методологія та організація наукових досліджень». Х. : ХНАМГ, 2011. 76 с.

Методи є упорядкованою системою, в якій визначається їх місце відповідно до конкретного етапу дослідження, використання технічних прийомів і проведення операцій із теоретичним і фактичним матеріалом у заданій послідовності.

*Методика* дослідження – це система правил використання методів, прийомів для проведення будь-якого дослідження. Іншими словами, у методиці дослідження на основі загальнометодичних принципів визначаються конкретні методи, процедури, які обумовлені закономірностями розвитку досліджуваного явища, предмета. В одній науковій галузі може бути декілька методик (комплексів методів), які постійно вдосконалюються під час наукових пошуків.

Основне призначення методики дослідження полягає в тому, щоб на основі відповідних принципів (вимог, умов, обмежень, приписів тощо) забезпечити успішне вирішення визначених завдань, практичних проблем і досягнення мети наукового дослідження. Потрібно зауважити, що обираючи методику важливо виходити із конкретних завдань дослідження та специфіки об'єктів, явищ і процесів, які вивчаються.

Особливістю наукових досліджень є їхня багатоваріантність, тобто можливість вирішення поставлених завдань різними методами й шляхами. До того ж, часто вони не рівноцінні як по величині витрат, так і за часом, який потрібний для досягнення поставленої мети.

У методології наукових досліджень виділяють два *рівні пізнання*:

• теоретичний – висунення і розвиток наукових гіпотез і теорій, формулювання законів та виведення з них логічних наслідків, зіставлення різних гіпотез і теорій;

• емпіричний – спостереження і дослідження конкретних явищ, експеримент, а також групування, класифікація та опис результатів дослідження[11].

Кожному дослідженню обов'язково передує постановка *проблеми*, що означає:

• визначення того, що є невідомим, що потребує доведення;

• формулювання питання, що відображає основний зміст проблеми, та обґрунтування його важливості для науки;

• виділення окремих завдань, встановлення послідовності їх вирішення та обрання методів[12].

Перед тим, як розпочинати дослідження необхідно обміркувати формулювання наукової проблеми та окреслити шляхи її розв'язання. Така робота може здійснюватися, для прикладу, наступним чином:

1) виявлення нових, незрозумілих для існуючих теорій, фактів та явищ, а також практичних проблем, які потребують вирішення та наукового обґрунтування. Попередній аналіз характеру та обсягу нової інформації спонукатиме до наукового пошуку та створення нових теорій;

2) висування, обґрунтування й оцінювання гіпотез, спрямованих на розв'язання проблеми. При цьому варто розробляти не одну конкретну гіпотезу, а декілька; порівнювати ці гіпотези щодо ступеня їх емпіричного і теоретичного обґрунтування;

3) визначення мети вирішення і типу проблеми, її зв'язок з іншими питаннями. Повне і комплексне розв'язання проблеми передбачає наявність достатнього обсягу якісної емпіричної

---

[11] Рассоха І. М. Конспект лекцій з навчальної дисципліни «Методологія та організація наукових досліджень». Х. : ХНАМГ, 2011. 76 с.

[12] Дубасенюк О. А. Методологія та методи науково-педагогічного дослідження : навч.-метод. посіб. Вид. 2-ге допов. Житомир : Полісся, 2019. С. 19.

інформації, а також необхідного рівня теоретичних уявлень. У разі, якщо їх недостатньо, досліднику доводиться обмежуватися частковим варіантом;

4) попередній опис та інтерпретація проблеми.

Наукове дослідження рухається від емпірики до теорії, а від теорії спрямовується до практики, яка її перевіряє. Отримане знання, нова теорія стимулює висунення наступної гіпотези, нової ідеї, формулювання новітніх положень, що знову спричинює рух наукового пізнання.

Сучасна наука володіє потужним арсеналом методів, які призначені для розв`язування наукових задач на кожному із рівнів пізнання.

На *теоретичному рівні* проводяться логічні дослідження зібраних фактів, розробка понять, суджень та виконуються умовиводи. Тут створюються системи знань, теорії, в яких розкриваються зв`язки, формулюються закони в їх системній єдності та цілісності У процесі такої роботи співвідносяться попередні наукові уявлення з тими, що з`являються.

Для теоретичного рівня дослідження характерні наступні методи: *сходження від абстрактного до конкретного, аналіз, синтез, індукція, дедукція, аналогія, порівняння, абстрагування, моделювання, ідеалізація, формалізація, екстраполяція, аксіоматичний метод, системний аналіз та ін*.

*Сходження від абстрактного до конкретного (конкретизація)* – це загальна форма руху наукового пізнання, закон відображення дійсності і мислення. У відповідності до цього методу мислення бере свій початок від конкретного в дійсності до абстрактного в мисленні і від нього – до конкретного в мисленні. Мета методу – виокремити функціональні зв’язки між складниками досліджуваного об’єкта, явища чи процесу.

*Аналіз* (з грецького – *розкладання*) – метод пізнання, який дозволяє роздрібнювати предмети дослідження на складові, кожну із

яких розглядають окремо у межах єдиного цілого. Серед форм аналізу можна виділити *класифікацію* (поділ на класи, групи, типи тощо) предметів і явищ.

*Синтез* (*поєднання, з'єднання, складання*) – метод вивчення об'єкта у його цілісності, в єдності та взаємозв'язку його складових. На противагу аналізу, даний метод дозволяє з'єднувати окремі частини, характеристики об'єкта і розглядати їх як одне ціле.

Аналіз і синтез тісно переплетені в будь-якому науковому дослідженні, що забезпечує об'єктивне, виважене пізнання дійсності і, разом із тим, відображає єдність протилежностей у відношенні до взаємозв'язку одиничного (окремого) і загального.

*Індукція* – метод наукового пізнання, логіка якого розгортається від конкретного до загального. Тобто, загальне положення виводиться міркуваннями з одиничних суджень.

*Дедукція* – метод логічного умовиводу від загального до часткового, тобто спочатку досліджують стан об'єкта в цілому, а потім вивчають його окремі елементи. Сутністю дедукції як методу пізнання є використання загальнонаукових положень під час дослідження конкретних явищ.

Індуктивні умовиводи продукують частіше ймовірні знання, тому що вони ґрунтуються на емпіричних спостереженнях скінченного числа об'єктів. Дедуктивні міркування породжують нові, достовірні знання, тому що їх вихідні положення дійсні.

*Аналогія* являє собою метод, коли висновок про подібність предметів, явищ робиться на підставі їх схожості за окремими ознаками, а також за аналогією знання про один предмет переноситься на інший шляхом умовиводу.

*Порівняння* – це метод зіставлення досліджуваних об'єктів, явищ чи процесів і виявлення їх подібності та відмінності.

Сутність *абстрагування* як методу наукового пізнання полягає в уявному виділенні конкретних ознак та властивостей об'єкта, явища або процесу. Завдяки абстрагуванню стає можливим із всієї

сукупності ознак і властивостей виокремити загальні та найбільш важливі. Це певне відволікання від несуттєвих характеристик, співвідношень, що дає можливість реальним об'єктам уявно надати гіпотетичних, нереальних ознак. Це спрощує вирішення проблеми в цілому.

Розрізняють такі види абстрагування: ототожнювання (утворення понять шляхом об`єднання предметів, виділених за своїми властивостями, в особливий клас); ізолювання (відокремлення властивостей, щільно пов`язаних із предметами); конструктивізація (не береться до уваги невизначеність між реальними об`єктами); припущення можливого здійснення.

*Моделювання* – це спосіб наукового пізнання, сутність якого полягає в дослідженні моделі об'єкта пізнання на основі абстрактно-логічного мислення за принципами наочності та об'єктивності. Під ***моделлю*** розуміють систему, що заміщує об'єкт пізнання і являє собою джерело інформації про неї.

Метод моделювання передбачає дослідження на моделі об'єкта пізнання, перенесення знань з моделі на оригінал завдяки суттєвій подібності і несуттєвій відмінності між ними.

Різноманітні моделі створюються в процесі наукових досліджень у різних галузях науки. Модель обов'язково повинна мати спільні риси з об'єктом дослідження. Вона відображає основні суттєві сторони явищ чи процесів, які відбуваються в об'єкті спостереження. Моделі поділяються на:

- інтуїтивні (виражені на папері);
- фізичні (макети, муляжі, що дозволяють унаочнювати, ілюструвати природні процеси);
- математичні (дозволяють кількісно досліджувати явища);
- знакові (вираженні символами, алгоритмами, графічно);
- функціональні (орієнтовані на функції систем);
- структурні (описують оргструктури систем);

• природні (являють собою змінені пропорційно об‘єкти дослідження, що дозволяє найбільш повно досліджувати процеси, які протікають у природних умовах).

Моделювання повинне відповідати наступним *вимогам*: по-перше, бути максимально простим, зручним, інформативним, сприяти вдосконаленню самого об`єкта; по-друге, сприяти визначенню або спрощенню характеристик об`єкта, допомагати встановленню раціональних способів побудови, управління або його вивчення.

Нині моделювання є одним з основних методів теоретичного дослідження. Моделювання дає можливість експериментувати над об'єктом вивчення (змінювати параметри, вхідні дані, умови й обмеження) з метою з’ясувати, до яких результатів призведе така зміна. Воно також виконує важливі евристичні функції: встановлює негативні тенденції, визначає позитивні шляхи вирішення проблем, пропонує альтернативні варіанти.

*Ідеалізація.* Суть методу – конструювання мисленнєво об’єктів, яких немає насправді або які практично нездійсненні. Метою ідеалізації є позбавити реальні об’єкти деяких властивостей, що їм притаманні, і наділити (умовно) ці об’єкти певними нереальними, гіпотетичними властивостями, завдяки чому приходять до розвитку певної властивості.

*Формалізація* – метод вивчення об'єктів шляхом відображення їхньої структури в знаковій формі за допомогою штучних мов, наприклад, мовою математики. Серед переваг: узагальненість підходів; стислість і чіткість фіксації значень; однозначність у трактуванні тощо.

Формалізацію можна розглядати як вид моделювання. Зокрема, йдеться про те, щоб виразити суттєві ознаки досліджуваного явища за допомогою символів й тоді вивчення змісту здійснювати на основі знакової моделі, згідно із формальними правилами.

*Екстраполяція* – це розповсюдження будь-яких закономірностей або тенденцій досліджуваного об'єкта, які спостерігаються у певному часовому інтервалі, на інший проміжок часу. Основна його функція – прогностична; він використовується на двох рівнях: якісному (описовому) та кількісному (статистичному).

*Аксіоматичний метод* – метод побудови наукової теорії, за якою деякі твердження приймаються без доведень, а всі інші знання виводяться з них відповідно до певних логічних правил.

*Системний аналіз* передбачає вивчення об'єкта дослідження як сукупності елементів, що утворюють систему. У наукових дослідженнях він передбачає оцінку поведінки об'єкта як системи з усіма факторами, які впливають на його функціонування[13].

До основних *особливостей* системного аналізу можна віднести такі:

– розглядаються всі теоретично можливі альтернативні шляхи і засоби досягнення мети, визначається оптимальна комбінація та сполучення різних методів і засобів;

– альтернативи оцінюються з позицій перспективи, зокрема для систем, які мають стратегічне значення;

– відсутні стандартні, строго детерміновані рішення;

– чітко розмежовуються різні точки зору на вирішення однієї проблеми;

– застосовується підхід до проблем, для яких не повністю визначені вимоги щодо термінів реалізації та вартості;

– визнається принципове значення організаційних і суб`єктивних чинників у процесі прийняття рішень і відповідно до цього розробляються процедури широкого застосування якісних (логічних) міркувань в аналізі й узгодженні різних точок зору;

[13] Романчиков В. І. Основи наукових досліджень : навч. посіб. К. : Центр учбової л-ри, 2007. С. 33.

– особлива увага приділяється факторам ризику і невизначеності, їх урахуванню й оцінці під час вибору оптимального рішення серед декількох можливих варіантів[14].

Потрібно зазначити, що нерідко наукове пізнання характеризується певною двоїстістю: з одного боку, це прагнення розглядати об'єкт цілісно, а з іншого, – до систематизації знання про об'єкт на основі використання певних конкретних, часткових бачень.

Характерні ознаки *емпіричного пізнання* – це збирання фактів (від лат. *factum* – зроблене або те, що відбулося), їх первинний опис, узагальнення і систематизація.

Будь-яке наукове дослідження розпочинається зі збору, систематизації та узагальнення фактів. Розрізняють: факти дійсності – це події, явища та процеси, які відбувалися або відбуваються реально; наукові факти – це відображені у свідомості дослідника факти дійсності, що перевірені, усвідомлені та зафіксовані мовою науки як емпіричні судження.

Основні етапи емпіричного дослідження:

I – процес отримання фактів із першоджерел, тобто з реальної дійсності (події, діяльність людей, соціальних груп, партій, держави в різних сферах суспільного життя, природні явища та процеси), а також із вторинних та третинних джерел (свідчення очевидців, документи, мемуари, доробки інших дослідників, дані статистики тощо).

II – первинна обробка, систематизація та оцінювання фактів у їх взаємозв'язку, тобто осмислення і чіткий опис здобутих фактів у термінах наукової мови, їх класифікація та виявлення залежностей між ними.

[14] Важинський С. Е., Щербак Т. І. Методика та організація наукових досліджень : навч. посіб. Суми : СумДПУ імені А. С. Макаренка, 2016. С. 62.

Отже, на емпіричному рівні пізнавальною функцією є описова характеристика явищ, а результатом дослідження – одержання наукових фактів.

До емпіричних методів наукового дослідження відносять: *спостереження, експеримент, порівняння, вимірювання, інвентаризація, контрольні заміри, метод експертних оцінок, документалістика, опитування, тестування, анкетування тощо.*

*Спостереження* – це спосіб пізнання об'єктивного світу на основі безпосереднього сприйняття предметів і явищ за допомогою відчуттів. Воно дозволяє отримати первинний матеріал для вивчення. Спостереження ведеться за планом і підпорядковується певній тактиці. Це один із складних і трудомістких методів, що зумовлене специфікою суб'єкта та об'єкта спостереження; його проведення вимагає певної формалізації процедур, розробки інструментарію, який забезпечить надійність вихідних даних.

*Експеримент* – це система операцій, впливу або спостережень, спрямованих на одержання інформації про об'єкт у дослідницьких випробуваннях, які можуть проводитися в природних та штучних умовах. Це активний спосіб отримання нових знань, у процесі якого експериментатор керує ходом подій.

Проведення експерименту базується на знаннях про об'єкт, які дозволяють структурно визначити фактори, що впливають на перебіг процесу; передбачають висунення і доведення гіпотез дослідження, контроль за ходом процедур, забезпечення його чистоти та можливості повторень.

На підготовчому етапі розробляється програма експерименту, створюються умови, за яких можливе проведення експериментального дослідження; визначаються фактори впливу, можливості змін; виділяються види об'єктів дослідження та об'єктів, що контролюються; складається план експериментальних робіт; розробляються засоби контролю, регулювання, реєстрації даних; готуються засоби обробки та аналізу інформації.

Завершується експеримент обробкою емпіричних даних, логічними узагальненнями, аналізом і теоретичною інтерпретацією фактичного матеріалу, отриманого дослідним шляхом.

*Порівняння* – це процес встановлення подібності або відмінності предметів та явищ дійсності, а також знаходження спільних властивостей декількох об'єктів. За допомогою цього методу виявляються кількісні та якісні характеристики об'єкта дослідження, класифікується, впорядковується та оцінюється зміст явищ і процесів. Шляхом порівняння встановлюються співвідношення тотожності та відмінності.

Уточнимо, по-перше, порівняння має здійснюватися за наявності об'єктивних спільностей між об'єктами, явищами та процесами, а по-друге – порівнювати потрібно за найважливішими, суттєвими ознаками.

*Вимірювання* – це визначення числового значення певної величини за допомогою одиниць вимірювання, систем фіксації та реєстрації кількісних показників об'єкта. Його результати виражаються числами, що уможливлює їхню статистичну та математичну обробку. Одержані кількісні дані є основою для наукового аналізу якісних характеристик досліджуваного об'єкта, виявлення його суттєвих властивостей і зв'язків, закономірностей поведінки та розвитку.

*Інвентаризація* – метод перевірки об'єктів дослідження, які є в наявності, кількісними прийомами. Суть його в тому, що перевірка наявності і стану об'єктів здійснюється оглядом, підрахунками, зважуванням, вимірюванням.

*Контрольні заміри* – прийоми фактичного контролю, що застосовуються при перевірці достовірності даних про обсяги виконаних робіт, наданих послуг тощо.

Метод *експертних оцінок* – це спосіб передбачення та оцінки майбутніх результатів дій на основі прогнозів фахівців. У ході застосування методу проводиться опитування спеціальної групи

експертів (5–7 осіб) із метою визначення певних змінних величин, необхідних для оцінки досліджуваного питання. Експертизи різних видів застосовуються в технологічних, судових, бухгалтерських, криміналістичних, товарознавчих, соціальних та інших дослідженнях. До експертиз вдаються тоді, коли у складі дослідників немає відповідних фахівців, або за виявленими результатами виникла потреба в експертних висновках.

*Документалістика* – інформаційне моделювання, вивчення документів, нормативно-правове регулювання та ін. У такий спосіб досліджується нормативно-правова, договірна, облікова, звітна та інша інформація про об’єкти. При цьому розглядаються об’єкти основних фондів, товарно-матеріальні цінності, кошти та інші засоби, відображені в системі планової, договірної, нормативно-правової та облікової інформації. При формуванні інформаційної моделі необхідно забезпечити повноту характеристики об’єкта дослідження, вибір істотних змінних і представлення їх у формі інформаційного образу.

*Опитування* – метод збору соціальної інформації про досліджуваний об'єкт під час безпосереднього (інтерв'ю, бесіда) чи опосередкованого (анкетування) соціально-психологічного спілкування дослідника і респондента шляхом реєстрації відповідей респондентів на сформульовані запитання. До опитування вдаються, коли необхідним, а буває і єдиним, джерелом інформації є людина, яка сама бере участь, представляє, провадить досліджувані явища чи процеси.

*Бесіда* – вид опитування, який застосовують з метою одержати додаткову інформацію чи роз'яснення щодо того, що залишилося недостатньо зрозумілим після спостереження. Бесіду проводять відповідно до заздалегідь окресленого плану з переліком питань, які потрібно з'ясувати. Вона відбувається довільним чином, запис відповідей співрозмовця, зазвичай, не практикується.

*Інтерв'ю* передбачає викладення точок зору у заздалегідь визначеній послідовності. Відповіді можна записувати за допомогою різних гаджетів. Нині теорія і практика масових опитувань у своєму арсеналі має численні види організації інтерв'ю (групові, інтенсивні, пробні, стандартизовані, нестандартизовані тощо).

*Тестування* також можна вважати різновидом опитування. Його застосовують, коли масове опитування за рядом причин провести складно. Тестування часто проводять двічі: 1) діагностичне, на початку дослідження; 2) веріфікаційне, при завершенні. Тести потрібно складати так, щоб вони однозначно виявляли певні властивості опитуваних. Важливим для досягнення мети дослідження є дотримання *принципу репрезентативності*, тобто достатності фактичного матеріалу

*Анкетування* є одним з найбільшпоширених видів опитування, яке передбачає самостійне заповнення анкети респондентом. Використовуючи роздаткову, надіслану поштою чи опубліковану в соціальних мережах анкету, дослідник за досить короткий час може зібрати первинну інформацію від великої кількості респондентів. Проте важливо враховувати, що результат дослідження значною мірою залежить від змісту анкети, від того, як сформульовані запитання, скільки є заповнених анкет, як добиралися респонденти.

Для одержання достовірної інформації необхідно, щоб учасник опитування: а) сприйняв потрібну інформацію; б) правильно зрозумів її; в) зміг згадати, за необхідності, події минулого; г) обрав достеменну відповідь на поставлене запитання; ґ) зміг адекватно висловитися; д) мав бажання щиро відповісти на запитання.

Теоретичні та емпіричні методи взаємопов`язані й взаємно зумовлюють один одного у цілісній структурі наукового пізнання. Емпіричне дослідження, виявляючи за допомогою власних методів новітні дані, стимулює теоретичне пізнання, визначає для нього наступні завдання. Теоретичне ж дослідження відкриває нові обрії для емпіричних вишукувань, орієнтує та спрямовує на пошук нових

фактів, сприяє вдосконаленню методів і засобів практичної реалізації. Отже, теоретичний та емпіричний рівні пізнання слід розглядати в їх діалектичному взаємозв'язку і взаємообумовленості.

## 1.2. Методологічні основи статистичного моделювання

У проведенні наукових досліджень суттєву роль відіграють статистичні моделі. *Статистика* – наука, що вивчає методи кількісного охоплення і дослідження масових, зокрема суспільних, явищ і процесів, а також кількісний облік масових явищ. Ця наука використовує інформацію практичної діяльності господарських організацій, узагальнює її і розробляє методи проведення статистичних досліджень. У свою чергу, підприємства, організації й установи використовують теоретичні розробки і положення статичної науки для розв'язання конкретних управлінських завдань.

Статистика встановлює кількісні характеристики різноманітних показників суспільного життя, підтверджує або спростовує їх розвиток, визначає співвідношення між окремими показниками, дає цифрову оцінку закономірностям, які проявляються в них.

Ця наука також вивчає вплив природних та технічних факторів на зміну кількісних характеристик соціально-економічних явищ і процесів. У ході статистичних досліджень встановлюються загальні властивості, виявляються схожі та відмінні риси, певні елементи групуються, явища типізуються тощо.

До речі, терміном «статистика» також називають: 1) сукупність цифрових даних, які характеризують ті чи інші явища суспільного життя або їх сукупність (наприклад, статистика посівних площ, статистика урожайності тощо); 2) процес збирання, зберігання й обробки даних про масові явища, тобто галузь практичної діяльності.

Вивчення кількісної сторони явищ і процесів тісно пов'язане з їх якісним змістом. За допомогою чисел статистика розкриває ступінь

розвитку суспільних явищ, їх напрям і швидкість змін, щільність зв'язків і взаємозалежностей. У багатьох науках також використовується статистична інформація для перевірки, обґрунтування та ілюстрації своїх теоретичних розробок і положень[15].

Наукові дослідження у галузі статистики можуть спричинити:

1) науково-технічний ефект, який проявляється у підвищенні науково-технічного рівня, поліпшенні параметрів техніки і технологій;

2) економічний ефект (зростання національного доходу, скорочення грошових витрат на виробництво продукції, зниження витрат на наукові дослідження й т.ін.). Економічна ефективність має відображати вплив на розвиток економіки країни в цілому, а також регіонів, галузей, організацій і підприємств, що беруть участь у реалізації технологічних нововведень;

3) соціально-економічний ефект (підвищення продуктивності праці, покращення санітарно-гігієнічних, психологічних, організаційних умов праці, захист природи тощо);

4) маркетинговий ефект.

У своїх дослідження статистика розробляє і використовує комплекс методів, прийомів і засобів, які в сукупності формують *статистичну методологію*. Застосування під час статистичних досліджень певних методів визначається поставленими завданнями і залежить від вихідної інформації, якою володіє науковець.

Статистична методологія ґрунтується на загальнофілософських і загальнонаукових принципах. Статистичну методологію виділяє, по-перше, точні вимірювання та кількісний опис масових явищ; по-друге, використання універсальних показників для характеристики неупереджених статистичних закономірностей.

[15] Ткач Є. І., Сторожук В. П. Загальна теорія статистики : підручник. К. : Центр учбов. л-ри, 2009. 442 с.

Виділяють три *етапи статистичного дослідження*:

І – збір масових первинних даних шляхом реєстрації фактів чи опитування респондентів;

ІІ – систематизація та групування даних; відбувається перехід від характеристик одиничного до загального;

ІІІ етап – аналіз матеріалів зведення та групування за допомогою узагальнюючих статистичних показників (абсолютні, відносні та середні величини, показники варіації, індекси, критерії, статистичні коефіцієнти тощо)[16].

Статистичні методи тісно пов'язані з математикою. Однак, математична статистика вивчає закономірності масових явищ і процесів в абстрактній формі, а статистика, як суспільна наука, характеризує суспільні явища в конкретних умовах їх існування і розвитку. У статистичних дослідження широко послуговуються наступними математичними методами: аналіз варіаційних рядів, кореляційний і регресійний аналіз, методи теорії ймовірностей та ін.

У сучасних умовах значення математики для розвитку статистичних методів значно зростає у зв'язку із широким використанням обчислювальної, комп'ютерної техніки, використання можливостей програмування.

Статистичний аналіз даних стає невід'ємним атрибутом наукового дослідження у багатьох наукових сферах. Серед найбільш ефективних засобів пізнання законів і закономірностей навколишнього світу виділяють метод *моделювання*.

Статистичне моделювання – молодий і перспективний науковий напрямок, він постійно розвивається із середини двадцятого століття у зв'язку із розширенням можливостей обчислювальної техніки.

*Статистичні моделі* – це моделі математичні, вони виражаються у формі рівнянь, нерівностей, функціональних

[16] Ткач Є. І., Сторожук В. П. Загальна теорія статистики : підручник. К. : Центр учбов. л-ри, 2009. 442 с.

залежностей, алгоритмів; при їх розв'язуванні поєднуються логіко-алгебраїчні та ймовірнісні методи. Статистичні моделі використовують для діагностики стану об'єктів дослідження, при вивченні причинно-наслідкових зв’язків та динаміки явищ і процесів, для їх моніторингу, при прогнозуванні та прийнятті оптимальних рішень щодо перебігу дослідження та його результатів тощо.

Особливого значення набувають статистичні моделі в наступних випадках:

- об’єкт недоступний для безпосереднього дослідження;

- об’єкт настільки складний, що дослідження його втрачає сенс через складність самого дослідження або ж через наявність великої кількості побічних для даного дослідження факторів;

- дослідження на реальному об’єкті неможливі з інших міркувань (моральних, фінансових або конкурентних)[17].

Процес моделювання складається з трьох основних елементів: суб’єкта, об’єкта дослідження та моделі, за допомогою якої суб’єкт пізнає об’єкт.

Модель встановлює відповідність між сукупністю фактів і гіпотезами, імітує механізм формування закономірностей. Експериментують на моделях, а результати розповсюджуються на практику. Основна вимога до моделі – подібність, адекватність її реальному процесу.

Існують різні *класифікації моделей*:

- за цільовим призначенням моделі поділяють на *теоретико-аналітичні* (призначені для науково-теоретичного дослідження різних процесів) і *прикладні* (для розв’язування практичних задач різної складності);

[17] Руська Р. В., Іващук О. Т. Методи економіко-статистичних досліджень : навч. посіб. : Тернопіль : Тайп, 2014. С. 15-16.

- у залежності від часу моделі поділяють на *статичні* (усі співвідношення охоплюють один період часу) й *динамічні* (описують процес зміни об'єкту чи процесу в часі);

- за характером відображення причинно-наслідкових зв'язків моделі поділяють на *детерміновані* (характеризуються тим, що виходи однозначно визначаються множиною входів і саму модель можна подати як певну функцію невипадкових параметрів і змінних) й *імовірнісні* (виділяються тим, що умови функціонування і характеристики станів змодельованого об'єкта є випадковими величинами, які описуються теорією ймовірностей);

- за характером взаємозв'язків між параметрами, які характеризують досліджуваний об'єкт, моделі поділяють на *лінійні*, які описуються лінійними математичними залежностями, і *нелінійні*, які визначаються нелінійними математичними співвідношеннями;

- за ступенем повноти охоплення об'єкту дослідження моделі поділяють на *макро-* і *мікромоделі*;

- за співвідношенням вхідних (екзогенних) і вихідних (ендогенних) параметрів розрізняють моделі *закриті* та *відкриті*[18].

Зв'язок між математичною моделлю і реальним процесом забезпечується поєднанням у моделі інформації двох типів: 1) гіпотез, які мають логічні обґрунтування щодо природи та характеристик процесу, співвідношень і взаємозв'язків між ними; 2) дослідних даних, які описують ці властивості.

У зв'язку зі значною інформатизацією суспільства та швидкими темпами розвитку інформаційно-комунікаційних технологій робота з статистичними моделями сьогодні передбачає роботу з комп'ютером та зі спеціальним програмним забезпеченням, яке дозволяє швидко опрацьовувати різні математичні співвідношення.

[18] Каламбет С. В. Методолія наукових досліджень : навч. посіб. / С. В. Каламбет, С. І. Іванов, Ю. В. Півняк Ю. В. Дн-вськ : Вид-во Маковецький, 2015. 191 с.

Комп'ютерна модель (модель, реалізована на комп'ютері, за допомогою інформаційно-комунікаційних технологій), допомагає спостерігати та досліджувати певні явища та процеси; проводити багаторазові випробування; отримувати кількісні показники, для знаходження яких необхідні великі за обсягом та складні математичні розрахунки, та аналізувати їх; здійснювати прогнозування досліджуваних явищ чи процесів.

Важливим є створення робочої гіпотези, яка аргументує ймовірну причину існування фактів, які спостерігаються. Специфікою гіпотези є те, що в ній припускаються нові положення, які виходить за межі існуючих знань, пропонуються ідеї, які носять вірогідний характер, на базі яких ведеться пошук нових даних. Головне завдання гіпотези – розкрити ті об'єктивні зв'язки та співвідношення, що є визначальними для досліджуваного явища.

Робоча гіпотеза – це головний методологічне знаряддя, що започатковує процес дослідження, спрямовує та визначає його логіку. Для вирішення питання про прийняття чи відкидання певної гіпотези, її потрібно порівняти з альтернативними гіпотезами. Це необхідно завдяки характерній неоднозначності гіпотези, неможливості бути цілковито впевненим в її істинності.

Формулюючи гіпотезу, кожен дослідник сподівається, що вона виявиться істинною. Але гіпотеза не завжди знаходить підтвердження, а тому доводиться висувати нову. Отже, головні вимоги до гіпотези: можливість її перевірки; певна прогнозованість; логічна несуперечливість.

Забезпечення перевірки гіпотези є логічною умовою, дотримання якої дає право на висунення. Якщо гіпотезу не можна перевірити, вона ніколи не приведе до істинного знання. Прогнозованість – це, власне зміст гіпотези, що перевіряється, а логічна несуперечливість означає, що гіпотеза не вступає у протиріччя з існуючими фактами про дане явище або клас явищ.

Можна виділити наступні *етапи статистичного моделювання*:

1) Характеристика мети та об'єкта моделювання.

2) Розвідувальний аналіз даних.

3) Математична формалізація моделі.

4) Оцінювання параметрів моделі.

5) Перевірка адекватності моделі.

6) Аналіз та інтерпретація результатів[19].

*Мета* – це остаточне призначення моделі. В залежності від мети дослідження для опису одного і того самого процесу можливе створення різних моделей.

Об’єктом наукового дослідження є певна частина дійсності – досить конкретний предмет чи явище, на яке спрямована наукова діяльність дослідника з метою пізнання його сутності, закономірностей розвитку і можливостей використання в практичній діяльності. Це процес або явище, яке породжує проблемну ситуацію й обране для дослідження. Під предметом дослідження розуміється те, що знаходиться в межах об’єкта і завжди співпадає з темою дослідження. Об’єкт і предмет дослідження співвідносяться між собою як загальне і часткове[20].

*Об'єктом* моделювання виступає статистична сукупність, в якій реалізується закономірність. Характеристика об'єкта моделювання включає: вибір одиничного елемента сукупності; визначення просторових і часових меж; формування сукупності основних характеристик моделі.

*Розвідувальний аналіз* даних передбачає: статистичний опис об'єкта (визначення середніх, стандартних відхилень, інших характеристик розподілу); уніфікацію типів ознак; тестування

[19] Єріна А. М. Статистичне моделювання та прогнозування : навч. посіб. К. : КНЕУ, 2001. С. 6.

[20] Краус Н. М. Методологія та організація наукових досліджень : навч.-метод. посіб. Полтава : Оріяна, 2012. 183 с.

сукупності на однорідність, ідентифікацію аномальних спостережень; відтворення пропущених даних; оцінювання взаємозв'язків між ознаками.

На етапі *математичної формалізації* моделі обґрунтовується алгебраїчна форма розрахунків, відношення між властивостями процесу описуються символами та знаками, у певних випадках, будуються блок-схеми. *Оцінювання параметрів* моделі нині відбувається із використанням новітніх комп'ютерних технологій, спеціальних програмних продуктів.

*Перевірка адекватності* моделі означає оцінювання ступеня відповідності параметрів моделі характеристикам об'єкта. На цьому етапі використовують різні процедури порівняння висновків, одержаних у ході досліджень моделі; перевірки статистичних гіпотез за допомогою статистичних критеріїв.

Щодо етапу *аналізу та інтерпретації результатів*, то інтерпретація має узгоджуватися з первинними гіпотезами. Основні висновки формулюються в змістовних термінах: зміст параметрів моделі, правильність перевірюваних гіпотез, оцінювання ступеня їх вірогідності.

Безпосередня реалізація статистичного моделювання передбачає чіткий план дій, в якому умовно можна виділити три кроки: *математична модель – алгоритм розрахунку – комп'ютерна програма.*

Спочатку статистичну модель та її фрагменти досліджують теоретичними методами, що дає змогу отримати важливі (концептуальні) нові знання про досліджуваний об'єкт вивчення. На другому етапі розробляють чи вибирають алгоритм реалізації моделі за допомогою комп'ютерної техніки. Алгоритми не повинні спотворювати основні властивості моделі, бути економічними та адаптивними щодо особливостей розв'язування різновидів задач та використання комп'ютерних засобів. Далі створюється комп'ютерна програма реалізації алгоритму розв'язування статистичної задачі за

допомогою використання мов систем програмування чи мов конкретних прикладних пакетів програм. Безпосередньо в процесі моделювання неодмінно відбувається вдосконалення та уточнення всіх складових процесу.

Можна сформулювати два *принципи статистичного моделювання*:

1) підпорядкованість меті дослідження на всіх етапах моделювання;

2) забезпечення адекватності моделі[21].

Ґрунтується метод статистичного моделювання на багатократному проведенні випробувань побудованої моделі із наступною статистичною обробкою одержаних даних з метою визначення характеристик процесу у вигляді статистичних оцінок його параметрів. Завдяки переходу від окремих фактів до масових можна визначити загальну закономірність, позбавлену впливу випадкових причин.

Отже, для одержання достатньо надійних результатів необхідно забезпечувати велике число реалізацій досліду, до того ж, якщо відбудеться зміна хоча б одного з вихідних параметрів задачі, доведеться проводити нову серію з багатьох випробувань. З іншого боку, якщо сама модель є доволі складною, то невиправдано велика кількість випробувань може спричинити затримання одержання результату. Тому важливо правильно оцінити необхідну кількість результатів випробувань.

Теоретичною основою методу статистичного моделювання є закон великих чисел. У теорії ймовірностей закон великих чисел базується на доведенні низки теорем для різних умов збіжності за ймовірністю середніх значень результатів (на підставі великої

[21] Єріна А. М. Статистичне моделювання та прогнозування : навч. посіб. К. : КНЕУ, 2001. С. 8.

кількості спостережень) до деяких величин, як от, теореми Чебишева та Бернуллі.

Розв'язування задач методом статистичного моделювання включає:

- опрацювання й побудову структурної схеми процесу, виділення основних взаємозв'язків;
- формалізоване відтворення процесу;
- моделювання випадкових явищ (подій, величин, функцій), що властиві досліджуваній системі;
- моделювання процесу функціонування системи (на підставі використання даних, що отримані на попередньому етапі) – відтворення процесу відповідно до розробленої структурної схеми і формалізованого опису (імітаційні прогони);
- акумулювання результатів моделювання, статистичне опрацювання, аналіз та їх інтерпретацію[22].

Більшість проблем статистики потребують застосування спеціальних методів вирішення, які визначаються відповідно до характеру досліджуваного об'єкта.

Серед спеціальних методів статистичного аналізу можна виділити: кореляційний, факторний аналіз, метод імплікаційних шкал, контент-аналіз та ін.

*Кореляційний аналіз* – це процедура для вивчення співвідношення між незалежними змінними. Зв'язок між цими величинами виражається в узгодженості змін, що спостерігаються. Обчислюється коефіцієнт кореляції між двома змінними. Чим він більшим, тим точніше можна спрогнозувати значення однієї з них за значенням інших.

*Факторний аналіз* дає можливість встановити багатомірні зв'язки змінних величин за декількома ознаками. На основі парних

[22] Вітлінський В. В. Моделювання економіки : навч. посіб. К. : КНЕУ, 2003. 408 с.

кореляцій, отриманих у результаті кореляційного аналізу, одержують набір нових, укрупнених ознак – факторів. У результаті послідовної процедури отримують фактори другого, третього та інших рівнів. Факторний аналіз дозволяє узагальнити отримані результати.

*Метод імплікаційних шкал* є наочною формою вимірювання та оцінки отриманих даних, які градуюються за кількістю або інтенсивністю ознак. Шкали класифікуються за типами або рівнем виміру. Прості шкали дають однозначну оцінку тієї чи іншої ознаки. Серію шкал («батарею») можна перетворити в єдину шкалу значень окремих ознак. Ця процедура називається шкалюванням.

*Контент-аналіз* посідає особливе місце в системі методів, оскільки допомагає дати інтерпретацію змісту інформації через кількісні показники. Останнім часом контент-аналіз розуміють як якісно-кількісний аналіз змісту сукупності текстового масиву. Контент-аналіз на доповнення до традиційних методів логіко-аналітичного аналізу застосовують переважно до текстових масивів (опублікованих і неопублікованих), а не конкретних текстів.

Сутність методу полягає в знаходженні і виділенні в тексті певних змістових понять, одиниць аналізу, що цікаві досліднику, а також визначенні частоти їх появи в документі у відповідності до змісту. Ретельний підрахунок кожної одиниці спостереження з обов'язковим урахуванням частоти її появи в тексті дає змогу виявити закономірності, присутні в документі, які традиційними методами вивчити не можна.

Специфічною особливістю статистичних методів є їх комплексність, що спричинено як багатоманітністю форм статистичних закономірностей, так і безпосередньою складністю здійснення статистичного дослідження. Оскільки природа явищ, які досліджує статистика, доволі складна і непередбачувана, то й вивчати їх треба у взаємозв'язку і взаємообумовленості.

Філософський підхід, який закладено в методологічну основу статистичної науки, вимагає розгляду явищ і процесів в їх русі, постійних змінах. Для цього розроблена певна система показників, які дають змогу охарактеризувати варіації змін рівнів явищ, визначити тенденції і закономірності їх розвитку. Важливим моментом є встановлення границь переходу кількісних явищ у якісні форми їх прояву.

Характеристика явищ і процесів, які досліджуються статистикою, забезпечується завдяки використанню відповідної інформації, котру одержують за результатами статистичного спостереження.

*Статистичне спостереження* – це науково організований, систематизований, плановий збір статистичної інформації за допомогою якої оцінюється економічна, екологічна, соціальна та інші найважливіші складові діяльності суспільства на всіх рівнях[23].

Організація статистичного спостереження вимагає використання адекватних методів і статистичної методології, адже лише завдяки правильному, науково-обґрунтованому їх застосуванню й науковій обробці можна забезпечити досягнення неупередженості в процесі збору необхідної інформації. Основне завдання статистичного спостереження полягає в отриманні новітньої, повної і, головне, об‘єктивної статистичної інформації.

Організацію статистичного спостереження можна розділити на такі етапи: 1) підготовка до проведення спостереження; 2) реєстрація статистичних даних та формування інформаційної бази; 3) контроль результатів спостереження.

На першому етапі важливим є визначення *мети спостереження***,** тобто основного очікуваного результату

[23] Щурик М. В., Ключенко А. В. Статистика : навч. посіб. для студ. вищ. навч. закл. усіх рівнів акредит. Івано-Франківськ : НАІР, 2016. 274 с.

статистичного дослідження. Відповідно до поставленої мети визначається *об'єкт* – сукупність одиниць досліджуваного явища, про яке зібратимуться статистичні дані. Якість будь-якого спостереження залежить від *програми спостереження* – це перелік чітко сформульованих запитів, на які необхідно одержати відповіді.

Упорядкування, систематизація та наукова обробка даних статистичного дослідження називається *статистичним зведенням* Воно буває просте (підсумовують лише одиниці спостереження) і складне (здійснюється поділ сукупності на групи або підгрупи; при цьому проводять також підрахунок групових і загальних ознак, а результати дослідження подають у вигляді статистичних таблиць).

Одним із найбільш поширених інструментів наукових досліджень й обробки інформації за результатами статистичного спостереження є *метод групувань* – розподіл статистичної сукупності на групи, підгрупи за істотними ознаками з метою встановлення зв'язку та закономірностей розвитку досліджуваних явищ і процесів предмета статистики.

Функції методу групування: 1) виявлення однорідних явищ, шляхом поділу їх на однорідні групи і підгрупи; 2) статистична оцінка структури явищ та структурних зрушень; 3) встановлення взаємозв'язку та взаємозалежності між досліджуваними ознаками.

Таке групування здійснюється у методичній послідовності: вибір групувальної ознаки, визначення кількості груп та встановлення розміру інтервалу. Взаємозв'язані ознаки, котрі використовуються в процесі проведення статистичного групування, поділяються на факторні та результативні (цей поділ є дещо умовним та значною мірою залежить від завдання дослідження).

Розрізняють: просте – групування за однією ознакою; складне – за декількома.

У тих випадках, коли проведення суцільного спостереження неможливе або його вважають недоцільним застосовують вибіркове спостереження. *Вибіркове спостереження* – це вид спостереження,

коли досліджується не вся сукупність, а лише її частина, що відібрана за встановленими правилами відбору, завдяки яким забезпечується репрезентативність сукупності в цілому[24]. Таке спостереження є більш оперативним, оскільки значно скорочує терміни проведення робіт; дозволяє суттєво зекономити завдяки скороченню обсягу робіт; його результати часто точніші, так як для його проведення можна підібрати більш кваліфікованих виконавців, простіше організувати контроль.

Частина одиниць, яка відібрана для спостереження за певними встановленими правилами, називається *вибірковою сукупністю*, а вся сукупність одиниць, із якої проводиться відбір, – *генеральною*. Застосування вибіркового методу передбачає поширення результатів на всю сукупність.

Попередження систематичних помилок досягається в результаті застосування науково обґрунтованих способів формування вибіркової сукупності.

При формуванні вибіркової сукупності мають бути забезпечені дві умови: 1) рівні можливості для кожної одиниці генеральної сукупності потрапити у вибірку (принцип рівноможливості); 2) досить представницька чисельність вибіркової сукупності.

Розрізняють: 1) індивідуальний відбір, коли у вибірку потрапляють окремі одиниці генеральної сукупності; 2) груповий відбір – до вибірки дістаються якісно однорідні групи або серії досліджуваних одиниць; 3) комбінований відбір як поєднання індивідуального і групового відбору.

Однак, при проведенні вибіркового спостереження можуть виникати погрішності, похибки. Різниця між показниками вибіркової й генеральної сукупності називається похибками вибірки, котрі

[24]Щурик М. В., Ключенко А. В. Статистика : навч. посіб. для студ. вищ. навч. закл. усіх рівнів акредит. Івано-Франківськ : НАІР, 2016. С. 236.

поділяються на похибки реєстрації і похибки репрезентативності (представництва).

Похибки реєстрації виникають у результаті використання помилкових або неточних даних. Забезпечення репрезентативності вибірки в статистиці досягається завдяки дотриманню принципу випадковості відбору одиниць. Його суть у тому, що можливість залучення у вибіркову сукупність або навпаки вилучення одиниці з неї не може вплинути будь-який інший фактор, крім випадкового.

Для виявлення і усунення помилок запроваджується чіткий контроль статистичних матеріалів. Розрізняють *арифметичний контроль* – це обчислювальна перевірка одержаних даних і зіставлення показників, які взаємопов'язані між собою або випливають один з іншого; *логічний контроль,* який полягає у взаємному зіставленні одержаних відповідей на питання програми, виходячи з їх логічного зв'язку, або їх співставлення з іншими джерелами та виявлення невідповідностей у цих відповідях.

Отже, об‘єктивність результатів вибіркового спостереження досягається завдяки розв‘язанню таких завдань:

1) визначення із заданою ймовірністю можливих границь похибки репрезентативності, за умови, що фактичні границі більші заданих, результатами такої вибірки користуватись не можна;

2) визначення ймовірності того, що можливі границі похибки репрезентативності не перевищать заданих величин;

3) встановлення необхідного обсягу вибірки.

Поширення вибіркових даних на генеральну сукупність здійснюють шляхом порівняння характеристик генеральної сукупності з показниками вибірки. Найчастіше використовують такі способи поширення вибіркових даних:

1) *прямого перерахунку* (середні величини або частки, одержані в результаті дослідження вибіркової сукупності, множать на кількість одиниць генеральної сукупності);

2) *поправочних коефіцієнтів*. Його застосуванню передує порівняння даних вибіркового спостереження та генеральної сукупності, у результаті чого встановлюють відсоток недообліку. Коефіцієнти, що одержані за результатами порівняння, використовують для внесення змін в формуляри суцільного обліку.

Величини, значення, залежності, одержані в ході статистичних досліджень, зручно зображати в графічний спосіб. Це робить їх більш наочними, спрощує розгляд, аналіз відповідних матеріалів і розуміння результатів наукових досліджень. За допомогою графіків краще вдається виявити і презентувати характеристику структури, динаміки, загальну картину закономірностей розвитку досліджуваного об'єкта, явища чи процесу у часі і в просторі. Для унаочнення використовують різне забарвлення, штрихування, фони тощо.

Найбільш поширеним способом графічного зображення результатів наукових досліджень є діаграми (рис. 1)[25].

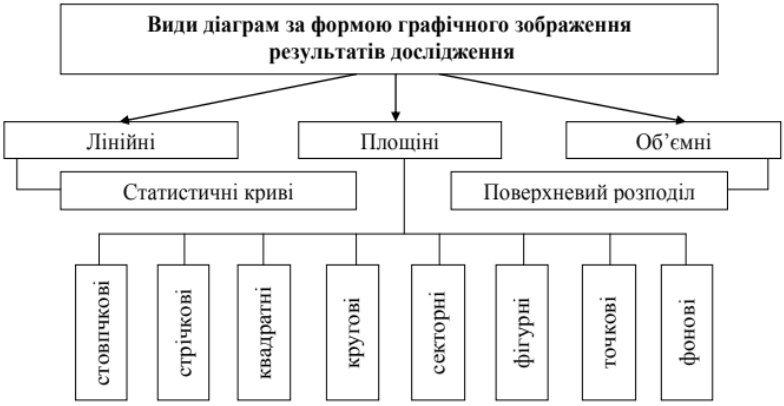



*Рис. 1.* **Види статистичних діаграм**

[25] Каламбет С. В. Методолія наукових досліджень : навч. посіб. / С. В. Каламбет, С. І. Іванов, Ю. В. Півняк Ю. В. Дн-вськ : Вид-во Маковецький, 2015. 191 с.

Отже, статистичне мислення засноване на узагальненні, умінні в розрізнених фактах побачити закономірне та випадкове, що є необхідним як для науки, так і для повсякденного життя. Протягом тривалого часу статистику розглядали лише як джерело необхідних даних для управлінських, науково-дослідних, практичних потреб різних структур, організацій. Нині термін «статистика» широко вживається в побуті, статистичні дані використовуються в повсякденній практиці, коли мовою цифр яскраво відображаються різні аспекти суспільного життя, науки тощо.

Однак, на сьогодні теорія статистики ще не є остаточною, зокрема, нині здійснюються ґрунтовні дослідження безпосередньо щодо методології статистики. У статистиці, як і в інших науках, ще існує багато невідомого, вона потребує удосконалення, але її майбутнє очевидне.

### *Список питань до самоконтролю:*

**1.** Методологія – це:

а) сукупність визначених правил, прийомів, способів, норм пізнання певного суб'єкта чи явища;

б) вчення про систему методів наукового пізнання та перетворення реальної дійсності;

в) сукупність методів дослідження;

г) конкретна методика, за якою проводиться наукове дослідження;

д) правильної відповіді немає.

**2.** Наукова проблема – це:

а) висунута наукова гіпотеза;

б) форма наукового мислення щодо дослідження нового, котре виникло у процесі пізнання або практичної діяльності;

в) виявлення нових фактів та явищ;
г) усі відповіді правильні;
д) правильної відповіді немає.

**3.** Обєкт дослідження – це:
а) галузь;
б) підприємство;
в) структури органів управління;
г) усі відповіді правильні;
д) немає правильної відповіді.

**4.** Предмет дослідження – це:
а) певна сфера діяльності;
б) підприємство або группа підприємств;
в) структури органів управління;
г) усі відповіді правильні;
д) немає правильної відповіді.

**5.** Методика дослідження – це:
а) основний метод дослідження;
б) основні принципи дослідження;
в) сукупність методів і прийомів дослідження;
г) усі відповіді правильні;
д) немає правильної відповіді.

**6.** Метод дослідження – це:
а) спосіб пізнання, дослідження явищ природи і суспільного життя;
б) система поглядів, система опису певного предмета або явища, стосовно його побудови, функціонування, що сприяє його розумінню, тлумаченню, вивченню головних ідей;

в) система методологічних і методичних принципів і прийомів, операцій і форм побудови наукового знання;

г) процес вивчення певного об'єкта (предмета або явища) з метою встановлення закономірностей його виникнення, розвитку і перетворення в інтересах раціонального використання у практичній діяльності людей;

д) немає правильної відповіді.

**7.** Розкриття теорії питання, вдосконалення існуючих та здобуття нових знань, це відноситься до рівня пізнання:

а) теоретичного;

б) практичного;

в) первинного;

г) емпіричного;

д) немає правильної відповіді.

**8.** До емпіричних методів дослідження відносяться:

а) абстрагування;

б) узагальнення;

в) формалізація;

г) спостереження;

д) правильної відповіді немає.

**9.** Першою стадією статистичного дослідження слугує:

а) розрахунок і аналіз узагальнюючих зведених показників;

б) зведення і групування;

в) статистичне спостереження;

г) збір статистичних даних;

д) правильної відповіді немає.

**10.** Другою стадією статистичного дослідження слугує:

а) збір статистичних даних;

б) статистичне спостереження;

в) розрахунок і аналіз узагальнюючих зведених показників;

г) зведення і групування;

д) правильної відповіді немає.

**11.** Основне завдання статистичного спостереження полягає в:

а) отриманні новітньої, повної та об‘єктивної інформації;

б) зборі та систематизації статистичної інформації;

в) оцінці мінливості значень варіюючої ознаки;

г) підрахунку та аналізі результатів статистичного спостереження;

д) правильної відповіді немає.

**12**. За характером часової залежності моделі поділяють на:

а) детерменовані й імовірносні;

б) статичні й динамічні;

в) лінійні й нелінійні;

г) закриті й відкриті;

д) правильної відповіді немає.

**13**. За ступенем повноти охоплення об’єкту дослідження моделі поділяють на:

а) детерменовані й імовірносні;

б) статичні й динамічні;

в) лінійні й нелінійні;

г) теоретико-аналітичні й прикладні;

д) мікромоделі й макромоделі.

**14**. Статистичним зведенням називається:

а) пошук та використання узагальнюючих показників результатів спостереження;

б) визначення наочних способів подання результатів спостереження;

в) встановлення групувальних ознак та кількості груп;

г) упорядкування, систематизація та наукова обробка статистичних даних;

д) правильної відповіді немає.

**15**. Групування – розподіл статистичної сукупності на:

а) групи за певними ознаками з метою встановлення їх зв'язку;

б) групи, підгрупи за істотними ознаками з метою встановлення зв‘язку та закономірностей розвитку досліджуваних явищ і процесів;

в) однорідні соціально-економічні явища;

г) групи за певною типовою ознакою;

д) правильної відповіді немає.

### *Список рекомендованої літератури по темі:*

1. Важинський С. Е., Щербак Т. І. Методика та організація наукових досліджень : навч. посібник Суми : СумДПУ імені А. С. Макаренка, 2016. 260 с.
2. Вітлінський В. В. Моделювання економіки : навч. посіб. К. : КНЕУ, 2003. 408 с.
3. Дубасенюк О. А. Методологія та методи науково-педагогічного дослідження : навч.-метод. посіб. Вид. 2-ге допов. Житомир : Полісся, 2019. 256 с.
4. Єріна А. М. Статистичне моделювання та прогнозування : навч. посіб. К. : КНЕУ, 2001. 170 с.
5. Каламбет С. В. Методолія наукових досліджень : навч. посіб. / С. В. Каламбет, С. І. Іванов, Ю. В. Півняк Ю. В. Дн-вськ : Вид-во Маковецький, 2015. 191 с.

6. Краус Н. М. Методологія та організація наукових досліджень: навч.-метод. посіб. Полтава : Оріяна, 2012. 183 с.
7. Кустовська О. В. Методологія системного підходу та наукових досліджень : курс лекцій. Тернопіль : Економічна думка, 2005. 124 с.
8. Рассоха І. М. Конспект лекцій з навчальної дисципліни «Методологія та організація наукових досліджень». Х. : ХНАМГ, 2011. 76 с.
9. Романчиков В. І. Основи наукових досліджень : навч. посіб. К.: Центр учбової л-ри, 2007. 254 с.
10. Руська Р. В., Іващук О. Т. Методи економікостатистичних досліджень : навч. посіб. : Тернопіль : Тайп, 2014. 190 с.
11. Ткач Є. І., Сторожук В. П. Загальна теорія статистики : підручник. К. : Центр учбов. л-ри, 2009. 442 с.
12. Шейко В. М., Кушнаренко Н. М. Організація та методика науково-дослідницької діяльності : підруч. К. : Знання-Прес, 2002. 295 с.
13. Щурик М. В., Ключенко А. В. Статистика : навч. посіб. для студ. вищ. навч. закл. усіх рівнів акредит. Івано-Франківськ : НАІР, 2016. 274 с.
14. Цехмістрова Г.С. Основи наукових досліджень : навч. посіб. К. : Вид. дім «Слово», 2004. 240 с.
15. Юринець В. Є. Методологія наукових досліджень : навч. посіб. Львів : ЛНУ ім. І. Франка, 2011. 178 с.

# Розділ 2. МОДЕЛІ СТАТИСТИЧНИХ КЛАСИФІКАЦІЙ

## 2.1. Сутність багатовимірних класифікацій

Різноманітність методів багатовимірного статистичного аналізу даних визначається різнобічністю інформації та завданнями, які вирішуються у процесі її обробки та проведенні досліджень. Порівняння об'єктів аналізу між собою за певними ознаками та розподілення їх сукупності на однорідні групи за допомогою процедури класифікації є найпоширенішою практикою.

Сутність поняття класифікації розкривається через її трактування як системного розподілу множини явищ та процесів за істотними характеристиками та розташування їх у певному порядку за ступенем подібності або відмінності між ними.

Найпростішим методом поділу елементів сукупності є групування, в якому за найбільш інформативною ознакою відбувається поділ об'єктів залежно від мети аналізу. При потребі класифікації за декількома властивостями, використовують комбінаційне або багатовимірне групування.

У випадках унеможливлення упорядкування класифікаційних ознак застосовується найпростіший метод багатовимірного групування шляхом визначення інтегрального показника або індексу та проведення класифікації об'єктів за ним. Подальша реалізація цього підходу можлива за декількома узагальнюючими показниками методом головних компонент на підставі факторного аналізу.

Класифікація об'єктів пов'язана з розпізнаванням образів, яке вирішуються шляхом продукування поняття о класі об'єктів. Одним із завдань розпізнавання образів є розбиття сукупності одиниць дослідження на однорідні класи за умови, що об'єкти одного класу схожі між собою або близькі за деяким критерієм відповідності; об'єкти з різних класів повинні бути досить різними, далекими друг від друга. Образ – це множина об'єктів, подібних один до одного у деякому фіксованому відношенні; процедура розпізнавання образу

об’єкту – класифікація об’єктів за певними принципами (за усвідомленими групами). Розпізнати образ об'єкту – значить вказати до якого образу, класу схожих об'єктів він належить.

До методів багатовимірної класифікації, які дозволяють поділити об’єкти сукупності на однорідні групи та класи, комплексно врахувавши їх якісні та кількісними характеристики, відносять кластерний та дискримінантний аналіз.

Альтернативна назва дискримінантного аналізу «класифікація з вчителем», який включає методи розпізнавання образів у ситуації, якщо дослідник має так звані навчальні вибірки. Відмінність кластерного аналізу у відсутності навчальної вибірки, що характеризує метод як класифікацію без навчання.

Значною перевагою кластерного аналізу є прості та логічні методи розпізнавання образів, алгоритми рішень яких легко формалізуються у комп’ютерних програмах, та можливість скорочення, стискання великих за обсягом масивів соціально-економічної інформації до компактного та наочного сприйняття. Відмінною характеристикою методу є комплексне вивчення ознак, відсутність обмежень щодо їх виду. Також кластерний аналіз дозволяє сфокусуватися на змістовній складовій багатофакторних об’єктів. Недоліком аналізу є залежність складу та кількості кластерів від обраних критеріїв класифікації. Також можуть викривлятися або втрачатися індивідуальні риси окремих об’єктів в процесі процедури зведення вихідного масиву даних до більш компактного вигляду за рахунок заміни узагальнюючими значеннями параметрів кластеру їх характеристик.

У дискримінантному аналізу, на відміну від кластерного, нові кластери не утворюються, а лише формулюється правило, за яким об'єкти класифікації відносяться до одного з вже існуючих (навчальних) підмножин (класів) на основі порівняння величини дискримінантної функції вивчаємого об'єкту з деякою константою дискримінації (наприклад, з'являється новий об'єкт з подібними

досліджуваними ознаками та його потрібно класифікувати до відповідного класу з найбільшою імовірністю). Обмеженість використання методу полягає у відсутності необхідної бази даних, вимоги про нормальність багатовимірного розподілу дискримінантних змінних та рівності коваріаційних матриць, скептичному ставленні до можливих результатів[1].

### 2.2. Кластерний аналіз

Кластеризація вважається найважливіший компонентом та ознакою інтелекту даних. Кластерний метод – багатовимірна статистична процедура впорядкування об'єктів у порівнянно однорідні групи за інформацією про вибірку об'єктів. У результаті проведення кластерного аналізу та поділу множини досліджуваних об'єктів за сукупністю ознак на однорідні групи одержають класи, які називають кластерами (від англ. cluster – група, скупчення, кисть, рій, гроно; група елементів, що характеризуються загальною властивістю) та таксонами (від англ. taxon – систематизована група) або образами. Одночасно з вирішенням задачі класифікації даних, метод дозволяє виявляти відповідні структури у множині елементів, а саме виокремлювати компактні, віддалені одну від іншої групи об'єктів або здійснювати пошук природнього розбиття сукупності на області скупчення[2].

Кластерний аналіз відноситься до напрямів статистичного дослідження соціально-економічних та науково-технічних процесів, має широке застосування у сферах макроекономіки, маркетингу, менеджменту, фінансах, медицині, біології, геології, метеорології, документалістиці, соціології. Метод нараховує більше ста

[1] Klecka, WR. (1980). Discriminant analysis. Sage Publications, Beverly Hills.

[2] Бізнес-аналітика багатовимірних процесів : навчальний посібник [Електронний ресурс] / Т. С. Клебанова, Л. С. Гур'янова, Л. О. Чаговець та ін. – Харків : ХНЕУ ім. С. Кузнеця, 2018. – 272 с.

алгоритмів реалізації та має витоки розробки з антропології завдяки Драйверу та Крьоберу (1932 р.), психології – Зубіну (1938 р.) та Тріону (1939 р.), набувши популяризації через класифікацію ознак у теорії особистості Кеттеллем (1943 р.).

Виділення груп повинно відповідати двом принципам: ступінь подібності між одиницями сукупності, що включені до однієї групи, вищий порівняно зі схожістю одиниць, що належать до різних класів.

Звичайною формою представлення вихідних даних для проведення кластерного аналізу слугує прямокутна таблиця «об'єкт – ознака». Кожен рядок матриці містить характеристику про n – об'єктів за m – ознак. Об'єктами є конкретні предмети дослідження, що потребують класифікації; ознаками - певні властивості об'єктів:

$$X = \begin{pmatrix} x_{11} \dots x_{1j} \dots x_{1m} \\ \dots \dots \dots \dots \dots \\ x_{j1} \dots x_{ij} \dots x_{im} \\ \dots \dots \dots \dots \dots \\ x_{n1} \dots x_{nj} \dots x_{nm} \end{pmatrix}$$

Різні властивості можуть виражатися як числовими, так і нечисловими значеннями. Наприклад, обсяг виробництва може вимірюватися в кілограмах або тонах, ціна житла - в тисячах гривень (доларів) і т. п. Такі ознаки називаються кількісними (безперервними). Над ними можна проводити арифметичні операції.

На відміну від числових характеристик ряд ознак може мати дискретні значення. У свою чергу, дискретні ознаки діляться на дві групи. Перша група - порядкові (рангові) змінні. Таким ознаками властива впорядкованість значень. До них можна віднести вік, поверх будинку, рік випуску та ін. Значення рангових змінних представляються натуральними числами. Друга група дискретних ознак не має такої впорядкованості і носить назву номінальних змінних. Це змінні, що приймають два значення (дихотомічні) або

більше. Цим значенням можна поставити у відповідність деякі числа, які, однак, не будуть відображати будь-якої впорядкованості значень змінної. Прикладом таких ознак є стать респондента, тип будинку, вид транспортного засобу і т. п. Ці ознаки відносяться до шкали найменувань. Їх можна вважати якісними характеристиками об'єктів.

Для кількісної оцінки подібності у кластерному аналізі вводиться поняття «відстань між об'єктами», що також має назву метрика або міра (назва залежить від методу обчислення).

Якщо кожен об'єкт описується m-ознаками, то він може бути представлений як точка у n-мірному просторі, і схожість з іншими об'єктами буде визначатися як відповідна відстань.

Відстанню (метрикою) між *i-м* та *j-м* об'єктами в просторі ознак називається така величина $d_{ij}$, яка задовольняє наступним аксіомам[3]:

1. $d_{ij} \geq 0$ невід'ємність;

2. $d_{ij} = d_{ji}$ симетрія;

3. $d_{ij} + d_{jq} \geq d_{iq}$ нерівність трикутника характеризує відстань між трьома об'єктами та має назву метрична нерівність ($q$ - номер об'єкту);

4. якщо $d_{ij} \neq 0, \text{то } i = \neq j$ відмінність (розбіжність) нетотожних об'єктів;

5. якщо $d_{ij} = 0, \text{то } i = j$ нерозрізненість тотожних об'єктів (відстань між ними дорівнює нулю).

Міру близькості або подібності об'єктів тлумачать як величину, зворотну відстані між об'єктами: має межу та зростаюча зі зростанням близькості об'єктів: $\mu_{ab}$неперервна; $\mu_{ab} = \mu_{ba}$; $1 \leq \mu_{ab} \leq 0$.

Перехід від відстані до міри близькості здійснюється за формулою:

[3] Klecka, WR. (1980). Discriminant analysis. Sage Publications, Beverly Hills.

$$\mu = \frac{1}{1+d} \tag{2.1}$$

Вибір способу обчислення відстані між об'єктами залежить від мети дослідження, фізичної та статистичної природи спостережень, апріорної інформації про характер імовірнісного розподілу. У свою чергу саме застосування певного методу розрахунку відстані впливає на остаточний результат розподілення об'єктів на класи.

Найбільш поширені способи визначення відстані між об'єктами представлені у табл. 2.1.

Використання описаних метрик відстаней підходить до об'єктів, які можуть бути представлені у вигляді множини крапок у *k-мірному* просторі. У випадку неможливості представлення деяких об'єктів соціології та економіки в описаному вигляді, доцільно у якості відстаней застосовувати різницю між одиницею та коефіцієнтом кореляції.

Загальний алгоритм реалізації кластерного аналізу можна представити у вигляді послідовності процедур:

*Крок 1:* значення вихідних змінних нормуються одним з наступних способів:

1. $z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$; 2. $z_{ij} = \frac{x_{ij}}{x_{\max j}}$; 3. $z_{ij} = \frac{x_{ij}}{\bar{x}_j}$; 4. $z_{ij} = \frac{x_{ij}}{x_{\min j}}$

де $x_{ij}$ – значення *j-ої* ознаки *i-ого* об'єкту,

$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_i$ – середнє арифметичне значення *j-ої* ознаки,

$\sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2$ – дисперсія *j-ої* ознаки,

$\sigma_j = \sqrt{\sigma_j^2}$ – середнє квадратичне відхилення.

*Таблиця 2.1.* **Способи обчислення відстані між об'єктами**

| Способи розрахунку | Формула | Характеристика використання |
|---|---|---|
| Лінійна відстань *[Хемінгова відстань, City-block Manhattan distances – мангеттенська відстань міських кварталів)* | $d_{ij} = \sum_{k=1}^{m} \lvert x_{ik} - x_{jk} \rvert$ | Використовується як міра відмінності об'єктів, що задаються дихотомічними (мають лише два значення) якісними ознаками номінальної шкали; зменшує вплив окремих великих викидів на величину. Відстань дорівнює кількості розбіжностей (незбігів) значень відповідних ознак для *i*-го та *j*-го об'єктів. |
| Евклідова відстань *[Euclidean distances]* | $d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$ | ✓ спостереження беруться з генеральних сукупностей, що мають багатовимірний нормальний розподіл з коваріаційною матрицею виду $\sigma^2 E_k$ (компоненти X взаємно незалежні та мають однакову дисперсію);<br>✓ компоненти вектору спостереження однорідні за фізичною суттю і однаково важливі для класифікації;<br>✓ ознаковий простір співпадає з геометричним простором. |

Продовження таблиці 2.1

| Способи розрахунку | Формула | Характеристика використання |
|---|---|---|
| Відстань Чебишева *[Chebychev distance metric]* | $d_{ij} = max\|x_{ik} - x_{jk}\|$ | Застосовується у випадках необхідності виявлення несхожих об'єктів за якоюсь однією координатою |
| Відстань Мінковського *[Power metric]* | $d_{ij} = \left(\sum_{k=1}^{m}\|x_{ik} - x_{jk}\|^p\right)^{1/p}$ | Узагальнює попередні відстані та дозволяє збільшити або зменшити вагу розмірності, для якої відповідні об'єкти дуже відрізняються:<br>✓ якщо $p = 1$, це Манхеттенська відстань;<br>✓ якщо $p = 2$, це Евклідова відстань;<br>✓ якщо $p = 3$, це відстань Чебишева |
| Відстань Махаланобіса *[Mahalanobis distance]* | $d_{ij} = (x_{ik} - x_{jk})R^{-1}(x_{ik} - x_{jk})^T$ | Використовують якщо компоненти $x_1, \dots, x_n$ вектору спостережень X мають різну значимість. Розрахунок спирається не лише на вектори точок у багатовимірному просторі, а й на кореляційну матрицю, яка відображає парні кореляційні зв'язки усіх ознак ($R^{-1}$ – матриця обернена до матриці коефіцієнтів парної кореляції, $(x_{ik} - x_{jk})^T$ – вектор-стовпчик, транспонований відносно векторів-рядка $(x_{ik} - x_{jk})$. |

*Джерело*: узагальнено автором на основі [1, 2, 3, 4, 5]

*Крок 2:* розраховується матриця відстаней або матриця мір подібності.

*Крок 3:* знаходиться пара найближчих кластерів, за обраним алгоритмом вони об'єднуються, новому кластеру присвоюється найменший з номерів об'єднаних кластерів.

*Крок 4:* процедури 2, 3 та 4 повторюються до поки усі об'єкти об'єднаються в єдиний кластер або до досягнення поставленого порогу схожості.

*Крок 5:* інтерпретація отриманих результатів та перевірка їх якості.

Алгоритм ієрархічного кластерного аналізу при використанні програмного забезпечення реалізується за допомогою формули оцінювання схожості між кластерами Ланса–Вільямса (2.8.).

Методи кластерного аналізу можна узагальнити за допомогою наступної класифікації (рис.2.1):


*Рис.2.1.* **Класифікація методів кластерно аналізу**
*Джерело*: власні розробки автора

Кожний метод характеризується власним алгоритмом процедури та відмінностями у результатах кластеризації. Найбільш розповсюдженими методами кластерного аналізу є ієрархічні агломеративні методи (від латинського agglomero — приєдную, накопичую), за якими на першому кроці кожний об'єкт вибірки розглядається як окремий кластер, далі – процес об'єднання кластерів відбувається за матрицею відстаней або відстаней подібності [дерево утворюється від листя до стовбура]. При об'єднанні кластерів, що знаходяться на найбільший відстані один від одного, процес припиняється. Дендрограма результату кластеризації представлена на рис. 2.2.

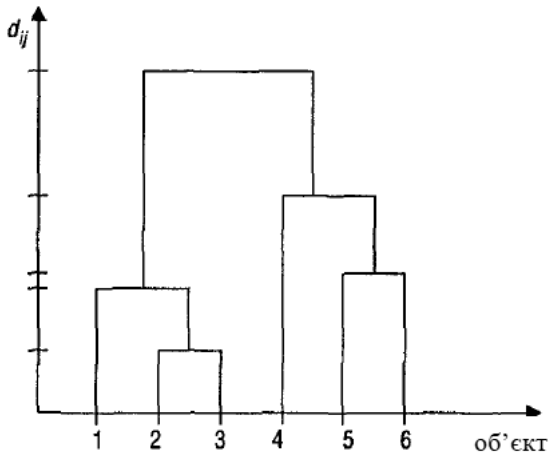



*Рис. 2.2.***Дендрограма ієрархічного агломеративного кластерного аналізу**

На відмінну від агломеритивних методів ієрархічні дивізимні (divisive - роз'єднуючи) методи [дерево формується від стовбура до листя] базуються на первинному припущенні, що усі об'єкти належать до одного кластеру. На наступних етапах від кластеру

від'єднаються групи подібних між собою об'єктів, у зв'язку з чим, кількість кластерів зростає, а відстань між ними зменшується (рис. 2.3.).

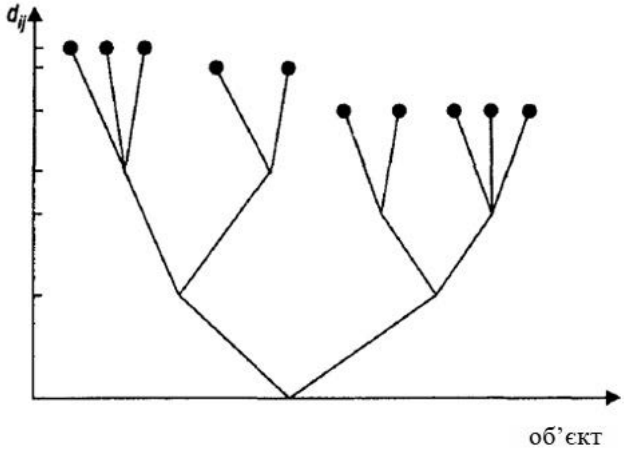
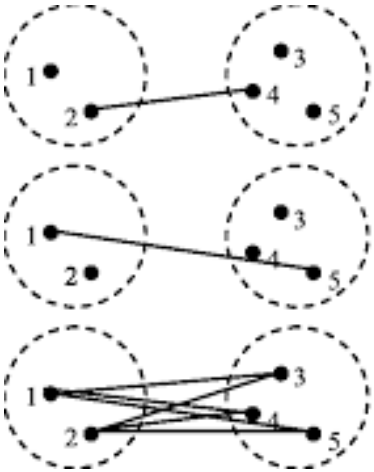



*Рис. 2.3.* **Дендрограма ієрархічного дивізимного кластерного аналізу**

При великий кількості спостережень використовують ітераційні методи дроблення вихідної сукупності У процесі процедури нові кластери формуються до моменту виконання початково заданої умови (кількість утворених кластерів, поріг завершення кластеризації, інше), підпорядкування правилу закінчення процесу. Існують два підходи: перший полягає у визначенні меж кластерів як найбільш щільних ділянок у багатовимірному просторі вихідних даних, тобто визначення кластера там, де є велике "згущення точок"; другий підхід полягає у мінімізації міри відмінності об'єктів. Зазначені методи вимагають від дослідника професійної інтуїції при виборі типу класифікаційних процедур та постановки початкових умов розподілу сукупності у зв'язку з їх чутливістю до змін у заданих параметрах. Застосування ітераційних методів, на відмінну від ієрархічних, може призводити до утворення кластерів, що

перетинаються: один об'єкт одночасно належить до декількох кластерів.




$d_{24}$

$d_{15}$

$$\frac{d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25}}{6}$$

*Рис. 2.4.* **Відстань між класами об'єктів за принципом найближчого сусіда, далекого сусіда та середнього зв'язку**

Проведення процедури класифікації ґрунтується на певних правилах вимірювання відстані між групами об'єктів або міри подібності двох груп об'єктів, які реалізуються за допомогою наступних методів (рис.2.4.):

1) за принципом найближчого сусіда відповідно до правилу одиночного зв'язку (*Single Linkage*, звужує простір - *класи об'єднуються за найближчий межі*):

$$d_{min}(S_l, S_m) = \min_{X_i \in S_l, X_j \in S_m} d\left(X_i, X_j\right) \tag{2.2}$$

За правилом на першому кроці кожне спостереження розглядається як окремий кластер та об'єднуються два найбільш близьких об'єкта, що мають максимальну міру подібності. На наступному до них приєднується об'єкт з максимальною мірою схожості з одним з об'єктів кластеру, тобто для його включення до кластеру потрібна

максимальна подібність лише з одним членом кластеру. Перераховується матриця відстаней, розмірність якої знижується на одиницю. Реалізація алгоритму закінчується, коли усі вихідні спостереження об'єднані в клас. Оскільки відстань між будь-якими двома кластерами в цьому алгоритмі дорівнює відстані між двома найближчими елементами, що представляють свої класи, то отримані в результаті кластери можуть мати досить складну форму, але не повинні бути опуклими. Одночасно перевагою та недоліком алгоритму є потрапляння в один кластер елементів за ланцюжком, що їх з'єднує, близьких між собою.

2) за принципом далекого сусіда до правилу повних зв'язків (*complete Linkage*, розтягує простір - класи поєднуються по дальній межі):

$$d_{max}(S_l, S_m) = \max_{X_i \in S_l, X_j \in S_m} d\left(X_i, X_j\right) \tag{2.3}$$

Відповідно до правилу два об'єкта, що належать до однієї групи, мають коефіцієнт подібності, який менше деякого порогового значення. Саме це порогове значення визначає максимально допустимий діаметр підмножини кластеру. Метод далекого сусіда усуває недоліки притаманні першому методу завдяки визначенню відстані між двома кластерами через відстань між двома найвіддаленішими один від одного представниками своїх кластерів.

3) за принципом середнього сусіда відповідно до правилу середнього зв'язку визначається середня арифметична усіх попарних відстаней між представниками аналізованих груп (*pair-group average*, не змінює простір - об'єкти об'єднуються у відповідності з відстанню до центру класу):

$$d_{max}(S_l, S_m) = \frac{1}{n_i n_m} \sum_{X_l \in S_l, X_j \in S_m} \sum d\left(X_i, X_j\right) \tag{2.4}$$

Правило попарної середньої використовується у незваженому та зваженому вигляді. Перший варіант є доречним у випадку формування об'єктами дійсно відмінних, різних груп у вигляді протяжних кластерів ланцюгових типу; другий – при формуванні нерівних розмірів кластерів, де їх розміри виступають в якості вагових коефіцієнтів.

4) центроїдний метод за яким відстань між двома кластерами визначається через відстань між їхніми центрами тяжіння (*pair-group centroid,* незважена та зважена форми застосування яких ідентично попередньому поясненню):

$$d_{max}(S_l, S_m) = d(\bar{X}_i, \overline{X_m}) \qquad (2.5)$$

5) метод Уорда (*Ward's method*), за яким в якості цільової функції застосовують внутрішньогрупову суму квадратів відхилень, що є сумою квадратів відстаней між кожною точкою (об'єктом) та середньою за кластером, де міститься цей об'єкт. На кожному кроці поєднуються такі два кластери, які призводять до мінімального збільшення цільової функції, тобто внутрішньогрупової суми квадратів відхилень. Завдяки методу об'єднуються близько розташовані кластери за формою подібні до гіперсфер.

Універсальним варіантом розглянутих методів розрахунку є формула відстані між класами, запропонована академіком А. Колмогоровим. Узагальнена відстань ґрунтується на степеневій середній та розраховується за формулою:

$$d_{gen}(S_l, S_m) = \left[\frac{1}{n_l n_m}\sum_{X_i \in S_l}\sum_{X_j \in S_m} d^r\left(X_i, X_j\right)\right]^{1/r} \qquad (2.6)$$

При $r \to \infty$:
$d_{gen}(S_l, S_m) = d_{max}(S_l, S_m)$,

при $r \to -\infty$:

$d_{gen}(S_l, S_m) = d_{min}(S_l, S_m)$,

при $r = 1$:

$d_{gen}(S_l, S_m) = d(S_l, S_m)$.

Якщо $S_{(m,q)} = S_m \cup S_q$ – група елементів, отримана шляхом об'єднання кластерів $S_m$ та $S_q$, узагальнена відстань між кластерами $S_l$ та $S_{(m,q)}$ визначається за формулою:

$$d_{gen}\left(S_l, S_{(m,q)}\right) = \left\{\frac{n_m[d_{gen}(S_l,S_m)]^r + n_q[d_{gen}(S_l,S_q)]^r}{n_m+n_q}\right\}^{1/r} \qquad (2.7)$$

Використання різних методів призводить до різних кластерних структур та вагомо впливає на якість проведення кластеризації. У зв'язку з цим, алгоритм повинен обиратися з врахуванням наявної інформації про існуючу структуру сукупності об'єктів або зважати на вимоги оптимізації математичних критеріїв.

Розрахунок відстані між класами $S_l$ та $S_{(m,q)}$, що отримані шляхом об'єднання двох інших кластерів $S_m$ та $S_q$, здійснюється за формулою Ланса–Вільямса:

$$d_{l,(m,q)} = \rho\left(S_l, S_{m,q}\right) = \alpha d_{lm} + \beta d_{lq} + \gamma d_{mq} + \delta\left|d_{lm} - d_{lq}\right|, \qquad (2.8)$$

де: $d_{lm} = d(S_l, S_m)$, $d_{lq} = d\left(S_l, S_q\right)$, $d_{mq} = d\left(S_m, S_q\right)$ – відстані між класами $S_l, S_m, S_q$,

$\alpha, \beta, \gamma, \delta$ – параметри, значення яких визначає специфіку процедури кластеризації, її алгоритм.

Якщо $\alpha = \beta = \frac{1}{2}, \gamma = 0, \delta = -\frac{1}{2}$, то відстань визначається за методом найближчого сусіда та при проведенні класифікації за коефіцієнтом подібності отримуємо оцінку за самим несхожим об'єктом.

Якщо $\alpha = \beta = \delta = \frac{1}{2}$, $\gamma = 0$, то відстань отримують за методом далекого сусіда та за коефіцієнтом подібності виявляється найбільш схожий об'єкт.

Також використовують наступні варіанти:

- ✓ медіанний, якщо $\alpha = \beta = \frac{1}{2}, \gamma = \frac{1}{4}, \delta = 0$;
- ✓ звичайної середньої, якщо $\alpha = \beta = \frac{1}{2}, \delta = -\frac{1}{2}, \gamma = 0$ ;
- ✓ групової середньої, якщо $\alpha = \frac{n_m}{n_m+n_q}$ , $\beta = \frac{n_q}{n_m+n_q}$ , $\gamma = \delta = 0$ ($n_m$ та $n_q$ – кількість первинних об'єктів у відповідних кластерах);
- ✓ центроїдний, якщо $\alpha = \frac{n_m}{n_m+n_q}$, $\beta = \frac{n_q}{n_m+n_q}, \gamma = \frac{-n_m \cdot n_q}{(n_m+n_q)^2}$, $\delta = 0$.

Відмічається, що найбільш стійким до обмежень є алгоритм звичайного та групового середнього зв'язку.

*Задача 2.1:* провести класифікацію 6 регіонів, кожен з яких характеризується двома ознаками, за методом найближчого сусіда.

*Таблиця 2.2.* **Структура ВДВ за регіонами**

| № регіону | Валова додана вартість, % | |
|---|---|---|
| | Секція В | Секція С |
| 1 | 6,9 | 11,5 |
| 2 | 11,1 | 20,1 |
| 3 | 21,3 | 34,2 |
| 4 | 20,5 | 22,1 |
| 5 | 9,7 | 13,4 |
| 6 | 18,2 | 29,4 |

За допомогою формули Евклідової відстані розрахуємо відстані між об'єктами. Між першим та другим об'єктами відстань визначається:

$$d_{12} = \sqrt{(6{,}9 - 11{,}1)^2 + (11{,}5 - 20{,}1)^2} = 9{,}57$$

між першим та третім об'єктами:

$$d_{13} = \sqrt{(6{,}9 - 21{,}3)^2 + (11{,}5 - 34{,}2)^2} = 26{,}88$$

Звісно, що $d_{11} = d_{22} = \cdots = d_{66} = 0$.

Відповідно обчисляються відстані між іншими регіонами, на підставі яких будується матриця відстаней, що має наступний вигляд:

$$R_1 = \begin{matrix} (1) \\ 0 \\ 9{,}57 \\ 26{,}88 \\ 17{,}24 \\ 3{,}38 \\ 21{,}17 \end{matrix} \begin{matrix} (2) \\ 9{,}57 \\ 0 \\ 17{,}4 \\ 9{,}61 \\ 6{,}85 \\ 11{,}7 \end{matrix} \begin{matrix} (3) \\ 26{,}88 \\ 17{,}4 \\ 0 \\ 12{,}13 \\ 23{,}82 \\ 5{,}71 \end{matrix} \begin{matrix} (4) \\ 17{,}24 \\ 9{,}61 \\ 12{,}13 \\ 0 \\ 13{,}87 \\ 7{,}65 \end{matrix} \begin{matrix} (5) \\ 3{,}38 \\ 6{,}85 \\ 23{,}82 \\ 13{,}87 \\ 0 \\ 18{,}12 \end{matrix} \begin{matrix} (6) \\ 21{,}17 \\ 11{,}7 \\ 5{,}71 \\ 7{,}65 \\ 18{,}12 \\ 0 \end{matrix}$$

На цьому етапі є шість класів та кожний місить по одному об'єкту. На кожному наступному кроці об'єднуємо в один клас складові, що мають найменшу відстань між друг другом. Найбільш близькими є першій та п'ятий об'єкти $d_{15} = 3{,}38$, тому об'єднуємо їх в один кластер. Після об'єднання утворюються п'ять кластерів:

$$S_2, S_3, S_4, S_{1,5}, S_6.$$

Розраховується відстань між за принципом найближчого сусіда за формулою … між кластерами $S_2$ та $S_{1,5}$:

$$d_{2(1,5)} = d(S_2, S_{(1,5)}) = \frac{1}{2} d_{21} + \frac{1}{2} d_{25} - \frac{1}{2} |d_{21} - d_{25}| =$$

$$= \frac{1}{2}(9{,}57 + 6{,}85) - \frac{1}{2}|9{,}57 + 6{,}85| = 6{,}85.$$

Таким чином, відстань $d_{2(1,5)}$ дорівнює відстані від другого об'єкту до найближчого до нього об'єкту, що входить до кластеру $S_{(1,5)}$, тобто $d_{2(1,5)} = d_{25} = 6{,}85$. Матриця відстаней набуває вигляду:

$$R_2 = \begin{array}{c} \begin{matrix} (2) & (3) & (4) & (1{,}5) & (6) \end{matrix} \\ \begin{pmatrix} 0 & 17{,}4 & 9{,}61 & 6{,}85 & 11{,}7 \\ 17{,}4 & 0 & 12{,}13 & 23{,}82 & 5{,}71 \\ 9{,}61 & 12{,}13 & 0 & 13{,}87 & 7{,}65 \\ 6{,}85 & 23{,}82 & 13{,}87 & 0 & 18{,}12 \\ 11{,}7 & 5{,}71 & 7{,}65 & 18{,}12 & 0 \end{pmatrix} \end{array}$$

На цьому кроці об'єднуємо третій та шостий об'єкти між якими найменша відстань $d_{36} = 5{,}71$ та отримуємо чотири кластера:

$$S_2, S_4, S_{(1,5)}, S_{(3,6)}.$$

Знов шукаємо матрицю відстаней та для розрахунку відстані до кластеру $S_{(1,5)}$ скористаємось матрицею відстаней $R_2$, тому що, наприклад, відстань між кластерами $S_{(1,5)}$ та $S_{(3,6)}$ дорівнює:

$$d_{(1,5),(3,6)} = d(S_{(1,5)}, S_{(3,6)}) = \\ = \frac{1}{2}\, d_{(1,5),3} + \frac{1}{2}\, d_{(1,5),6} - \frac{1}{2}\, \left| d_{(1,5),3} - d_{(1,5),6} \right| =$$

$$= \frac{1}{2}(23{,}82 + 18{,}12) - \frac{1}{2}|23{,}82 + 18{,}12| = 18{,}12.$$

Нова матриця відстаней має наступний вигляд:

$$R_3 = \begin{array}{c} \begin{matrix} (2) & (4) & (1{,}5) & (3{,}6) \end{matrix} \\ \begin{pmatrix} 0 & 9{,}61 & 6{,}85 & 11{,}7 \\ 9{,}61 & 0 & 13{,}87 & 7{,}65 \\ 6{,}85 & 13{,}87 & 0 & 18{,}12 \\ 11{,}7 & 7{,}65 & 18{,}12 & 0 \end{pmatrix} \end{array}$$

За найменшою відстанню $d_{(1,5),2} = 6{,}85$ об'єднуємо кластери $S_{(1,5)}$ та $S_2$ та формуємо три кластера:

$$S_{(2,1,5)}, S_4, S_{(3,6)}.$$

На підставі:

$$d_{(2.1,5),(3,6)} = d(S_{(2,1,5)}, S_{(3,6)}) = \frac{1}{2}\, d_{2(3,6)} + \frac{1}{2}\, d_{(1,5),(3,6)} - \\ -\frac{1}{2}\left|d_{2(3,6)} - d_{(1,5),(3,6)}\right| = 11{,}7,$$

будується нова матриця:

$$R_4 = \begin{matrix} (2,1,5) \\ (4) \\ (3,6) \end{matrix} \begin{pmatrix} 0 & 9{,}61 & 11{,}7 \\ 9{,}61 & 0 & 7{,}65 \\ 11{,}7 & 7{,}65 & 0 \end{pmatrix}$$

Останній крок дозволяє об'єднати кластери $S_{(4)}$ та $S_{(3,6)}$ за мінімальною відстанню $d_{(3,6),4} = 7{,}65$ та отримати два кластера, відстань між якими становить:

$$d_{(2.1,5),(4,3,6)} = d(S_{(2,1,5)}, S_{(4,3,6)}) = \frac{1}{2}\, d_{(2,1,5),4} + \frac{1}{2}\, d_{(2,1,5),(3,6)} - \\ -\frac{1}{2}\left|d_{(2,1,5),4} - d_{(2,1,5),(3,6)}\right| = 9{,}61,$$

а матриця виглядає наступним чином:

$$R_5 = \begin{matrix} (2,1,5) \\ (4,3,6) \end{matrix} \begin{pmatrix} 0 & 9{,}61 \\ 9{,}61 & 0 \end{pmatrix}.$$

Результат ієрархічної класифікації об'єктів для візуалізації представлені на дендрограмі, побудованої у пакеті STATISTICA (рис.2.5):

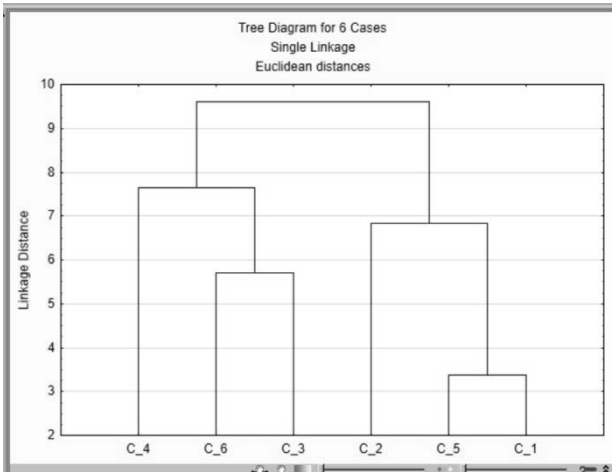



*Рис. 2.5.* **Дендрограма регіонів у програмному середовищі STATISTICA**

*Задача 2.2:* провести класифікацію 5 підприємств, кожне з яких характеризується трьома ознаками, методом далекого сусіда, здійснивши попередню стандартизацію даних (табл. 2.3):

*Таблиця 2.3.* **Показники ефективності підприємств**

| № | Ознаки | | |
|---|---|---|---|
| | Продуктивність праці одного робітника, тис. грн | Прибуток від реалізації продукції, тис. грн. | Капіталоозброєність одного робітника, тис. грн. |
| 1 | 130 | 90 | 65 |
| 2 | 80 | 75 | 90 |
| 3 | 140 | 80 | 100 |
| 4 | 70 | 76 | 80 |
| 5 | 60 | 66 | 110 |
| Середнє значення | 96 | 77,4 | 109 |
| Середнє квадратичне відхилення | 36,5 | 8,7 | 33,2 |

Стандартизуємо первинні дані за формулою:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

У результаті отримуємо матрицю $Z$ нормованих значень змінних:

$$\begin{pmatrix} 0{,}932 & 1{,}447 & -1{,}374 \\ -0{,}438 & -0{,}275 & 0{,}057 \\ 1{,}206 & 0{,}298 & 0{,}629 \\ -0{,}712 & -0{,}160 & -0{,}515 \\ -0{,}987 & -1{,}309 & 1{,}202 \end{pmatrix}$$

Для побудови матриці відстаней використовуємо Евклідову метрику, зокрема розрахунок для першого та другого об'єкту має вигляд:

$$d_{12} = \sqrt{(0{,}932 - (-0{,}438))^2 + (1{,}447 - (-0{,}275))^2 + \big((-1{,}374) - 0{,}057\big)^2} = 2{,}63.$$

На першому кроці матриця характеризує відстані між окремими об'єктами, кожний з яких є окремим кластером:

$$\begin{matrix} (1) & (2) & (3) & (4) & (5) \end{matrix}$$

$$R_1 = \begin{pmatrix} 0 & 2{,}63 & 2{,}33 & 2{,}46 & 4{,}23 \\ 2{,}63 & 0 & 1{,}83 & 0{,}65 & 1{,}64 \\ 2{,}33 & 1{,}83 & 0 & 2{,}28 & 2{,}78 \\ 2{,}46 & 0{,}65 & 2{,}28 & 0 & 2{,}08 \\ 4{,}23 & 1{,}64 & 2{,}78 & 2{,}08 & 0 \end{pmatrix}$$

Проаналізувавши елементи матриці, приходимо до висновку про найбільшу близькість об'єктів другого та четвертого ($d_{24} = 0{,}65$). Об'єднуємо вказані об'єкти в один кластер та перераховуємо відстані між іншими об'єктами та утвореним кластером, якому надаємо номер $S_4$:

$$R_2 = \begin{pmatrix} 0 & 2{,}33 & 2{,}63 & 4{,}23 \\ 2{,}33 & 0 & 2{,}28 & 2{,}78 \\ 2{,}63 & 2{,}28 & 0 & 2{,}08 \\ 4{,}23 & 2{,}78 & 2{,}08 & 0 \end{pmatrix}$$

Зокрема, відстань між першим об'єктом та новим кластером $S_4$ за алгоритмом далекого сусіда визначається наступним чином:

$$d_{S1,S4} = \max(d_{12}, d_{14}) = \max(2{,}63; 2{,}46) = 2{,}63$$

У побудованій матриці $R_2$ за мінімальною відстанню $d_{45} = 2{,}08$ знов знаходимо найближчі кластери, якими є $S_4$ та $S_5$. Об'єднуючи складові, утворюємо новий кластер з об'єктами другим, четвертим та п'ятим, присвоюємо йому $S_4$. В результаті виконаних кроків отримуємо три кластера $S_1(1)$, $S_2(2,4,5)$ та $S_3(3)$.

Перераховуємо відстані $d_{12}$ та $d_{23}$:

$$d_{1,4} = \max(d_{14}, d_{15}) = \max(2{,}63; 4{,}23) = 4{,}23$$

$$d_{3,4} = \max(d_{34}, d_{35}) = \max(2{,}28; 2{,}78) = 2{,}78$$

та отримуємо матрицю:

$$R_3 = \begin{pmatrix} 0 & 2{,}33 & 4{,}23 \\ 2{,}33 & 0 & 2{,}78 \\ 4{,}23 & 2{,}78 & 0 \end{pmatrix}$$

Виходячи зі значень відстаней об'єднуємо кластери $S_1$та $S_3$ ($d_{13}$=2,33) та надаємо йому номер $S_1$. В результаті процедури утворені два кластера: $S_1$, що об'єднує перший та третій об'єкти, та

$S_4$, у складі якого другий, четвертий та п'ятий об'єкти. На останньому етапі:

$$d_{1,4} = \max(d_{14}, d_{34}) = \max(4{,}23;\ 2{,}78) = 4{,}23$$

$$R_4 = \begin{pmatrix} 0 & 4{,}23 \\ 4{,}23 & 0 \end{pmatrix}$$

Останній крок дозволяє об'єднати кластери $S_1$ та $S_4$ на відстані 4,23 та представити підсумок класифікації у вигляді дендрограми, вигляд якої свідчить про більшу однорідність кластера $S_4$. Це пояснюється процесом об'єднання об'єктів у четвертий кластер за меншими відстанями, ніж у перший.

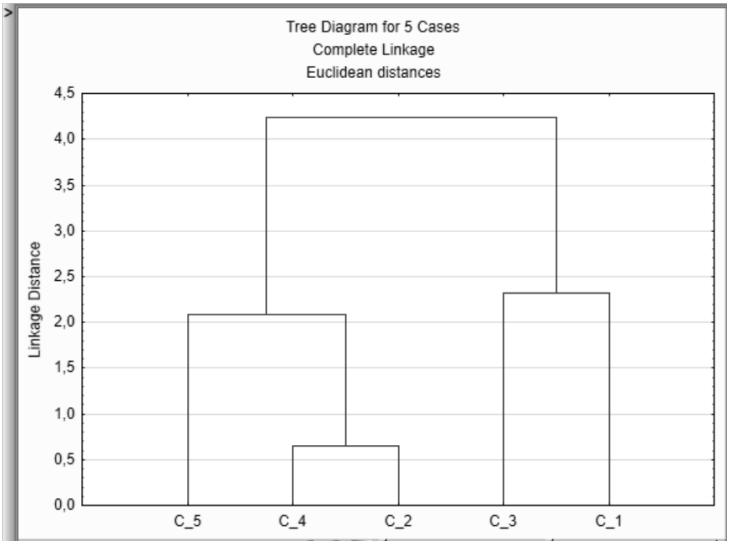



*Рис. 2.6.* **Дендрограма підприємств**

*Задача 2.3:* класифікувати об'єкти за методом *Уорда*, виділити три групи. Десять спостережень характеризуються ознаками (0 – ні, 1 –так): save – наявність заощаджень, income – наявність доходів, property – наявність власності, house – наявність житла, cattle – наявність худоби, land – наявність землі у власності:

| № п/п | save | income | property | house | cattle | land |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 3 | 0 | 1 | 1 | 0 | 0 | 0 |
| 4 | 0 | 1 | 0 | 0 | 0 | 0 |
| 5 | 1 | 1 | 1 | 1 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 1 | 1 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 1 | 0 | 0 |
| 10 | 0 | 1 | 1 | 0 | 0 | 1 |

Усі змінні характеризуються ознаками номінальної шкалою, для проведення кластерного аналізу побудуємо матрицю подібностей об'єктів за допомогою міри у вигляді коефіцієнту асоціативності *Жаккара* за формулою:

$$J = \frac{a}{a+b+c} \qquad (2.9)$$

Коефіцієнт Жаккара визначається на підставі таблиць взаємної спряженості, де 1 вказує на наявність змінної, 0 – відсутність. Коефіцієнт змінюється від 0 до 1 та приймає до уваги лише ознаки, які характерні хоча б для одного з об'єктів:

| | 1 | 0 |
|---|---|---|
| 1 | a | b |
| 0 | c | d |

Наприклад, для 1-го та 2-го об'єктів:

| | 1 | 0 |
|---|---|---|
| 1 | 2 | 0 |
| 0 | 4 | 0 |

$$J_{12} = \frac{2}{2+0+4}$$

для 2-го та 5-го об'єктів:

| | 1 | 0 |
|---|---|---|
| 1 | 4 | 2 |
| 0 | 0 | 0 |

$$J_{25} = \frac{4}{4+2+0}$$

Матриця подібності усіх об'єктів має наступний вигляд:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0,33 | 1 | 0,5 | 0,5 | 0 | 0 | 1 | 0 | 0,67 |
| 2 | 0,33 | 0 | 0,33 | 0,17 | 0,67 | 0 | 0 | 0,33 | 0,17 | 0,5 |
| 3 | 1 | 0,33 | 0 | 0,5 | 0,5 | 0 | 0 | 1 | 0 | 0,67 |
| 4 | 0,5 | 0,17 | 0,5 | 0 | 0,25 | 0 | 0 | 0,5 | 0 | 0,33 |
| 5 | 0,5 | 0,67 | 0,5 | 0,25 | 0 | 0 | 0 | 0,5 | 0,25 | 0,4 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 1 | 0,33 | 1 | 0,5 | 0,5 | 0 | 0 | 0 | 0 | 0,67 |
| 9 | 0 | 0,17 | 0 | 0 | 0,25 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0,67 | 0,5 | 0,67 | 0,33 | 0,4 | 0 | 0 | 0,67 | 0 | 0 |

Якщо кожний елемент матриці відняти від 1, отримуємо матрицю відстаней між об'єктами:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0,67 | 0 | 0,5 | 0,5 | 1 | 1 | 0 | 1 | 0,67 |
| 2 | 0,67 | 1 | 0,67 | 0,83 | 0,33 | 1 | 1 | 0,67 | 0,83 | 0,5 |
| 3 | 1 | 0,67 | 1 | 0,5 | 0,5 | 1 | 1 | 0 | 1 | 0,67 |
| 4 | 0,5 | 0,83 | 0,5 | 1 | 0,75 | 1 | 1 | 0,5 | 1 | 0,33 |
| 5 | 0,5 | 0,33 | 0,5 | 0,75 | 1 | 1 | 1 | 0,5 | 0,75 | 0,4 |
| 6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 8 | 0 | 0,67 | 0 | 0,5 | 0,5 | 1 | 1 | 1 | 1 | 0,67 |
| 9 | 1 | 0,83 | 1 | 1 | 0,75 | 1 | 1 | 1 | 1 | 1 |
| 10 | 0,33 | 0,5 | 0,33 | 0,67 | 0,6 | 1 | 1 | 0,33 | 1 | 1 |

Кластеризацію проведемо за допомогою пакету STATISTICA за матрицею відстаней методом Уорда. Наведена дендрограма наочно демонструє розділення об'єктів на заданих три кластера: 1-й кластер об'єднує 1, 3 , 4, 8, 10; 2-й – 6, 7; 3-й – 2, 5, 9.

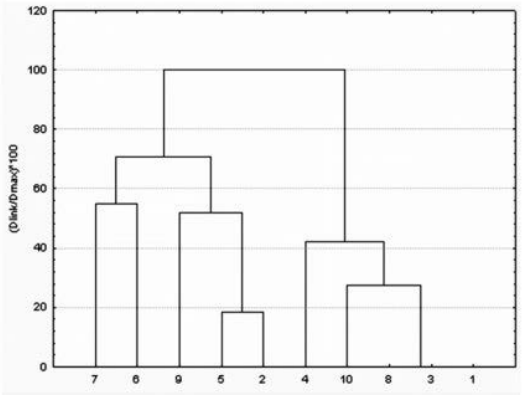



*Рис.2.7.* **Дендрограма об'єктів за методом Уорда**

*Задача 2.4:* за матрицею відстаней між п'ятьма об'єктами $X_1, \dots, X_5$ провести класифікацію дивізимним методом.

$$R = \begin{array}{c} \begin{matrix} (1) & (2) & (3) & (4) & (5) \end{matrix} \\ \begin{pmatrix} 0 & 4{,}49 & 2{,}16 & 3{,}53 & 3{,}24 \\ 4{,}49 & 0 & 3{,}26 & 1{,}92 & 1{,}93 \\ 2{,}16 & 3{,}26 & 0 & 2{,}68 & 2{,}74 \\ 3{,}53 & 1{,}92 & 2{,}68 & 0 & 0{,}71 \\ 3{,}24 & 1{,}93 & 2{,}74 & 0{,}71 & 0 \end{pmatrix} \end{array}$$

Відповідно до процедури проведення кластеризації за дивізимним методом аналізуємо відстані між собою та звертаємо увагу на найбільш віддалені об'єкти $X_1$та $X_2$ ($d_{12} = 4{,}49$).Оцінюємо відстань інших об'єктів до першого та другого:

$d_{31} < d_{32}$ – об'єкт $X_3$ ближче до $X_1$;

$d_{41} > d_{42}$ – об'єкт $X_4$ ближче до $X_2$;

$d_{51} < d_{52}$ – об'єкт $X_5$ ближче до $X_2$.

Виходячи з порівняння, утворюємо два кластера: $S_1(1,3)$ та $S_2(2,4,5)$. У кожному з них також аналізуємо відстань між об'єктами та на певному кроці відбувається розподілення того кластеру, де досягається максимум відстані між об'єктами:

$$d_{13} = 2{,}16 \quad d_{25} = 1{,}93 \quad d_{24} = 1{,}92 \quad d_{45} = 0{,}71$$

Найбільша відстань $d_{13} = 2{,}16$, відповідно об'єкти $X_1$та $X_3$ відокремлюємо в єдиний клас. У кластері $S_2(2,4,5)$ шукаємо максимальну відстань $\max(d_{25}, d_{24}, d_{45}) = 1{,}93$. Наступний крок дозволяє відокремити об'єкт $X_2$, останній – розділити кластер $S_4(4,5)$ на два кластера на відстані 0,71 (рис. 2.8).

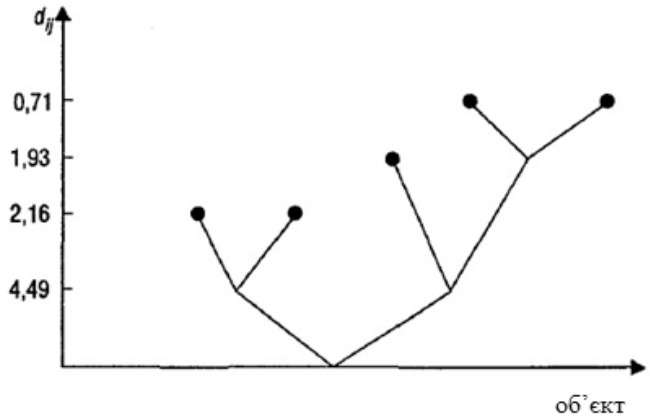



*Рис.2.8.* **Дендрограма кластеризації об'єктів дивізимним методом**

До ітераційної кластеризації відносять метод *k-середніх,* який базується на принципі мінімізації внутрикластерної дисперсії. Цей метод відноситься до процедур еталонного типу та процес його

реалізації починається зі встановлення певних початкових умов (кількості утворених класів, порогу завершення класифікації, інше). Метод *k-середніх* на відмінну від ієрархічних процедур не вимагає розрахунку та збереження матриці відстаней або подібностей між об'єктами, його алгоритм передбачає використання лише первинних, вихідних значень змінних. Для початку кластеризації повинні бути задані $k$ обраних об'єктів, що і будуть слугувати еталонами, тобто центрами кластерів. Метод еталонного типу характеризується як зручний та швидкодіючий.

Саме найбільш відповідальним та принциповим є вибір початкових умов, які впливають на тривалість процесу та результати класифікації. Його застосування доречно для обробки великих статистичних сукупностей. В той же час, метод *k-середніх* чутливо реагує на зміну заданих параметрів та вимагає від дослідника інтуїції при виборі типу класифікаційних процедур та вихідних параметрів. Ускладнення процедури може посилитися через невдалий вибір випадково обраної кількості кластерів, також можливо утворення розмитих чи мало заповнених кластерів. Доцільним вважається попередня кластеризація за допомогою одного з методів ієрархічного аналізу або на підставі експертних оцінок, а лише після цього вирішується питання про початкове розділення та статистичні критерії ітераційного алгоритму. Виконання декількох алгоритмів зі зміною кількості кластерів або порогу близькості для об'єднання об'єктів, створює можливість для вирішення питання про структуру вивчаємої сукупності та дозволяє обрати найкращий варіант розбиття за заданим критерієм.

Відповідно до алгоритму, запропонованого Дж. Мак-Куїном, математичний його опис наступний: маємо $n$ спостережень, кожне з яких характеризується $p$ ознаками $X_1, X_2, \dots, X_p$, які необхідно розділити на $k$ кластерів. Випадковим чином або дослідником з пеаних апріорних міркувань задаються $k$ точки об'єктів з $n$ крапок сукупності. Такі точки приймаються за еталон та кожному еталону

присвоюється порядковий номер, який одночасно слугує номером кластеру. На першому кроці з тих, що залишилися *(n-k)* об'єктів вилучається точка $X_i$ з координатами $x_{i1}, x_{i2}, \dots, x_{ip}$ та перевіряється до якого з еталонів (центрів) вона знаходиться найближче. Для цього використовується одна з розглянутих метрик (наприклад, евклідова відстань).

Об'єкт, що перевіряється, приєднується до того центру (еталону), якому відповідає мінімальна з відстаней. Еталон замінюється новим, перерахованим з урахуванням приєднаної точки, та кількість об'єктів, що входять до кластеру, збільшується на одиницю. Якщо зустрічаються дві або більше мінімальні відстані, то *і-ий* об'єкт приєднують до центру з найменшим порядковим номером. На наступному кроці обираємо точку $X_{i+1}$ та для неї повторюються всі процедури. Таким чином, через *(n-k)* кроків всі точки (об'єкти) сукупності виявляться віднесеними до одного з кластерів, але на цьому процес розбиття не закінчується. Для того щоб досягти стійкості розбиття за тим самим правилом, усі точки $X_1, X_2, \dots, X_n$ знову приєднують до отриманих кластерів, при цьому ваги продовжують накопичуватися. Нове розбиття порівнюється із попереднім. Якщо вони збігаються, робота алгоритму завершується; інакше цикл повторюється. Остаточне розбиття має центри тяжіння, які не збігаються з еталонами, їх можна позначити $C_1, C_2, \dots, C_k$, при цьому кожна точка $X_i (i = 1, 2, \dots, n)$ буде відноситися до такого кластера (класу) $l$, для якого відстань мінімальна.

Можливі дві модифікації методу *k-середніх*. Перша передбачає перерахунок центру тяжіння кластеру після кожної зміни його складу, а друга – лише після завершення перегляду всіх даних. В обох випадках ітеративний алгоритм цього методу мінімізує дисперсію всередині кожного кластера, хоча в явному вигляді такий критерій оптимізації не використовується.

Перевірка якості кластеризації за методом *k-середніх* базується на перерахунку середніх значень кожного кластеру та вибір варіанту з найбільшими відмінностями їх величини. До переваг методу відносять простоту та швидкість використання, зрозумілість та прозорість процедури; до недоліків – чутливість до викидів, які можуть викривляти середню (рекомендовано використання модифікації алгоритму у вигляді *k-медіани*), та повільна реалізація на великих масивах інформації (прискорення досягається при використанні на вибірках даних).

Загальний алгоритм більшості ітеративних методів класифікації реалізується наступним чином:

*Крок 1.* Вибір кількості кластерів на які розбивається сукупність, постановка завдання початкового розподілення об'єктів та визначення центрів тяжіння кластерів. Відбір початкових центроїдів базується на одному з наступних варіантах встановлення *k*-спостережень: випадковий, максимізація початкової відстані або вибір перших *k*-спостережень.

*Крок 2.* Визначення нового складу кожного кластеру відповідно до обраних мір подібності.

*Крок 3.* Перерахунок центрів тяжіння кластерів після перегляду усіх об'єктів та розподілу їх за кластерами.

*Крок 4.* Кроки 2 та 3 повторюються доки завдяки наступної ітерації не отримують склад кластерів, як на попередній або відповідають максимальній кількості ітерацій.

*Задача 2.5:* класифікувати об'єкти за чотирма змінними методом *k-середніх*, виділити дві групи.

| Об'єкти | A | B | C | D |
|---|---|---|---|---|
| $x_1$ | 5 | -1 | 1 | -3 |
| $x_2$ | 3 | 1 | -2 | -2 |

Довільно на першому кроці розподіляємо об’єкти на дві групи AB та CD та визначаємо координати центрів кластерів:

| Кластер | Координати центру | |
|---|---|---|
| | $\bar{x}_1$ | $\bar{x}_2$ |
| (AB) | $\frac{5+(-1)}{2}=2$ | $\frac{3+1}{2}=2$ |
| (CD) | $\frac{1+(-3)}{2}=-1$ | $\frac{-2+(-2)}{2}=-2$ |

На другому кроці за допомогою евклідової метрики визначаємо відстань кожного об’єкту до центру кластерів та перерозподіляємо їх у найближчу групу:

$$d^2\big(A,(AB)\big)=(5-2)^2+(3-2)^2=10,$$

$$d^2\big(A,(CD)\big)=(5+1)^2+(3+2)^2=61.$$

За результатами розрахунку відстань від об’єкту A до центру кластеру AB найменша, тому цей об’єкт залишається у ньому.

Для об’єкту B:

$$d^2\big(B,(AB)\big)=(-1-2)^2+(1-2)^2=10,$$

$$d^2\big(B,(CD)\big)=(-1+1)^2+(1+2)^2=9.$$

Об’єкт B ближче до кластеру CD та переміщається у цей кластер. Для груп (A) та (BCD) перераховуємо координати центрів:

| Кластер | Координати центру | |
|---|---|---|
| | $\bar{x}_1$ | $\bar{x}_2$ |
| (A) | 5 | 3 |
| (BCD) | $\frac{-1+1+(-3)}{3} = -1$ | $\frac{1-2+(-2)}{3} = -1$ |

Знов повторюємо процедуру перерахунку евклідової відстані кожного об'єкту до центу кластерів:

$$d^2\big(A,(BCD)\big) = (5-(-1))^2 + (3-(-1))^2 = 52,$$

$d^2\big(A,(A)\big) = 0 \rightarrow$ об'єкт A залишається у групі A;

$$d^2\big(B,(A)\big) = (5-(-1))^2 + (3-1))^2 = 40,$$

$$d^2\big(B,(BCD)\big) = (-1-(-1))^2 + (1-(-1))^2 = 4$$

→ об'єкт B залишається у групі BCD;

$$d^2\big(C,(A)\big) = (5-1)^2 + (3-(-2))^2 = 41,$$

$$d^2\big(C,(BCD)\big) = \big(-1-(-1)\big)^2 + (-2-(-1))^2 = 5$$

→ об'єкт C залишається у групі BCD;

$$d^2\big(D,(A)\big) = (5-(-3))^2 + (3-(-2))^2 = 89,$$

$$d^2\big(D,(BCD)\big) = \big(-3-(-1)\big)^2 + (-2-(-1))^2 = 5$$

→ об'єкт D залишається у групі BCD.

Отримана класифікація співпадає з отриманою на попередньому етапі, оскільки жоден з об'єктів не перемістився, таким чином ітеративний процес є завершеним та фінальна класифікація має дві групи: A та BCD.

Алгоритм подвійного об'єднання (Two-way joining) або блокової кластеризації базується на одночасному угрупуванні, як об'єктів, так і змінних. Теоретична сторона методу подвійного об'єднання розроблена Дж. А. Хартиганом[4]. Центральне місце в процедурі займає граничне значення параметра подвійного об'єднання, яке визначає, коли алгоритм одночасно врахує близькість в обох характеристиках досліджуваної сукупності. Воно задається дослідником та визначає кількість блоків: якщо граничне значення велике, то буде сформований тільки один блок у просторі «об'єкти-змінні». Якщо ж воно мале, то, навпаки, виділяється велика кількість таких блоків. Зазвичай система рекомендує величину граничного значення подвійного об'єднання на рівні половини стандартного відхилення всіх спостережень, тобто приблизно 0,5. Метод блокової кластеризації може бути корисний як допоміжний інструмент багатовимірного угрупування, коли основна процедура класифікації об'єктів уже виконана (наприклад, на базі ієрархічних алгоритмів, методу k-середніх і под.) та додатково необхідно з'ясувати, як у кожному з утворених кластерів розподілені значення чинників-симптомів. Алгоритм подвійного об'єднання надає можливість ідентифікувати типи, класи об'єктів за рівнем прихованої ознаки економічних об'єктів на основі аналізу величини тих чинників-симптомів, що складають інформаційну базу дослідження. У ході використання алгоритму подвійного об'єднання слід орієнтуватися на число кластерів, отримане на стадії основної процедури класифікації, підбираючи за допомогою граничного значення параметра відповідне число блоків у просторі «об'єкти-змінні»[5].

Метод пошуку згущень, який також відноситься до ітераційних методів, на відмінні від *k-середніх*, не вимагає

[4] Янковий О.Г. Латентні ознаки в економіці: монографія. Одеса: Атлант, 2015. 168 с.
[5] Там же.

попередньої визначеності кількості класів. Найчастіше в методі згущення об'єктів використовується плинна середня, що ґрунтується на центрі тяжіння.

Реалізація алгоритму відбувається за наступними кроками:

*Крок 1.* На підставі матриці відстаней або подібності обирається об'єкт, який приймається за початковий, умовний центр тяжіння першого класу. Вибір об'єкту ґрунтується на попередньому аналізі точок та їх оточення або здійснюється довільно.

*Крок 2.* Навколо обраного центру в межах заданого радіусу відбирається сукупність елементів, що потрапили всередину гіперсфери, та за ними визначається фактичний центр тяжіння шляхом обчислення його координат за середньою арифметичною простою.

*Крок 3.* Обираються елементи в межах радіуса навколо нового центру тяжіння та обчислюється середня для нової сукупності.

*Крок 4.* Якщо значення середніх за процедурою збігаються, то кластер вважається сформованим. Повторення результатів перерахунку свідчить про припинення переміщення центу сфери та завершення етапу утворення чергового кластеру.

*Крок 6.* Відібрані точки вилучаються з подальшого процесу кластеризації, для інших вибирають нову довільну точку та процедура формування нового кластера повторюється за описаними етапами.

Кількість кластерів залежить від радіусу, тому класифікацію за методом згущень доцільно проводити з використанням різних радіусів. Також для підвищення якості кластеризації оцінюється наповненість груп та контролюється внутрішньогрупова і міжгрупова варіація[6].

[6] Яровий А.Т., Страхов Є.М. Багатовимірний статистичний аналіз: навчально-методичний посібник для студентів математичних та економічних фахів. Одеса: Астропринт, 2015. 132 с.

*Задача 2.6:* провести класифікацію методом пошуку згущень.

| Елементи | 1-ша ознака | 2-га ознака |
|---|---|---|
| 1 | 1,8 | 5,3 |
| **2** | **2,7** | **5,5** |
| 3 | 1,8 | 4,8 |
| 4 | 2,5 | 4,8 |
| 5 | 3,0 | 5,0 |

Визначаємо радіус $R$=1 та умовним центром тяжіння обирається другий елемент. Обчислюємо відстані до усіх об'єктів за формулою мангеттенської відстані:

$$d_{ij} = \sum_{k=1}^{m} |x_{ik} - x_{jk}|$$

$c_{12} = |1{,}8 - 2{,}7| + |5{,}3 - 5{,}5| = 1{,}1$ $\quad c_{12} > R$ $\quad 1{,}1 > 1$
$c_{32} = |1{,}8 - 2{,}7| + |4{,}8 - 5{,}5| = 1{,}6$ $\quad c_{32} > R$ $\quad 1{,}6 > 1$
$c_{42} = |2{,}5 - 2{,}7| + |4{,}8 - 5{,}5| = 0{,}9$ $\quad c_{42} < R$ $\quad 0{,}9 < 1$
$c_{52} = |3{,}0 - 2{,}7| + |5{,}0 - 5{,}5| = 0{,}8$ $\quad c_{52} < R$ $\quad 0{,}8 < 1$

Встановленому радіусу відповідають лише четвертий та п'ятий елементи. Розраховуємо фактичний центр тяжіння:

$$\bar{z}_1 = \frac{2{,}7 + 2{,}5 + 3{,}0}{3} = 2{,}7 \qquad \bar{z}_2 = \frac{5{,}5 + 4{,}8 + 5{,}0}{3} = 5{,}1$$

Перераховуємо відстань від фактичного центру тяжіння до усіх об'єктів:

$c_{12} = |1{,}8 - 2{,}7| + |5{,}3 - 5{,}1| = 1{,}1$ $\quad c_{12} > R$ $\quad 1{,}1 > 1$
$c_{32} = |1{,}8 - 2{,}7| + |4{,}8 - 5{,}1| = 1{,}2$ $\quad c_{32} > R$ $\quad 1{,}6 > 1$
$c_{42} = |2{,}5 - 2{,}7| + |4{,}8 - 5{,}1| = 0{,}5$ $\quad c_{42} < R$ $\quad 0{,}9 < 1$
$c_{52} = |3{,}0 - 2{,}7| + |5{,}0 - 5{,}1| = 0{,}4$ $\quad c_{52} < R$ $\quad 0{,}8 < 1$

За результатом приходимо до висновку, що у сферу потрапили ті ж самі об'єкти 2, 4 і 5, які і формують перший кластер.

У зв'язку з великою кількістю різних способів розділення сукупності елементів на класи та отримання різних варіантів кластеризації, доцільним вважається порівняльний аналіз їх якості. У процедурі застосовується поняття функціонал якості розбиття $Q(S)$, який відрізняється відповідно для кожного застосованого методу класифікації. Принциповим у виборі є досягнення екстремуму (максимум або мінімум) обраного функціоналу, та, як правило, базується на емпіричних міркуваннях.

На практиці будь-яку сукупність у просторі $X$ спостережень (об'єктів) $X_1, X_2, \dots, X_n$ та $S = (S_1, S_2, \dots, S_p)$ за допомогою метрики $d$ можна розділити на задану кількість $p$ класів $S_1, S_2, \dots, S_p$. До найбільш розповсюджених характеристик фунціоналів якості класифікації відносяться (рекомендується вибір варіанту з мінімальним значенням описаних критеріїв):

- ✓ сума внутрішньокласових дисперсій:

$$Q_1(S) = \sum_{l=1}^{p} \sum_{x_i \in S_l} d^2(x_i, \bar{x}_l) \tag{2.10}$$

де $l$ – номер класу, $\bar{x}_l$ – центр *l-го* класу, $x_i$ – вектор значень змінних для *i*-го об'єкту, що входить до *l-го* класу, $\rho(x_i, \bar{x}_l)$ – відстань між *i-им* об'єктом та центром *l-го* класу.

- ✓ сума попарних внутрішньокласових дисперсій між елементами:

$$Q_2'(S) = \sum_{l=1}^{p} \sum_{x_i, x_j \in S_l} d^2(x_i, x_j) \tag{2.11}$$

$$Q_2{}'(S) = \sum_{l=1}^{p} \frac{1}{n_l} \sum_{x_i, x_j \in S_l} d^2(x_i, x_j) \tag{2.12}$$

- ✓ загальна внутрішньокласова дисперсія

$$Q_3(S) = \det(\sum_{l=1}^{p} n_l W_l) \tag{2.13}$$

де: $\det A$ - визначник матриці А;

$W_l$ – вибіркова коваріаційна матриця класу $S_l$, елементи якої визначаються за формулою:

$W_{qm}(l) = \frac{1}{n_l} \sum_{x_i \in S_l} (x_{iq} - \bar{x}_q)(x_{im} - \bar{x}_m), \quad q, m = 1, 2, \dots, k,$

де: $x_{iq}$ – *q-ma* компонента багатомірного спостереження $x_i$,

$\bar{x}_q$ – середнє значення *q-oї* компоненти, розрахована за спостереженнями *l-го* класу.

Якість класифікації характеризують також іншим видом узагальненої дисперсії, у якій операція сумування $W_l$ замінена операцією множення:

$$Q_4(S) = \prod_{l=1}^{p} (\det W_l)^{n_l} \tag{2.14}$$

Функціонали $Q_3(S)$та $Q_4(S)$ зазвичай використовують при вирішенні питання щодо зосередженості спостережень, які розподілені на класи, у просторі з розмірністю меншою за $k$.

*Задача 2.7.* Порівняти результати кластеризації 6 об'єктів за ієрархічними методами: *k-середніх* та далекого сусіда.

За методом *k-середніх* отримані три кластера:

кластер $S_1$ - об'єкт $n_1$,

кластер $S_2$ - об'єкти $n_2$ та $n_6$,

кластер $S_3$ - об'єкти $n_3$, $n_4$ та $n_5$.

Отримані дисперсії для кожної змінної у кожному кластері $\sigma_{ij}^2$:

| | | |
|---|---|---|
| $\sigma_{11}^2 = 0$ | $\sigma_{21}^2 = 0,004225$ | $\sigma_{31}^2 = 0,008422$ |
| $\sigma_{12}^2 = 0$ | $\sigma_{22}^2 = 0,25$ | $\sigma_{32}^2 = 0,666$ |
| $\sigma_{13}^2 = 0$ | $\sigma_{23}^2 = 0,2$ | $\sigma_{33}^2 = 0,1866$ |
| $\sum_{j=1}^{3} \sigma_{1j}^2 = 0$ | $\sum_{j=1}^{3} \sigma_{2j}^2 = 0,454225$ | $\sum_{j=1}^{3} \sigma_{3j}^2 = 0,861022$ |

Сумарна дисперсія усіх змінних за трьома кластерами становить:

$$Q_1^{k-\text{середніх}}(S) = \sum_{j=1}^{3} \sigma_{1j}^2 + \sum_{j=1}^{3} \sigma_{2j}^2 + \sum_{j=1}^{3} \sigma_{3j}^2 = 1,315247$$

За методом далекого сусіда утворилось також три кластера, але з іншим поділом елементів:

кластер $S_1$ - об'єкт $n_1$ та $n_4$,

кластер $S_2$ - об'єкти $n_2$ та $n_6$,

кластер $S_3$ - об'єкти $n_3$ та $n_5$.

У цьому випадку сумарна дисперсія усіх змінних:

$$Q_1^{\text{далекого сусіда}}(S) = \sum_{j=1}^{3} \sigma_{1j}^2 + \sum_{j=1}^{3} \sigma_{2j}^2 + \sum_{j=1}^{3} \sigma_{3j}^2 = 0,5016 + 0,4542 + 0,8617 = 1,8175.$$

Порівнюючи значення дисперсій, приходимо до висновку про кращу якість кластеризації за методом *k-середніх,* ніж за далеким сусідом.

У розглянутому прикладі функціонал $Q_i$ оцінює міру однорідності усіх кластері у цілому, а окремі значення $\sigma_1^2$ та $\sigma_2^2$ - кожного з кластерів.

Оцінювання якості кластеризації визначає проблематику порівняння результатів аналізу у наступних аспектах:

- ✓ евристичний – характеризується відсутністю формальної моделі для співставлення різних варіантів, алгоритм обривається на підставі інтуїтивних міркувань;
- ✓ екстремальний – задається критерій, який визначає параметри розділення на класи;
- ✓ статистичний – завдання аналізу реалізуються на основі імовірнісної моделі процесу дослідження.

Властивості кластерів дозволяють візуально зіставляти результати кластеризації. До них відносяться наступні характеристики:

1. *Щільність розподілу спостережень у середині кластеру* – дає можливість визначити наповненість кластеру або його ненасиченість, порожність. У зв'язку з відсутністю певного способу обчислення цього параметру компактності кластеру, використовують дисперсію відстані від його центру до окремих точок за принципом: чим менша дисперсія відстані, тим ближче до центру знаходяться одиниці спостереження, тим більша щільність кластеру; чим більша дисперсія відстані, тим, навпаки, менш наповнений кластер та об'єкти розміщуються як близько до центру, так і далеко від нього.
2. *Розмір кластеру* – відображає фактичну величину кластеру за допомогою радіусу, якщо він має круглу форму або є гіперсферою у багатовимірному просторі. У випадку подовженої форми, радіус або діаметр не відображає справжні розміри кластерів.
3. *Локальність, відокремленість кластерів* – характеризує ступінь перекриття та взаємної віддаленості кластерів один від одного у багатовимірному просторі. Як властивість створює умови для перетворення кластерів: або об'єднання найближчих кластерів та їх частин, що перекриваються, або відокремлення від кластера елементів, що більш віддалені від його центру, та інше.

Розглянуті методи відносяться до класичних, але у зв'язку з появою великих масивів даних, змінюються вимоги до алгоритмів кластеризації (масштабованість алгоритму). У нових розробках ієрархічні методи інтегровані з іншими, до них відносяться BIRCH, CURE, WaveCluster, CLARA, Clarans.

Алгоритм BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) реалізується у два етапу: на першому – формується попередній набір кластерів; на другому – до утворених кластерів застосовуються інші методи кластеризації. Аналогія алгоритму описується наступним чином: якщо кожен елемент даних уявити собі як намистину, що лежить на поверхні столу, то кластери намистин можна "замінити" тенісними кульками та перейти до детальнішого вивчення кластерів тенісних кульок: кількість намистин може бути досить велика, проте діаметр тенісних кульок підбирається таким чином, щоб на другому етапі при застосуванні традиційних алгоритмів кластеризації, визначити дійсну складну форму кластерів. Метод об'єднує збалансоване ітераційне скорочення та ієрархічну кластеризацію [7].

WaveCluster є алгоритмом кластеризації на основі хвильових перетворень. На початку роботи дані узагальнюються шляхом накладання простору даних багатовимірної решітки. На подальших кроках алгоритму аналізуються не окремі точки, а узагальнені їх характеристики, що потрапили до одного осередку решітки. У результаті такого узагальнення необхідна інформація міститься в оперативній пам'яті. На наступних кроках визначення кластерів застосовується хвильове перетворення до узагальнених даних. Основні особливості WaveCluster: складність реалізації; алгоритм може виявляти кластери довільних форм; алгоритм нечутливий до

[7] Nayyar, A., Puri, V. 2017. Comprehensive analysis & performance comparison of clustering algorithms for big data. Review of Computer Engineering Research, Vol. 4, No 2, p. 54-80.

шумів; застосовується лише до даних низької розмірності; ефективен з точки зору компромісу часової складності[8].

Алгоритм CLARA (Clustering LARge Applications) отримує безліч зразків з бази даних. Кластеризація застосовується до кожного зразків, на виході алгоритму пропонується найкраща кластеризація. Ефективність алгоритму залежить від обраного як зразка набору даних. Якість кластеризації на вибраному наборі не завжди дає хорошу кластеризацію на всій кількості даних.

Алгоритм Clarans (Clustering Large Applications based upon RANdomized Search) формулює завдання кластеризації як випадковий пошук у графах. В результаті роботи цього алгоритму сукупність вузлів графа є розбиттям безлічі даних на число кластерів, визначених користувачем. Якість отриманих кластерів визначається за допомогою критеріальної функції. Алгоритм Clarans сортує всі можливі розбиття множини даних у пошуках прийнятного рішення. Пошук рішення зупиняється у тому вузлі, де досягається мінімум серед певного числа локальних мінімумів.

Серед нових масштабованих алгоритмів можна також відзначити алгоритм CHAMELEON – ієрархічний метод на основі динамічного моделювання, CURE – має перевагу по надійності та виявленню скупчень, що мають несферичну форму та відхилення у розмірі, ROCK – розгортає зв'язки, а не відстані при виконанні об'єднань кластерів, Echidna працює з на усіх видах параметрів мережевого трафіку, DBScan, у якому поняття кластера формулюється з допомогою концепції щільності (density) та дозволяє ідентифікувати шум у просторовій базі даних, DENCLUE об'єднує функції впливу всіх точок та щільності [9].

[8] Nayyar, A., Puri, V. 2017. Comprehensive analysis & performance comparison of clustering algorithms for big data. Review of Computer Engineering Research, Vol. 4, No 2, p. 54-80.
[9] Там же.

## 2.3. Модель дискримінантного аналізу

Дискримінантний аналіз є потужним інструментом розпізнавання образів з навчанням. Процедури аналізу можна розділити на методи інтерпретації міжгрупових відмінностей – дискримінації та методи класифікації спостережень за групами.

При інтерпретації оцінюється можливість використання даного набору змінних, їх внесок, інформативність для характеристики несхожості, відмінності груп, тобто дискримінації об'єктів за певними ознаками[10]. Методи класифікації пов'язані з отриманням однієї або кількох функцій, які забезпечують можливість віднесення нового об'єкта до існуючих класів. Сферою застосування дискримінантного аналізу є соціологія, психологія, економіка, політологія, медицина, біологія.

За кількістю навчальних вибірок виділяють два види методів дискримінантного аналізу:

- ✓ для двох груп *(two-group descriminant analysis)* - будується лише одна дискримінантна функція з однією змінною;
- ✓ для трьох і більше груп застосовується множинний дискримінантний аналіз *(multiple descriminant anaysis)* - будується кілька дискримінантних функцій (за кількістю груп мінус одиниця).

Залежно від правил дискримінації в літературі розглядається три види аналізу:

- ✓ лінійний дискримінантний аналіз Фішера (запропонований Фішером) – правила дискримінації представлені у вигляді лінійної комбінації дискримінантних змінних;
- ✓ канонічний дискримінантний аналіз - правила дискримінації представлені у вигляді дискримінантних функцій;
- ✓ лінійний дискримінантний аналіз - правила дискримінації

[10] Бізнес-аналітика багатовимірних процесів : навчальний посібник [Електронний ресурс] / Т. С. Клебанова, Л. С. Гур'янова, Л. О. Чаговець та ін. – Харків : ХНЕУ ім. С. Кузнеця, 2018. – 272 с.

представлені сукупністю характеристик (групова коваріаційна матриця, груповий вектор середніх, визначник коваріаційної матриці).

Узагальнено задачу розрізнення (дискримінацію) можна сформулювати наступним чином: результат спостереження за об'єктом реалізований *k-вимірним* випадковим вектором $x = (x_1, x_2, \dots, x_k)^T$; потрібно встановити правило, відповідно до якого за спостереженим значенням вектору $x$ об'єкт може бути віднесений до однієї з можливих сукупностей $\pi_i \;\; i = 1, 2, \dots, l$. Для побудови правила дискримінації увесь вибірковий простір $R$ значень вектору $x$ розбивається на області $R_i \;\; i = 1, 2, \dots, l$ таким чином, що з потраплянням $x$ в $R_i$, об'єкт відносять до сукупності $\pi_i$.

Правило дискримінації ґрунтується на відповідності певному принципу оптимальності на основі апріорної інформації про сукупність та об'єкт. Інформація може представлятися у вигляді даних про функції *k -вимірного* розподілу ознак у кожній сукупності, або вибірок із цих сукупностей; апріорні ймовірності $p_i$ можуть задаватися або ні. Від повноти вихідної інформації залежить якість рекомендацій.

У процесі дискримінації відбувається перехід від вектору ознак, що характеризують об'єкт, до лінійної функції від них, дискримінантної функції як гіперплощини, що найкращим чином розділяє сукупність вибіркових точок.

Для практичної реалізації дискримінантного аналізу необхідно знати апріорні ймовірності та функції їх щільності. Вони можуть бути відомими з теоретичних міркувань або попередніх досліджень. Якщо ж вони невідомі, то їх замінюють статистичними оцінками, отриманими на основі наявних навчальних вибірок[11].

[11] Яровий А.Т., Страхов Є.М. Багатовимірний статистичний аналіз: навчально-методичний посібник для студентів математичних та економічних фахів. Одеса: Астропринт, 2015. 132 с.

Застосовують два підходи оцінювання функцій щільності ймовірності У першому (параметричний дискримінаційний аналіз) розподіл векторів ознак у кожній сукупності є нормальним, але відсутня інформація про параметри цього розподілу. Доцільним вважається заміна невідомих параметрів дискримінантної функції їх найкращими оцінками на базі вибіркових точок (правило дискримінації може будуватися на відношеннях правдоподібності; найчастіше використовують нормальний розподіл). У другому, непараметричному підході, методи дискримінації не потребують знання про функціональний вигляд розподілу та виконуються за незначною апріорною інформацією про сукупність (має практичну цінність).

Головна мета дискримінації у знаходженні такої лінійної комбінації змінних (дискримінантних змінних), яка оптимально поділить сукупність на групи та дозволить ідентифікувати класи. Найчастіше в якості такої комбінації певної множини ознак виступає дискримінантна функція.

Для кожного *k-го класу* визначають дискримінантну функцію $f_k$[12]:

$$f_k = a_{k0} + a_{k1}X_1 + a_{k2}X_2 + \ldots + a_{kp}X_p \qquad (2.15)$$

де $a_{ki}$ – коефіцієнти класифікуючої функції для *i-ої* змінної *k-го* класу (не інтерпретуються у зв'язку з тим, що вони нестандартизовані та кожному класу відповідає власна функція) та розраховуються за формулою:

$$a_{ki} = (n - \mathrm{g}) \sum_{j=1}^{p} b_{ij} X_{jk} \qquad (2.16)$$

[12] Klecka, WR. (1980). Discriminant analysis. Sage Publications, Beverly Hills.

де $b_{ij}$ – елемент матриці, оберненої до внутрішньогрупової матриці сум попарних добутків $W_{ij}$:

$$W_{ij} = \sum_{k=1}^{\mathrm{g}} \sum_{m=1}^{n_k} (X_{ikm} - X_{ik\cdot}) \left(X_{jkm} - X_{jk\cdot}\right) \tag{2.17}$$

$$a_{k0} = -0.5 \sum_{j=1}^{p} a_{kj} X_{jk}, \tag{2.18}$$

де $g$ – кількість класів, $n_k$ – кількість спостережень у *k-му* класі, $n_{\cdot}$ – загальна кількість спостережень усіх класів, $X_{ikm}$ – величина *i-оїї* змінної для *m-го* спостереження у *k-му* класі, $X_{ik\cdot}$– середня величина *i-оїї* змінної у *k-му* класі, $X_{jk\cdot}$– середня величина *i-оїї* змінної для усіх класів (загальна середня).

Об'єкт відноситься до класу з найбільшим значенням $f_k$. При геометричній інтерпретації дискримінантні змінні – осі $p$-вимірного евклідового простору, кожний об'єкт – точка простору з координатами, що являють собою значення спостереження кожної змінної. Для характеристики положення класу визначають його центроїд у вигляді уявної точки, координати якої є середніми значеннями змінних у класі та відповідають центрам їх тяжіння. Центроїд можна використовувати для вивчення відмінностей між класами у зв'язку з тим, що він займає положення типових спостережень відповідного класу. Якщо розміщення класів дійсно різняться (їх центроїди не співпадають), то ступінь розкиду спостережень в межах класів буде менша за загальний розкид.

При реалізації процедури класифікації використовують відстань між об'єктами та кожним з центроїдних класів для подальшого віднесення об'єкту до найближчого класу у вигляді узагальненої міри Махаланобіса:

$$D^2(X|G_k) = (n - \mathrm{g}) \sum_{i=1}^{p} \sum_{j=1}^{p} b_{ij} (X_i - X_{ik\cdot}) \left(X_j - X_{jk\cdot}\right) \tag{2.19}$$

де $D^2(X|G_k)$ – квадрат відстані від точки $X$ (даного об'єкту) до центроїду класу $k$.

Після розрахунку $D^2$ для кожного класу об'єкт класифікують у групу з найменшим значенням відстані. Цей клас, для якого типовий профіль за дискримінантними змінними більш схожий на профіль для досліджуваного об'єкту. Якщо відстань до найближчого класу велика, то узгодженість між профілями буде поганою, але порівняно з будь-яким іншим класом добра[13].

Імовірність приналежності об'єкту $X$ до *k-го* класу визначається за формулою:

$$Pr(G_k|X) = \frac{Pr(X|G_k)}{\sum_{i=1}^{g} Pr(X|G_k)} \tag{2.20}$$

Сума імовірностей для усіх класів дорівнює одиниці та має назву апостеріорна імовірність. Класифікація найбільшої з цих величин еквівалентна використанню найменшої відстані. Апостеріорна величина $Pr(G_k|X)$ оцінює імовірність приналежності об'єкту до *k-го* класу, величина $Pr(X|G_k)$ – частку об'єктів у цьому класі, які розташовані від центроїду далі за $X$.

Мірою якості розпізнавання класів при процедурі послідовного відбору є $\lambda$-статистика Уілкса. Критерій характеризує відмінності між класами за декількома дискримінантними змінними та однорідність кожного з них:

$$\lambda = \prod_{i=k+1}^{g} \frac{1}{1+\lambda_i} \tag{2.21}$$

де $\lambda_i$- власні значення матриці коваріаїцій.

Величина $\lambda$ є оберненої, тому чим ближче її значення до нуля, тим більша висока відмінність класів (центроїди добре розділені та

[13] Klecka, WR. (1980). Discriminant analysis. Sage Publications, Beverly Hills.

принципово відрізняються один від іншого по відношенню до ступеню розкиду всередині класів); збільшення $\lambda$ до максимального її значення (до одиниці) вказує на поступове погіршення відмінностей (центроїди груп співпадають, відсутні групові розходження). На основі $\lambda$-статистика Уілкса можна отримати тест істотності, апроксимуючи розподіл деякої функції від неї розподілом $\chi^2$(хі-квадрат) або *F*-критерієм.

Для використання алгоритму лінійного дискримінантного аналізу Фішера для двох класів за нормального закону розподілу показників необхідно виконання умов[14]:

- ✓ обсяг вибірки має бути більшим, ніж кількість змінних;
- ✓ кластери, серед яких здійснюють дискримінацію, підпорядковані багатовимірному нормальному розподілу;
- ✓ класи можуть перетинатися, але їх центри мають бути достатньо віддаленими один від одного;
- ✓ різниця між коваріаційними матрицями цих кластерів є статистично незначущою; кількість навчальних вибірок у кластері є меншою, ніж кількість дискримінантних функцій.

Алгоритм реалізації дискримінантного аналізу при нормальному законі розподілу показників стосовно лінійної дискримінантної функції для двох класів розглянемо на практичному прикладі (зад.2.8) та за допомогою програми STATISTICA (п.2.4).

*Задача 2.8:* за показниками, що характеризують виробничу діяльність, підприємства поділені на дві групи – передові та відсталі. За допомогою процедури дискримінантного аналізу потрібно класифікувати нове підприємство з характеристиками: вартість основних виробничих засобів – 55,541; чисельність промислово-виробничого персоналу 9,592 тис. осіб, балансовий прибуток 12,840.

[14] Бахрушин В. Є. Методи аналізу даних : навч. посіб. для студен-тів / В. Є. Бахрушин. – Запоріжжя : КПУ, 2011. – 268 с.

*Таблиця 2.5.* **Показники діяльності підприємств**

| Групи підприємств | Вартість основних виробничих засобів (ОВЗ) | Чисельність промислово-виробничого персоналу (ПВП) | Балансовий прибуток (Прибуток) |
|---|---|---|---|
| Передові | 224,228 | 17,115 | 22,981 |
| | 151,827 | 14,904 | 21,481 |
| | 147,313 | 13,627 | 28,669 |
| | 152,253 | 10,545 | 10,199 |
| Відсталі | 46,757 | 4,428 | 11,124 |
| | 29,033 | 5,510 | 6,091 |
| | 52,134 | 4,214 | 11,842 |
| | 37,050 | 5,227 | 11,875 |
| | 63,979 | 4,211 | 12,860 |

1. Вихідні дані представлені у табличній формі, яка містить дві підмножини об'єктів навчальних вибірок, що підлягають дискримінації Заносимо інформацію про них у матриці X та Y:

$$X = \begin{pmatrix} 224{,}228 & 17{,}115 & 22{,}981 \\ 151{,}827 & 14{,}904 & 21{,}481 \\ 147{,}313 & 13{,}627 & 28{,}669 \\ 152{,}253 & 10{,}545 & 10{,}199 \end{pmatrix}$$

$$Y = \begin{pmatrix} 46{,}757 & 4{,}428 & 11{,}124 \\ 29{,}033 & 5{,}510 & 6{,}091 \\ 52{,}134 & 4{,}214 & 11{,}842 \\ 37{,}050 & 5{,}227 & 11{,}875 \\ 63{,}979 & 4{,}211 & 12{,}860 \end{pmatrix}$$

де: $n_1 = n_X = 4$, $n_2 = n_Y = 4$,

та показники нового об'єкту у рядок матриці Z: $Z^T = (55{,}451 \quad 9{,}592 \quad 12{,}840)$. Метою дискримінантного аналізу є віднесення нового спостереження (рядка матриці Z) до навчальної підмножини X або Y.

2. Для розв'язання задачі визначаємо середні значення по кожній *j-ій* ознаці для *i–х* об'єктів всередині *k-оїї* підмножини (*k*=1,2):

$$\bar{x}_j^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ij}^{(k)} \qquad j = \overline{1,p} \tag{2.22}$$

$$\bar{y}_j^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k} y_{ij}^{(k)} \qquad j = \overline{1,p} \tag{2.23}$$

Результати розрахунку по кожній підмножині представляємо у вигляді вектору-стовпця $\bar{X}$ та $\bar{Y}$:

$$\bar{X}^{(k)} = \begin{pmatrix} \bar{x}_1^{(k)} \\ \bar{x}_2^{(k)} \\ \dots \\ \bar{x}_p^{(k)} \end{pmatrix} \qquad \bar{Y}^{(k)} = \begin{pmatrix} \bar{y}_1^{(k)} \\ \bar{y}_2^{(k)} \\ \dots \\ \bar{y}_p^{(k)} \end{pmatrix} \tag{2.24}$$

Вектори середніх мають вигляд:

$$\bar{X} = \begin{pmatrix} 168{,}92025 \\ 14{,}04775 \\ 20{,}8325 \end{pmatrix}$$

$$\bar{Y} = \begin{pmatrix} 45{,}7926 \\ 4{,}778 \\ 10{,}758 \end{pmatrix}.$$

3. Для кожної навчальної підмножини розраховуються коваріаційні матриці $S_X$ та $S_Y$ (розмірність *p x p*). Елементи кожної матриці розраховуються за формулою:

$$S_{jl}^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k}\left[\left(x_{ij}^{(k)} - \bar{x}_j^{(k)}\right)\left(x_{il}^{(k)} - \bar{x}_l^{(k)}\right)\right] \quad j,l = \overline{1,p} \tag{2.25}$$

Визначаємо оцінку коваріаційних матриць:

$$S_X = \begin{pmatrix} 1025{,}61 & 55{,}66575 & 28{,}94475 \\ 55{,}66575 & 5{,}6468625 & 10{,}27365 \\ 28{,}94475 & 10{,}27365 & 44{,}879675 \end{pmatrix}$$

$$S_Y = \begin{pmatrix} 145{,}8666 & -6{,}60952 & 22{,}78694 \\ -6{,}60952 & 0{,}371782 & -0{,}902484 \\ 22{,}78694 & -0{,}902484 & 5{,}750302 \end{pmatrix}$$

4. Розраховуємо об'єднану коваріаційну матрицю $\hat{S}$:

$$\hat{S} = \frac{1}{n_1 + n_2 - 2}(n_1 S_X + n_2 S_Y) \qquad (2.26)$$

Отримуємо незміщену оцінку сумарної коваріаційної матриці:

$$\hat{S} = \frac{1}{4 + 5 - 2}(4S_X + 5S_Y)$$

$$\hat{S} = \begin{pmatrix} 690{,}25328 & 27{,}087914 & 32{,}816242 \\ 27{,}087914 & 3{,}4923371 & 5{,}2260257 \\ 32{,}816242 & 5{,}2260257 & 29{,}752887 \end{pmatrix}.$$

5. Визначаємо матрицю $\hat{S}^{-1}$, обернену до об'єднаної коваріаційної матриці $\hat{S}$ , за формулою:

$$\hat{S}^{-1} = \frac{1}{|\hat{S}|}\hat{\hat{S}} \qquad (2.27)$$

де $|\hat{S}|$ – визначник матриці $\hat{S}$ ($|\hat{S}| \neq 0$);

$\hat{\hat{S}}$ –приєднана матриця, елементи якої є алгебраїчним доповненням елементів матриці $\hat{S}'$

Обернена матрицю до $\hat{S}$ матиме вигляд:

$$\hat{S}^{-1} = \begin{pmatrix} 0{,}0020945371 & -0{,}0173449116 & 0{,}00073714 \\ -0{,}0173449116 & 0{,}53214303 & -0{,}07433441 \\ 0{,}00073714 & -0{,}07433441 & 0{,}04565381 \end{pmatrix}$$

6. Визначаємо вектор-стовпчик $a = \begin{pmatrix} A_1 \\ A_2 \\ \dots \\ A_p \end{pmatrix}$ дискримінантних множників з врахуванням усіх елементів навчальної підмножини за формулою:

$$a = \hat{S}^{-1}(\bar{X} - \bar{Y}) \qquad (2.28)$$

Наведена формула отримана методом найменших квадратів виходячи з умови забезпечення найбільшої відмінності між дискримінантними функціями. Найкраще розділення двох навчальних підмножин забезпечується поєднанням мінімальної внутрішньогрупової варіації та максимальної міжгрупової варіації.

Знайдемо вектор оцінок коефіцієнтів дискримінації:

$$a = \hat{S}^{-1}(\bar{X} - \bar{Y}) = \hat{S}^{-1}\left(\begin{pmatrix} 123{,}12765 \\ 9{,}26975 \\ 10{,}0745 \end{pmatrix} - \begin{pmatrix} 0{,}10449979 \\ 2{,}0475006 \\ -0{,}13634981 \end{pmatrix}\right)$$

7. На наступному етапі знаходимо оцінку векторів значень дискримінантної функції для матриці вихідних даних:

$$\widehat{U}_X = X_a = \begin{pmatrix} 55{,}346433 \\ 43{,}457381 \\ 39{,}3990544 \\ 36.113833 \end{pmatrix}$$

$$\widehat{U}_Y = Y_a = \begin{pmatrix} 55{,}346433 \\ 43{,}457381 \\ 39{,}3990544 \\ 36{,}113833 \end{pmatrix}$$

8. Обчислимо середні значення оцінок дискримінантної функції:

$$\hat{\bar{u}}_X = \frac{1}{n_1}\sum_{i=1}^{n_1} \widehat{U}_X = 43{,}577047$$

$$\hat{\bar{u}}_Y = \frac{1}{n_2}\sum_{i=1}^{n_2} \hat{U}_Y = 13{,}102648$$

9. Шукаємо константу:

$$\hat{C} = \frac{1}{2}(\hat{\bar{u}}_X + \hat{\bar{u}}_Y) = \frac{1}{2}(43{,}577047 + 13{,}102648) = 23{,}339847$$

10. Дискримінантну функцію для *z-го* спостереження, яке підлягає дискримінації, отримуємо розв'язав рівняння:

$$\hat{u}_Z = a_1 z_1 + a_2 z_2 + a_3 z_3 \qquad (2.29)$$

Якщо $\hat{u}_Z \geq \hat{\mathrm{C}}$, то *z*-є спостереження потрібно віднести до сукупності X; якщо $\hat{u}_Z < \hat{\mathrm{C}}$, то *z*-є спостереження класифікується до сукупності Y.

Визначаємо можливість включення підприємства Z до групи передових. Матриця Z представлена одним рядком, тому $\hat{U}_Y$ позначимо $\hat{U}_Z$:

$$\hat{u}_Z = 0{,}10449979 \cdot 55{,}451 + 2{,}0478006 \cdot 9{,}952 - 0{,}13634981 \cdot \\ \cdot 12{,}840 \approx 23{,}69$$

Середнє значення дискримінантної функції $\hat{u}_Z$ менше, ніж константа $\hat{\mathrm{C}}$:

$$\hat{u}_Z < \hat{\mathrm{C}}$$
$$23{,}69 < 28{,}34,$$

Тому підприємство z с характеристиками $Z^T$ не може бути віднесеним до групи передових підприємств.

## 2.4. Реалізація кластерного та дискримінантного методів у програмному середовищі STATISTICA

Пакет аналітичного програмного забезпечення STATISTICA (розроблений StatSoft, зараз підтримується TIBCO Software Inc.) пропонує реалізацію ***кластерного аналізу*** в модулі *Cluster Analysis* за допомогою перемикача модулів системи *STATISTICA Module Switcher* або меню *Statistics*/*Multivariate exploratory/Cluster*.

Основні можливості та етапи проведення кластерного аналізу розглянемо на наступному прикладі: двадцять регіонів описуються показниками, що характеризують демографічну ситуацію ($x_1$-загальний коефіцієнт народжуваності, ‰; $x_2$-загальний коефіцієнт смертності, ‰; $x_3$-коефіцієнт фемінізації, ‰; $x_4$-кількість шлюбів; $x_5$-загальний коефіцієнт механічного приросту, ‰; Потрібно класифікувати регіони на однорідні групи (табл. 2.6).

*Таблиця 2.6.* **Основні показники відтворення населення країни**

| № регіону | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| 1 | 12,5 | 10,0 | 1218 | 1892 | 2,3 |
| 2 | 10,5 | 10,3 | 1145 | 345 | 2,6 |
| 3 | 13,9 | 10,1 | 1142 | 572 | 0,7 |
| 4 | 10,8 | 12,4 | 1173 | 402 | 0,9 |
| 5 | 14,0 | 18,1 | 1105 | 125 | 1,3 |
| 6 | 12,8 | 16,7 | 1134 | 292 | -2,7 |
| 7 | 11,8 | 14,7 | 1099 | 148 | 1,2 |
| 8 | 17,9 | 12,7 | 1055 | 392 | 0 |
| 9 | 16,4 | 14,9 | 1078 | 381 | -1,1 |
| 10 | 16,4 | 16,4 | 1135 | 162 | -1 |
| 11 | 12,8 | 16,5 | 1101 | 119 | -2 |
| 12 | 15,1 | 12,1 | 1095 | 360 | 7,7 |
| 13 | 16,5 | 12,8 | 1018 | 217 | -1,7 |
| 14 | 14,7 | 16,0 | 1084 | 253 | 1 |
| 15 | 14,8 | 13,8 | 1059 | 381 | -3,1 |
| 16 | 16,1 | 13,1 | 1071 | 335 | -1,2 |
| 17 | 14,0 | 16,2 | 1117 | 267 | -4 |
| 18 | 15,8 | 17,1 | 1060 | 173 | 1 |
| 19 | 14,4 | 17,3 | 1088 | 160 | 3,8 |
| 20 | 13,0 | 13,9 | 1066 | 96 | 0,8 |

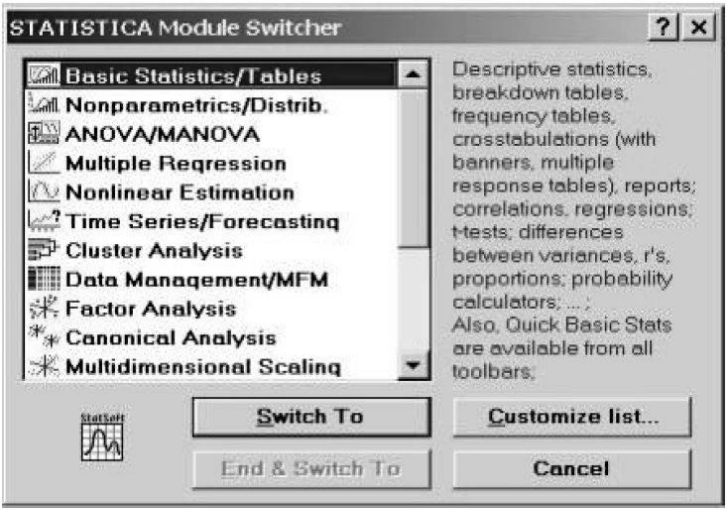



*Рис. 2.9.* **Вибір модуляCluster Analysis**

Можливість стандартизації первинних даних, з метою уникнення відмінностей в одиницях вимірювання, закладена в позиції *Data*, вибір *Standardize*, шо активується попереднім виділенням первинного масиву інформації. Здійснення стандартизації даних представлено на рис. 2.10.

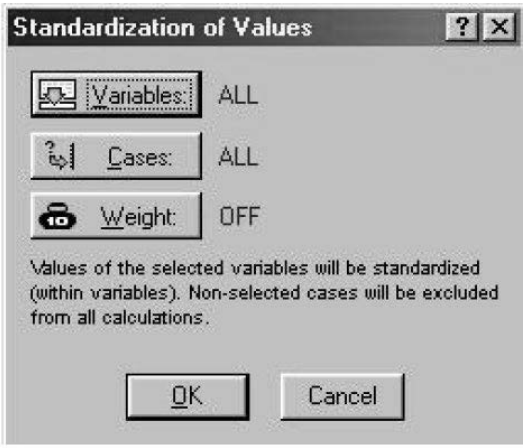



*Рис. 2.10.* **Виконання стандартизації даних в програмному середовищі STATISTICA**

За замовчуванням відбираються усі змінні *(Variables)* та спостереження *(Cases)*, але існує можливість нормування частини ознак або одиниць сукупності. Також за замовчуванням приймається однаковий внесок усіх спостережень в розрахунок середніх значень та стандартних відхилень. За допомогою позиції *Weight* можна коригувати вагу змінних у вікні *Define Weight*.

Після стандартизації масив має наступний вигляд (рис.2.11):

| | 1<br>Var1 | 2<br>Var2 | 3<br>Var3 | 4<br>Var4 | 5<br>Var5 |
|---|---|---|---|---|---|
| 1 | -0,8565907 | -1,6895293 | 2,50670468 | 4,01969113 | 0,74319718 |
| 2 | -1,8584511 | -1,5704085 | 0,92716699 | -0,022471 | 0,85608789 |
| 3 | -0,1552884 | -1,6498224 | 0,86225448 | 0,57065818 | 0,14111339 |
| 4 | -1,7081721 | -0,7365633 | 1,53301706 | 0,12646454 | 0,21637386 |
| 5 | -0,1051953 | 1,52673092 | 0,06166688 | -0,5973098 | 0,36689481 |
| 6 | -0,7063116 | 0,9708341 | 0,68915446 | -0,1609549 | -1,1383147 |
| 7 | -1,2072418 | 0,17669578 | -0,0681581 | -0,537213 | 0,32926457 |
| 8 | 1,84843252 | -0,6174425 | -1,0202083 | 0,1003355 | -0,1222983 |
| 9 | 1,09703718 | 0,25610961 | -0,5225457 | 0,07159356 | -0,5362309 |
| 10 | 1,09703718 | 0,85171335 | 0,71079196 | -0,5006324 | -0,4986006 |
| 11 | -0,7063116 | 0,89142027 | -0,0248831 | -0,6129872 | -0,874903 |
| 12 | 0,4458279 | -0,855684 | -0,1547081 | 0,01672258 | 2,77522997 |
| 13 | 1,14713021 | -0,5777356 | -1,8207958 | -0,3569227 | -0,7620123 |
| 14 | 0,24545581 | 0,69288569 | -0,3927207 | -0,2628581 | 0,2540041 |
| 15 | 0,29554883 | -0,1806665 | -0,9336582 | 0,07159356 | -1,2888356 |
| 16 | 0,94675812 | -0,4586149 | -0,6740082 | -0,0486 | -0,5738611 |
| 17 | -0,1051953 | 0,77229952 | 0,32131691 | -0,2262775 | -1,6275077 |
| 18 | 0,79647905 | 1,12966176 | -0,9120207 | -0,4718904 | 0,2540041 |
| 19 | 0,09517674 | 1,2090756 | -0,3061707 | -0,5058582 | 1,30765073 |
| 20 | -0,6061256 | -0,1409596 | -0,7821957 | -0,673084 | 0,17874363 |

*Рис.2.11.* **Масив статистичних даних після їх стандартизації**

Модуль *Cluster Analysis* пропонує наступні методи (рис. 2.12):

- ✓ *Joining (tree clustering)* - Об'єднання (деревовидна клатеризація),
- ✓ *K – means clustering* - кластеризация методом *k– середніх,*
- ✓ *Two-way joining* – двохходове обєднання.

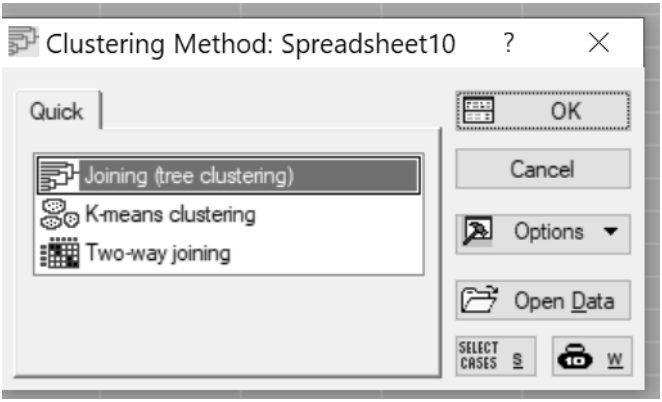



*Рис.2.12.* **Вибір методу кластеризації**

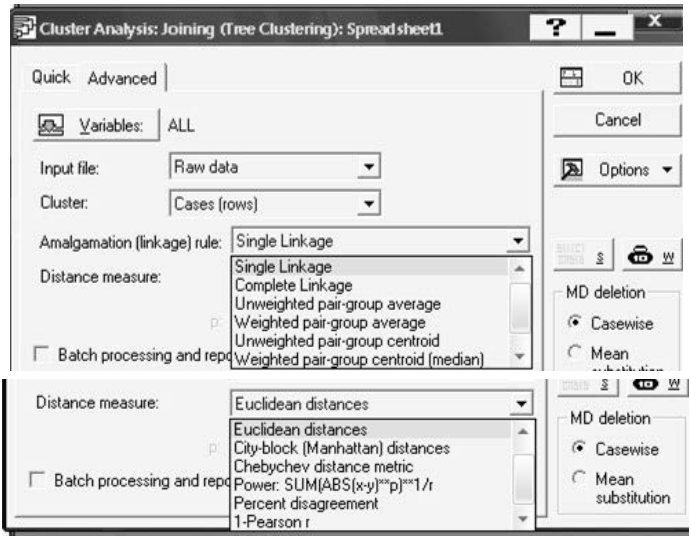



*Рис. 2.13.* **Визначення вихідних параметрів кластеризації**

Після підтвердження вибору методу (вибрано метод *Joining tree clustering*) обираємо *Variable* (Змінні), *Input File* (у вигляді *Raw Data-первинні дані* або *Distance Matrix-матриці відстаней*), *Cluster* (об'єкти кластеризації за *Variables [Columns]-змінни у стовпцях* або

*Cases [rows]-спостереження у рядках*), *Amalgamation rule* (правила обєднання), *Distance measure* (варіанти міри відстані) (рис.2.13).

Рядок *Amalgamation rule* пропонує наступні правила ієрархічного об'єднання:

- ✓ *Single linkage* – метод одиночного зв'язку за принципом найближчого сусіда;
- ✓ *Complete linkage – метод* повних зв'язків за принципом далекого сусіда;
- ✓ *Unweighted pair-group average* – незваженої попарної середньої;
- ✓ *Weighted pair-group average* – зваженої попарної середньої;
- ✓ *Unweighted pair-group centroid* – незваженої центроїдної;
- ✓ *Weighted pair-group centroid* – зваженої центроїдної;
- ✓ *Ward's method* – метод Уорда.

*Distance measure* представляє в якості метрики подібності такі відстані:

- ✓ *Square Euclidean distances* – квадрат евклідової відстані;
- ✓ *Euclidean distances* – евклідова метрика;
- ✓ *City-block (Manhattan) distances* – Мангеттенська відстань;
- ✓ *Chebychev distances metric* – відстань Чебишева;
- ✓ *Power metric* – степенева відстань Мінковського;
- ✓ *Percent disagreement* – відсоток незгоди (для категоріальних даних);
- ✓ *(1 – Personal r)* – (1 – коефіцієнт кореляції Пірсона).

У системі STATISTICA передбачено два способи обробки некомплектних спостережень завдяки вибору *Missing data [MD*

*deletion - відсутні дані]*: *Casewise deleted* (*построкове видалення*) або *Substituted by means* (*заміна середніми значеннями*).

Після активації обраних параметрів (процедура *Ward's method,* звичайна е*вклідова метрика* (*Euclidean distances),* пропущені дані відсутні) клавішою (*OK*) отримуємо результат кластеризації (рис. 2.14). Панель містить інформацію про кількість змінних та спостережень, обрані способи та правила класифікації, варіанти представлення підсумку процедури у розширеному вигляді *Advanced*: *Horizontal hierarchical tree plot* (горизонтальна деревоподібна діаграма); *Vertical icicle plot* (вертикальна деревоподібна діаграма – дендрограма); *Amalgamation schedule* (правило об'єднання в кластери); *Graf of amalgamation schedule* (графік порядку об'єднання); *Distance matrix* (матриця відстаней); *Descriptive statistics* (описові статистики).

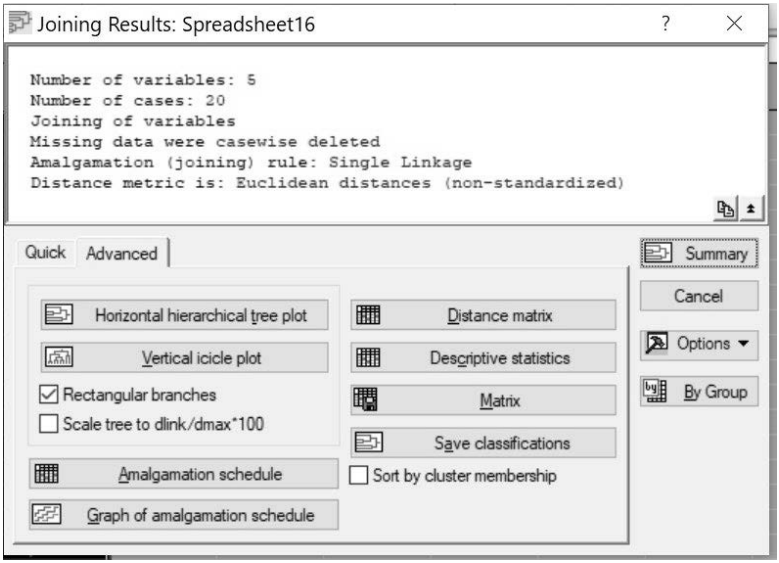



*Рис. 2.14.* **Вікно результатів кластеризації**

На дендрограмі класифікації *Vertical icicle plot* (рис. 2.15) горизонтальна вісь відображає об’єкти (регіони), вертикальна –

відстані між ними. Візуально виокремлюються три кластера: перший з найвищими показниками відтворення населення – 8, 9, 13, 15, 16 регіони; другий об'єднує десять регіонів з найнижчими значеннями; третій кластер охоплює 1, 2, ,3, 4 та 12 спостереження з середніми значеннями показників.

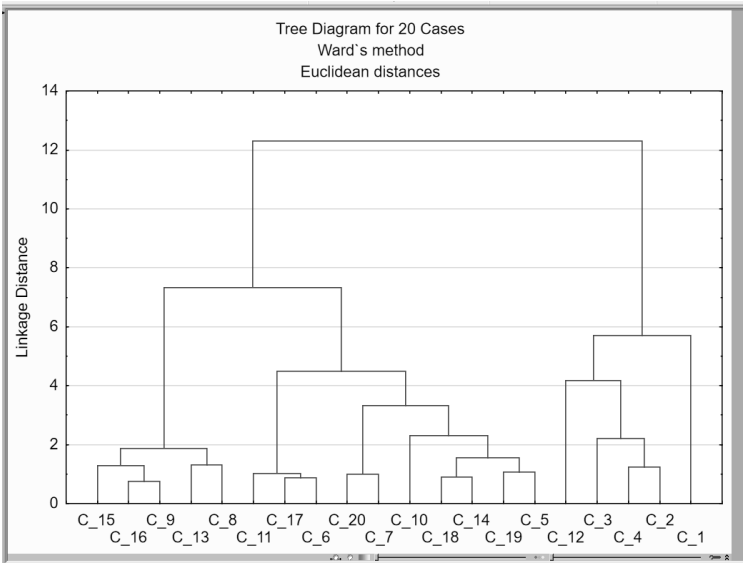



Рис. 2.15. **Дендрограма регіонів за результатами класифікації**

Обравши опцію *Amalgamation schedule,* отримуємо таблицю результатів поступового об'єднання регіонів у кластери, у якій перший стовпчик містить відстані для відповідних кластерів на кожному кроці класифікації (рис. 2.16). На першому кроці об'єднались 9 та 16 регіони, на другому – 6 та 17 і так далі.

Ініціювавши *Graph of amalgamation schedule,* є можливість аналізу процедури деревоподібної класифікації у графічному лінійному вигляді за поступовою зміною меж кластерів (рис. 2.17).

Amalgamation Schedule (Spreadsheet80)
Ward`s method
Euclidean distances

| linkage distance | Obj. No. 1 | Obj. No. 2 | Obj. No. 3 | Obj. No. 4 | Obj. No. 5 | Obj. No. 6 | Obj. No. 7 | Obj. No. 8 | Obj. No. 9 | Obj. No. 10 | Obj. No. 11 | Obj. No. 12 | Obj. No. 13 | Obj. No. 14 | Obj. No. 15 | Obj. No. 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ,7564511 | C_9 | C_16 | | | | | | | | | | | | | | |
| ,8829712 | C_6 | C_17 | | | | | | | | | | | | | | |
| ,8987586 | C_14 | C_18 | | | | | | | | | | | | | | |
| 1,006585 | C_7 | C_20 | | | | | | | | | | | | | | |
| 1,031916 | C_6 | C_17 | C_11 | | | | | | | | | | | | | |
| 1,081547 | C_5 | C_19 | | | | | | | | | | | | | | |
| 1,231402 | C_2 | C_4 | | | | | | | | | | | | | | |
| 1,280222 | C_9 | C_16 | C_15 | | | | | | | | | | | | | |
| 1,323881 | C_8 | C_13 | | | | | | | | | | | | | | |
| 1,551877 | C_5 | C_19 | C_14 | C_18 | | | | | | | | | | | | |
| 1,873190 | C_8 | C_13 | C_9 | C_16 | C_15 | | | | | | | | | | | |
| 2,200987 | C_2 | C_4 | C_3 | | | | | | | | | | | | | |
| 2,305593 | C_5 | C_19 | C_14 | C_18 | C_10 | | | | | | | | | | | |
| 3,336661 | C_5 | C_19 | C_14 | C_18 | C_10 | C_7 | C_20 | | | | | | | | | |
| 4,173023 | C_2 | C_4 | C_3 | C_12 | | | | | | | | | | | | |
| 4,496887 | C_5 | C_19 | C_14 | C_18 | C_10 | C_7 | C_20 | C_6 | C_17 | C_11 | | | | | | |
| 5,709340 | C_1 | C_2 | C_4 | C_3 | C_12 | | | | | | | | | | | |
| 7,338730 | C_5 | C_19 | C_14 | C_18 | C_10 | C_7 | C_20 | C_6 | C_17 | C_11 | C_8 | C_13 | C_9 | C_16 | C_15 | |
| 12,31549 | C_1 | C_2 | C_4 | C_3 | C_12 | C_5 | C_19 | C_14 | C_18 | C_10 | C_7 | C_20 | C_6 | C_17 | C_11 | C_8 |

*Рис. 2.16.* **Матриця послідовного об'єднання регіонів**

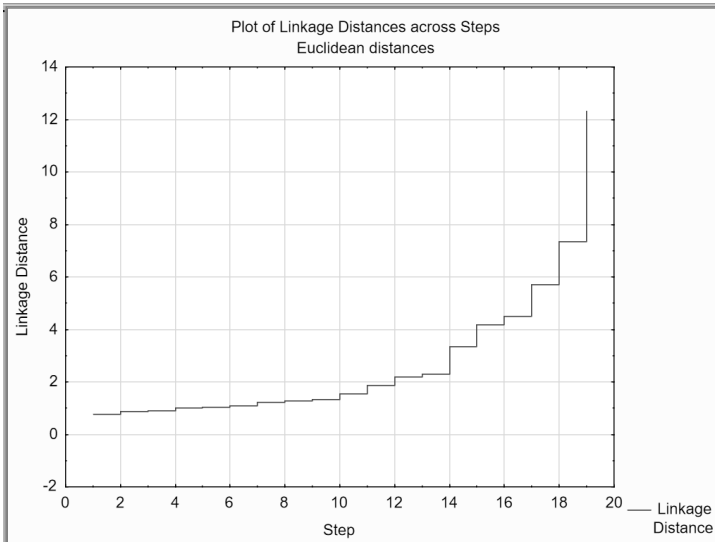



*Рис. 2.17.* **Графік послідовного об'єднання регіонів**

Перегляд матриці відстаней здійснюється через *Distance matrix,* фрагмент якої наведено на рис. 2.18, яка зберігається за допомогою опції *Save distance matrix (Зберегти матрицю відстаней).*

| Case No. | Euclidean distances (Spreadsheet80) | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | C_1 | C_2 | C_3 | C_4 | C_5 | C_6 | C_7 | C_8 | C_9 | C_10 | C_11 | C_12 | C_13 | C_14 | C_15 | C_16 | C_17 | C_18 | C_19 | C_20 |
| C_1 | 0,00 | 4,46 | 3,93 | 4,24 | 6,19 | 5,61 | 5,58 | 6,08 | 5,83 | 5,96 | 6,10 | 5,44 | 6,74 | 5,82 | 5,93 | 5,76 | 5,92 | 6,54 | 6,17 | 5,97 |
| C_2 | 4,46 | 0,00 | 1,94 | 1,23 | 3,74 | 3,44 | 2,24 | 4,41 | 4,01 | 4,09 | 3,41 | 3,27 | 4,51 | 3,42 | 3,83 | 3,70 | 3,89 | 4,28 | 3,67 | 2,72 |
| C_3 | 3,93 | 1,94 | 0,00 | 1,97 | 3,49 | 3,06 | 2,56 | 2,99 | 2,80 | 3,07 | 3,16 | 3,04 | 3,42 | 2,82 | 2,81 | 2,43 | 3,15 | 3,59 | 3,48 | 2,59 |
| C_4 | 4,24 | 1,23 | 1,97 | 0,00 | 3,23 | 2,56 | 2,03 | 4,39 | 3,69 | 3,46 | 2,80 | 3,75 | 4,54 | 3,12 | 3,56 | 3,56 | 3,14 | 4,01 | 3,47 | 2,75 |
| C_5 | 6,19 | 3,74 | 3,49 | 3,23 | 0,00 | 1,88 | 1,75 | 3,21 | 2,16 | 1,76 | 1,52 | 3,49 | 3,30 | 1,07 | 2,69 | 2,60 | 2,18 | 1,40 | 1,08 | 1,95 |
| C_6 | 5,61 | 3,44 | 3,06 | 2,56 | 1,88 | 0,00 | 1,94 | 3,62 | 2,38 | 1,95 | 0,89 | 4,55 | 3,51 | 2,03 | 2,24 | 2,64 | 0,88 | 2,62 | 2,79 | 2,33 |
| C_7 | 5,58 | 2,24 | 2,56 | 2,03 | 1,75 | 1,94 | 0,00 | 3,39 | 2,58 | 2,66 | 1,49 | 3,18 | 3,23 | 1,60 | 2,47 | 2,54 | 2,38 | 2,38 | 1,94 | 1,01 |
| C_8 | 6,08 | 4,41 | 2,99 | 4,39 | 3,21 | 3,62 | 3,39 | 0,00 | 1,32 | 2,49 | 3,30 | 3,34 | 1,32 | 2,23 | 1,99 | 1,09 | 3,15 | 2,15 | 3,05 | 2,65 |
| C_9 | 5,83 | 4,01 | 2,80 | 3,69 | 2,16 | 2,38 | 2,58 | 1,32 | 0,00 | 1,48 | 2,12 | 3,57 | 1,62 | 1,29 | 1,25 | 0,76 | 1,92 | 1,39 | 2,39 | 2,05 |
| C_10 | 5,96 | 4,09 | 3,07 | 3,46 | 1,76 | 1,95 | 2,66 | 2,49 | 1,48 | 0,00 | 1,99 | 3,88 | 2,92 | 1,61 | 2,32 | 1,97 | 1,72 | 1,84 | 2,33 | 2,57 |
| C_11 | 6,10 | 3,41 | 3,16 | 2,80 | 1,52 | 0,89 | 1,49 | 3,30 | 2,12 | 1,99 | 0,00 | 4,26 | 2,98 | 1,57 | 1,90 | 2,32 | 1,10 | 2,10 | 2,37 | 1,66 |
| C_12 | 5,44 | 3,27 | 3,04 | 3,75 | 3,49 | 4,55 | 3,18 | 3,34 | 3,57 | 3,88 | 4,26 | 0,00 | 4,00 | 2,99 | 4,20 | 3,45 | 4,76 | 3,35 | 2,61 | 3,04 |
| C_13 | 6,74 | 4,51 | 3,42 | 4,54 | 3,30 | 3,51 | 3,23 | 1,32 | 1,62 | 2,92 | 2,98 | 4,00 | 0,00 | 2,35 | 1,46 | 1,22 | 2,96 | 2,22 | 3,30 | 2,31 |
| C_14 | 5,82 | 3,42 | 2,82 | 3,12 | 1,07 | 2,03 | 1,60 | 2,23 | 1,29 | 1,61 | 1,57 | 2,99 | 2,35 | 0,00 | 1,88 | 1,62 | 2,04 | 0,90 | 1,21 | 1,32 |
| C_15 | 5,93 | 3,83 | 2,81 | 3,56 | 2,69 | 2,24 | 2,47 | 1,99 | 1,25 | 2,32 | 1,90 | 4,20 | 1,46 | 1,88 | 0,00 | 1,05 | 1,69 | 2,16 | 3,07 | 1,88 |
| C_16 | 5,76 | 3,70 | 2,43 | 3,56 | 2,60 | 2,64 | 2,54 | 1,09 | 0,76 | 1,97 | 2,32 | 3,45 | 1,22 | 1,62 | 1,05 | 0,00 | 2,18 | 1,86 | 2,72 | 1,87 |
| C_17 | 5,92 | 3,89 | 3,15 | 3,14 | 2,18 | 0,88 | 2,38 | 3,15 | 1,92 | 1,72 | 1,10 | 4,76 | 2,96 | 2,04 | 1,69 | 2,18 | 0,00 | 2,46 | 3,05 | 2,40 |
| C_18 | 6,54 | 4,28 | 3,59 | 4,01 | 1,40 | 2,62 | 2,38 | 2,15 | 1,39 | 1,84 | 2,10 | 3,35 | 2,22 | 0,90 | 2,16 | 1,86 | 2,46 | 0,00 | 1,41 | 1,91 |
| C_19 | 6,17 | 3,67 | 3,48 | 3,47 | 1,08 | 2,79 | 1,94 | 3,05 | 2,39 | 2,33 | 2,37 | 2,61 | 3,30 | 1,21 | 3,07 | 2,72 | 3,05 | 1,41 | 0,00 | 1,96 |
| C_20 | 5,97 | 2,72 | 2,59 | 2,75 | 1,95 | 2,33 | 1,01 | 2,65 | 2,05 | 2,57 | 1,66 | 3,04 | 2,31 | 1,32 | 1,88 | 1,87 | 2,40 | 1,91 | 1,96 | 0,00 |

*Рис. 2.18.* **Матриця відстаней**

Рядок Descriptive statistics (Описові характеристики) відкриває таблицю результатів із середніми значеннями та стандартними відхиленнями для кожного об'єкта, включеного до кластерного аналізу, тобто для кожного спостереження та змінної залежно від параметрів, обраних у стартовій панелі *Cluster*.

| Case No. | Means and Standard Deviations ( | |
|---|---|---|
| | **Mean** | Std.Dev. |
| **C_1** | 0,944695 | 2,351587 |
| C_2 | -0,333615 | 1,318764 |
| C_3 | -0,046217 | 0,977803 |
| C_4 | -0,113776 | 1,205052 |
| C_5 | 0,250558 | 0,794241 |
| C_6 | -0,069119 | 0,896397 |
| C_7 | -0,261331 | 0,622263 |
| C_8 | 0,037764 | 1,101891 |
| C_9 | 0,073193 | 0,672313 |
| C_10 | 0,332062 | 0,771693 |
| C_11 | -0,265533 | 0,721595 |
| C_12 | 0,445478 | 1,384271 |
| C_13 | -0,474067 | 1,066459 |
| C_14 | 0,107353 | 0,438907 |
| C_15 | -0,407204 | 0,676176 |
| C_16 | -0,161665 | 0,663695 |
| C_17 | -0,173073 | 0,902908 |
| C_18 | 0,159247 | 0,851865 |
| C_19 | 0,359975 | 0,848911 |
| C_20 | -0,404724 | 0,407759 |

*Рис. 2.19.* **Описові статистики результатів кластеризації**

Метод *k-середніх (k-means clustering)* суттєво відрізняється від ієрархічних агломеративних методів процедурою попереднього визначення кількості кластерів та дозволяє побудувати їх на якомога більших відстанях один від іншого. За попереднім масивом регіонів та показниками відтворення населення реалізуємо метод *k-середніх.* Налаштування параметрів кластеризації матиме наступний вигляд (рис. 2.20), що включає наступні характеристики аналізу: змінні (*Variables)*, об'єкти *(Cluster)*, число кластерів *(Number of clusters)*, число ітерацій *(Number of iterations)*, початкові центри кластерів – опції (*Initial cluster centers*).

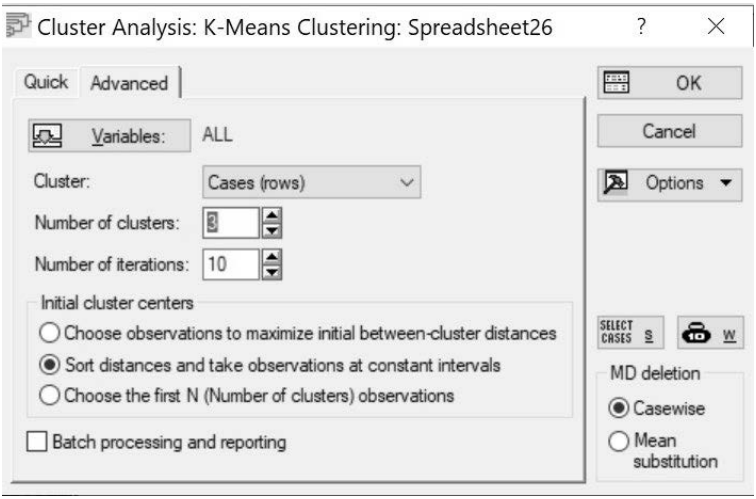



*Рис. 2.20.* **Діалогове вікно методу k-середніх**

В результаті проведення процедури на кожній ітерації об’єкти переміщуються у різні кластери, тому задається максимально бажана кількість таких спроб. Коригування параметрів початкових центрів кластерів впливає на кінцеві результати конфігурації. Опція *Choose observations to maximize initial between-cluster distances (Вибрати набл, максиміз. початкові відстані між кластерами)* обирає перші *k* відповідно до кількості кластерів, спостережень, які є центрами кластерів. Наступні спостереження замінюють раніше обрані центри у разі, якщо найменша відстань до будь-якого з них

більша, ніж найменша відстань між кластерами. Як результат початкові відстані між кластерами максимізуються. За опцією *Sort distances and take observations at constant intervals (Сортувати відстані та вибрати спостереження на постійних інтервалах)*, спочатку сортуються відстані між усіма об'єктами, а потім як початкові центри кластерів вибираються спостереження на незмінних інтервалах. *Choose the first N (Number of cluster) (Вибрати перші N [кількість кластерів]спостережень)* бере перші N (кількість кластерів) спостережень як початкові центри кластерів.

Результати аналізу методом *k-середніх* включають характеристику кількості змінних (5), кількості спостережень (20), назву методу, кількості кластерів (3) та реалізованих ітерацій (3) (рис. 2.21). Також є можливість розширеного представлення підсумків за інформацією *Advanced* у наступних напрямах.

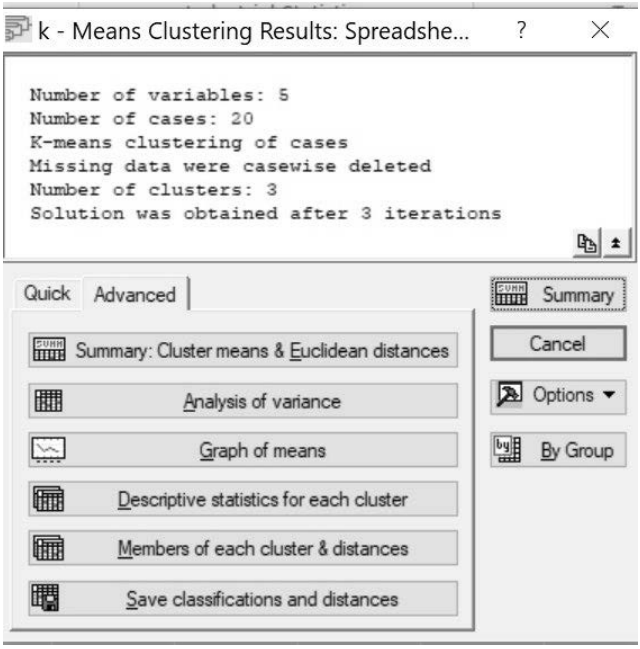



*Рис. 2.21.* **Опції результатів методом *k*-середніх**

Ефективність класифікації оцінюється параметрами

*Analysis of variance (дисперсійний аналіз*) (рис. 2.22): міжгрупова дисперсія *(between SS)*, внутрішньогрупова дисперсія *(within SS)*, відповідні значення числа ступенів свободи для кожної *(df)*, значення *F*-критерію для перевірки гіпотези про нерівність дисперсій між кластерами та всередині них, *p-level*. Якщо гіпотеза відхиляється, отримана класифікація не має сенсу, тому що дані є статистично однорідними та не можуть бути розділені на різні групи. Чим менше значення внутрішньогрупової дисперсії та більше значення міжгрупової, тим краще ознака характеризує приналежність об'єктів до кластера. Параметри *F* і *p* визначають внесок ознаки в класифікацію.

У наведеному прикладі розраховані значення рівня значущості дозволяють прийняти гіпотезу про нерівність дисперсій (значення менше за 0,05), тому розділення сукупності регіонів на три кластера є обґрунтованим.

| | Analysis of Variance (Spreadsheet26) | | | | | |
|---|---|---|---|---|---|---|
| Variable | **Between SS** | df | Within SS | df | F | signif. p |
| **Var1** | 9,25258 | 2 | 9,74742 | 17 | 8,06849 | 0,003437 |
| Var2 | 15,48268 | 2 | 3,51732 | 17 | 37,41558 | 0,000001 |
| Var3 | 11,43189 | 2 | 7,56811 | 17 | 12,83955 | 0,000400 |
| Var4 | 6,51348 | 2 | 12,48652 | 17 | 4,43394 | 0,028206 |
| Var5 | 6,84419 | 2 | 12,15581 | 17 | 4,78583 | 0,022452 |

*Рис. 2.22.* **Результати дисперсійного аналізу**

*Summary:Cluster Means & Euclidean Distances (середні значення у кластерах та евклідові відстані)* складається з двох таблиць: перша містить середні значення показників у кожному кластері; друга - Евклідові відстані (під головною діагоналлю) та квадрат евклідових відстаней (над діагоналлю) між центрами кластерів (рис. 2.23).

| Variable | Cluster Means (Spreadsheet26) | | |
|---|---|---|---|
| | **Cluster No. 1** | Cluster No. 2 | Cluster No. 3 |
| **Var1** | -0,120223 | 1,066981 | -0,82653 |
| Var2 | 0,808036 | -0,315670 | -1,30040 |
| Var3 | -0,070322 | -0,994243 | 1,13489 |
| Var4 | -0,454907 | -0,032400 | 0,94221 |
| Var5 | -0,144876 | -0,656648 | 0,94640 |

| Cluster Number | Euclidean Distances between Clusters<br>Distances below diagonal<br>Squared distances above diagonal | | |
|---|---|---|---|
| | **No. 1** | No. 2 | No. 3 |
| **No. 1** | 0,000000 | 0,793244 | 1,907948 |
| No. 2 | 0,890643 | 0,000000 | 2,521586 |
| No. 3 | 1,381285 | 1,587950 | 0,000000 |

*Рис. 2.23.* **Евклідові відстані між кластерами та середні значення кластерів**

Лінійний графік середніх значень кластерів відображається у *Graf of means (графік середніх значень),* за яким простежуються міжкласові відмінності середніх за усіма ознаками (рис. 2.24).

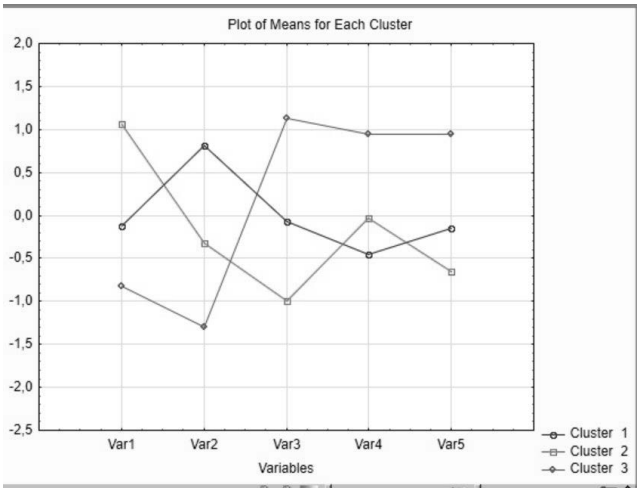



*Рис. 2.24.* **Середні значення ознак для кожного кластера**

*Descriptive Statistics for each cluster (Описова статистика кожного кластера*) виводить три вікна (за кількістю кластерів), у кожному з яких представлені їх характеристики у вигляді середньої, середньоквадратичного відхилення, дисперсії (рис. 2.25).

Descriptive Statistics for Cluster 1 (Sprea
Cluster contains 10 cases

| Variable | Mean | Standard Deviation | Variance |
|---|---|---|---|
| **Var1** | -0,120223 | 0,715491 | 0,511927 |
| Var2 | 0,808036 | 0,486471 | 0,236654 |
| Var3 | -0,070322 | 0,551221 | 0,303845 |
| Var4 | -0,454907 | 0,176440 | 0,031131 |
| Var5 | -0,144876 | 0,873223 | 0,762519 |

Descriptive Statistics for Cluster 2 (Sprea
Cluster contains 5 cases

| Variable | Mean | Standard Deviation | Variance |
|---|---|---|---|
| **Var1** | 1,066981 | 0,553975 | 0,306889 |
| Var2 | -0,315670 | 0,362400 | 0,131334 |
| Var3 | -0,994243 | 0,503136 | 0,253146 |
| Var4 | -0,032400 | 0,190280 | 0,036206 |
| Var5 | -0,656648 | 0,423570 | 0,179412 |

Descriptive Statistics for Cluster 3 (Sprea
Cluster contains 5 cases

| Variable | Mean | Standard Deviation | Variance |
|---|---|---|---|
| **Var1** | -0,82653 | 0,989004 | 0,978129 |
| Var2 | -1,30040 | 0,464248 | 0,215527 |
| Var3 | 1,13489 | 0,977359 | 0,955231 |
| Var4 | 0,94221 | 1,736485 | 3,015379 |
| Var5 | 0,94640 | 1,069520 | 1,143873 |

*Рис. 2.25.* **Описові статистики для кожного кластеру**

*Members for each cluster & distances (Члени кожного кластеру та їх відстані)* демонструють наповненість кожної групи регіонами та евклідову відстань від центру класу до цього спостереження (змінної). Центр класу – середні величини за всіма змінними (спостереженнями) для цього класу (рис. 2.26).

| Case No. | Members of Cluster Number 1 ( and Distances from Respective Cluster contains 10 cases<br>**Distance** |
|---|---|
| C_5 | 0,404069 |
| C_6 | 0,635631 |
| C_7 | 0,601962 |
| C_10 | 0,666481 |
| C_11 | 0,426719 |
| C_14 | 0,298968 |
| C_17 | 0,693592 |
| C_18 | 0,601937 |
| C_19 | 0,689244 |
| C_20 | 0,599298 |

| Case No. | Members of Cluster Number 2 ( and Distances from Respective Cluster contains 5 cases<br>**Distance** |
|---|---|
| C_8 | 0,448454 |
| C_9 | 0,339308 |
| C_13 | 0,418260 |
| C_15 | 0,453317 |
| C_16 | 0,170031 |

| Case No. | Members of Cluster Number 3 ( and Distances from Respective Cluster contains 5 cases<br>**Distance** |
|---|---|
| C_1 | 1,519630 |
| C_2 | 0,651103 |
| C_3 | 0,535450 |
| C_4 | 0,700299 |
| C_12 | 1,239426 |

*Рис. 2.26.* **Склад кластерів та їх відстані до центру кластера**

*Save classifications and distances* дозволяє зберегти результати.

Результати наповнення класів за ієрархічним методом і методом *k-середніх* співпадають.

Метод кластерізації *Two-way joining (Двовходове об'єднання)* одночасно класифікує як спостереження, так і змінні. Вибір змінних аналізу через *Variablema* та груп операцій *Threshold Value (Значення порога)* з режимами *User defined (Задане користувачем)* або *Computed from data (Std.Dev./2) (Обчислене за даними)* дозволяє отримати кластери з характеристиками *Descriptive statistics for cases (rows) (Описові статистики [рядок])* та *Descriptive statistics for variables (Описові статистики для змінних)*, *Reordered statistics*

*for variables (Переупорядкована матриця даних)* та візуальним їх представленням *Summary: Two-way joining graph*. Складності в інтерпретації отриманих результатів, що по'вязані з неоднорідністю отриманих кластерів, не сприяють широкій практиці використання процедури *Two-way joining.*

Для апробації можливостей програми *Statistica* у ***дискримінантному аналізі*** скористаємось умовою зад. *2.8* (табл.*2.5*). Сформуємо таблицю вихідних даних, у якій залишимо вільний рядок без заповнення для нового об'єкту дискримінації. Обираємо функцію *Discriminant Analysis (Дискримінантний аналіз),* у *Variables* призначаємо залежну *Grouping* **(**група) та незалежні *Independent* (ОВЗ, ПВП, Прибуток) змінні (рис. *2.27*).

| | 1<br>ОВЗ | 2<br>ПВП | 3<br>Прибуток | 4<br>Група |
|---|---|---|---|---|
| 1 | 224,228 | 17,115 | 22,981 | П |
| 2 | 151,827 | 14,904 | 21,481 | П |
| 3 | 147,313 | 13,627 | 28,669 | П |
| 4 | 152,253 | 10,545 | 10,199 | П |
| 5 | 46,757 | 4,428 | 11,124 | В |
| 6 | 29,033 | 5,51 | 6,091 | В |
| 7 | 52,134 | 4,214 | 11,842 | В |
| 8 | 37,05 | 5,227 | 11,875 | В |
| 9 | 63,979 | 4,211 | 12,86 | В |
| 10 | | | | |

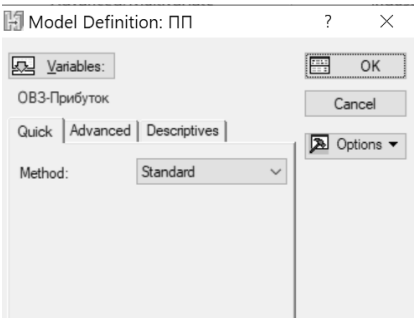



*Рис. 2.27.* **Стартова панель дискримінантного аналізу**

Позиція *Method* дозволяє задавати метод включення змінних у модель: *Standart (Стандартний), Forward stepwise (Покроковий із включенням) та Backward stepwise (Покроковий із винятком)*, що активують відповідну реалізацію процедури.

Результати проведення аналізу містять інформацію про *Number of variables in the model (число змінних у моделі)* – 3; *Wilks' Lambda (значення лямбди Уілкса)* = 0,0928107; *Approx. F* (3, 5) *(наближене значення F – статистики, пов'язаної з лямбдою Уілкса*) = 16,29104; $p < 0,052$ – *рівень значущості F – критерію* для значення 16,29104 (табл. 2.28).

Значення статистики Уілкса в інтервалі [0,1] та наближається

до 0 свідчить про хорошу дискримінацію (якщо значення ближчі к 1→погана дискримінація); показники *Wilks' Lambda* та *F- критерію* вказують на коректність класифікації (фактичне значення більше за табличне $F_{0,05}(3,5) = 5,41$, нульова гіпотеза про належність спостережень до одного класу відхиляється, тому дискримінантний аналіз можливий).

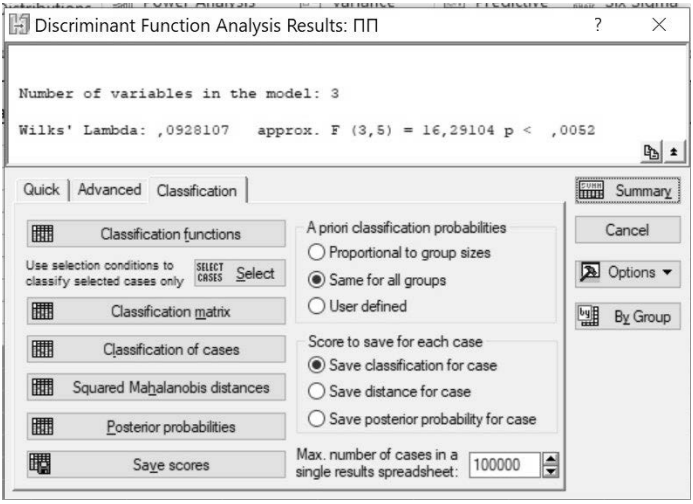



*Рис. 2.28.* **Результати дискримінантного аналізу**

Для перевірки коректності навчальних вибірок подивимося на результати класифікаційної матриці з кількістю спостережень у кожному класі та ймовірністю їх потрапляння у групу, натиснувши кнопку *Classification matrix (Матриця класифікації)* (рис. 2.29), попередньо обравши *Same for all groups* у правій частині вікна *Discriminant Function Analisis Results* (рис. 2.28). Зі значень матриці можна зробити висновок, що об'єкти були вірно віднесені експертним шляхом до виділених груп. Для контролю підприємств, які неправильно віднесені до відповідних груп, доцільно переглянути *Classification of cases (Класифікація спостережень)* (рис. 2.30).

У таблиці класифікацій спостережень некоректно віднесені об'єкти помічаються зірочкою (*). Таким чином, задача отримання

коректних навчальних вибірок у тому, щоб виключити з них об'єкти, які за своїми показниками не відповідають характеристикам однорідності групи.

Classification Matrix (ПП)
Rows: Observed classifications
Columns: Predicted classifications

| Group | Percent Correct | П p=,44444 | В p=,55556 |
|---|---|---|---|
| **П** | 100,0000 | 4 | 0 |
| В | 100,0000 | 0 | 5 |
| Total | 100,0000 | 4 | 5 |

*Рис. 2.29.* **Матриця класифікацій**

З цією метою за допомогою метрики Махаланобіса визначається відстань від усіх *n*-об'єктів до центру тяжіння кожної групи (вектор середніх) за навчальною вибіркою. Віднесення експертом *i-го* об'єкта в *j-ю* групу вважається помилковим, якщо відстань Махаланобіса від об'єкта до центру його групи значно вища порівняно з відстанню від нього до центру інших груп, а апостеріорна ймовірність потрапляння у свою групу нижча за критичне значення. В цьому випадку об'єкт вважається некоректно віднесеним та повинен бути виключений з вибірки.

Classification of Cases (ПП)
Incorrect classifications are marked with *

| Case | Observed Classif. | 1 p=,44444 | 2 p=,55556 |
|---|---|---|---|
| **1** | П | П | В |
| 2 | П | П | В |
| 3 | П | П | В |
| 4 | П | П | В |
| 5 | В | В | П |
| 6 | В | В | П |
| 7 | В | В | П |
| 8 | В | В | П |
| 9 | В | В | П |

*Рис. 2.30.* **Класифікація спостережень**

Процедура виключення об'єктів з навчальних вибірок та повторення процесу тестування триває до досягнення 100% загального коефіцієнту коректності у матриці класифікації та правильного віднесення усіх спостережень до відповідних груп. У наведеному прикладі не спостерігається випадків неправильної класифікації та підтверджується гарна апроксимація дискримінантних функцій[15] (рис. 2.30).

Завдяки *Classification functions (Функції класифікації)* визначаються коефіцієнти дискримінантних функцій для двох класів (рис. 2.31) та ймовірність зарахування підприємства до тієї чи іншої групи.

| | Classification Functions; grouping: Група | |
|---|---|---|
| Variable | П<br>p=,44444 | В<br>p=,55556 |
| ОВЗ | 0,1214 | 0,02047 |
| ПВП | 3,0683 | 0,94621 |
| Прибуток | 0,0324 | 0,17434 |
| Constant | -32,9494 | -4,22645 |

*Рис. 2.31.* **Коефіцієнти дискримінантних функцій**

Лінійні дискримінантні функції для двох виділених класів матимуть наступний вигляд:

$$f_1 = -32{,}9494 + 0{,}1214 \cdot x_1 + 3{,}0683 \cdot x_2 + 0{,}0324 \cdot x_3$$

$$f_2 = -4{,}22454 + 0{,}02047 \cdot x_1 + 0{,}94621 \cdot x_2 + 0{,}17434 \cdot x_3$$

Для реалізації класифікації нового об’єкту, не закриваючи *Discriminant Function Analisis Results*, до таблиці з первинними даними у залишений рядок, вносяться його параметри. Активувавши *Posterior probabilities (Апостеріорні ймовірності)* та

[15] Лабораторний практикум з навчальної дисципліни "Статистичні методи оцінки регіонального розвитку" для студентів напряму підготовки 6.030506 "Прикладна статистика" денної форми навчання : / уклад. О. В. Раєвнєва, І. В. Аксьонова, Г. І. Свидло. Харків : ХНЕУ ім. С. Кузнеця, 2016. 68 с.

обравши один зі способів визначення ймовірності *(Proportional to group sizes - пропорційні розмірам груп, Same for all groups - однакові для всіх груп, User defined - задані користувачем),* отримуємо результат процедури розпізнавання образу з учителем (табл. 2.32).

| Case | Posterior Probabilities (Spreadsheet69) Incorrect classifications are marked with<br>**Observed Classif.** | П<br>p=,44444 | В<br>p=,55556 |
|---|---|---|---|
| 1 | П | 1,000000 | 0,000000 |
| 2 | П | 1,000000 | 0,000000 |
| 3 | П | 0,999983 | 0,000017 |
| 4 | П | 0,999484 | 0,000516 |
| 5 | В | 0,000000 | 1,000000 |
| 6 | В | 0,000000 | 1,000000 |
| 7 | В | 0,000000 | 1,000000 |
| 8 | В | 0,000000 | 1,000000 |
| 9 | В | 0,000000 | 1,000000 |
| 10 | --- | 0,021203 | 0,978797 |

*Рис. 2.32.* **Апостеріорні ймовірності**

Нове підприємство доцільно приєднати до класу відсталих за значенням максимальної ймовірності, Такий результат збігається з розрахунками та висновками зад. *2.8.* Для визначення приналежності нових об'єктів до виділених класів доречно також залучити опції *Classification of cases (Класифікація спостережень), Squared Mahalanobis distances (Квадрати відстаней Махаланобіса)* та результати *Classification functions (Функції класифікації)* у вигляді лінійних дискримінантних функцій.

За допомогою *Scatterplot of canonical scores* в опції *Perform canonical analysis (Канонічний аналіз)* на вкладці *Canonical scores* можна побудувати діаграму розсіву об'єктів у просторі канонічних коренів.

***Список питань до самоконтролю:***

**1.** Сформулюйте сутність процесу класифікації.

**2.** У чому полягає відмінність класифікації з вчителем та розпізнавання образів без вчителя?

**3.** Поясність завдання кластерного аналізу.

**4.** Які міри відстані використовують при проведенні кластерного аналізу?

**5.** Як оцінюється якість кластеризації?

**6.** Сформулюйте алгоритм та принцип використання ієрархічних процедур класифікації.

**7.** Алгоритм методу *k- середніх*.

**8.** Поясність особливості ітеративного методу пошуку згущення об’єктив.

**9.** У чому полягає сутність та процедура дискримінантного аналізу?

**10.** Дайте характеристику нових розробок методів кластеризації.

**11.** Провести класифікацію ієрархічним агломеративним методом домогосподарств, кожне з яких характеризується двома ознаками:

1) з використанням евклідової метрики: а) методом найближчого сусіда; б) методом далекого сусіда; в) методом середнього сусіда; г) центроїдним методом;

2) з використанням зваженої евклідової метрики (вага 0,25 та 0,75) методом найближчого сусіда.

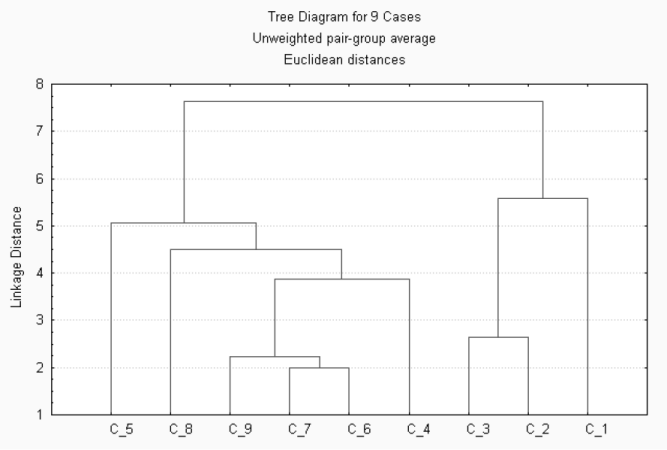
| № п/п | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| споживання фруктів, кг/місяць ($x_1$) | 22,3 | 17,8 | 9,9 | 19,4 | 7,7 | 9,2 | 7,1 |
| споживання молока, л/місяць ($x_2$) | 9,2 | 5,6 | 5,2 | 10,1 | 8,4 | 6,3 | 8,9 |

**12.** За результатами дослідження успішності діяльності команди-малої групи, що орієнтована на вирішення ділових завдань та складається з молодих інженерів-програмістів, провести класифікацію спеціалістів за наведеними ознаками. Для кластеризації використати евклідову відстань за методом далекого сусіда.

| № п/п | Залежність від групових стандартів | Відповідальність | Трудова активність | Працездатність | Розуміння мети | Мотивація |
|---|---|---|---|---|---|---|
| 1 | 2 | 7 | 9 | 8 | 10 | 3 |
| 2 | 4 | 2 | 8 | 8 | 8 | 1 |
| 3 | 2 | 3 | 9 | 7 | 8 | 1 |
| 4 | 7 | 3 | 5 | 6 | 4 | 0 |
| 5 | 2 | 2 | 5 | 3 | 7 | 2 |
| 6 | 4 | 3 | 5 | 5 | 5 | 2 |
| 7 | 5 | 4 | 4 | 5 | 5 | 3 |
| 8 | 6 | 1 | 4 | 4 | 7 | 0 |
| 9 | 5 | 3 | 3 | 5 | 4 | 2 |

**13.** Прокоментуйте кластеризацію молодих інженерів-програмістів за показниками успішності діяльності команди-малої

групи, реалізованої за результатами попереднього завдання у системі STATISTICA.




**14.** Класифікувати об'єкти за чотирма змінними методом *k-середніх*, виділити дві групи.

| Об'єкти | A | B | C | D |
|---|---|---|---|---|
| $x_1$ | 6 | 1 | -1 | 4 |
| $x_2$ | 4 | -2 | 1 | 2 |

**15.** Провести класифікацію ієрархічними агломеративними методами країн за показниками, які характеризують демографічну ситуацію. Кластеризацію реалізувати за допомогою пакета STATISTICA за допомогою евклідової відстані методом найближчого сусіда без стандартизації. Результати проаналізувати у розрізі автоматично отриманої інформації, дати характеристику складу кожного кластеру, його статистичним параметрам та представити у вигляді дендрограми.

| Країна | Коефіцієнт народжуваності (на 1000 осіб населення), ‰ | Коефіцієнт смертності (на 1000 осіб населення), ‰ |
|---|---|---|
| 1 | 14,9 | 5,9 |
| 2 | 9,6 | 6,3 |
| 3 | 9,3 | 14,2 |
| 4 | 8,9 | 8,8 |
| 5 | 14,0 | 9,7 |
| 6 | 21,4 | 6,8 |
| 7 | 10,6 | 11,3 |
| 8 | 8,3 | 14,7 |
| 9 | 18,6 | 4,2 |
| 10 | 18,5 | 5,4 |
| 11 | 22,3 | 5,3 |
| 12 | 7,8 | 14,8 |

**16.** Провести класифікацію промислових підприємств методом *k-середніх*, поділив сукупність на три групи (передові, середні, відстаючі). Кластеризацію реалізувати за допомогою пакета STATISTICA здійснивши попередню стандартизацію даних. Результати проаналізувати у розрізі автоматично отриманої інформації дати характеристику складу кожного кластеру, його статистичним параметрам та представити у вигляді дендрограми.

| № п/п | Продуктивність праці, тис. грн./чел. | Заробітна плата одного робочого, грн. | Капіталовіддача, грн. |
|---|---|---|---|
| 1 | 170,4 | 60200 | 1,43 |
| 2 | 165.7 | 58900 | 1,30 |
| 3 | 169,3 | 61002 | 1,37 |

| | | | |
|---|---|---|---|
| 4 | 172,5 | 68200 | 1,65 |
| 5 | 179,7 | 68305 | 1,91 |
| 6 | 179,4 | 67900 | 1,68 |
| 7 | 181,5 | 68211 | 1,94 |
| 8 | 180,5 | 68500 | 1,89 |
| 9 | 188,9 | 67900 | 1,94 |
| 10 | 198,1 | 68909 | 2,06 |
| 11 | 191,0 | 69400 | 1,96 |
| 12 | 164,6 | 52280 | 1,02 |
| 13 | 177,3 | 67980 | 1,85 |
| 14 | 156,5 | 57403 | 0,88 |
| 15 | 143,7 | 57002 | 0,62 |

**17.** Провести класифікацію ієрархічними агломеративними методами регіонів за їх соціальними характеристиками. Кластеризацію реалізувати за допомогою пакета Statistica за допомогою евклідової відстані методом середнього зв'язку, попередньо стандартизував дані. Результати проаналізувати у розрізі автоматично отриманої інформації, дати характеристику складу кожного кластеру, його статистичним параметрам та представити у вигляді дендрограми. За класифікацію визначте оптимальну кількість груп та реалізуйте кластеризацію методом *k-середніх*. Порівняйте отримані варіанти між собою.

| Регіон | Рівень бідності, % | Відношення середнього грошового доходу до прожиткового мінімуму, разів | Рівень безробіття, % |
|---|---|---|---|
| 1 | 43,3 | 1,74 | 12,7 |
| 2 | 23,8 | 3,48 | 14,3 |
| 3 | 28,2 | 2,44 | 12,1 |

| | | | |
|---|---|---|---|
| 4 | 20,0 | 3,04 | 12,2 |
| 5 | 52,6 | 1,44 | 13,7 |
| 6 | 24,9 | 2,75 | 8,0 |
| 7 | 78,0 | 0,99 | 21,2 |
| 8 | 49,0 | 2,21 | 18,9 |
| 9 | 42,0 | 1,94 | 18,0 |
| 10 | 68,3 | 1,22 | 20,9 |
| 11 | 47,6 | 1,90 | 16,5 |

**18.** Діяльність дванадцяти машинобудівних підприємств характеризується показниками рентабельності (x1,%) та продуктивності праці (x2, тис. грн. / чол.). Перші чотири підприємства мають високий рівень організації управління, а п'ять підприємств – низький. Потрібно обґрунтувати та реалізувати відповідний метод класифікації останніх трьох підприємств.

| № п/п | Групи підприємств | $x_1$ | $x_2$ |
|---|---|---|---|
| 1 | Високий | 23,4 | 9,1 |
| 2 | | 19,1 | 6,6 |
| 3 | | 17,5 | 5,3 |
| 4 | | 17,2 | 10,0 |
| 5 | Низький | 5,4 | 4,3 |
| 6 | | 6,6 | 5,5 |
| 7 | | 8,0 | 5,7 |
| 8 | | 9,7 | 5,5 |
| 9 | | 9,1 | 6,6 |
| 10 | Потребують класифікації | 9,9 | 7,4 |
| 11 | | 14,2 | 9,4 |
| 12 | | 12,9 | 6,7 |

**19.** Використовуючи алгоритм кластерного аналізу, сформуйте з перших десяти спостережень дві повчальні вибірки. На підставі отриманих вибірок проведіть класифікацію п'яти фірм, що залишилися. Дайте економічну інтерпретацію результатів.

| № п/п | Капіталоозброєність праці, тис. грн./чел. | Капіталовіддача основних засобів, грн./грн. | Питома вага робочих у складі персоналу |
|---|---|---|---|
| 1 | 4,82 | 1,67 | 0,46 |
| 2 | 3,85 | 1,78 | 0,72 |
| 3 | 4,75 | 1,23 | 0,67 |
| 4 | 5,35 | 1,32 | 0,71 |
| 5 | 8,7 | 0,75 | 0,66 |
| 6 | 7,3 | 1,15 | 0,72 |
| 7 | 6,4 | 1,26 | 0,7 |
| 8 | 5,9 | 1,43 | 0,75 |
| 9 | 6 | 1,28 | 0,63 |
| 10 | 8,95 | 0,95 | 0,76 |
| 11 | 7,41 | 1,18 | 0,69 |
| 12 | 4,71 | 1,9 | 0,71 |
| 13 | 5,03 | 1,81 | 0,72 |
| 14 | 6,94 | 1,29 | 0,73 |
| 15 | 7,95 | 0,98 | 0,7 |

***Список рекомендованої літератури по темі:***

1. Бахрушин В.Є. Методи аналізу даних: навчальний посібник для студентів / В.Є. Бахрушин. – Запоріжжя: КПУ, 2011. – 268 с.

2. Бізнес-аналітика багатовимірних процесів : навчальний посібник / Т. С. Клебанова, Л. С. Гур'янова, Л. О. Чаговець та ін. – Харків : ХНЕУ ім. С. Кузнеця, 2018. 272 с.
3. Єріна А.М. Статистичне моделювання та прогнозування: підручник /А.М. Єріна, Д.Л. Єрін. К.: КНЕУ, 2014. 348 с.
4. Лабораторний практикум з навчальної дисципліни "Статистичні методи оцінки регіонального розвитку" для студентів напряму підготовки 6.030506 "Прикладна статистика" денної форми навчання : / уклад. О. В. Раєвнєва, І. В. Аксьонова, Г. І. Свидло. Харків : ХНЕУ ім. С. Кузнеця, 2016. 68 с.
5. Пістунов І.М., Антонюк О.П., Турчанінова І.Ю. Кластерний аналіз в економіці: Навч. Посібник. Дніпропетровськ: Національний гірничий університет, 2008. 84 с.
6. Янковий О.Г. Латентні ознаки в економіці: монографія. Одеса: Атлант, 2015. – 168 с.
7. Яровий А.Т., Страхов Є.М. Багатовимірний статистичний аналіз: навчально-методичний посібник для студентів математичних та економічних фахів. Одеса: Астропринт, 2015. 132 с.
8. Everitt, B. S., Landau, S., Leese, M., Stahl, D. (2011). Cluster Analysis. 5th ed, John Wiley & Sons, Ltd.
9. Hennig, C., Meila, M., Murtagh, F., Rocci, R. (2016). Handbook of Cluster Analysis. CRC Press, Taylor & Francis Group.
10. Manly, Bryan F.J. (2005). Multivariate Statistical Methods: A primer, Third edition, Chapman and Hall.
11. Rencher, A.C., Christensen, W.F. (2002). Methods of Multivariate Analysis, Second edition, Wiley.

# Розділ 3. МОДЕЛЮВАННЯ РЯДІВ РОЗПОДІЛУ

## 3.1. Випадкові величини як джерело статистичних даних у моделюванні рядів розподілу

Розподіл більшості реальних явищ і процесів описуються ламаними кривими з неявною закономірністю. Для виявлення закономірностей емпіричного розподілу одиниць сукупності за значенням варіанти здійснюють його моделювання. Моделювання ряду розподілу полягає у доборі теоретичної функції розподілу, яка б найкращим чином описувала емпіричний розподіл. Теоретичний розподіл описується плавною лінією, тобто кривою розподілу, що характеризує залежність між варіантами і частотами.

Моделюванням рядів розподілу вирішуються такі задачі:

– перевірка гіпотези щодо відповідності емпіричного розподілу нормальному, що є необхідною умовою застосування статистичних методів аналізу взаємозв'язку, тенденції розвитку тощо;

– згладжування випадкових відхилень емпіричного ряду;

– визначення відсутніх проміжних частот в інтервалах емпіричного ряду;

– прогнозування розподілів взаємопов'язаних явищ та процесів (наприклад, аналізуючи розподіл населення за рівнем доходів і структурою витрат можна прогнозувати обсяг споживання товарів і послуг або імітувати його зміну залежно від рівня інфляції).

Випадкова величина – це функція, яка описує кількісний результат випадкових подій у вибірковому просторі. Слід звернути увагу, що випадкова величина відноситься до випадкового процесу, тоді як значення спостереження – це фіксована величина, яка фіксується при спостереженні цього процесу. Правило визначення

імовірностей значень випадкової величини називається розподілом імовірностей.

Існує два типи випадкових величин:

– дискретні, для яких існує дискретний набір числових значень та можна перерахувати всі можливі значення такої величини в результаті спостережень;

– неперервні, приймають будь-яке значення з певного інтервалу та можливі значення такої величини неможливо перерахувати.

За умови відсутності інформації про імовірність настання певного результату відносного іншого – розподіли імовірностей обох подій буде однаково імовірними. Такий рівномірний розподіл можна застосовувати для суб'єктивної оцінки ймовірностей для дискретних або неперервних величин.

Значення дискретної випадкової величини та пов'язані з нею ймовірності можуть бути виражені через функцію ймовірностей. Неперервні випадкові величини та їх ймовірності описуються функцією щільності ймовірностей.

Функція ймовірностей для дискретного рівномірного розподілу за умови мінімального ($a$) та максимального ($b$) значення результату, кількості однаково ймовірних результатів ($n$) буде мати наступний вигляд: $f(x) = 1/n = 1/(b – a + 1)$. Слід зазначити, що для будь якої функції ймовірність заданого значення ($x$) має бути в діапазоні від 0 до 1, а сума ймовірностей всіх значень має дорівнювати 1[1].

Функція щільності ймовірностей для неперервних випадкових величин за умови мінімального ($a$) та максимального ($b$) значення результату, буде мати наступний вигляд: $f(x) = 1/n = 1/(b – a)$[1].

Графічне зображення функцій для дискретних та неперервних випадкових величин зображено на рис. 3.1:

[1] Pinder, Jonathan P. Introduction to Business Analytics using Simulation / Jonathan P. Pinder — 2017. – P. 434.

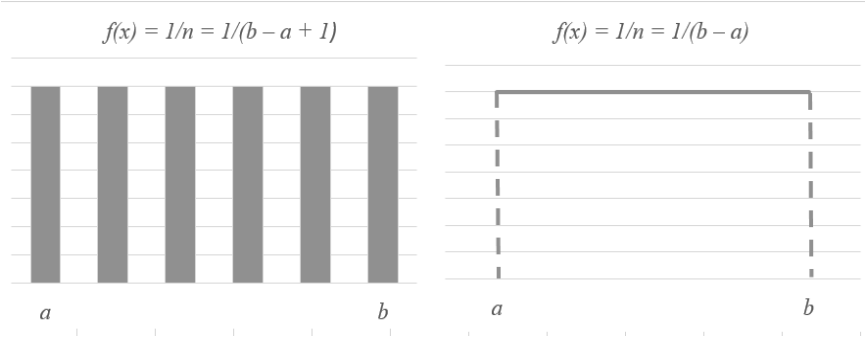



*Рисунок 3.1.* **Функція ймовірності та щільності ймовірності для дискретного та неперервного рівномірного розподілу**

Дискретні випадкові величини можна охарактеризувати за допомогою:

– очікуваного (середнього) значення випадкової величини – $E(x) = \sum x \cdot p(x)$;

– стандартного відхилення випадкової величини (рівень ризику у бізнесі) – $\sigma = \sqrt{\sum (x - E(x))^2 \cdot p(x)}$.

На підставі даних проєктів про можливі обсяги виробництва, необхідно оцінити рівень ризику та надати рекомендації щодо обрання одного з проєктів. Вихідні дані подано у табл. 3.1:

*Таблиця 3.1.* **Обсяги виробництва та їх ймовірності**

| Проєкт А | | Проєкт Б | |
|---|---|---|---|
| Обсяг виробництва, тис шт. | Імовірність обсягу виробництва | Обсяг виробництва, тис шт. | Імовірність обсягу виробництва |
| 1500 | 0,1 | 1800 | 0,2 |
| 2000 | 0,6 | 2200 | 0,6 |
| 2500 | 0,3 | 2000 | 0,2 |

Середній очікуваний обсяг виробництва та стандартне відхилення кожного з проектів буде становити:

- проект А:

$E(x) = 1500 \cdot 0{,}1 + 2000 \cdot 0{,}6 + 2500 \cdot 0{,}3 = 2100$ тис шт.

$\sigma = \sqrt{(1500 - 2100)^2 \cdot 0{,}1 + (2000 - 2100)^2 \cdot 0{,}6 + (2500 - 2100)^2 \cdot 0{,}3} = 300$ тис шт.

- проект Б:

$E(x) = 1800 \cdot 0{,}2 + 2200 \cdot 0{,}6 + 2000 \cdot 0{,}2 = 2080$ тис шт.

$\sigma = \sqrt{(1800 - 2080)^2 \cdot 0{,}2 + (2200 - 2080)^2 \cdot 0{,}6 + (2000 - 2080)^2 \cdot 0{,}2} = 160$ тис шт.

Наведені розрахунки не дають підстави для остаточних висновків щодо обрання одного із запропонованих проєктів. Серед двох проєктів більший обсяг виробництва можна очікувати за проєктом А, де рівень ризику буде більше або менше середнього значення на 300 тис шт. Тоді як за проєктом Б, можливі межі коливання відносно очікуваного обсягу виробництва будуть становити 2080±160 тис шт.

Цікавим за практичною спрямованістю у розподілі величин є застосування Закону Бенфорда. Астроном Саймон Ньюкомб, працюючи з книгою таблиць логарифмів, виявив, що сторінки на початку книги заслинені сильніше, ніж інші сторінки. Така дивність спостерігалася не лише на одному конкретному екземплярі але і на більшості інших. Розмірковуючи на цю тему, Ньюкомб дійшов висновку, що це не може бути простим збігом обставин, що призвело до відкриття закону розподілу чисел. Згідно закону, якщо випадково вибрати будь-яке число з таблиці, яка містить фізичні значення або статистичні дані, вірогідність того що воно буде починатися з одиниці, приблизно складає 0,301.

Американським фізиком Френком Бенфордом було продовжено дослідження розподілу чисел, під час якого було

проаналізовано близько 20 таблиць різних за природою явищ. Результатом став висновок про те, що має місце одна і та ж сама закономірність: чисел, що розпочинаються з одиниці, значно більше, чим чисел, які розпочинаються з будь-якої іншої цифри, яка може бути виражена математично у такий спосіб:

$$p(d) = lg(d+1) - lg(d) \tag{3.1}$$

Аналіз чисел показав, що одиниця є першою значущою цифрою з імовірністю не 1/9, як слід було чекати, а близько 1/3. Згодом закон Бенфорда отримав своє пояснення – він притаманний безлічі чисел, які можуть зростати експоненціально (іншими словами, темп зростання величини пропорційний її поточному значенню). Наприклад, до них входять рахунки за електрику, залишки товарів на складах, ціни на акції, чисельність населення, смертність, довжини річок, площі країн, висоти найвищих споруд у світі.

Закон, виявлений Бенфордом, виглядає так: якщо основа системи числення $b$ ($b > 2$), то для цифри $d$ ($d \in \{1, \ldots, b-1\}$) імовірність бути першою значущою цифрою складає:

$$p(d) = log_b(d+1) - log_b(d) = log_b\left(1 + \frac{1}{d}\right) \tag{3.2}$$

Це в точності відстань між d і d+1 на логарифмічній шкалі.

Для рівномірного розподілу, якщо ми маємо послідовність цифр 1, 2, 3, 4, 5, 6, 7, 8, 9, 0 (=10), тоді у нас є 10 відрізків (від 0 до 1, ..., від 8 до 9, від 9 до 10). Зверніть увагу, всі відрізки лежать в інтервалі [0, 10]. Для відрізку [$d$, $d+1$] рівномірний розподіл має бути пропорційний його довжині, тобто довжині відрізку [$d$, $d+1$], тобто $(d+1) - d$, поділене на довжину відрізку [0, 10], яка складає 10–0:

$$\frac{(d+1)-d}{10-0}=\frac{1}{10} \tag{3.3}$$

Якщо логарифми неперервно розподілені, то слід застосовувати функцію логарифма до числа до того, як ви розглядаємо відрізки. Для логарифмів розглядаємо відрізки від 1 до 10 (оскільки *log100* не має сенсу). У цьому випадку матимемо інтервали від *lg1* до *lg2*, ..., від *lg8* до *lg9*, від *lg9* до *lg10*. Усі відрізки лежать в інтервалі [*lg1, lg10*]=[*0, 1*]. Довжина останнього складає 1–0. Отже, відрізок [*d, d+1*] на нормальній шкалі, в логарифмічній шкалі рівномірний розподіл буде пропорційний його довжині, тобто:

$$\frac{log_{10}(d+1)-log_{10}(d)}{1-0}=log_{10}(d+1)-log_{10}(d) \tag{3.4}$$

Отже, чим менше цифра, тим більше імовірність того, що з неї починатиметься випадковий десятковий дріб (табл. 3.2):

*Таблиця 3.2.* **Імовірність появи першої цифри у випадковому десятковому дробі відповідно до закону Бенфорда**

| Цифра | Частота появи першої цифри |
|---|---|
| 1 | 0,3010300 |
| 2 | 0,1760910 |
| 3 | 0,1249390 |
| 4 | 0,0969100 |
| 5 | 0,0791812 |
| 6 | 0,0669468 |
| 7 | 0,0579919 |
| 8 | 0,0511525 |
| 9 | 0,0457575 |

Багато математиків мають сумніви у справедливості висновків Бенфорда, оскільки значення чисел на шкалі від нуля до нескінченості зростає рівномірно. Виникає питання, чому в записі даних тих або інших підрахунків цифри беруть по-різному.

Річ у тім, що при підрахунку беруться не математичні абстракції, а абсолютно конкретні предмети. На абстрактній шкалі натуральних чисел двійка знаходиться від одиниці не далі, ніж шестірка від п'ятірки, але для реальних речей, перелічених, виміряних або зважених, шлях від одного до двох дуже довгий. Щоб його прокласти, кількість (вага, розмір тощо) речей повинна зрости удвічі. П'ятірці, для того, щоб перетворитися на шістку, треба додати усього лише п'яту частину вже наявного. Ще менший крок від 9000 до 10000. Але потім, для того, щоб в першому порядку одиниця змінилася на двійку, знову знадобиться багато часу і сил. Таким чином, все, що зростає в обсязі, розмірі, вазі або ціні, достатньо довго залишається в «області одиниці».

Точна форма Закону Бенфорда може бути пояснена якщо припустити, що рівномірно розподілені логарифми чисел, наприклад, імовірність знаходження числа між 100 і 1000 (логарифм між 2 і 3) буде такий самий, як і між 10000 і 100000 (логарифм між 4 і 5). Для багатьох послідовностей чисел, які описуються експоненціальним зростанням, таких як доходи або ціни на біржі, це розумне припущення.

Цей закон може бути альтернативно пояснений тим фактом, що якщо перша цифра дійсно має особливий розподіл, то він не повинен залежати від величин, в яких він вимірюється. Це означає, що при переводі, наприклад, футів в ярди (множення на константу), розподіл повинен залишитися незмінним. Така властивість носить назву масштабна інваріантність, і єдиний неперервний розподіл, який відповідає цій вимозі – той, в якому логарифм рівномірно розподілений.

Наприклад, перша (не нульова) цифра довжини або відстані об'єкту повинна мати такий же розподіл незалежно від того чи проводиться вимір у футах, ярдах або чомусь іншому. Але в ярді є три фути, тому вірогідність, що перша цифра довжини в ярдах буде 1, має бути такою ж, як вірогідність, що перша цифра довжини у футах 3, 4 або 5. Застосовуючи це до усіх можливих шкал вимірів дає логарифмічний розподіл, і враховуючи що *log10(1) = 0* і *log10(10) = 1* дає закон Бенфорда. Тобто якщо є розподіл першої цифри, який не залежить від одиниць виміру, єдиним розподілом першої цифри може бути той, яке підкоряється закону Бенфорда.

Для чисел, що взяті з певного розподілу, наприклад, значення IQ, росту людей або інших змінних, які підкоряються нормальному розподілу, закон не виконується. Проте, якщо «перемішати» числа з цих розподілів, наприклад, взявши числа з газетних статей, закон Бенфорда знов проявиться. Це також може бути доведено математично: якщо неодноразово і випадково вибирати розподіл ймовірності, а потім випадково вибрати число згідно цього розподілу, послідовність, яка утвориться, підкорятиметься закону Бенфорда.

Бенфорд назвав цю закономірність «законом аномальних чисел», а закон, відкритий Бенфордом, отримав назву «Закон Бенфорда». Однак практичного застосування він так і не знайшов, залишаючись в розряді математичних курйозів.

Закону Бенфорда відповідають послідовності, які є результатом природних подій:

– номери платіжних доручень від різних покупців (уся сукупність);
– суми платежів від покупців;
– суми в авансових звітах;
– залишки товарів на складах;
– номери будинків в адресах клієнтів.

Не відповідають Закону Бенфорда неприродні, штучні системи:

– поштові індекси;

– номери телефонів (перші цифри – номер АТС);

– виграшні номери в лото і рулетку (тут цифри – лише символи, їх легко можна замінити, наприклад, на літери);

– будь-які об'єми даних розмір яких недостатній для застосування статистичних методів;

– суми платежів від покупців і об'єми замовлень, якщо продається декілька позицій однієї номенклатури.

Існують спеціальні тести, які можуть проводитися як на відповідність Закону Бенфорда, так і навпаки. Для проведення тестів набір даних повинен відповідати таким умовам:

По-перше, наявність геометричного розподілу. Дискретна випадкова величина *X* має геометричний розподіл, якщо вона приймає значення *1, 2, .., m, ..* (нескінченна, але лічильна послідовність значень) з імовірністю: $P(X = m) = p \cdot q^{m-1}$, де $0<p<1$, $q = 1 - p$, $m = 1, 2, ..$

Очікуване середнє випадкової величини *X*, яка має геометричний розподіл з параметром *p*, дорівнює:

$$M(x) = \frac{1}{p}, \tag{3.5}$$

а її дисперсія складає

$$\sigma^2(x) = \frac{q}{p^2}, \tag{3.6}$$

де $q = 1 - p$.

Так, імовірність ураження цілі дорівнює 0,6. Здійснюється стрільба по мішені до першого попадання (кількість патронів не обмежена). Треба скласти ряд розподілу кількості зроблених пострілів, знайти очікуване середнє і дисперсію цієї випадкової

величини. Визначити імовірність того, що для ураження цілі знадобиться не більше трьох патронів.

Випадкова величина $x_i$ – кількість зроблених пострілів – має геометричний розподіл з параметром $p$=0,6. Ряд розподілу $x_i$ має вигляд:

| $x_i$ | *1* | *2* | *3* | *...* | *m* |
|---|---|---|---|---|---|
| $p_i$ | *0,6* | *0,24* | *0,096* | *...* | $0,6 \cdot 0,4^m$ |

Отже, очікуване середнє становить – $M(x) = \frac{1}{p} = \frac{1}{0,6} = 1,67$, а дисперсія буде складати – $\sigma^2(x) = \frac{q}{p^2} = \frac{0,4}{(0,6)^2} = 1,11$.

Імовірність того, що для ураження цілі знадобиться не більше трьох патронів дорівнює:

$$p(x \le 3) = p + p \cdot q + p \cdot q^{3-1} = 0,6 + 0,6 \cdot 0,4 + 0,6 \cdot 0,4^2 =$$
$$= 0,6 + 0,24 + 0,096 = 0,936.$$

По-друге, дані повинні належить до однакових об'єктів. Не можна змішувати дані платіжних доручень і дані адрес клієнтів.

По-третє, не повинно бути обмежень для чисел за максимумом і мінімумом. Якщо є деяка межа (наприклад, граничний розмір розрахунків готівкою), то така сукупність даних вже може не бути ідеальною Бенфорд-послідовністю.

По-четверте, відсутність системи нумерації. Числа не мають бути штучними системами. Наприклад, набір цифр в ідентифікаційному номері платника податку (ІНПП) не буде Бенфорд-послідовністю, оскільки перші дві цифри в ІНПП – код регіону, другі дві – код інспекції, а остання цифра - контрольна – обчислюється з усіх попередніх.

Ідея застосування всіх тестів одна: якщо в результаті дослідження і побудови послідовності цифр емпіричних даних виявлені значні розбіжності з еталонними значеннями, то це є

сигналом для проведення спеціального дослідження, яке виявить причину появи таких розбіжностей. Існують декілька видів тестів:

1) аналіз «першої цифри» і «другої цифри»: набір даних аналізується на частоту появи різних цифр – від 1 до 9 в якості першої цифри в числі і від 0 до 9 як другої цифри в числі. Результуючі дані відображаються у таблиці або у формі графіку. За наявності значних розбіжностей з еталонними значеннями проводиться спеціальне дослідження, яке повинне дати відповідь на питання про причину таких розбіжностей.

2) аналіз «першої і другої цифри»: досліджується частота появи цифр від 10 до 99 на початку чисел. Результат представляється у вигляді графіку, який містить порівняння експериментальних даних з еталонними. Комбінації початкових цифр, які не відповідають еталонному значенню, вважаються аномальними. Цей тест ефективний для виявлення штучних обмежень після досягнення деякого встановленого ліміту.

3) аналіз «з першою по третю цифру»: визначає частоту появи комбінацій цифр з 100 до 999 в перших трьох знаках набору даних. Аналогічний попередньому аналізу, за винятком того, що вимагає значно більшого обсягу початкових даних. На малих обсягах (менше 10000 значень) можливі помилкові випадкові викиди. Проте цей метод дозволяє точніше проводити аналіз та виявляти тенденції, які залишилися б непоміченими при використанні попередніх тестів.

4) аналіз «округлення»: проводиться для того, щоб оцінити частоту появи різних цифр в останніх знаках. Він дозволяє виявити випадки систематичного округлення в тих наборах даних, де округлення бути не може за визначенням (наприклад, пробіг автомобілів, свідчення лічильників витрати електроенергії або кількості зроблених копій у копіювального апарату).

5) аналіз «дублікатів»: спочатку знаходяться числа з однаковими значеннями, потім визначається частота появи кожного

з цих чисел. Після чого виводиться таблиця результатів, які упорядковані за спаданням кількості збігів. Аналіз дозволяє виявити випадки аномальної присутності в наборі даних цих однакових значень.

Однак слід заначити, що використання тестів вимагає наявності значних обсягів інформації та великих масивів даних.

## 3.2. Біноміальний розподіл і розподіл Пуассона для дискретних випадкових величин

Біноміальний (дискретний) розподіл має місце у випадку, коли кількість наступів події певної випадкової величини $X$ є відсотком (ймовірністю) від загальної кількості можливих результатів $n$ спроб (обсяг сукупності). Біноміальний розподіл застосовують в теорії та практиці статистики контролю якості продукції, при моделюванні систем масового обслуговування та інших сферах діяльності.

Біноміальний розподіл має місце за таких умов:

– якщо в кожній з $n$ спроб ймовірність настання події $\pi$ (фіксована величина, яка визначає імовірність настання події) однакова;

– якщо всі спроби незалежні одна від одної.

Окрім очікуваного середнього та стандартного відхилення, біноміальний розподіл характеризується біноміальною часткою ($p$):

$$p = \frac{X}{n} \tag{3.7}$$

Для кількість настання події випадкової величини $X$ в разі біноміального розподілу очікуване середнє обчислюється як добуток кількості можливих результатів $n$ спроб на ймовірність їх настання:

$$E(x) = n \cdot \pi \tag{3.8}$$

Тоді стандартне відхилення буде виражатися формулою:

$$\sigma = \sqrt{n \cdot \pi(1-\pi)} \tag{3.9}$$

Очікуване середнє для біноміальної частки буде дорівнювати імовірності настання самої події:

$$E\left(\frac{X}{n}\right) = E(p) = \pi \tag{3.10}$$

Формула для обчислення стандартного відхилення буде розраховано за формулою:

$$\sigma = \sqrt{\frac{\pi(1-\pi)}{n}} \tag{3.11}$$

Представимо, що під час виробництва 2000 пакувальних матеріалів відсоток бракованої продукції становить 6,3 %, який має випадковий характер. Розрахунок показників для біноміального розподілу наведено у табл. 3.3:

*Таблиця 3.3.* **Розрахунок показників для біноміального розподілу виробництва пакувальних матеріалів**

| Показник | Кількість виробленої продукції, $X$ | Частка бракованої продукції, $p$ |
|---|---|---|
| Очікуване середнє | $E(x) = n \cdot \pi = 2000 \cdot 0{,}063 = 126$ | $E(x) = \pi = 0{,}063$ |
| Стандартне відхилення | $\sigma = \sqrt{n \cdot \pi(1-\pi)} = \sqrt{2000 \cdot 0{,}063 \cdot (1-0{,}063)} = 10{,}865$ | $\sigma = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0{,}063 \cdot (1-0{,}063)}{2000}} = 0{,}0054$ |

Таким чином, слід очікувати, що в середньому при виробництві 2000 пакувальних матеріалів очікується 126 бракованих виробів. Величина цієї невизначеності становить 10,865 виробів. Очікується також, що за умови наявності бракованої продукції відсоток браку буде в межах 6,3±0,54 %.

У біноміальному розподілі, у випадку малих і середніх значень *n* та відомої величини $\pi$, додатні ймовірності величини *X* при заданому значення *a*, визначаються за формулою:

$$P(X=a)=\binom{n}{a}\cdot\pi^{a}(1-\pi)^{n-a}=\frac{n!}{a!\cdot(n-a)!}\cdot\pi^{a}(1-\pi)^{n-a} \qquad (3.12)$$

При великих *n* можна використовувати наближення основане на нормальному розподілі для визначення ймовірностей.

Надлишкові резерви банку в для надання кредитів у поточному місяці становлять 100 тис грн. До банку за позикою звернулося семеро клієнтів. Ймовірність банкрутства кожного з яких складає 15 %. Формування резервів для покриття можливих втрат вимагає інформацію про можливість дефолту одночасно 3 і менше позичальників.

Дана величина є дискретною (величина банкрутств клієнтів може бути лише цілочисельною). Припустимо, що вона має біноміальний розподіл. Середнє значення $E(x)=n\cdot\pi=7\cdot0{,}15=1{,}05$, тобто очікується, що збанкрутує 1 позичальник. Стандартне відхилення $\sigma=\sqrt{n\cdot\pi(1-\pi)}=0{,}945$, що задає ризик у сенсі того, наскільки результат конкретного спостереження буде відхилятися від середнього значення.

Ймовірності настання подій розрахуємо за формулою:

$$P(X=0)=\binom{7}{0}\cdot0{,}15^{0}(1-0{,}15)^{7-0}=\frac{7!}{0!\cdot(7-0)!}\cdot0{,}15^{0}(1-0{,}15)^{7-0}=0{,}321$$

$$P(X=1)=0{,}396$$

$$P(X=2)=0{,}210$$

$$P(X=3)=0{,}062$$

Отже, імовірність того, що відбудеться 3 дефолти становить 6,2 %. Проте ймовірність того, що збанкрутує менше 3 клієнтів (тобто 0, 1 і 2), складає 97,2 %. Розрахунок ймовірностей настання конкретних подій за наявності інформації про ставку відсотку за кредитами і величини позики кожного з клієнтів дає можливість визначити розмір збитків при настанні дефолту, який банк зможе покрити власними коштами, а який необхідно віднести в премію за ризик.

Розподіл Пуассона, подібний біноміальному розподілу, але залежить від очікуваного середнього кількості настання події за певний проміжок часу. Основна відмінність полягає в тому, що в ньому немає заданого числа можливих спроб *n*. Якщо деяка подія відбувається випадково і незалежно в кожній зі спроб, при цьому середня кількість настання події зі зростанням кількості спроб не змінюється, тоді кількість настання події у фіксованій кількості спроб буде підкорятися розподілу Пуассона.

До основних характеристик даного розподілу можна віднести наступні[2]:

– стандартне відхилення буде дорівнювати кореню квадратному із середнього значення;

– зі збільшенням середніх значень розподіл Пуассона буде близьким до нормального (рис. 3.2.);

– ймовірність того, що випадкова величина *X* із середнім значенням μ=a, що має розподіл Пуассона, виражається формулою:

$$P(X=a)=e^{-\bar{x}}\left(\frac{\bar{x}^{a}}{a!}\right). \qquad (3.13)$$

[2] Siegel, Andrew F. (2017), Practical business statistics / Andrew F. Siegel. – 8th ed., – P. 641.

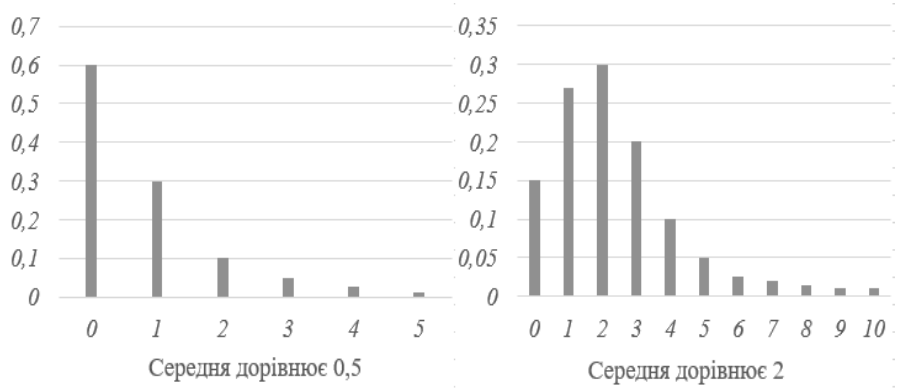



*Рисунок 3.2.* **Графічне зображення розподілу Пуассона із різною середньою очікуваною кількістю настання подій**

Розглянемо тижневу статистику кількості поданих скарг відвідувачів ресторану:

*Таблиця 3.4.* **Динаміка кількості скарг відвідувачів ресторану**

| День тижня | Кількість поданих скарг, шт. |
|---|---|
| Понеділок | 0 |
| Вівторок | 1 |
| Середа | 1 |
| Четвер | 0 |
| П’ятниця | 2 |
| Субота | 2 |
| Неділя | 3 |

Дана величина є дискретною і підпорядковується розподілу Пуассона. За умови, що очікується перевірка, необхідно визначити, яка ймовірність того, що наступного дня буде подано 2 заяви або менше.

Знайдемо середню кількість поданих скарг за тиждень:

$$\bar{x} = \frac{0+1+1+0+2+2+3}{7} = 1{,}286.$$

Стандартна похибка є коренем квадратним із середнього значення:

$$\sigma = \sqrt{\bar{x}} = 1{,}134.$$

Імовірність того, що випадкова величина, яка має розподіл Пуассона, розраховується за формулою:

$$P(X=2) = e^{-\bar{x}}\left(\frac{\bar{x}^2}{2!}\right) = 0{,}229;$$

$$P(X=1) = e^{-\bar{x}}\left(\frac{\bar{x}^1}{1!}\right) = 0{,}355;$$

$$P(X=0) = e^{-\bar{x}}\left(\frac{\bar{x}^0}{0!}\right) = 0{,}276.$$

Отже, імовірність того, що наступного дня надійде 2 скарги, становить 22,9 %. А ймовірність того, що скарг буде менше за 2, складає 63,1 % (сума ймовірностей 1 і 0 скарг).

## 3.3. Нормальний розподіл. Несиметричні розподіли та перетворення на основі інтегрального рівняння кривої нормального розподілу

Ідеально статистичним законом розподілу є закон нормального розподілу. Нормальний розподіл утворюється за умови, що на величину *x* впливає велика кількість незалежних випадкових причин і жодна з них не має пріоритетного впливу, який математично описаний функцією нормального розподілу Лапласа.

Нормальний розподіл має вигляд теоретичної гладкої кривої у формі дзвону без випадкових відхилень. Така крива представляє ідеальний набір даних, в якому більшість чисел сконцентровано в середній частині діапазону значень, а решта значення із загасанням, симетрично розташовані по обидві сторони від вершини дзвону. Фактично існує багато різних кривих нормального розподілу, які відрізняються розташуванням центру і масштабом (шириною дзвону):

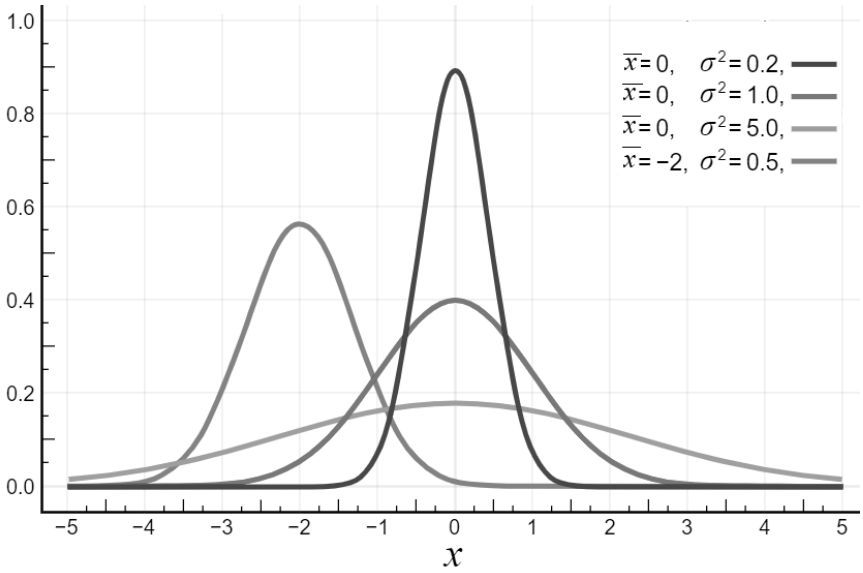



*Рисунок 3.3.* **Графічне зображення кривих нормального розподілу з різними центрами**

Крива нормального розподілу задається функцією щільності імовірностей, де $\overline{x}$ – середнє значення випадкової величини (центр, що визначає горизонтальне положення найвищої точки), $\sigma$ – середнє квадратичне відхилення випадкової величини (визначає ширину дзвону (мінливість), $e$=2,718 – основа натурального логарифма (число Ейлера), $\pi$ = 3,14:

$$f(x)=\frac{1}{\sqrt{2\pi\sigma}}e^{\frac{-(x-\bar{x})^2}{2\sigma^2}} \tag{3.14}$$

Функцію нормального розподілу Лапласа використовують для визначення ймовірності потрапляння нормально розподіленої випадкової величини у певний інтервал:

$$f(x)=\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{t^2}{2}} \tag{3.15}$$

Будь-яку випадкову величину, що має нормальний розподіл, можна представити через стандартний нормальний розподіл, якщо використовувати не реальні одиниці виміру, а стандартні відхилення:

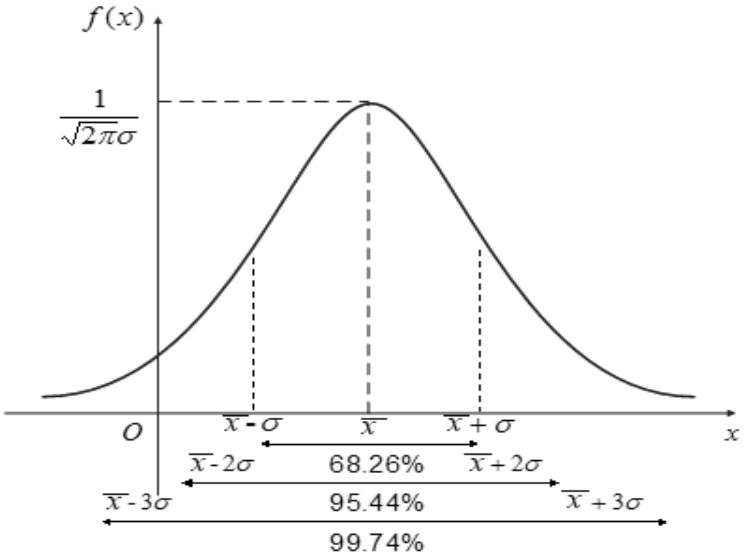



*Рисунок 3.4.* **Графічне зображення нормального розподілу випадкової величини з різними стандартними відхиленнями**

Крива нормального розподілу відповідає таким параметрам:

1) за симетричного розподілу – $\bar{x} = Mo = Me$, розподіл має правосторонню асиметрію – $\bar{x} > Mo > Me$, розподіл має лівосторонню асиметрію – $\bar{x} < Mo < Me$;

2) бічні сторони асимптотично наближуються до осі абсцис, тобто чим більше значення відхиляються від середньої, тим рідше вони зустрічаються;

3) крива має дві точки згину, що відстають від середньої на $\pm\sigma$ (правило «трьох сігм»): у проміжку $\pm\sigma$ знаходиться $\approx$68,26 % усіх значень ознаки; у проміжку $\pm 2\sigma$ знаходиться $\approx$95,44 % усіх значень ознаки; у проміжку $\pm 3\sigma$ знаходиться $\approx$99,74 % усіх значень ознаки;

4) чим більшого значення набуває $\sigma$ – тим більш гостровершинний розподіл, чим меншого значення набуває $\sigma$ – тим розподіл більш плосковершинний.

Набори даних, які характеризують реальні соціально-економічні явища або бізнес процеси, не завжди можуть описуватись функціями нормального розподілу. Більшість мають асиметричний (скошений) розподіл, в якому значення даних на одній стороні кривої затухають швидше, ніж на інший. Такий розподіл характеризує величини, які виражені додатними числами та не мають обмежень зверху і з одного боку. В результаті на гістограмі багато значень даних сконцентровано навколо нуля, і кількість значень стає все меншим при руху по горизонтальній осі:

Одна з проблем, пов'язаних з асиметрією даних, полягає в тому, що подальший аналіз передбачає використання стандартних статистичних методів за наявності принаймні нормального розподілу даних. Якщо ці методи застосовувати до несиметричного розподілу, то отриманий результат може бути неточним або просто невірним.

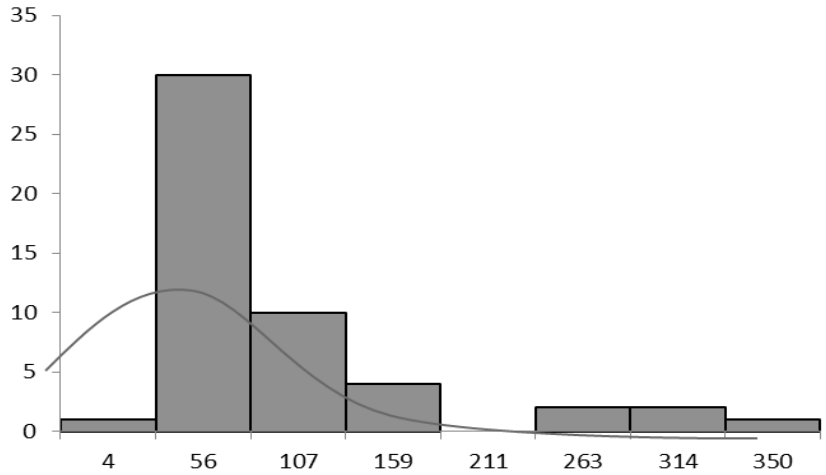



*Рисунок 3.5.* **Гістограма асиметричного розподілу**

Для вирішення проблеми з асиметрією застосовують перетворення (апроксимація), яке наближає несиметричний розподіл в більш симетричний. Апроксимація полягає в заміні частот емпіричного розподілу на частоти близького йому теоретичного розподілу, що описується відповідною функцією.

Апроксимація включає такі етапи:

1) добір функції кривої розподілу, яка б найточніше описувала емпіричний розподіл; зазвичай використовується набір відомих функцій: нормального, логнормального, біноміального, експоненційного розподілів;

2) обчислення параметрів функції кривої розподілу;

3) розрахунок теоретичних частот на підставі функції кривої розподілу;

4) перевірка гіпотези узгодженості емпіричного і теоретичного розподілу частот з певним рівнем істотності.

Апроксимація за функцією нормального розподілу зводиться до розрахунку теоретичних частот за формулою, де $\Sigma f$ – сума усіх частот або обсяг сукупності, $h$ - ширина інтервалу:

$$f'' = \sum f \frac{h}{\sigma} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \tag{3.16}$$

При розрахунку теоретичних частот використовувались лише два параметри – $\bar{x}$ та σ, а для їхнього обчислення беруться середини інтервалів. Тобто припускається, що щільність розподілу в межах інтервалу однакова, що насправді не є дійсним або не невідомо. Тому більш точною буде апроксимація розподілу за допомогою інтегрального рівняння кривої нормального розподілу, де *dt* – ширина інтервалу, а розрахунок ***t*** виконується на підставі верхніх меж інтервалів:

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \tag{3.17}$$

Значення *F(x)* визначаються за відповідною таблицею за підстановкою $t>0$. Якщо $t<0$, то за умови симетричності нормального розподілу $F(-t) = 1 - F(t)$. Тобто імовірність того, що ознака *x* набуватиме значень в інтервалі від $x_i$ до $x_{i+1}$ дорівнюватиме різниці $F(t_{i+1}) - F(t_i)$.

Послідовність розрахунку теоретичних частот за допомогою інтегрального рівняння кривої нормального розподілу буде наступною:

1) обчислюються параметри ряду розподілу $\bar{x}$ та σ, на підставі середин інтервалу;

2) розраховується значення *t*, на підставі верхніх меж інтервалів;

3) за даними *t* знаходяться табличні значення функції $F(t_i)$;

4) визначається імовірність $F(t_{i+1}) - F(t_i)$;

5) обчислюються теоретичні частоти за формулою:

$f''_{i+1} = n\,[F(t_{i+1}) - F(t_i)]$, а для $f_i' = n\,F(t_i)$ (3.18)

Перевірка гіпотези про закон розподілу здійснюється в такій послідовності:

1) формулюється нульова гіпотеза про відсутність розбіжностей між емпіричними і теоретичними частотами розподілу;

2) обирається статистичний критерій узгодженості обох розподілів та встановлюється прийнятний рівень істотності;

3) обчислений критерій узгодженості порівнюється з його критичним значенням. Якщо фактичне значення критерію менше критичного, розбіжності між емпіричним і теоретичними розподілами вважаються випадковими, що свідчить про адекватність емпіричного розподілу обраному теоретичному.

Нульова гіпотеза формулюється щодо випадковості розбіжностей між фактичними і теоретичними частотами розподілу. Альтернативна гіпотеза буде стверджувати, що розбіжності обумовлені неправильним добором функції розподілу.

Для перевірки гіпотез на відповідність теоретичному розподілу використовуються такі статистичні критерії узгодженості – Пірсона, Романовського, Колмагорова-Смірнова.

За допомогою критерія Пірсона $\chi 2$ перевіряється нульова гіпотеза про випадковість відхилень частот ($\chi^2=0$) та висувається альтернативна гіпотеза про відсутність випадковості відхилень ($\chi 2>0$). Формула розрахунку критерія Пірсона, де $f_i$ та $f'_i$ – відповідно емпіричні і теоретичні частоти розподілу:

$$\chi^2 = \sum_{i=1}^{m} \frac{(f_i - f_i')^2}{f_i'} \qquad (3.19)$$

Фактичні значення χ2 порівнюються з критичними $\chi^2_{1-\alpha}(df)$, з імовірністю $1-\alpha$ та числом ступенів свободи $df = m-q-1$, де $m$ – кількість груп, $q$ – число параметрів функції (для нормального розподілу $q=2$ ($\bar{x}$ та σ). Оскільки умовою перевірки гіпотези є $\Sigma f = \Sigma f'$, то число $df$ зменшується на 1.

Критичні значення критерію знаходяться за таблицею. Для зменшення ймовірності помилки II роду використовують високі рівні істотності α=0,05 (0,10 та 0,2). Нульова гіпотеза про випадковість відхилень, тобто узгодженість розподілів, приймається за умови $\chi^2 < \chi^2_{1-\alpha}(df)$. Якщо $\chi^2 > \chi^2_{1-\alpha}(df)$, тоді приймається альтернативна гіпотеза про невідповідність фактичного розподілу теоретичному.

За відсутності таблиць критичних значень застосовують критерій Романовського:

$$R = \frac{\left|\chi^2 - df\right|}{\sqrt{2df}} \qquad (3.20)$$

Якщо розраховане значення критерію $R<3$, то відхилення вважаються випадковими. Якщо значення критерію $R>3$, відхилення невипадкове, тобто фактичний розподіл не узгоджується з теоретичним.

Критерій Колмагорова-Смірнова являє собою максимальне відхилення між кумулятивними частотами або частками фактичного і теоретичного розподілів:

- для часток: $\lambda = D\sqrt{n} = \max\left|cum\ d_i - cum\ d_i'\right|\sqrt{n}$ (3.21)

- для частот: $\lambda = \frac{D}{\sqrt{n}} = \frac{max\left|cum\ f_i - cum\ f_i'\right|}{\sqrt{n}}$ (3.22)

На підставі значень λ за таблицею значень функції знаходять імовірність, з якою можна стверджувати чи є відхилення фактичних і теоретичних частот випадковими.

Найбільш поширеним типом перетворення даних є логарифмування змінної *x*, яке можна використовувати тільки для додатних чисел. Логарифмування перетворює скошені (асиметричні) дані в симетричні, оскільки відбувається розтягування шкали навколо нуля. Логарифмування стягує разом великі значення – зменшуючи різницю між ними та іншими значеннями в наборі даних, і розтягує малі значення – збільшуючи різницю між ними й іншими значеннями.

Логарифмічну шкалу можна інтерпретувати скоріше як мультиплікативну, ніж як адитивну. Найчастіше використовують десятковий та натуральний логарифми:

*Таблиця 3.5.* **Порівняльна таблиця результатів логарифмування значень**

| Число | Десятковий логарифм | Натуральний логарифм |
|---|---|---|
| 0,001 | -3 | -6,9 |
| 0,01 | -2 | -4,6 |
| 0,1 | -1 | -2,3 |
| 1 | 0 | 0,0 |
| 2 | 0,301 | 0,7 |
| 5 | 0,699 | 1,6 |
| 10 | 1 | 2,3 |
| 100 | 2 | 4,6 |
| 10000 | 4 | 9,2 |
| 100000 | 5 | 11,5 |
| 100000000 | 8 | 18,4 |

При побудові логарифмічно нормального розподілу замість значень $x_i'$ (середина інтервалу) використовуються $ln\, x_i$. Теж саме стосується обох параметрів середнього логарифму $x_i$ та дисперсії:

$$\overline{\ln x} = \frac{\sum \ln x_i \; f_i}{\sum f_i} \tag{3.23}$$

$$\sigma_{\ln x} = \sqrt{\frac{\sum (\ln x_i - \overline{\ln x})^2 \; f_i}{\sum f_i}} \tag{3.24}$$

$$t = \frac{ln\, x_i' - \overline{ln\, x}}{\sigma} \tag{3.25}$$

Перевірка гіпотези здійснюється за допомогою критерію Пірсона з числом ступенів свободи $df = m–q–1$.

У випадку дискретного ряду розподілу, де із збільшенням значень ознаки $x_i$ стрімко зменшуються частоти (частки) розподілу і значення середньої наближується до значення дисперсії $\bar{x} \rightarrow \sigma^2$, апроксимація здійснюється за допомогою функції кривої Пуассона:

$$P_x = \frac{\bar{x}^{x_i} e^{-\bar{x}}}{x!} \tag{3.26}$$

Де $e = 2{,}718$, $P_x$ – імовірність появи окремих значень $x_i$, на підставі якої обчислюються теоретичні частоти: $f_i` = n \bullet P_x$. Перевірка гіпотез здійснюється за будь-яким критерієм (Пірсона, Романовського, Колмогорова-Смірнова).

Розглянемо розподіл молодих сімей за кількістю дітей, в якому $\bar{x}$=1,4 та $e^{-1,4}$=0,2466. Вихідна умова та допоміжні розрахунки наведено в таблиці 3.6:

*Таблиця 3.6.* **Розрахункова таблиця розподілу сімей за кількістю дітей для перевірки гіпотези за законом Пуассона**

| Число дітей | Кількість сімей | $x \bullet f$ | $x!$ | $\frac{-}{x}^{x_i} e^{-\bar{x}}$ | $P_x$ | $f''$ | $(f - f'')^2$ | $\frac{(f - f'')^2}{f''}$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 28 | 0 | 1 | 0,246 | 0,246 | 25 | 9 | 0,364 |
| 1 | 32 | 32 | 1 | 0,345 | 0,345 | 35 | 9 | 0,260 |
| 2 | 20 | 40 | 2 | 0,483 | 0,241 | 24 | 16 | 0,661 |
| 3 | 14 | 42 | 6 | 0,676 | 0,112 | 11 | 9 | 0,797 |
| 4 | 4 | 16 | 24 | 0,947 | 0,039 | 4 | 0 | 0 |
| 5 | 2 | 10 | 120 | 1,326 | 0,011 | 1 | 1 | 1 |
| 6 | 0 | 0 | 720 | 1,857 | 0,002 | 0 | 0 | 0 |
| x | 100 | 140 | x | x | x | 100 | x | 3,085 |

За результатами розрахунків $\chi^2=3,085$. Порівняємо отримане значення із критичним з рівнем істотності $\alpha=0,05$ та $\alpha=0,10$, числом ступенів свободи $df=m–q–1=4$:

$\chi^2(3,085) < \chi^2_{1-0,05}(4)=9,5$ та $\chi^2(3,085) < \chi^2_{1-0,10}(4)=7,8$.

Таким чином, гіпотеза про випадковість відхилень відхиляється, тобто з імовірністю 0,95 і 0,90 можна стверджувати, що розподіл сімей за кількістю дітей можна вважати стандартним і він узгоджується із розподілом Пуассона.

## 3.4 Розрахунок ймовірностей для стандартного нормального розподілу. Апроксимація різних видів розподілу нормальним

У стандартному нормальному (неперервному) розподілі із середнім значенням $\bar{x}=0$ та стандартним відхиленням $\sigma=1$, для обчислення ймовірності випадкової величини здійснюють нормування за формулою:

$$z=\frac{x-\bar{x}}{\sigma} \qquad (3.27)$$

Будь-яку величину, що має нормальний розподіл, можна представити через стандартний нормальний розподіл, якщо використовувати не реальні одиниці виміру, а стандартні відхилення. При розрахунку ймовірностей для нормального розподілу використовують таблиці ймовірностей для стандартного нормального розподілу (див. Додатки). Таблиця містить ймовірності випадкової величини *z*, яка приймає значення менше деякого заданого значення *x*.

Правила розрахунку ймовірностей для нормального розподілу[3]:

1. Ймовірність того, що випадкова величина *z* менше значення *x* буде дорівнювати табличному значенню імовірності *z*.

2. Ймовірність того, що випадкова величина *z* більше значення *x* буде дорівнювати 1 мінус табличне значення імовірності *z*.

3. Ймовірність того, що випадкова величина *z* буде знаходитись між $z_1$ та $z_2$ буде дорівнювати різниці більшого і меншого табличного значення імовірності *x*.

4. Ймовірність того, що випадкова величина *z* буде знаходитись за межами $z_1$ та $z_2$ буде дорівнювати 1 мінус різниця більшого і меншого табличного значення імовірності *x*.

Виберемо неперервну випадкову величину «Рівень виплат страхових компаній», який показує яка частка обсягу страхових виплат у обсязі премій. Припустимо, що розглядається можливість інвестування у страхову компанію. Одним із важливих показників для оцінки ризику при інвестуванні є саме рівень виплат, оскільки цей показник має бути від 30 до 60%. Якщо значення показника буде менше, можливо отримати виплати буде дуже складно. Якщо

[3] Smith, Gary (2015), Essential Statistics, Regression, and Econometrics, Second Edition / Gary Smith. – P. 396.

значення показника більше, можливо політика компанії занадто ризикова, і вона незабаром збанкрутує. Також це може натякати на те, що через цю страхову просто відмивають кошти.

Для прийняття рішення щодо інвестування у страхову компанію розглядається список Топ-50 страхових компаній за 2019 р., що функціонують на ринку України (табл. 3.7). Варто зазначити, що вибірка не розширювалась за межі топ-50 компаній, оскільки інші – не є репрезентативними, бо не займають значну частку ринку, а їх середнє значення рівня страхових виплат становить 17,71%, що для нашого аналізу зовсім не буде мати сенсу.

*Таблиця 3.7.* **Рівень виплат Топ-50 страхових компаній України у 2019 р., %**

| № | Назва страхової | Рівень виплат | № | Назва страхової | Рівень виплат |
|---|---|---|---|---|---|
| 1 | Уніка | 78,48 | 26 | НІКО страхування | 38,71 |
| 2 | Нафтогазстрах | 69,95 | 27 | Саламандра | 37,97 |
| 3 | Мега-Полис | 64,62 | 28 | Європейський страховий альянс | 37,43 |
| 4 | Провідна | 63,64 | 29 | ТЕКОМ | 36,90 |
| 5 | Ю.Ес.Ай. | 61,14 | 30 | УКРФІНСТРАХ | 36,30 |
| 6 | Раритет | 55,17 | 31 | Експрес страхування | 35,84 |
| 7 | Граве Україна | 54,84 | 32 | Оранта | 35,73 |
| 8 | Універсальна | 54,46 | 33 | АСКО-ДОНБАС ПІВНІЧНИЙ | 34,27 |
| 9 | Омега | 49,85 | 34 | ВЕЛТЛІНЕР | 32,79 |
| 10 | ВИДИ страхування | 49,40 | 35 | Міжнародна страхова компанія | 32,42 |
| 11 | Країна | 49,18 | 36 | Арсенал страхування | 30,68 |

Продовження таблиці 3.7

| № | Назва страхової | Рівень виплат | № | Назва страхової | Рівень виплат |
|---|---|---|---|---|---|
| 13 | Просто-Страхування | 45,4 | 38 | Альфа страхування | 30,58 |
| 14 | Євроінс Україна | 45,09 | 39 | UTICO | 29,92 |
| 15 | Коллонейд Україна | 43,94 | 40 | ALLIANZ Україна | 29,22 |
| 16 | ARX | 43,05 | 41 | ГАРАНТІЯ СО | 28,91 |
| 17 | Інго Україна | 43,05 | 42 | ВУСО | 28,84 |
| 18 | Брокбізнес | 43,00 | 43 | UPSK | 28,78 |
| 19 | Перша | 42,77 | 44 | АЛЬФА-ГАРАНТ | 26,36 |
| 20 | Ван Клик | 41,91 | 45 | Оберіг | 26,03 |
| 21 | Харківська муніципальна СК | 40,86 | 46 | Глобал Гарант ПрАТ | 23,83 |
| 22 | PZU Україна | 40,11 | 47 | Еталон | 23,19 |
| 23 | ТАС СГ | 40,05 | 48 | АСКА | 21,69 |
| 24 | Княжа | 39,39 | 49 | Мега-Гарант | 21,42 |
| 25 | Крона | 39,25 | 50 | Кредо | 19,06 |

*Джерело*: визначено за даними сайту FORTUNE https://forinsurer.com/ratings/nonlife/19/12/10

За формулою середньої арифметичної простої визначимо середній рівень виплат – $\bar{x} = 40{,}04\%$, та їх стандартне відхилення – $\sigma = 12{,}79\%$.

Тепер розглянемо три можливих ситуації:

- поганий сценарій №1: рівень страхових виплат менше за 30%;

- хороший сценарій №2: рівень страхових виплат становить від 30% до 60%;

- поганий сценарій №3: рівень страхових виплат більше за 60% (не включаючи).

За формулою 3.27, отримуємо нормовані значення та визначимо відповідні імовірності настання подій для кожного зі сценаріїв:

- поганий сценарій №1: $Z = \frac{x - \bar{x}}{\sigma} = \frac{30 - 40{,}04}{12{,}79} = -0{,}78$,

$p(z)$= 0,2177;

- хороший сценарій №2: $z_1$=–0,78, для якого $p(z_1)$=0,2177; $z_2$=1,56, для якого $p(z_2)$=0,9406. Тоді, загальна імовірність буде становити: $p(z_2\text{-}z_1)$=0,9406-0,2177=0,7229;

- поганий сценарій №3: $z$=1,56, для якого $p(z)$=0,9406. Для такого значення z, імовірність буде дорівнювати $p(1–z)$=1–0,9406=0,0594.

Таким чином, з імовірністю 21,77% рівень виплат обраної компанії буде становити менше за 30%, що є ризиковим для інвестора. З імовірністю 5,94% рівень виплат обраної страхової буде становити більше за 60%, що також є небезпечним та ризиковим для інвестора, оскільки такий високий рівень показника може свідчити про ризикову політику діяльності страхової, досить низьку прибутковість та можливе банкрутство.

Рівень виплат обраної компанії буде в межах від 30% до 60% з імовірністю 72,29%, що є цілком безпечним для інвестора. В цілому, така ситуація є дещо ризиковою для інвестора, але прийнятною.

За допомогою нормального розподілу можна апроксимувати біноміальний розподіл при достатньо великих значеннях $n$ і ймовірностях наближених до 0 або 1. Це спрощує процедуру обчислення ймовірностей того, що деяка величина менше певного значення, перевищує його, знаходиться між двома значеннями або поза інтервалом.

Для цього необхідно обрати такий нормальний розподіл, який досить близький до даного біноміальному розподілу. В такому розподілі значеннями середнього і стандартного відхилення будуть

подібними до значень у біноміального розподілу, який апроксимується.

Розподіл ймовірностей, обчислених з використанням формули для біноміального розподілу, показано на рис. 3.6. Розподіл має властиву нормальному розподілу колоколоподібну форму. Незважаючи на те, що розподіл все ще залишається дискретним, що не є його головною властивістю.

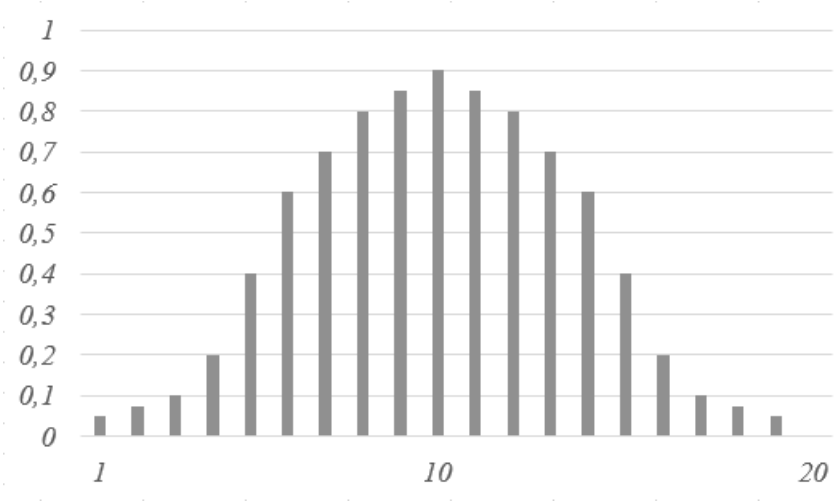



*Рисунок 3.6.* **Розподіл ймовірностей для біноміального розподілу при n =100 і $\pi$ =0,10**

Для того щоб апроксимувати біноміальний розподіл (з дискретними цілочисловим значеннями) за допомогою нормально розподіленої випадкової величини (неперервної), відкладемо від кожного значення вправо і вліво 0,5, щоб включити в розгляд всі числа, розташовані навколо цілих чисел. Таке розширення необхідно у зв'язку з тим, що для будь-якої нормально розподіленої випадкової величини ймовірність її точної рівності 3 дорівнює нулю, водночас всі значення нормально розподіленої випадкової величини в інтервалі від 2,5 до 3,5 округлюються до цілого числа 3.

Аналогічно ймовірність того, що біноміально розподілена випадкова величина прийме значення в інтервалі від 6 до 9, відповідає імовірності того, що нормально розподілена (з такими же значеннями середнього і стандартного відхилення) величина потрапить в проміжок від 5,5 до 9,5. Імовірність того, що значення буде знаходитися поза межами інтервалу, дорівнює одиниці мінус ймовірності потрапляння в цей інтервал.[4]

Провідна будівельна компанія «EuroHouse» планує реалізувати новий будівельний проект: житловий комплекс преміум-класу із власними місцями для паркування автомобілів, дитячим садком, хімчисткою, парком, а також басейнами на даху кожного будинку. Звичайно, це досить затратно, і керівник цієї компанії, переймаюсь, чи дійсно варто будувати ЖК із басейном, чи дійсно майбутні мешканці готові переплатити значну суму за можливість мати власний басейн на даху будинку. У відділі роботи з клієнтами знаходиться база даних потенційних майбутніх мешканців цього ЖК, тобто людей, які виявили бажання купити квартиру у подібному за усіма характеристиками будинку. У цій базі наявно 1000 клієнтів. Працівниками відділу продажів упродовж тижня проводилось телефонне опитування потенційних клієнтів з приводу доцільності розміщення басейнів на дахах будинків. Після 2-годинних обговорень імовірність схвалення розміщення басейнів клієнтами була оцінена як 10%.

Необхідно оцінити імовірність таких подій:

А – із усіх опитаних клієнтів погодяться на розміщення басейну 110 клієнтів;

Б – клієнтів, що погодяться на розміщення басейну на дахах будинків буде від 90 до 120.

Розрахуємо середнє значення кількості потенційних клієнтів, які погодяться щодо розміщення басейнів на дахах будинків:

[4] Siegel, Andrew F. (2017), Practical business statistics / Andrew F. Siegel. – 8th ed., – P. 641.

$E(X) = n\pi = 1000 \times 0{,}10 = 100$ потенційних клієнтів. Якщо говорити про долю або відсоток, то $E(X) = \pi = 0{,}10$ або 10%

Тоді, стандартне відхилення дорівнює: $\sigma = \sqrt{n\pi \times (1-\pi)} = \sqrt{1000 \times 0{,}10 \times (1-0{,}10)} = \sqrt{90} = 9{,}49$ потенційних клієнтів або $\sqrt{\frac{\pi \times (1-\pi)}{n}} = \sqrt{\frac{0{,}10 \times 0{,}90}{1000}} = \sqrt{0{,}00009} = 0{,}0095$ або 0,95%.

Таким чином, при заданій імовірності кожного окремого клієнта схвально оцінити розміщення басейна на даху (0,1) очікується, що в середньому близько 100 клієнтів погодяться на це. Слід очікувати, що отриманий результат (кількість клієнтів, що погодяться переплатити за басейн на даху) буде більше або менше середнього значення на 9,49 клієнта або на 0,95%.

Є підозри, що розподіл імовірності кількості потенційних клієнтів, що погодяться переплатити за басейн на даху, наближається до нормального. Була здійснена перевірка розподілу на нормальність, шляхом розрахунків відповідних імовірностей різних значень біноміального розподілу при заданій кількості потенційних клієнтів та імовірності згоди на розміщення басейну на даху для кожного з них та побудові гістограми розподілу імовірностей (рис.3.7).

Як бачимо, у даній ситуації біноміальний розподіл досить близький до нормального, тому для з'ясування імовірності настання подій А і Б можемо застосувати апроксимацію біноміального розподілу нормальним.

Оскільки, $a = 110$, визначимо інтервал для нормально розподіленої імовірності: $110 - 0{,}5;\ 110 + 0.5$, тобто необхідно знайти імовірність того, що від 109,5 до 110,5 клієнтів погодяться на басейн на даху.

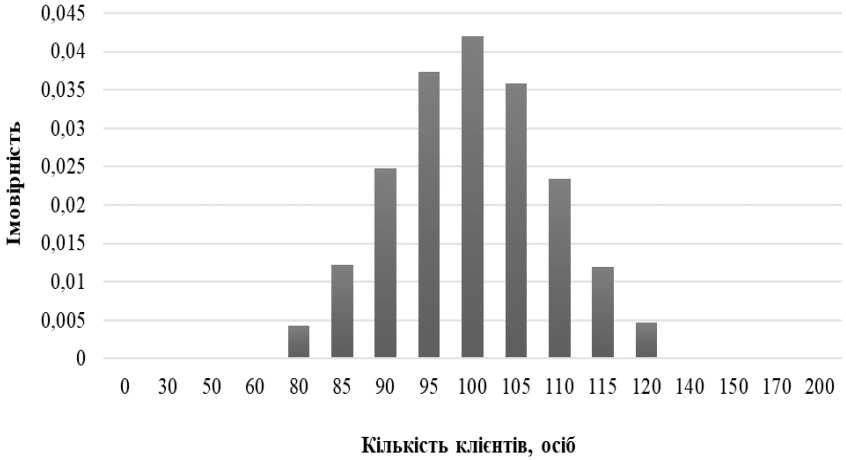



*Рисунок 3.7.* **Гістограма розподілу імовірностей кількості потенційних клієнтів, що готові переплатити за басейн на даху будинку**

Для цього проведемо нормування цих значень:

$$z_1 = \frac{109{,}5 - 100}{9{,}49} = 1{,}00$$

$$z_2 = \frac{110{,}5 - 100}{9{,}49} = 1{,}11$$

Тепер знайдемо імовірності, які відповідають значенням $z_1$ та $z_2$: 0,8413 та 0,8665. Тоді імовірність того, що рівно 110 клієнтів погодяться переплатити складає: $0{,}8665 - 0{,}8413 = 0{,}0252$. Тобто у 2,52 % випадків кількість клієнтів, що погодяться переплати значну суму коштів за басейн на даху, складе рівно 110 осіб (тобто від 109,5 до 110,5, якщо говорити термінами нормального розподілу).

б) За аналогією до а): інтервал для нормально розподіленої величини дорівнює: $90 - 0{,}5; 120 + 0{,}5$, тобто [89,5; 120,5].

Знаючи середнє значення та середньоквадратичне відхилення проведемо нормування цих значень:

$$z_1 = \frac{89,5 - 100}{9,49} = -1,11$$
$$z_2 = \frac{120,5 - 100}{9,49} = 2,16.$$

Тепер знайдемо імовірності, які відповідають значенням $z_1$ та $z_2$ – 0,1335 та 0,9846. Тоді імовірність того, що від 90 до 120 клієнтів погодяться переплатити за наявність басейну складає: $0,9846 - 0,1335 = 0,8511$. Тобто у 85,11% випадків кількість клієнтів, що погодяться переплати значну суму коштів за басейн на даху, становить від 90 до 120 осіб (тобто від 89,5 до 120,5 у термінах нормального розподілу).

Якщо застосовувати не апроксимацію біноміального розподілу нормальним, а просто порахувати зазначені імовірності для відповідних подій та провести з ними відповідні дії, отримаємо майже ті ж самі результати (різниця у 1-2%).

За допомогою нормального розподілу, також можна здійснювати апроксимацію розподілу Пуассона. Розглянемо приклад надходження звернень до відділу продажу щодо купівлі нерухомості. Кількість звернень за день в середньому становить 460 телефонних дзвінків. Визначимо ймовірність того, що завтрашній день виявиться перевантаженим – кількість звернень буде становити 500 або більше.

Стандартне відхилення становить $\sqrt{460} = 21,44761$. Оскільки середнє значення достатньо велике, то для даного розподілу можна як наближення використовувати нормальний розподіл. Будь-яке значення, що перевищує 499,5 – буде заокруглено до 500. Таким чином, нормована кількість звернень дорівнює:

$$Z = \frac{499,5 - 460}{21,44761} = -1,84$$

За таблицею стандартного нормального розподілу визначимо імовірність такої події: 1 – 0,967 = 0,033. Таким чином, імовірність того, що завтрашній день виявиться перевантаженим, складає лише близько 3%.

### *Список питань до самоконтролю:*

**1.** Щоденний запит на номер у готелі з підвищеним рівнем комфортності варіює від 10 до 20 запитів. У готелі налічується 17 номерів зазначеного типу класу. Результати однакових запитів на бронювання номерів з підвищеним рівнем комфортності мають наступну послідовність: 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20. Чому буде дорівнювати ймовірність відмови клієнту у розміщені такого типу класу номері, тобто яка ймовірність дефіциту таких номерів?

**2.** У регіональному відділені салону з прокату автомобілів в наявності залишилось шість автомобілів для їх оренди на вихідні. Із 20 попередніх записів відомо, що кількість автомобілів наданих в оренду у вихідні дні становить: 7, 9, 8, 7, 10, 8, 7, 8, 9, 10, 7, 6, 8, 9, 8, 8, 6, 7, 7, 9. Прийнято рішення замовити додаткову кількість автомобілів з іншого салону, щоб не понести додаткові втрати від простою салону. Визначте, яку кількість автомобілі необхідно додатково замовити?

**3.** Залежно від кількості проданих квитків на захід, запит на придбання рекламної продукції становить 5-10%. Чому буде дорівнювати ймовірність купівлі менше 7% рекламної продукції залежно від кількості проданих квитків на захід?

**4.** Компанія бере участь у торгах на поставку деталей виробнику електроніки. Вартість заявки конкурентів на 10 попередніх аналогічних контрактів становить: 369,800; 387,300; 401,400; 403,200; 404,800; 380,300; 401,800; 389,700; 387,600; 407,700. Якщо торги будуть виграні, загальна вартість на виконання контракту становить 350 000 гр. од. Визначте оптимальну вартість заявки?

**5.** Існує можливість вкладення коштів у один з чотирьох проєктів, пов'язаних з використанням земельної ділянки що перебуває у власності самої компанії. Виплати за проектами представлені у вигляді дискретного розподілу:

| Проект | Виплата, тис гр. од. | Імовірність виплат |
|---|---|---|
| 1 | 60,0 | 1 |
| 2 | 130,0 | 0,6 |
| | 70,0 | 0,4 |
| 3 | 100,0 | 0,6 |
| | 60,0 | 0,4 |
| 4 | 500,0 | 0,1 |
| | 0,0 | 0,9 |

Завдання:

- визначте очікувані платежі для кожного з проєктів та їх пріоритетність, спираючись на очікувані платежі;
- розрахуйте стандартне відхилення для кожного з проєктів та їх пріоритетність, спираючись на рівень ризику;
- чи можна будь-який проєкт (або проєкти) повністю виключити з розгляду, якщо враховувати очікувані платежі та їх рівень ризику;
- чи можна обрати один з проектів як найкращій за всіма характеристиками?

**6**. Визначте такі значення А та Б, при яких 95% звернень до відділу продажів в перші дві години робочого дня будуть знаходиться в межах цих значень (тобто в межах двох стандартних відхилень). Ймовірності кількості звернень від 0 до 6 наступні: 0,129; 0,264; 0,271; 0,185; 0,095; 0,039; 0,017.

**7.** Модель випадкового коливання цін акцій стверджує, що прибуток фондового ринку не залежить від прибутку в інші періоди. Якщо щомісячний прибуток акцій є незалежним, у 60 % випадків буде позитивним і негативним приблизно у 40% випадків, якою буде ймовірність: 12 послідовних позитивних результатів (отримання прибутку); 12 послідовних негативних результатів (отримання збитку); отримати прибуток, якщо у попередньому місяці був збиток.

**8.** Визначте ймовірність народження 2 хлопчиків з 5 немовлят, при чому ймовірність народження хлопчика становить 0,51 при кожному народженні, а стать немовлят є незалежною випадковою величиною?

**9.** У п’яти регіональних відділеннях служби доставки було проведено перевірку на наявність неукомплектованості замовлень. Дані перевірки:

| Номер відділення | Кількість перевірених замовлень | Кількість бракованих замовлень | Ймовірність виявлення браку |
|---|---|---|---|
| 1 | 3741 | 30 | 0,080 |
| 2 | 1864 | 31 | 0,026 |
| 3 | 5006 | 11 | 0,022 |
| 4 | 3596 | 1 | 0,003 |
| 5 | 6501 | 8 | 0,012 |

Якщо перевірити 500 готових замовлень у кожному відділені, то якою буде: точна біноміальна імовірність 5 виявлених недоукомплектованих замовлень у кожному з відділень та ймовірність наявності щонайменше 5 недоукомплектованих замовлень у кожному з них.

**10.** Страхова компанія запровадила поліс, який поширюється на специфічні травми працівників у виробничих підприємствах. Ймовірність того, що працівник отримає таку травму, становить від 1 до 1000. На підприємстві працює 12 працівників. Визначте: чому нормальне наближення розподілу не підходить для даної ситуації; яка ймовірність того, що буде травма; яка ймовірність отримати більше двох травм.

**11.** Відділом аналітик компанії з виробництва і продажу пропану проаналізовано попит останніх 5 років на цистерни пропану. Було виявлено сильне зростання попиту протягом місяців, впродовж несприятливих погодних умов на узбережжі (сильні шторми, урагани, торнадо тощо). Було визначено, що ймовірність щомісячного сильного шторму становить приблизно 10%. Завдання:

- за умови що даний розподіл є біноміальним, якою буде ймовірність відсутності штормів та наявності однієї бурі за рік;
- за умови що це розподіл Пуассона, якою буде ймовірність відсутності штормів та наявності однієї бурі за рік.

**12.** З попереднього досвіду в наступну суботу очікується надходження товару в середньому на суму 2353,25 гр. од., зі стандартним відхиленням в 291,63 гр. од. Розподіл нормальний. Знайдіть ймовірності наступних подій: 1) це буде звичайна субота з надходженням товару в межах від 2000 грн до 2500 гр. од.;
2) це буде успішна субота з надходженням товару понад 2500 гр. од.; 3) це буде посередня субота з надходженням товару менше 2000 гр. од.

### *Список рекомендованої літератури по темі:*


1. Бізнес-статистика: навч. посібник / [Матковський С.О., Гринкевич О.С., Вдовин М.Л., Вільчинська О.М., Марець О.Р., Сорочак О.З.] – К.: Алерта. 2016. – 280 с.
2. Василенко О.А., Сенча І.А. Математично-статистичні методи аналізу у прикладних дослідженнях: навч. посіб. – Одесса: ОНАЗ ім.. О.С. Попова, 2011. – 166 с.
3. Єріна А.М., Єрін Д.Л. Статистичне моделювання та прогнозування: підручник. – К.: КНЕУ, 2014 – 348, [4] с.
4. Захожай В.Б., Чорний А.Ю. Статистика якості: Підруч. Для студ. Вищ. Навч. закл. – К.: МАУП, 2005. – 576 с.
5. Лупан І.В., Авраменко О.В. Компютерні статистичні пакети: навчально-методичний посібник. – Кіровоград, 2010. – 218 с.
6. Статистика підприємств: навч. посібник / [С.О. Матковський, О.С. Гринькевич, О.З. Сорочак та ін.]; за ред. С.О. Матковського – К.: АЛЕРТА, 2013. -560 с.
7. Статистичні спостереження: переписи, моніторинги, вибіркові обстеження : навчальний посібник / [А.М. Єріна, З.О. Пальян]. – К.: 2019 – 308, [4] с.
8. Pinder, Jonathan P. (2017), Introduction to Business Analytics using Simulation / Jonathan P. Pinder. – P. 434.
9. Siegel, Andrew F. (2017), Practical business statistics / Andrew F. Siegel. – 8th ed., – P. 641.
10. Smith, Gary (2015), Essential Statistics, Regression, and Econometrics, Second Edition / Gary Smith. – P. 396.

# Розділ 4. МЕТОДИ І ЛІНІЙНІ МОДЕЛІ БАГАТОВИМІРНОГО РЕГРЕСІЙНОГО АНАЛІЗУ

## 4.1. Регресія, критерії регресії

Регресія походить від латинського слова regressio – обернений рух, відхід. В реальності “прямий рух” визначає вплив подій (факторів) процесу на результуючу подію (результуючий фактор). Обернений рух – вплив результуючої події (результуючого фактору) на інші події (фактори) процесу. На противагу кореляції, яка відображає взаємний зв'язок подій (факторів), регресія – односторонній зв'язок подій (факторів). Прикладом кореляції може слугувати взаємний зв'язок двох макроекономічних факторів інфляції в країні і безробіття .Інфляція впливає на безробіття і навпаки, безробіття впливає на інфляцію. Тільки односторонню залежність за змістом, регресію можна виявити на прикладі двох факторів – виробітку продукції робітником протягом зміни і рівнем його освіти. Вплив рівня освіти на рівень виробітки в реальності можливий, зворотній вплив не можливий. В теорії ймовірності і математичної статистики одностороння стохастична залежність факторів, іменується як регресія фактору *Y* на фактори *X*.

Широке застосування регресії відбувається в обробці статистичних даних .Обробка статистичних даних передбачає використання методів і моделей статистичного аналізу даних. До методів аналізу даних відносять кореляційний аналіз, регресійний аналіз, факторний, дискримінантний аналіз, кластерний аналіз, канонічний аналіз, методи порівняння середніх, частотний аналіз, крос табуляція, аналіз відповідальності, дерева класифікації, багатовимірне шкалювання, моделювання структурними рівняннями, методи аналізу виживання, часові ряди, нейронні мережі, планування експерименту, карти контролю якості.

Основним дослідженням складних систем і процесів за їх статистичними даними є встановлення і перевірка виду зв'язку між незалежними змінними (факторами, предикатами), статистику яких дослідник може змінювати, та залежною змінною із заданою похибкою. Розв'язанням таких задач займається регресивний аналіз. Завданням регресивного аналізу є необхідним підібрати по можливості функцію, яка найкраще апроксимує статистичні дані результуючого фактору. Регресивний аналіз допускає, що регресія (результуючий фактор *Y*) є лінійна комбінація незалежних базисних функцій $f_i$ (*x)*, що описують головні фактори з невідомими коефіцієнтами $\beta_i$ . і помилок *U*: $Y = \beta_1 + \beta_2 f_2(x) + ... + \beta_n f_n(x)$.

Таким чином, отримано лінійну залежність результуючого фактору (відгуку) *Y* від нових факторів $f_i$ (*x)*, що є функціями факторів *X* за перерахованими статистичними даними. Як висновок: нелінійна регресія може бути зведена до лінійної. Остання повинна бути предметом першочергового вивчення.

Інтерпретація результатів статистичного аналізу (зокрема, регресивного аналізу), що важливо, здійснюється за трьома структурами статистичних даних, часовою, просторовою і змішаною (просторово-часовою). Формалізація (математичний опис) просторово-часового характеру статистичних (змішана структура статистичних) даних відбувається за функціональною матрицею об'єкт-властивість, ознака. Вона має вигляд:

$$\begin{vmatrix} x_{1,1}(t), & x_{1,2}(t), ...., x_{1,p}(t) \\ x_{2,1}(t), & x_{2,2}(t), ...., x_{2,p}(t) \\ .... & .... & .... & .... \\ x_{n,1}(t), & x_{n,2}(t), ...., x_{n,p}(t) \end{vmatrix} \quad (4.1)$$

де: $x_{i,j}(t)$ – значення *j*-тої ($j$ = 1, . . . ,*p*) аналізованої ознаки, що характеризує *i*-тий статистичний об'єкт ($i$=1, . . . ,*n*), в момент часу *t*.

Відображення статистичних даних за матрицею (4.1) в різні моменти часу називають просторово-часовою структурою статистичних даних. Прикладом просторово-часової структури статистичних даних може слугувати реалізація продукції (x) на ринку товарів і послуг групою із п'яти підприємств (об'єктів, $i$=1, . . . ,5 ) за наступними ознаками (факторами) – ціна реалізації, доходи споживачів продукції, затрати на її рекламу($j$ = 1, . . . ,3) в різні періоди часу (t).

Просторова форма статистичних даних одержується із матриці (4.1) за одномоментним зрізом статистичних даних, фіксацією часу ( параметра t) і ознаки (номера $j$) – характеризується вектором ($x_{ij}(t)$, $x_{2,j}(t), \ldots x_{ipj}(t)$). Приклад – реалізація продукції ($x$) рядом підприємств за місяць ($t$) за фіксованою ціною (за $j$- ознакою).

Значення $j$-тої ознаки у різних об'єктах (тобто послідовність $x_{1,j}(t)$, $x_{2,j}(t), \ldots, x_{n,j}(t)$) у фіксований момент часу $t$ називають *варіацією* ознаки. Упорядкований розподіл статистичних одиниць (об'єктів) за зростаючими або спадними значеннями ознаки називають *варіаційним рядом* (*ранжованим* рядом). Якщо $i$-тий об'єкт розглядати за однією $j$-тою ознакою в послідовні моменти часу $t_1, t_2, \ldots, t_N$, то послідовність $x_{i,j}(t_1)$, $x_{i,j}(t_2)$, . . . $x_{i,j}(t_m)$ називають *часовим рядом*. Для соціально-економічних процесів найбільш типова ситуація, коли моменти часу $t_1, t_2, \ldots, t_m$ рівновіддалені. У цьому випадку часовий ряд називають *динамікою* (*динамічним рядом*). Для нього прийнято позначення часу $t_k = k$. Приклад динаміки – помісячна реалізація продукції підприємством певного виду.

На практиці використовують дві структури статистичних даних просторова і часова. Просторово-часова вважається вимушеною структурою. Інтерпретація показників регресії в такій формі втрачає практичний зміст. Тому розмежування простору і часу відбувається завдяки введенню в модель регресії такого універсального фактору як час.

Статистична модель регресії і її побудова виникає на понятті регресії в математичній статистиці. Необхідність виникнення

регресії можна пояснити на реальному прикладі. Можна мати одну або декілька ділянок посіву сільськогосподарської культури. Стосовно структури статистичних даних, маючи тільки одну ділянку, статистику урожайності культури можна представити по роках вектором $Y=(y_1, y_2, \ldots, y_T)$ (часова структура статистичних даних), або по ділянках (просторова структура статистичних даних).

Завдяки тому, що на урожайність впливає багато факторів наприклад, кількість внесених добрив, затрати живої праці, атмосферні опади і т. д. серед яких обов'язково будуть випадкові фактори наприклад, атмосферні опади, розмір урожайності буде завжди випадкова величина.

Для багатофакторних процесів, моделювання полягає в спрощеному їх описі - у виділенні головних факторів ($X=X_1,,X_2, \ldots ,X_k$), що впливають істотно на результуючий фактор (Y), а також неголовних не ідентифікованих (неконтрольованих) факторів ($U$), вплив яких за рядом причин на результуючий фактор вважається незначним. В соціально-економічних процесах не ідентифікованих фактори ($U$) носять назву факторів збурення, в галузях природничого і технічних напрямків – помилки. Урожайність ($Y$) результуючий фактор, головні фактори ($X$): внесені добрива, затрат живої праці, атмосферні опади, інші - фактори збурення.

Для оцінки урожайності і її параметрів, а також для прогнозних рішень урожайності потрібно мати багатофакторну функціональну модель залежності урожайності від усіх факторів:

$$Y = f(X_1, X_2, X_3, U\} \qquad (4.2)$$

$X_1$ – внесені добрива, $X_2$ – затрати живої праці, $X_3$ - атмосферні опади. За кількістю рівнянь (одне рівняння) функціональна модель (4,2) належить до простих моделей. Урожайність є випадковою величиною, як функція від випадкових величин, тому модель (4.2) стохастична. Простішою є залежність головних факторів від факторів збурення - лінійна залежність, тобто адитивна модель:

$$Y = f(X) + U = f(X_1, X_2, X_3,) + U \qquad (4.3)$$

Функціональну модель (4.3) можна побудувати, коли є вибір виду функції ($f$) певного класу. Кожен клас функцій (наприклад клас лінійних функцій) характеризується рядом параметрів ($\beta$), що визначають внутрішню структуру моделі.

$$Y = f(X, \beta) + U = f(X_1, X_2, X_3, \beta) + U \qquad (4.4)$$

Завдання моделювання урожайності полягає в реалізації статистичної моделі: знаходження функціональної залежності за статистичними значеннями факторів.

Реально, оскільки фактори збурення можуть мати незначний вплив на результуючий фактор, або не можуть бути ідентифіковані, на практиці, за відповідним набором значень ($x$) головних факторів ($X$), орієнтація урожайності (Y), або її прогноз, відбувається за очікуваним її значенням, яким є середня урожайність ($Y_{ср}$). Тобто орієнтація відбувається за умовним середнім математичним сподіванням:

$$Y_{ср}(x) = M(Y|X=x) \qquad (4.5)$$

яке носить назву регресії випадкової величини $Y$ на величину $X$ за відповідним вибором значень ($x$) головних факторів (опадів і внесених добрив), Функція, яка виникає в результаті вибору $Y_{ср}(x)$ – називають функцією регресії.

За змістом урожайність $Y$ і його середнє $Y_{ср}$ випадкові величини. Під дією таких факторів, як кількість внесених добрив, затрати живої праці, атмосферні опади ($X$ ) і інших неконтрольованих факторів (факторів збурення, ($U$) латентних змінних) урожайність зазнає постійного відхилення (розсіювання) від його середнього значення, $Y - Y_{ср}(x) = U(x)$, звідки

$$Y = Y_{ср}(x) + U(x). \qquad (4.6)$$

Різниця $Y - Y_{ср}(x) = U(x)$ відноситься до помилки або значень факторів збурення, свідчить про випадкові відхилення $Y - y_{ср}(x)$ урожайності, з вибраними навмання значеннями головних факторів $X=x$, від середнього її значення $y_{ср}(x)$. Коли в (4.6) перейти до умовного математичного сподівання (середнього значення), отримаємо ланцюг очевидних рівностей:

$$M(Y|X=x)= M(y_{ср}(x) |X=x)+ M(U|X=x) ,$$

або $y_{ср}(x)= y_{ср}(x)+ M(U|X=x)$,

з якого слідує, що $M(U|X=x) = 0$ (для кожного набору значень головних факторів середнє значення фактору збурення рівне нулю). Якщо вважати випадкові величини $X$ і $U$ незалежними, то $M(U|X=x)= M(U) = 0$. Таким чином, рівність математичного сподівання нулю $M(U) = 0$ є критерієм незалежності головних факторів ($X$) і факторів збурення $(U)$, тобто критерієм регресії. Залежність цих факторів відноситься до побудови кореляційно- регресивних моделей. Прикладом такої залежності є зв'язок між продуктивністю мартенівських печей і процентним вмістом вуглецю в металі. Залежність такого типу взагалі характерна для опису ходу технологічних процесів.

Практично, наявність статистичної регресії перевіряється за тестом гіпотези $M(U) = 0$, або $M(Y - y_{ср}(x))=0$. $M(Y) = y_{ср}(x)$. Більшість параметричних тестів розроблені для нормально розподілених даних. Для деяких гіпотез існують параметричні тести для вибірок , що підлягають іншим законам розподілу. Генеральна сукупність в більшості випадків підлягає нормальному, або близькому до нього розподілу. За умови наявності нормального закону розподілу генеральної сукупності, з якої вибірка здійснювалась, маємо перевірити гіпотезу про рівність математичного сподівання урожайності середній урожайності за статистикою. Тест нормального закону розподілу можна провести за багатьма критеріями: Колмогорова-Смирнова, Ліплієфорса, Шапіро-Уілка, які присутні в багатьох пакетах статистичних методів обробки даних. Слід пам’ятати, що за характеристиками статистичних даних, тестування нормального закону розподілу здійснюється як за не згрупованими (дискретна ознака статистичних даних) так і згрупованими статистичними даними (інтервальна ознака статистичних даних).

Зазначимо, що в багатьох зокрема, соціально-економічних процесах названі критерії, як правило, не придатні для перевірки

наявності нормального закону розподілу результуючого показника, як надто жорсткі. З іншої сторони, нехтуючи цими критеріями, можна отримати хибні результати статистичних досліджень. Тому бажано мати справу із приблизно нормальним розподілом даних.

*Задача 4.1.* В таблиці 4.1 проведено опитування 20 домогосподарств за середнім місячним доходом, тис. грн. Виконати діагностику показника на предмет відповідності до нормального закону розподілу.

*Таблиця 4.1.* **Вихідні дані середнього місячного доходу домогосподарств, тис. грн.**

| № з/п | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Дохід | 20,5 | 31,5 | 47,7 | 26,2 | 44,0 |
| № з/п | 6 | 7 | 8 | 9 | 10 |
| Дохід | 8,28 | 30,8 | 17,2 | 19,9 | 9,96 |
| №з/п | 11 | 12 | 13 | 14 | 15 |
| дохід | 55,8 | 25,2 | 29 | 85,5 | 15,1 |
| № з/п | 16 | 17 | 18 | 19 | 20 |
| Дохід | 28,5 | 21,4 | 17,7 | 6,42 | 84,9 |

Вважається, що 20 домогосподарств складають вибірку з деякої їх (генеральної) сукупності. Потрібно за вибірковими даними (за навчальною вибіркою) здійснити статистику генеральної сукупності на предмет існування нормального закону розподілу, а за його відсутності – існування розподілу, близького до нормального.

Можна перевірити, що за критерієм Пірсона (за згрупованими статистичними даними) підстав прийняти основну гіпотезу про нормальний закон розподілу генеральної сукупності немає. Згрупувавши вихідні дані у 5 типових груп, визначимо середнє значення місячного доходу домогосподарств по кожній групі та представимо результат обчислень у вигляді графіка.

Серед програм комп'ютерної системи Excel пакету "*Аналіз даних"* є функція "*Гістограма'*, за якою (рис. 4.1) слідує, що

гістограма, побудована за статистичними даними прикладу 4.1, візуально не підтверджує наявність нормального розподілу генеральної сукупності (форма гістограми не дзвіноподібна).

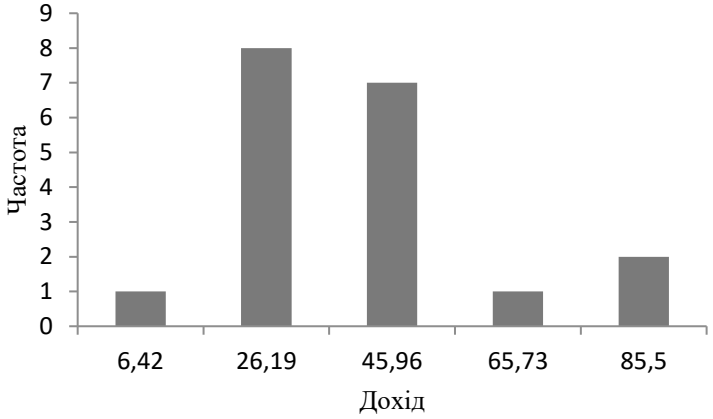



*Рисунок 4.1.* **Розподіл частот середнього місячного доходу домогосподарств, тис. грн**
*Джерело*: розрахунки авторів

Перевіримо за вибіркою наявність у генеральної сукупності розподілу, близького до нормального за рядом критеріїв, які гарантують тільки виконання достатніх умов, наявності розподілу близького до нормального. Якщо виявиться, що за межами інтервалу ( $\bar{x}$ -2 $\tilde{\sigma}$ ; $\bar{x}$ +2 $\tilde{\sigma}$ ) буде знаходитись більше 5% значень ознаки, то це не значить відсутність розподілу близького до нормального.(правило "двох сігм"). У даному прикладі 4.1: середній дохід домогосподарств по вибірці становить: $\bar{x}$ = 31,028 грн; квадратичне відхилення 2 $\tilde{\sigma}$ = 2×4,97= 9,94 грн; інтервал діагностики: ( $\bar{x}$ -2 $\tilde{\sigma}$ ; $\bar{x}$ +2 $\tilde{\sigma}$ ) = (31,028-9,94; 31,028+9,94) = (21,038; 40,968). Можна підрахувати, що за межами даного інтервалу міститься більше 5% вибірки (тобто гістограма має довгі хвости). Тому не можна стверджувати, що вибірка представляє розподіл, близький до нормального. Протилежний до даного критерію має наступний зміст: розподіл

вважається близький до нормального, коли встановлено, що від 50% до 85 % всіх значень ознаки розташовується в границях одного стандартного відхилення від середнього арифметичного (тобто в інтервалі ($\bar{x}$ -2$\tilde{\sigma}$; $\bar{x}$ +2$\tilde{\sigma}$)) і коефіцієнт ексцесу по абсолютній величині не перевищує значення рівного двом. Безпосередньо можна перевірити, що за протилежним критерієм розподіл, близький до нормального, теж не має місця.

Для діагностики за вибіркою наявності в генеральній сукупності закону розподілу близького до нормального використовують стандартні похибки коефіцієнтів асиметрії *A(T)*і ексцесу *E(T)* за формулами Блісса:

$$A(T)=\sqrt{\frac{6T(T-1)}{(T-2)(T+1)(T+3)}}\,,\ \mathrm{E}(T)=\sqrt{\frac{4(T^2-1)\,\mathrm{A}^2(T)}{(T-3)(T+5)}} \qquad (4.7)$$

де: *T*- об'єм вибірки. В прикладі 4.1: *A(20)*=0,512; *E(20)* = 0,992. Система нерівностей (за критерієм перевірки, показники повинні бути більші 0,05) *A(20)*> 0,05; *E(20)* = 0,992>0.05 не гарантує виконання закону розподілу залишків моделі, близького до нормального, тому слід переходити до діагностики інших законів розподілу результуючого показника.

Варто звернути увагу на відсутність в статистиці *Excel* стандартних похибок коефіцієнтів асиметрії *A(T)* і ексцесу *E(T*) і тому неможливо провести діагностику за схемою Блісса. Задані показники представлено в програмному середовищі *SPSS: Skewness Std. Error* (стандартна помилка коефіцієнта асиметрії) зі значенням 0,512 і *Kurtosis Std.Error* (стандартна помилка коефіцієнта ексцесу) зі значенням 0,992 (рис. 4.2).

```
DESCRIPTIVES VARIABLES=X1
  /STATISTICS=MEAN SUM STDDEV VARIANCE RANGE MIN MAX SEMEAN KURTOSIS
SKEWNESS.
```

**Descriptives**

```
[DataSet0]
```

**Descriptive Statistics**

| | N | Range | Minimum | Maximum | Sum | Mean | | Std. Deviation | Variance | Skewness | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic | Statistic | Statistic | Std. Error | Statistic | Std. Error |
| X1 | 20 | 79,08 | 6,42 | 85,50 | 625,56 | 31,2780 | 5,00339 | 22,37583 | 500,678 | 1,483 | ,512 | 1,770 | ,992 |
| Valid N (listwise) | 20 | | | | | | | | | | | | |

*Рисунок 4.2.* **Стандартні помилки коефіцієнтів асиметрії і ексцесу в середовищі SPSS**

*Джерело:* розрахунки авторів

Тому для подальшого розв'язання задачі 4.1 слід переходити до логнормального закону розподілу. Логарифм ознаки "Дохід" має наступні значення, представлені в таблиці 4.2.

*Таблиця 4.2* . **Статистичні дані логарифма доходу домогосподарств**

| № | *ln(x)* | № | *ln(x)* | № | *ln(x)* | № | *ln(x)* |
|---|---|---|---|---|---|---|---|
| 1 | 1,859 | 6 | 2,874 | 11 | 3,226 | 16 | 3,784 |
| 2 | 2,114 | 7 | 2,991 | 12 | 3,350 | 17 | 3,865 |
| 3 | 2,299 | 8 | 3,020 | 13 | 3,367 | 18 | 4,022 |
| 4 | 2,715 | 9 | 3,063 | 14 | 3,427 | 19 | 4,442 |
| 5 | 2,845 | 10 | 3,227 | 15 | 3,450 | 20 | 4,4894 |

*Джерело:* розрахунки авторів

Перевіримо причетність логарифмічної ознаки (таблиця 4.2) для перевірки на відповідність до нормального закону розподілу. В системі *SPSS* завантажимо дані логарифмічної ознаки (рис. 4.3). Серія кроків *Analyze* → *Nonparametric* → *Sample K.S* (приклад Колмогорова-Смирнова) призводить до появи вікна *One- Sample Kolmogorov-Smirnov Test* (приклад тесту Колмогорова-Смирнова)

(рис. 4.4). Після натискання кнопки ОК з'являються результати тесту у вигляді таблиці (рис. 4.5).

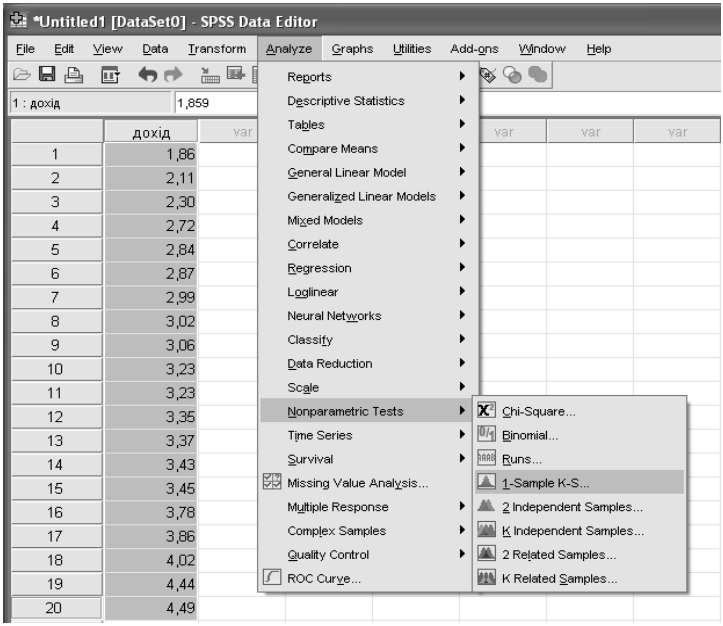



*Рисунок 4.3.* **Статистична оцінка логарифмічної ознаки в програмному середовищі SPSS**

*Джерело:* розрахунки авторів

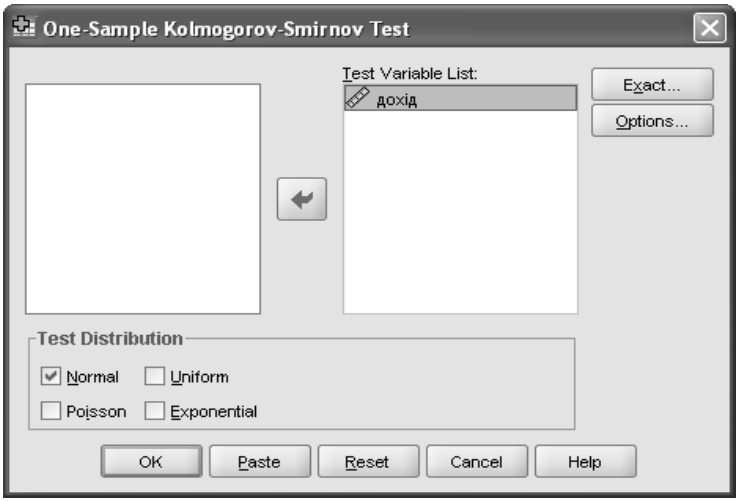



*Рисунок 4.4.* **Тест Колмогорова-Смирнова в програмному середовищі SPSS**

*Джерело:* розрахунки авторів

***One-Sample* Kolmogorov-Smirnov Test**

| | | Дохід |
|---|---|---|
| N | | 20 |
| Normal Parameters[a] | Mean | 3,2214 |
| | Std. Deviation | ,69477 |
| Most Extreme Differences | Absolute | ,121 |
| | Positive | ,121 |
| | Negative | -,094 |
| Kolmogorov-Smirnov Z | | ,542 |
| Asymp. Sig. (2-tailed) | | ,931 |

*Рисунок 4.5.* **Результати одновибіркового тесту Колмогорова-Смірнова в програмному середовищі SPSS**

*Джерело:* розрахунки авторів

У таблиці одновибіркового тесту Колмогорова -Смірнова (рис. 4.5) наведено результати наступних показників тесту нормального розподілу: середнє значення генеральної сукупності логарифмічної ознаки: Mean = 3,2214, середнє квадратичне відхилення значення генеральної сукупності логарифмічної ознаки Std. Deviation = 0,69447 ,

Найбільша екстремальна різниця (Most Extreme Differences):

Positive (позитивна) $D_n^+ = 0{,}121;$

Negative (негативна) $D_n^- = 0{,}094;$

Absolute (абсолютна) $D_n = \max\left(D_n^+, D_n^-\right) = 0{,}121;$

Kolmogorov-Smirnov Z (значення критерію)

$$D_n\left(\sqrt{20} - 0.01 + \frac{0.85}{\sqrt{20}}\right) = 0{,}542.$$

За рівнем значущості α= 0,05 із таблиці квантилів розподілу Колмогорова-Смирнова (табл. 4.3) знаходимо значення критерію 0,542, яке менше за квантиль 0,895.

*Таблиця 4.3.* **Таблиця квантилів розподілу Колмогорова-Смирнова**

| Рівень значущості | 0,15 | 0,10 | 0,05 | 0,03 | 0,01 |
|---|---|---|---|---|---|
| Квантіль розподілу | 0,775 | 0,819 | 0,895 | 0,955 | 1,035 |

Звідси слідує висновок – нульова гіпотеза про відповідність розподілу домогосподарств нормальному розподілу не відхиляється, оскільки присутній логнормальний закон розподілу. Перехід до логнормального закону виявився вдалим шляхом підбору розподілу

із законів розподілу, який присутній в багатьох пакетах обробки статистичних даних, зокрема і в SPSS.

## 4.2. Лінійні багатофакторні регресійні моделі та методи їх дослідження

Позитивна діагностика розподілу генеральної сукупності, близького до нормального, дає можливість за практичних потреб вибору виду функції регресії, тобто вибору специфікації регресії. Завдяки розвинутої схемі лінійного статистичного аналізу. вибір специфікації моделі слід починати в класі лінійних функцій. Реалізацію даної схеми розглянемо на прикладі 4.2.

*Задача 4.2.* Вивчається залежність обсягу реалізації на ринку товарів і послуг деякої продукції (результуючий фактор) від: ціни реалізації і середньодушових доходів споживачів (головні фактори). Всі інші фактори впливу (фактори збурення) на реалізацію продукції вважаємо неістотними. Статистичні дані показників приведені в таблиці 4.4.

Взагалі, за структурою лінійна специфікація моделі визначається неоднозначно (лінійних функцій існує нескінченно багато). Тому постає питання вибору найбільш кореляційного функціонального зв’язку між пояснювальною змінною і пояснюючими змінними. З практичних міркувань, для забезпечення такого кореляційного зв’язку вплив факторів збурення на результуючий фактор повинен бути мінімальним. Напевно, кращим кореляційним зв’язком буде той, в якого відхилення значень регресанта від функцій регресії буде найменшим.

*Таблиця 4.4.* **Вихідні дані обсягів реалізації продукції, ціни реалізації і душових доходів споживачів**

| Номер спостереження | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Обсяг реалізації, тон | 25 | 30 | 20 | 25 | 15 | 10 | 20 | 35 | 40 | 30 |
| Ціна, грн за 1 кг | 70 | 60 | 75 | 70 | 90 | 100 | 80 | 60 | 50 | 65 |
| Дохід споживачів на 1 особу, грн за період | 200 | 210 | 220 | 230 | 220 | 210 | 230 | 240 | 250 | 250 |

Тобто, за міру відхилення можна взяти:

1) суму модулів відхилень:

$$g(\beta_1,...,\beta_n)=\sum_{k=1}^{n}\left|y_t-\sum_{t=1}^{T}\beta_k x_{t,k}\right|=\sum_{t=1}^{T}|u_t| \qquad (4.8)$$

Це *метод найменших модулів*, що знаходить мінімум суми відхилень їх абсолютних значень. Він є найкращим (за максимальної вірогідності) у випадку, коли відхилення мають розподіл Лапласа. Стосовно випадкових значень цей метод значно менш чутливий, ніж метод найменших квадратів (МНК), проте він може мати більш, ніж один розв'язок і що важливо, для нього не існує простої формули визначення оцінки параметрів моделі.

За міру відхилення також можна взяти:

2) суму квадратів відхилень (за МНК):

$$f(\beta) = f(\beta_1, \beta_2, ..., \beta_n) = \sum_{t=1}^{T} \left[ y_t - (\beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_n x_{tn} \right]^2 = \sum_{t=1}^{T} u_t^2 \quad (4.9)$$

Функція в такому випадку $f(\beta_1, ..., \beta_n)$ називається цільовою. Для двох факторної моделі цільова функція матиме вигляд:

$$f(\beta) = f(\beta_1, \beta_2) = \sum_{t=1}^{T} \left[ y_t - (\beta_1 x_{t1} + \beta_2 x_{t2}) \right]^2 \quad (4.10)$$

Потрібно знайти такі значення параметрів $\beta_1, ..., \beta_n$, які дають мінімум цільової функції . Мінімум функції існує в обох випадках і він досягається при одних і тих же значеннях параметрів за формулою .

$$\tilde{\beta} = (X'X)^{-1} Y'X = (X'X)^{-1} X'Y; \quad (4.11)$$

де: *X* – матриця регресорів (матриця, складена з одиничного вектору і масиву векторів статистичних даних головних факторів – ціни і доходів споживачів),

*X′*- транспонована до матриці регресорів,

*Rang ( X X′)= n* (матриця *X X′* повинна мати повний ранг),

*Y*- вектор статистичних даних результуючого фактору (попиту на продукцію).

Стосовно набору інструментів досліджень, тобто використання математичного апарату, другий випадок оптимізації (формула 4.11) для розв'язання даної задачі підходить найкраще.

Одержаний розрахунковий вектор $\tilde{\beta}$ є оцінкою за методом найменших квадратів (МНК- оцінкою) вектору $\beta$ коефіцієнтів моделі. Вектор $\tilde{Y} = X\tilde{\beta}$ називають розрахунковим, він є МНК-оцінкою для регресанта *Y*. Реально, за змістом функції регресії, він визначає середній обсяг реалізації продукції за різними статистичними значеннями головних факторів - ціни реалізації і доходів споживачів. Різниця $\tilde{Y} - X\tilde{\beta} = Y - \tilde{Y}$, тобто різниця векторів фактичних значень показника і розрахункових називають

вектором залишків або помилок регресії. позначають $\tilde{U} = Y - \tilde{Y} = \tilde{Y} - X\tilde{\beta}$. Формула знаходження розрахункових коефіцієнтів моделі має ряд наслідків один із яких стосується суми залишків моделі, середнє їх значення завжди рівне нулю, що є підтвердженням факту відсутності систематичних помилок у вимірах результуючого і головних факторів. МНК - оцінкою вектору збурень *U* є вектор залишків $\tilde{U}$. У МНК оцінкою дисперсії вектору збурень є наступна величина:

$$\sigma_{\tilde{u}}^2 = \frac{\sum_{t=1}^{T} \tilde{U}_t^2}{T-n} \tag{4.12}$$

Рівняння $Y = X\tilde{\beta} + \tilde{U}$ , або його координатний запис:

$$y_t = \sum_{k=1}^{n} x_{tk}\tilde{\beta}_k + u_t \tag{4.13}$$

називають *емпіричним рівнянням кореляційної моделі*. Вираз $X\tilde{\beta}$, або координатний його запис:

$$\tilde{y}_t = \sum_{k=1}^{n} x_{tk}\tilde{\beta}_k \tag{4.14}$$

називають функціональною, регулярною, або систематичною частиною регресанта (результуючого показника), за якою приймається ряд важливих рішень практичного використання лінійної моделі регресії.

Крім вище названих методів побудови множинної регресії існують інші. Розглянемо їх детальніше.

*Метод максимальної вірогідності*. Використовується коли відомі всі розподіли відхилень для всіх спостережень. При класичній та узагальненій моделях лінійної регресії з умовою нормальності відхилень приводить до того ж результату, що і метод найменших квадратів та узагальнений метод найменших квадратів відповідно.

*Ортогональна регресія*. Застосовується у випадках коли серед значень змінних можуть бути випадкові складові й при оцінці враховують можливі відхилення по всіх змінних.

В залежності від сфер застосування лінійного регресійного аналізу лінійна регресія класифікуються з допомогою положень (передумов, припущень), які можуть бути накладені на лінійну регресію. Положення визначають такі поняття, як лінійна модель нормальної регресії, лінійна модель класичної регресії і узагальнена модель регресії або економетрична модель. Наступне положення 1 вимагає наявності регресії взагалі.

*Положення* 1**.** Математичне сподівання вектору збурень нульовий вектор, $MU$=0. Цю рівність потрібно розуміти як покоординатну рівність нулю, $Mu_j$=0, $j$=1,…,$T$. Якщо $MU \neq 0$, то йдеться про помилку специфікації моделі.

Зауважимо, що зміст положення 1 відповідає такому явищу як гомоскедастичність.

Наступне положення 2, вимагає розуміння понять математичної статистики. Якщо $U = (u_1, \ldots, u_T)$ випадковий вектор, то через $\Sigma(u)$ позначають його коваріаційну матрицю, вона складається із попарних коваріацій його компонент:

$$cov(u_j, u_i) = \begin{cases} \sigma^2_{u_i}, колиi = j \\ \sigma_{u_i,u_j}, колиi \neq j \end{cases} \quad (4.15)$$

Коваріаційна матриця збурень симетрична, з елементами по головній діагоналі, що відображають дисперсії вектору збурень.

*Положення* 2. Коваріаційна матриця вектору збурень діагональна з однаковими елементами по діагоналі. Якщо в діагональній матриці $\Sigma(u)$ по головній діагоналі стоять, взагалі різні елементи, то таке явище називають гетероскедантичністю. Практично її можна виявити, коли неоднакові дисперсії виникають для окремих груп незалежних змінних. Загальне положення коваріаційної матриці відповідає явищу автокореляції компонент вектору збурення.

*Положення* 3**.** Вектор збурень має нормальний закон розподілу.

*Положення* 4**.** Довжина рядів спостережень більша кількості регресорів .Це положення називають положенням про достатність степенів вільності моделі. Якщо $T$ – довжина (або об'єм вибірки) спостережень, $n$ – кількість регресорів моделі, то $T$- $n$ – число степенів вільності. Положення 4 стверджує, що $T$- $n$> 0, тобто наявність достатньої кількості степенів вільності.

*Положення* 5**.** Матриця регресорів має повний стовпчиковий ранг. Якщо $X$ – матриця регресорів, $dim\ X = T \times n$ її розмір, то положення 5 стверджує, що $rang\ X = n$, матриця має повний ранг, або незалежні змінні утворюють лінійно-незалежну систему векторів

*Положення* 6. Регресори – це детерміновані (нестохастичні, невипадкові) величини. Положення 6 носить методичний характер і слугує для полегшення побудови лінійної моделі регресії, воно може бути замінено, у випадку стохастичності регресорів, на умову їх незалежності, як випадкових величин, і незалежності від факторів збурення.

Якщо в лінійній моделі виконуються всі перші шість положень, то така регресія називається лінійною нормальною регресійною моделлю. Якщо якесь із положень, крім 1, 3 і 4 узагальнюється, то така лінійна модель називається узагальненою лінійною регресією, або економетричною моделлю. Якщо в лінійній моделі регресії виконуються положення 1, 2, 4, 5 і 6, то така лінійна модель вважається лінійною моделлю класичної регресії. Лінійна класична регресія знаходить своє застосування в природничих і технічних науках. Перевірка названих положень відбувається за певними статистичними тестами – гіпотезами, критеріями та розподілами.

### 4.3. Якість лінійної регресії

Вибору певного класу лінійної регресії повинно передувати дослідження якості регресії статистичним даним на предмет істотності лінійної залежності результативного показника від головних факторів.

В поняття якості (стійкості) лінійної регресії вкладається вибір критеріальних характеристик, які б підтверджували правильність вибору лінійної специфікації регресії. Якість нормальної лінійної регресії оцінюється наявністю лінійного зв'язку між результуючим показником і факторами. Відсутність лінійного зв'язку між ендогенними і екзогенними змінними в рівнянні однофакторної моделі: $y = \beta_1 + x_t\beta_2 + u_t$ означає рівність нулю другого коефіцієнта моделі, а в багатофакторній моделі: $y_t = \beta_1 x_{t1} + ... + \beta_k x_{tk} + ... + \beta_n x \quad + u_t$ рівність нулю всіх невідомих коефіцієнтів, починаючи із другого. Перевірка невідомих параметрів стосовно їх значень в математичній статистиці здійснюється, як відомо, шляхом перевірки гіпотез. Для однофакторної моделі потрібно перевірити гіпотезу $H_o : \beta_2 = 0$, за альтернативної гіпотези $H_a : \beta_2 \neq 0$. Для багатофакторної моделі перевіряється гіпотеза $H_o : \beta_2 = 0, \beta_3 = 0, ..., \beta_n = 0$ за альтернативної гіпотези – відмінного від нуля хоча б одного коефіцієнта: $H_a : \beta_k \neq 0 \ (k = 2, \ldots, n)$. Перевірка гіпотез здійснюється із застосуванням коефіцієнта детермінації за наступним *F*-критерієм:

$$F = \frac{R^2}{1 - R^2} \frac{T - n}{n - 1} \ , \tag{4.16}$$

Коефіцієнт детермінації ($R^2$ ) визначається як відношення варіації результуючого показника до пояснювальної його варіації.

Коефіцієнт детермінації $R^2$ , як величина, побудована за вибірковими даними, є випадковою, тому дане відношення, як частка випадкових величин, представляє собою теж випадкову величину, розподілену за законом Фішера (*F*- розподіл) із степенями вільності *T*- *n* і *n*-1. Рівень значимості гіпотез можна знайти за комп'ютерним розрахунком, наприклад, в програмному середовищі *Excel* за операторною формулою $\mu := FPACП(F, n-1, T-n)$, На першій позиції в цих формулах знаходиться значення *F*-крітерію, на другій і третій позиціях – число степенів вільності чисельника і знаменника *F*-крітерію. Значення параметра $\mu$ не більше п'яти відсотків ($\mu \le 0{,}05$) гарантує не менше 95% відсоткове відхилення нульової гіпотези, тобто не менше, ніж в 95% всіх вибірок маємо лінійну модель регресії. Випадок $\mu > 0{,}05$, за певних умов практики, говорить про непридатність лінійної специфікації, потрібно переходити до побудови нелінійної регресії. У задачі 4.2, для одно факторної моделі: $R^2$ = 0,977; *n*-1 = 2-1 = 1; *T*- *n* = 10 – 2 = 8; *F* = 170,667; $\mu := FPACП(170{,}667, 1, 8) = 1{,}11637 \times 10^{-6} < 0{,}005$; для двох факторної моделі: $R^2$ = 0,9708; *n*-1 = 3-1 = 2; *T*- *n* = 10 – 3 = 7; *R*= 0.985. ( рис. 4.6).

| *Регресійна статистика* | |
|---|---|
| Множинний R | 0,985314389 |
| R-квадрат | 0,970844444 |
| Нормований R-квадрат | 0,962514286 |
| Стандартна помилка | 1,767430203 |
| Спостереження | 10 |

*Рисунок 4.6.* **Розрахункові показники множинної регресійної моделі в програмному середовищі Excel**

*Джерело:* розрахунки авторів

В Excel функція “Регресія” пакету “Аналіз даних” четверта графа (*F*) в таблиці “Дисперсійний аналіз” (рис. 4.7) визначає статистику Фішера $F$ = 116,5457 , п’ята графа – рівень значимості $\mu = FРАСП(116{,}5457{,}2{,}7) = 4{,}238E-06$=4,238×10$^{-6}$. який менший 0.05 (рис. 4.7). Звідки слідує висновок: так як допустимий рівень значимості виявився меншим допустимого, то лінійна регресія є значимою, тобто специфікація регресії (моделі), а саме - вид функціонального зв’язку, вибрана вірно.

*Дисперсійний аналіз*

| Показник | *df* | *SS* | *MS* | *F* | *Значимість F* |
|---|---|---|---|---|---|
| Регресія | 2 | 728,1333 | 364,0667 | 116,5457 | 4,23E-06 |
| Остаток | 7 | 21,86667 | 3,12381 | | |
| Разом | 9 | 750 | | | |

*Рисунок 4.7.* **Показники дисперсійного аналізу множинної регресійної моделі в програмному середовищі Excel**
*Джерело:* розрахунки авторів

В двох факторній регресії, (задача 4.2) тест на її якість можна провести і за порівняльним аналізом розрахункового значення критерію і табличного за оператором Excel *FРАСПОБР* (ймовірність, степінь свободи 1, степінь свободи 2) = *FРАСПОБР*(0,05;10-3;3-1) =19,353. Так як $F_{розр} > F_{табл}$, 116,54 > 19, 354,то слідує висновок про те, що дана множинна лінійна регресія є значимою .

## 4.4. Адекватність множинної лінійної регресії статистичним даним

В процесі моделювання завжди виникає питання вибору специфікації моделі. Якщо для побудови моделі вибрано лінійну модель регресії, то за якими критеріями такий вибір зроблено? Яка похибка такого вибору? Похибка вибору повинна оцінюватись через відхилення від регресанта функції регресії. На перший погляд останнє повинно визначатись сумою квадратів залишків моделі. Однак ця величина має істотній недолік, який ускладнює порівняння степеня відповідності різних регресивних рівнянь – це відсутність її верхньої межі. Цей недолік усувається використанням такого поняття, як коефіцієнт детермінації. Коефіцієнт детермінації виник як результат порівняння варіацій фактичних значень регресанта і його розрахункових значень за лінійною моделлю. Цей результат випливає із наступної тотожності:

$$\sum_{t=1}^{T}(y_t-\bar{y})^2=\sum_{t=1}^{T}(\tilde{y}_t-\bar{\bar{y}})^2+\sum_{t=1}^{T}(y_t-\tilde{y}_t)^2= \sum_{t=1}^{T}(y_t-\tilde{y}_t)^2+\sum_{t=1}^{T}\tilde{u}_t{}^2 \qquad (4.17)$$

У формулі (4.17) - це друга графа (*SS*) в таблиці "Дисперсійний аналіз" (рис. 4.7), яка визначає суму квадратів відхилень розрахункових значень результуючого показника, отриманих за моделлю, від його середнього значення:

$$SS_{\text{R}}=\sum_{t=1}^{T}(\tilde{y}_t-\bar{y})^2=728{,}1333333,$$

а також визначає суму квадратів залишків моделі:

$$SS_{\text{P}}=\sum_{t=1}^{T}(y_t-\tilde{y}_t)^2=\sum_{t=1}^{T}\tilde{u}_t{}^2=21{,}86666667.$$

Загальна сума: $SS\ SS_{\text{M}} = SS_{\text{P}} + SS_{\text{R}} = \sum_{t=1}^{T}(y_t-\bar{y})^2 = 750.$

Ділення тотожності на варіацію фактичних значень регресанта приводить нас до поняття коефіцієнта детермінації:

$$R^2=\frac{\sum_{t=1}^{T}(\tilde{y}_t-\bar{\bar{y}})^2}{\sum_{t=1}^{T}(y_t-\bar{\bar{y}})^2}=\frac{\frac{\sum_{t=1}^{T}(\tilde{y}_t-\bar{y})^2}{T}}{\frac{\sum_{t=1}^{T}(y_t-\bar{y})^2}{T}}, \qquad (4.18)$$

діапазон зміни якого лежить у межах від нуля до одиниці і тому може бути виражений у відсотках. Формулу (4.18) називають *першою версією коефіцієнта детермінації*. Вона має як статистичний, так і практичний зміст стосовно тієї галузі знань, в якій модель регресії будується. Статистичний зміст першої версії коефіцієнта детермінації визначає долю поясненої вибіркової дисперсії (розсіювання): $\sum_{t=1}^{T}(\mathrm{y}_t-\overline{y})^2 \Big/ T$ в загальній пояснювальній дисперсії: $\frac{\sum_{t=1}^{T}(\tilde{y}_t-\bar{y})^2}{T}$.

Взагалі, практичний зміст коефіцієнта детермінації визначає, що зміна результуючого показника на $R^2$*100 % залежить від зміни головних факторів і на (1- $R^2$)100% залежить від зміни факторів збурення. Таке пояснення даного показника надто загальне і вимагає певного уточнення. Практичне пояснення коефіцієнта детермінації має дати відповідь на питання – за рахунок чого в лінійної моделі регресії забезпечується зміна результуючого показника в залежності від зміни факторів, наприклад в розрахунку на одиницю (або на декілька одиниць) свого виміру. Пояснення можна отримати із представлення коефіцієнта детермінації за формулою (4.18) і тотожності (4.19) у вигляді наступної системи рівностей:

$$\begin{cases} \sum_{t=1}^{T}(y_t-\bar{y})^2=\sum_{t=1}^{T}(y_t-\tilde{y}_t)^2+\sum_{t=1}^{T}\tilde{u}_t{}^2 \\ R^2=\frac{\sum_{t=1}^{T}(\tilde{y}_t-\bar{y})^2}{\sum_{t=1}^{T}(y_t-\bar{\tilde{y}})^2} \end{cases}, \qquad (4.19)$$

Із системи (4.18) слідує тотожність:

$$\sum_{t=1}^{T}(\mathrm{y}_t-\overline{\mathrm{y}})^2=\sum_{t=1}^{T}(\mathrm{y}_t-\overline{\mathrm{y}})^2R^2+\sum_{t=1}^{T}(\mathrm{y}_t-\overline{\mathrm{y}})^2(1-R^2). \qquad (4.20)$$

Практичний коментар даної рівності полягає в тому, що квадратична зміна ( $\sum_{t=1}^{T}(\mathrm{y}_t-\overline{\mathrm{y}})^2$ ) результуючого показника (y) на $R^2$ відсотків залежить від квадратичної зміни основних факторів і на *1-*

$R^2$ відсотків залежить від квадратичної зміни інших факторів. Тотожність (4.20) можна замінити іншою:

$$1 = 1R^2 + 1(1 - R^2) \ , \qquad (4.21)$$

структурний зміст якої наступний: в кожній квадратній одиниці зміни регресанта $R^2$ одиниць (відсотків) зміни приходиться за будь-якої лінійної зміни головних факторів і $1\text{-}\ R^2$ одиниць (відсотків) – за будь-якої лінійної зміни факторів збурення. Коротко, одиниця виміру зміни кг$^2$ (або просто кілограм, за лінійною мірою) структурується за головними факторами і факторами збурення.

Зауважимо, коефіцієнт детермінації, за своїм визначенням, інваріантний відносно лінійної зміни регресанта (зміна на декілька одиниць регресата, на стільки ж одиниць змінює його середнє значення). Виникає питання можливої зміни складових частин у формулі (4.20) в результаті зміни факторів (незалежних змінних) на одиницю , або декількох одиниць їх виміру. За такої зміни факторів складові "одиниці" у формулі (4.19) є інваріантними. Для перевірки даного факту, маючи запис розрахункових (середніх) значень результуючого фактору: $\tilde{y}_t = \tilde{\beta}_1 + \tilde{\beta}_2 x_{t2} + \tilde{\beta}_3 x_{t3}$, його середнього значення: $\bar{y} = \tilde{\beta}_1 + \tilde{\beta}_2 \bar{x}_2 + \tilde{\beta}_3 \bar{x}_3$, емпіричного рівняння моделі: $\tilde{y}_t = \tilde{\beta}_1 + \tilde{\beta}_2 x_{t2} + \tilde{\beta}_3 x_{t3} + \tilde{u}_t$, враховуючи, що залишки моделі мають нульову коваріацію зі спостережуваними значеннями регресорів та розрахунковим значенням регресанта, тобто:

$cov(X_k . \widetilde{U}) = 0, k = 1, \dots, n,\ cov(\widetilde{Y}, \widetilde{U}) = 0,$

можна представити коефіцієнт детермінації в наступному вигляді:

$$R^2 = \frac{\sum_{t=1}^{T}\left[\tilde{\beta}_2(x_{t2} - \bar{x}_2) + \tilde{\beta}_3(x_{t3} - \bar{x}_3)\right]^2}{\sum_{t=1}^{T}\left[\tilde{\beta}_2(x_{t2} - \bar{x}_2) + \tilde{\beta}_3(x_{t3} - \bar{x}_3) + \tilde{u}_t\right]^2} =$$

$$\frac{\sum_{t=1}^{T}\left[\tilde{\beta}_2(x_{t2} - \bar{x}_2) + \tilde{\beta}_3(x_{t3} - \bar{x}_3)\right]^2}{\sum_{t=1}^{T}\left[\tilde{\beta}_2(x_{t2} - \bar{x}_2) + \tilde{\beta}_3(x_{t3} - \bar{x}_3)\right]^2 + \sum_{t=1}^{T}\tilde{u}_t{}^2} \qquad (4.22)$$

Представлення коефіцієнта детермінації за даною формулою дає можливість переконатись, що сукупна зміна (в квадратичній метриці) головних факторів наприклад, на декілька одиниць свого

виміру кожного із головних факторів, не змінює коефіцієнта детермінації.

Інваріантність коефіцієнта детермінації, а також практичний зміст рівності (4.21) дозволяють дати йому практичний зміст в цілому, як структури одиниці виміру результуючого показника, відносно факторів за будь-якої їх зміни.

Зміст коефіцієнта детермінації в прикладі 4.2 (за просторової структури статистичних даних): за будь якої лінійної зміни всіх факторів (головних факторів і факторів збурення) в кожній тоні зміни реалізації продукції в середньому, за головними факторами ця зміна складала 971 кг (97,1%) даного продукту і 29 кг (29,1%) - за факторами збурення.

Слід пам'ятати, що зміна результуючого показника відбувається в квадратичній метриці. Наприклад, якщо зміна реалізації продукції становить 2 тони, то за рахунок головних факторів вона становитиме 4×971 = 3884 кг. і 4×29=118 кг- за іншими факторами. Сама зміна реалізації може мати двояку тенденцію, на підвищення попиту і на його зниження.

Якщо потрібно зробити порівняльний аналіз структур зміни реалізації однієї тони даного продукту на декількох ринках, то доцільно зробити перехід від натурального представлення складових в структурі (4.21) до відсоткової.

*Друга версія коефіцієнта детермінації* визначається за формулою:

$$R^2 = 1 - \frac{\sum_{t=1}^{T} \widetilde{u_t^2}}{\sum_{t=1}^{T} (y_t - \bar{y})^2} \qquad (4.23)$$

і слугує поясненням екстремальних значень коефіцієнта детермінації: $R^2 = 0$ і $R^2 = 1$, тобто, коли модель лінійної регресії описує стаціонарний процес (результуюча змінна і середнє значення результуючої змінної не залежить від головних факторів моделі, а залежить тільки від значення факторів, які не увійшли в модель) і коли є лінійна модель залежності результуючої змінної від всіх змінних моделі.

Практика виробила нормативне значення коефіцієнта детермінації 0,95. Якщо вибірковий коефіцієнт детермінації виявиться меншим нормативного 0,95 і його значення, з якихось міркувань не влаштовує дослідника, то виникає необхідність прийняти рішення про введення в модель додаткових головних факторів заданого об'єму даних. Як відомо, введення додаткових факторів може лише збільшити коефіцієнт детермінації, поліпшити апроксимацію статистичних даних. Але є інша проблема застосування вибіркового коефіцієнта детермінації в тому, що його значення не дозволяє порівняння моделей з різною кількістю факторів. Введення додаткових факторів, при заданому об'ємі вибірки, зменшує число степенів вільності моделі, що негативно впливає на статистичні оцінки параметрів моделі. Якщо додатково ставиться завдання мати заданий рівень прогнозних оцінок параметрів моделі, то для таких цілей використовують альтернативнй показник коефіцієнта детермінації, наприклад скорегований (нормований) коефіцієнт детермінації (за Тейлом) за формулою:

$$\bar{R}^2 = 1-(1-R^2)\frac{\mathrm{T}-1}{\mathrm{T}-n}, \qquad (4.24)$$

де: $T$ – об'єм вибірки, $n$ – кількість регресорів. В якості критерію вибору перевага надається моделі регресії з більшим скорегованим коефіцієнтом детермінації. У прикладі 4.2 для однофакторної моделі (вважається, що на попит впливає тільки фактор ціни):

$$\bar{R}^2 = 1\text{-}(1\text{-}0{,}977^2)9/8 = 0{,}949.$$

Для двофакторної моделі:

$$\bar{R}^2 = 1\text{-}(1\text{-}0{,}985^2)9/7 = 0{,}963.$$

Звісно, в практичних задачах перевага надається двофакторній моделі.

На практиці доцільно використовувати показники за лінійною мірою їх зміни. Тому корисно мати результат зміни результуючого показника в залежності від сукупної зміни факторів не в

квадратичній метриці, а в лінійній. Приведення коефіцієнта детермінації до лінійної метрики можна здійснити використанням множинного коефіцієнта кореляції між фактичними значеннями результуючої змінної і її розрахунковими значеннями за відомою формулою:

$$R = \frac{\sum_{t=1}^{T}(\tilde{y}_t - \bar{y})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{T}(y_t - \bar{y})^2}\sqrt{\sum_{t=1}^{T}(\tilde{y}_t - \bar{y})^2}}, \qquad (4.25)$$

Дана формула слугує відображенням третьої версії коефіцієнта детермінації лінійної моделі регресії - він є квадрат множинного коефіцієнта кореляції, тобто:

$$R = \frac{\sqrt{\sum_{t=1}^{T}(\tilde{y}_t - \bar{y})^2}}{\sqrt{\sum_{t=1}^{T}(y_t - \bar{y})^2}} \qquad (4.26)$$

Множинний коефіцієнт кореляції за даною формулою представлений в середньоквадратичному вимірі інваріантний відносно лінійної зміни факторів. Аналог тотожності типу (4.21), а також (4.20) (типу $1 = 1R + 1(1-R)$) для нього не має місця. Тому не можливо за факторами структурувати одиницю зміни результуючого показника, Можна показати лінійну зміну результуючого показника в середньоквадратичній метриці тільки за головними факторами. У прикладі 4.2 значення коефіцієнта кореляції $R$=0.985 визначає, що зміна реалізації однієї тони даного продукту зумовлена реалізацією 985 кг (98,5%) за головними факторами (ціни і доходів). Віднести 1,5 кг. (1,5%) зміни реалізації за іншими факторами, в середньоквадратичній метриці, неможливо.

Множинний коефіцієнт кореляції визначає ступінь лінійної залежності, якщо така існує між результуючим показником і факторами у вибірці. Вибірковий коефіцієнт множинної кореляції є середнім очікуваним значенням для істинного множинного

коефіцієнта кореляції ($R_p$) всієї сукупності. Останній з них – це невідома величина, його статистичну оцінку можна знайти за критерієм Стьюдента:

$$S = \frac{1-R}{\sqrt{T}} \tag{4.27}$$

де: *R*- множинний коефіцієнт кореляції,
*S* – помилка коефіцієнта кореляції,
*T* – об'єм вибірки.

Знаходження довірчого інтервалу вимагає незміщеної оцінки коефіцієнта детермінації. Незміщеність усувається переходом до незміщеного коефіцієнта детермінації $R_n^2$ за формулою:

$$R_n^2 = 1 - \frac{T-1}{T-n}(1-R^2), \tag{4.28}$$

де: $T$ - об'єм вибірки, $n$ –кількість регресорів.
Коефіцієнт кореляції:

$$R_n = \sqrt{1 - \frac{T-1}{T-n}(1-R^2)}\ . \tag{4.29}$$

В задачі 4.2 він становить:

$$R_n = \sqrt{1 - \frac{10-1}{10-3}(1-0.9708^2)} = 0{,}962,$$

значення помилки $S = \frac{1-R_n}{\sqrt{T}} = \frac{1-0.962}{\sqrt{10}} = 0{,}012$. Довірчий інтервал коефіцієнта кореляції має наступний вид:

$$R_n - t(\alpha/2,k)S \le R_p \le R_n + t(\alpha/2,k)S \tag{4.30}$$

де: $k$ – кількість степенів вільності,
$\alpha$ – рівень значимості,
$t(\alpha/2,k)$ - критична точка розподілу Стьюдента.

У задачі 4.2, $R_n$=0,962,

$t(\alpha/2,k) = t(0.05/2,7)$ = *СТЬЮДРАСПРОБР*(0,025; 7) =2,841. $t(\alpha/2,k)S$ =2,841×0,012 = 0,034.

Довірчий інтервал (0,962-0,034; 0,962+0,034) =(0,928;0,996). Отриманий довірчий інтервал є степінь статистичної оцінки лінійної залежності регресанта і факторів у всій генеральній сукупності. Істинний множинний коефіцієнт кореляції знаходиться в межах (покривається інтервалом ) від 0,928 до 0,996 . Практичний зміст множинного коефіцієнт кореляції визначає, що зміна на ринку реалізації в розмірі однієї тони даного продукту зумовлена зміною реалізації за головними факторами в межах від 928 кг (92,8%) до 996 кг (99,6%) продукції.

Зміну реалізації продукції в середньоквадратичній метриці можна ще трактувати, як порівняння ризиків відхилення регресанта від очікуваного (середнього) його значення за вибіркою (статистикою) і ризику відхилення його значень за моделлю. У прикладі 4.2 ризик зменшився на велдичину: 2,9% = 100% - 97,1%.

Вибірковий коефіцієнт детермінації випадкова величина, тому підлягає статистичній оцінці. Зв'язок вибіркових показників: коефіцієнта детермінації і множинного коефіцієнта кореляції дозволяє статистично оцінити коефіцієнт детермінації, тобто побудувати довірчий інтервал для істинного його значення шляхом піднесення кінців довірчого інтервалу множинного коефіцієнта до квадрату. У задачі 4.2 довірчий інтервал для коефіцієнта детермінації матиме вид ($0{,}928^2$, $0{,}996^2$) = (0,861,0,992). Практичний зміст довірчого інтервалу: зміна попиту в розмірі однієї тони продукції зумовлена зміною в межах від 861 кг (86,1%) до 992 кг (99,2%) продукції за головними факторами і від 108 кг. (10,8%) до 139 кг (13,9%) продукції за іншими факторами. Зазначена зміна продукції відбулась в лінійній метриці і за будь-якої можливої лінійної зміни всіх факторів.

Нормативне значення коефіцієнта детермінації 0,95 є орієнтовним. В залежності від дослідження він може відрізнятись від прийнятого.

Відхилення коефіцієнта детермінації від прийнятого в меншу (більшу сторону компенсується включенням (виключенням) в групу (із групи регресорів) додаткових факторів . Тому виникає питання, на яку величину збільшиться (зменшиться) коефіцієнт детермінації, коли буде включений (виключений) *k*- тий регресор. Формула для розрахунку граничного внеску $k$ – того регресора в коефіцієнт детермінації має вигляд

$$\Delta R_k^2 = \frac{(1-R^2)t_k^2}{T-k} \tag{4.31}$$

де: *T*- об'єм вибірки,

$k$ – кількість регресорів,

*T*- *k* – число степенів вільності,

$t_k^2$ - квадрат критичної точки розподілу Стьюдента, яка визначається як частка від ділення *k* – того розрахункового коефіцієнта регресії і його середньоквадратичного відхилення:

$$t_k = \frac{\tilde{\beta}_k}{\sigma_{\tilde{\beta}_k}} \tag{4.32}$$

де $\tilde{\sigma}_{\beta_k}$- стандартна помилка розрахункового коефіцієнта регресії, він є діагональним елементом матриці:

$\tilde{\sigma}_u^2(X^T X)^{-1}$, де $\tilde{\sigma}_u^2$ - МНК оцінка вектору збурень обчислена за формулою (4.9). Представимо необхідні розрахунки показників у задачі 4.2:

$$(X`X)^{-1} = \begin{vmatrix} 39.3 & -0,115429 & -0,136762 \\ -0,115429 & 0,00067 & 0,000297 \\ -0,136627 & 0,00029 & 0,00051 \end{vmatrix},$$

$$\sigma_{\tilde{u}}^{\ 2} = \frac{\sum_{t=1}^{T} \tilde{U}_t^{\ 2}}{T-n} = 3{,}124,$$

$$\tilde{\sigma}_{\tilde{u}}^2 (X`X)^{-1} = \begin{bmatrix} 122.832705 & -0{,}360599 & -0{,}427244 \\ -0{,}360599 & 0{,}002095 & 0{,}000928 \\ -0{,}427244 & 0{,}000928 & 0{,}001595 \end{bmatrix},$$

стандартні помилки розрахункових коефіцієнтів:

$$\sigma_{\tilde{\beta}_1} = \sqrt{122.832705} = 11.082995,\ \sigma_{\tilde{\beta}_2} = \sqrt{0.0020955} = 0.046,$$

$$\sigma_{\beta_3} = \sqrt{0.001595} = 0.0399$$

(в Excel знаходяться в третій колонці третьої таблиці пакету "Регресія"). Критерії розподілу Стьюдента:

$$t_1 = \frac{47.267}{11.082992} = 4.265,\ t_2 = \frac{|-0552|}{0.046} = 12,$$

$$t_3 = \frac{0.077}{0.04} = 1.925. \qquad (4.33).$$

В Excel значення критеріїв розподілу Стьюдента знаходяться в колонці "t-статистика" третьої таблиці пакету "Регресія". У прикладі 4.2 розрахунок граничного внеску $k$ – того регресора коефіцієнт детермінації становить:

$$\Delta R_2^2 = \frac{(1-0.985^2)12^2}{10-3} = 0.6125,$$

$$\Delta R_3^2 = \frac{(1-0.985^2)1.925^2}{10-3} = 0.0158$$

Отримані результати розрахунків $\Delta R_2^2$ і $\Delta R_3^2$ означають наступне: коли із даної регресії буде виключений фактор $x_{t2}$ (ціну реалізації продукції), то коефіцієнт детермінації $R^2$ зменшиться на $\Delta R_2^2 = 0{,}612$ і стане рівним 0,359. У рівнянні регресії лишиться тільки фактор $x_{t3}$ – середньо душові доходи споживачів. Виключення

фактора $x_{t3}$ (середньодушових доходів споживачів) приведе до зменшення $R^2$ на 0,0158, він стане рівним, 0,971 - 0,016=0,955. Зміна коефіцієнтів детермінації відбулась в квадратичній метриці Коментар зміни множинного коефіцієнта кореляції.

Практичний зміст проведених операцій полягає в наступному: виключений із регресії ціновий фактор свідчить, що в тоні реалізованої продукції реалізувалось, за будь-якої зміни всіх факторів, 359 кг. (35,9%) за фактором середньодушових доходів споживачів і 1000-359=641 кг.(64,1%) за іншими факторами. Виключений фактор $x_{t3}$ - середньодушових доходів споживачів свідчить, що в тоні реалізованої продукції реалізувалось, за будь-якої зміни всіх факторів, 955 кг (95,5%) за фактором ціни реалізації і 1000-955=45 кг (4,5%) за іншими факторами.

Крім відсоткового вкладу сукупної зміни головних факторів (ціни і доходи споживачів) і факторів збурення в зміну результуючого фактору (попиту), про що свідчить коефіцієнт детермінації, регресія забезпечує можливість отримати залежність зміни результуючого фактору (попиту) від зміни кожного із головних факторів (ціни реалізації і доходів споживачів) окремо. Характеристиками такої залежності є розрахункові коефіцієнти лінійної моделі регресії $\tilde{\beta}_2$ і $\tilde{\beta}_3$. Питання знаходиться в площині практичного змісту всіх розрахункових коефіцієнтів моделі регресії $\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3$.

Якщо в рівнянні регресії: $\tilde{y} = \tilde{\beta}_1 + \tilde{\beta}_2 x_{t2} + \tilde{\beta}_3 x_{t3}$ змінити (наприклад, збільшити) другий регресор на одиницю свого виміру, або на деяку величину δ, лишаючи всі інші регресори незмінними, то отримаємо нові значення регресанта:

$\tilde{\tilde{y}} = \tilde{\beta}_1 + \tilde{\beta}_2(x_{t2} + 1) + \tilde{\beta}_3 x_{t3}$ $(\tilde{\tilde{y}} = \tilde{\beta}_1 + \tilde{\beta}_2(x_{t2} + \delta) + \tilde{\beta}_3 x_{t3})$.

Віднімемо ці дві рівності, отримаємо:

$$\tilde{\tilde{y}} - \tilde{y} = \tilde{\beta}_2 \ (\tilde{\tilde{y}} - \tilde{y} = \tilde{\beta}_2 \delta),$$

звідки виплаває, що зміна другого регресора на одиницю свого виміру, або на декілька одиниць (δ), при незмінності іншого фактору,

викликає зміну регресанта в середньому на $\tilde{\beta}_2$ ($\tilde{\beta}_2\delta$) одиниць свого виміру. Для конкретного прикладу 4.2. практичний зміст коефіцієнта $\tilde{\beta}_2$= -0,552 повинен бути узгодженим із структурою статистичних даних. Нехай всі фактори представлені в динаміці – статистичні дані за десять послідовних місяців року. Точна інтерпретація $\tilde{\beta}_2$; виходячи із статистичних даних збільшення ціни реалізації даного продукту протягом десяти місяців року на 1 євро приводило до зменшення (знак "мінус") попиту в середньому на 552 грам. За умови, що доходи споживачів протягом аналізованого періоду були незмінними. Аналогічна інтерпретація для $\tilde{\beta}_2$, коли вибіркові дані стосуються реалізації продукції десяти підприємств ринку (просторова структура статистичних даних). Обробка за просторової структури статистичних даних засобами регресивного аналізу може підтвердити наявність закону попиту на даному ринку товарів і послуг – це знак мінус факторного коефіцієнта $\tilde{\beta}_2$. Відсутність знаку мінус означає можливу відсутність закону попиту на ринку, тобто симптом монополізації ринку, можливих регуляторних процесів ціни з боку держави і інших факторів.

Зміст факторного коефіцієнта $\tilde{\beta}_3$ доходів споживачів у випадку просторової структури даних: збільшення доходів споживачів ринку на 1 грн викликало збільшення попиту на ринку в середньому на 77 грам за умови незмінності ціни реалізації. Порівняльний аналіз реакції попиту відносно факторів ціни і доходів - для попиту чутливість вища відносно зміни ціни.

Коефіцієнти $\tilde{\beta}_2$і $\tilde{\beta}_3$ відносні показники виміру зміни результуючого фактору. Вільний розрахунковий коефіцієнт $\tilde{\beta}_1$ абсолютний показник. Його практичний зміст можна отримати із рівняння регресії - при $x_{t2} = 0$ і $x_{t3} = 0$, одержимо $\tilde{y}_t = \tilde{\beta}_1$. Це фактично прогнозне середнє значення результуючого показника, яке з'являється за нульових прогнозованих значень головних факторів. В задачі 4.2 середнє значення показника $\tilde{y}_t$ = 47,267 т виникло за нульових значень ціни реалізації платіжного попиту і доходів

споживачів, тобто це можна трактувати як середнє значення споживчого попиту на ринку даної продукції. Слід пам'ятати, що вільний розрахунковий коефіцієнт не завжди повинен мати практичний зміст. Наприклад, за матеріалами фінансової звітності підприємств побудовано парну лінійну модель регресії залежності рівня рентабельності (збитковості) (у) реалізованої продукції від фондоозброєності праці (х) на підприємствах регіону, визначено регресійне рівняння даної моделі, $\tilde{y} = 0{,}81 + 0{,}095x$. Зміст вільного коефіцієнта, який можна отримати із рівняння при *x*=0, відсутній за неможливості відсутності (нульових значень) основних фондів підприємства. Ще один приклад, за об'єму вибірки 30 фізичних осіб, за результатами побудови моделі регресії залежності рівня місячної заробітної плати ($\tilde{y}$) від стажу роботи (*x*) отримано рівняння регресії $\tilde{y} = 13168 + 648x$. Оскільки в умові задачі відсутні обмеження стосовно мінімального рівня заробітної плати, то при *x*=0, тобто відсутності стажу роботи, отримаємо з рівняння регресії середній очікуваний рівень заробітної плати в розмірі 13168 грн. Взагалі, практичний зміст вільного розрахункового коефіцієнта цілком залежить від умови задач обробки статистичних даних.

В інтерпретації коефіцієнтів $\tilde{\beta}_k$ приймались до уваги одиниці виміру ендогенних змінних. За розрахунковими коефіцієнтами не можна судити про ступінь впливу регресора на регресант. Не можна визначити, якщо така задача ставиться, в якому із двох збутових підприємств попит реагує, сильніше, наприклад на зміну ціни реалізації продукції. Для визначення ступеня впливу головних факторів на результуючий фактор застосовують середні коефіцієнти еластичності за формулою:

$$e_k = \tilde{\beta}_k \frac{\bar{X}_k}{\bar{Y}}\% \text{ (k=2,3)} \qquad (4.34)$$

де: $\tilde{\beta}_k$ - розрахунковий коефіцієнт,

$\bar{X}_k$ - середнє арифметична фактору,

$\bar{Y}$ - середнє значення регресанта.
У задачі 4.2:

$$e_2 = \tilde{\beta}_2 \frac{\bar{X}_2}{\bar{Y}} = -0{,}552 \frac{72}{25} \approx -1{,}6\%,\ e_3 = \tilde{\beta}_3 \frac{\bar{X}_3}{\bar{Y}} = 0{,}077 \frac{226}{25} \approx 0{,}7\%.$$

Зміст коефіцієнтів еластичності для часової структури статистичних даних (аналізований період – десять днів): збільшення ціни реалізації даного продукту протягом десяти днів на 1% приводило до падіння попиту в середньому на 1,6% за умови, що середньодушові доходи споживачів залишались незмінними.

У випадку просторової структури статистичних даних: зростання (падіння) ціни на ринку продукції на 1% по всіх суб'єктах господарської діяльності приводило до падіння (зростання) попиту в середньому на 1,6% за умови, що середньодушові доходи споживачів залишались незмінними.

Для виведення *формули середньої еластичності* для всіх послідовних проміжків статистичних даних за лінійною регресією потрібно визначити дане поняття з практичних міркувань. Потрібно визначити відсоткову зміну середнього значення, коли відповідний фактор збільшиться на всіх проміжках на один відсоток, яка визначається за формулою:

$$e_k = \frac{\bar{y}_k^* - \bar{y}}{\bar{y}}, \qquad (4.35)$$

де: $\bar{y}$ - середнє арифметичне результуючого показника,

$\bar{y}_k^*$ - середнє арифметичне значення результуючого показника за моделлю при збільшенні фактору на 1%.

## 4.5. Статистичні оцінки для істинних значень коефіцієнтів регресії

Методи регресійного аналізу не дозволяють знаходити значення істинних коефіцієнтів багатофакторної емпіричної моделі тільки значення розрахункових коефіцієнтів і встановити практичний зміст останніх. Застосування методів математичної статистики в регресійному аналізі дозволяють статистично оцінити значення істинних коефіцієнтів моделі і дати їм практичне тлумачення. Виявляється, що статистичною оцінкою (яка задовольняє умовам незміщеності, ефективності, обґрунтованості) для істинних коефіцієнтів моделі є розрахункові коефіцієнти. Незміщеність оцінок означає, що математичне сподівання розрахункового коефіцієнта збігається з його істинним значенням $\beta_k = M\tilde{\beta}_k$ . В класі всіх незміщених оцінок їх ефективність оцінок визначається теоремою Гауса-Маркова. Обґрунтованість оцінок за незміщеності і ефективності істиних коефіцієнтів моделі слідує із того факту, що дисперсійно-коваріаційна матриця збурень за збільшення об'єму вибірки зводиться до нульової матриці. Виходячи із обґрунтованості оцінок можна константувати, що ймовірність відхилення вектору істинних коефіцієнтів моделі від математичного сподівання вектору розрахункових коефіцієнтів моделі прямує до одиниці при зростанні кількості спостережень, тобто $\gamma_k = P(|\beta_k - \tilde{\beta}_k| < \delta_k) \to 1$ при $T \to \infty$ , $k = 1,\ldots n$ . Величину $\gamma_k$ - називають довірчою ймовірність (гарантією) оцінки відхилення, $\alpha_k = 1 - P(|\beta_k - \tilde{\beta}_k| < \delta_k)$ - рівнем значимості оцінки відхилення, $\delta_k$ - точність оцінки відхилення. Нерівність:

$$\tilde{\beta}_k - \delta_k < \beta_k < \tilde{\beta}_k + \delta_k$$

складає зміст довірчого інтервалу. Говорять, що параметр $\beta_k$ покривається інтервалом $(\tilde{\beta}_k - \delta_k; < \tilde{\beta}_k + \delta_k)$ з гарантією $\gamma_k$ % або

рівнем значимості $\alpha_k$ %. Приведемо технологію побудови параметра $\delta_k$. Структура параметра $\delta_k$ визначається як добуток середньоквадратичного відхилення $\sigma_{\tilde{\beta}_k}$ розрахункового коефіцієнта на критичну точку $t_k$ розподілу Стьюдента за стандартним рівнем значимості $\alpha_k$ = 0,05. Таким чином, довірчий інтервал має наступний вигляд

$$(\tilde{\beta}_k - \sigma_{\tilde{\beta}_k} t_k < \tilde{\beta}_k + \sigma_{\tilde{\beta}_k} t_k) \qquad (4.36)$$

В прикладі 4.2 критичну точку можна знайти за операторною системою *Excel* :

$t_k$= *СТЬЮДРАСПОБР*( $\gamma_k$ ;*T- n*) =

*СТЬЮДРАСПОБР*(0,05;7)=2,306.

Розрахунок довірчих інтервалів для факторних коефіцієнтів

$(\tilde{\beta}_2 - \sigma_{\tilde{\beta}_2} t_k; \tilde{\beta}_2 + \sigma_{\tilde{\beta}_2} t_2)$ = (-0,552-0,046×2,365; -0,552+0,046×2,365) =(-0,66;-0,443);

$(\tilde{\beta}_3 - \sigma_{\tilde{\beta}_3} t_3; \tilde{\beta}_3 + \sigma_{\tilde{\beta}_3} t_3)$ = (-0,77-0,04×2,365; -0,77-0,04×2,365)=(-0,018;0,172).

Розрахунок довірчого інтервалу для вільного коефіцієнта

$(\tilde{\beta}_1 - \sigma_{\tilde{\beta}_1} t_1; \tilde{\beta}_1 + \sigma_{\tilde{\beta}_1} t_1)$ = (47,267-11,083×2,365; 47,267+11,083×2,365)=(21,056;73,478).

Довірчі інтервали для коефіцієнтів моделі приведені в Excel в 6-тій і 7-мій колонках третьої таблиці програми “Регресія” (рис. 4.8).

Загальний практичний зміст довірчих інтервалів для факторних коефіцієнтів моделі. Загальна інтерпретація довірчого інтервалу у випадку часової структури статистичних даних: виходячи із статистичних даних , на 95% можна гарантувати, що збільшення *k*-того фактора протягом аналізованого періоду на одиницю свого виміру приводить до зміни результуючого показника в межах від $\tilde{\beta}_k - \sigma_{\tilde{\beta}_k} t$ до $\tilde{\beta}_k + \sigma_{\tilde{\beta}_k} t_k$ одиниць свого виміру.

| | *Коефіцієнти* | *Стандартна помилка* | *t-статистика* | *P-значення* | *Нижній 95%* | *Верхній 95%* |
|---|---|---|---|---|---|---|
| Y | 47,26666667 | 11,08265381 | 4,264923139 | 0,003723908 | 21,0603547 | 73,4729786 |
| Змінна X 1 | -0,552 | 0,045765051 | -12,06160565 | 6,14347E-06 | -0,6602171 5 | -0,44378285 |
| Змінна X 2 | 0,077333333 | 0,039932823 | 1,936585664 | 0,094006438 | -0,0170927 89 | 0,17175946 |

*Рисунок 4.8.* **Визначення коефіцієнтів множинної регресійної моделі та їх довірчих інтервалів в програмному середовищі Excel**

*Джерело:* розрахунки авторів

Загальна інтерпретація довірчого інтервалу у випадку просторової структури статистичних даних: виходячи із статистичних даних , на 95% можна гарантувати, що збільшення $k$-того фактору по всіх статистичних об'єктах на одиницю свого виміру приводить до зміни результуючого показника в межах від $\tilde{\beta}_1 - \sigma_{\tilde{\beta}_1} t_1$ до $\tilde{\beta}_1 + \sigma_{\tilde{\beta}_1} t_1$ одиниць виміру.

Практичний зміст довірчих інтервалів стосовно задачі 4.2, за просторовою структурою статистичних даних, об'єктами якої є філії торговельного підприємства.

Довірчий інтервал (-0,661;-0,443) має наступний зміст: виходячи із статистичних даних, на 95% можна гарантувати, що збільшення ціни реалізації по всіх філіях на 1 євро, приводить до падіння попиту в межах від 661 грам до 443 грам (не більше ніж на 661 грам і не менше ніж на 443 грам) за умови, що середньодушові доходи споживачів залишалися незмінними.

Зміст довірчого інтервалу (-0,018;0,172): виходячи із статистичних даних, на 95% можна гарантувати, що збільшення

середньодушових доходів споживачів на 1 євро , приводило, для однієї частини підприємств, до падіння попиту не більше, ніж на 18 грам, для другої частини підприємств зростання попиту не більше ніж на 172 грам, за умови, що ціна реалізації для всіх підприємств лишалась незмінною. Не дивлячись на те, що для однієї частини підприємств попит падав, для іншої зростав, в середньому попит для всіх підприємств зростав на (172+18)/2 = 77 грам, як зазначалось вище.

Зміст довірчого інтервалу (21,056; 73,478) полягає в тому, що виходячи із статистичних даних, на 95% (або з помилкою 5%) можна гарантувати коливніть споживчого попиту в межах від 21,056 тон до 73,478 тон.

Розрахункові коефіцієнти рівняння регресії, як відомо, є статистичними оцінками істинних коефіцієнтів регресії. Для прийняття рішень стосовно їх вагомості необхідно мати розмір їх похибок. Розмір їх похибок перевіряється за рівнем їх значимості. Використання формул знаходження критичних точок (4.33). дозволяє розрахувати в Excel рівні значимості коефіцієнтів регресії:

$$\alpha_1 = \textit{СТЬЮДРАСП}(4{,}265{,}7{,}2) = 0{,}003723,$$
$$\alpha_2 = \textit{СТЬЮДРАСП}(12{,}7{,}2) = 6{,}35831 \times 10^{-6},$$
$$\alpha_3 = \textit{СТЬЮДРАСП}(1{,}925{,}7{,}2) = 0{,}095625512.$$

Звідси слідує висновок: не всі коефіцієнти менші 0,05, це стосується першого і другого коефіцієнта, тому їх зміщення не спостерігається. Проте, третій коефіцієнт має зміщення. Таким чином, положення про рівність нулю математичного сподівання вектору збурення не виконується. У реальних дослідженнях ті факторні змінні, коефіцієнти яких є зміщеними, підлягають вилученню, якщо це доцільно за прийнятих рішень. Інакше, збільшенням об'єму вибірки здійснюють подальший аналіз регресії.

За потреб, стосовно методів включення і виключення факторів в регресії, програма SPSS пропонує на вибір один із наступних методів побудови регресії щодо кількості факторів.

*1. Метод включення* (його ще називають прямим методом) *Forward* – це покрокова процедура вибору факторів, за якої фактори послідовно включаються в модель. Критерієм включення є коефіцієнт частинної кореляції із результуючим показником. Спочатку включається фактор з найбільшим модульним значенням коефіцієнта, коли воно задовольняє прийняте порогове значення і т. д.

2. Метод виключення (його ще називають оберненим методом) *Backward* – це покрокова процедура, яка починається з моделі, що містить всі фактори, а потім відбувається поетапне виключення факторів з найменшим частинним коефіцієнтом кореляції до тих пір, доки відповідний коефіцієнт не виявиться незначущим.

3. Опція *Selection Variable* призначена для включення факторів у модель.

4. Опція *Case Labels* – для їх виключення, *WLS Weight* – вибір ваги факторів для проведення регресії (рис.4.9)

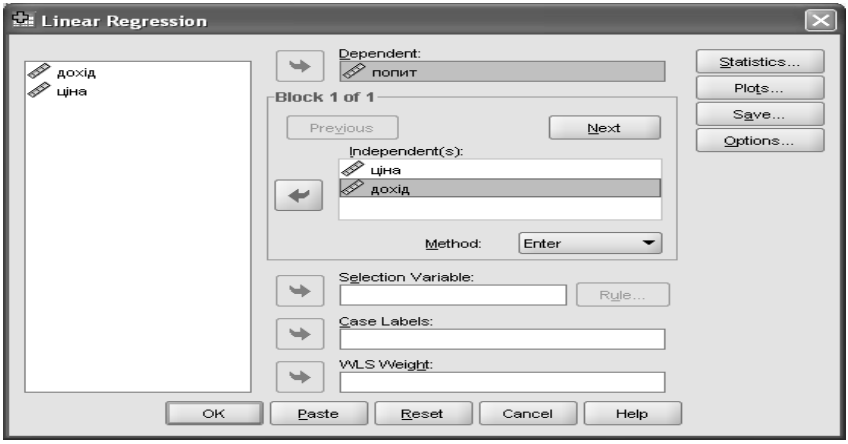



*Рисунок 4.9.* **Перевірка наявності факторів виключення в лінійній регресії в програмному середовищі SPSS**

*Джерело:* розрахунки авторів

## 4.6. Прогнозування в лінійних моделях регресії

Одне із завдань регресійного аналізу полягає в прогнозуванні результуючого фактору (показника). В соціально-економічних процесах прогноз виступає як економічна категорія – це наукове судження про стан об'єкта, системи, процесу в майбутньому і шляхи досягнення цього стану. Сам процес прогнозування пов'язаний з часом. В регресійному аналізі це поняття має більш широкий зміст. Статистичні дані можуть мати як часову так і просторову структуру. Тому в кількісному плані прогноз являє собою оцінку ендогенних змінних за деяким набором пояснюючих незалежних змінних. За способом кількісного відображення прогнози поділяються на точкові (передбачення) і інтервальні. За часової структури статистичних даних про істинні точкові прогнози (передбачення) регресанта говорять тоді, коли в часових рядах прогнозний період знаходиться після базового. За обробки просторових рядів даних істиний прогноз має місце, коли він відноситься не до елементів вибірки (статистичних об'єктів) регресанта, а до інших елементів, що не увійшли у вибірку, але які відносяться до тієї ж генеральної сукупності, з якої вибірка здійснювалась. Обчислюються точкові прогнози, виходячи із рівняння регресії заміною вектору $\left(1, x_{t2}, \mathrm{x}_{t3}\right)$ на вектор прогнозних значень $\left(1, x^*_{t2}, \mathrm{x}^*_{t3}\right)$ головних факторів. Прогнозні значення головних факторів можуть бути знайдені як на об'єктивній основі – у випадку часових рядів (шляхом побудови трендової моделі), так і на суб'єктивній основі – експертним методом. Таким чином, передбачуване значення $\tilde{y}^*$ регресанта (результуючого показника) знаходиться за формулою:

$$\tilde{y}^* = \sum_{k=1}^{3} \tilde{\beta}_k x_k^* \qquad (4.37)$$

Нехай у прикладі 4.2 статистичні дані мають часову структуру, регресант – обсяги помісячної реалізації продукції торговельним підприємством, фактори – ціна реалізації протягом місяця,

середньодушові доходи споживачів протягом місяця. Потрібно знайти прогнозний обсяг продажу в одинадцятому місяці за прогнозною ціною 110 грн за 1 кг. і прогнозованими доходами споживачів 225 грн. Прогнозне значення знаходиться за формулою:

$$\tilde{y}^* = 47,7 - 0,552 \times 110 + 0,077 \times 225 = 3,95 \text{ кг.}$$

Звідси слідує висновок: в одинадцятому місяці за ціною 110 грн за кг. і доходами споживачів 225 грн, торговельне підприємство передбачає реалізувати в середньому 3,95 кг даної продукції.

В даному випадку істинне значення реалізації продукції (точкове ), за заданими прогнозними даними факторів невідоме, відоме тільки середнє точкове значення прогнозу. Також невідоме середнє значення реалізації продукції (математичне сподівання генеральної сукупності) всіх учасників ринку, тобто невідомий середній сукупний попит. Наведені два прогнозні параметри можна статистично оцінити за прогнозним довірчим інтервалом. Структура довірчого інтервалу параметрів прогнозу має наступний вигляд: *статистична оцінка параметра прогнозу ± точність прогнозу*.

Статистичною оцінкою параметрів прогнозу є точкове значення прогнозу $\tilde{y}^*$. Якщо δ – точність прогнозу, то прогнозний інтервал має наступний вигляд: $(\tilde{y}^* - \delta; \tilde{y}^* + \delta)$. Точність прогнозу визначається як добуток критичної точки розподілу Стьюдента, знайденої за певним рівнем значимості, на похибку прогнозу, обчислену окремо для індивідуального значення регресанта і його середнього значення. Для прогнозів рівень значимості повинен складати (один відсоток) η = 0,01, або рівень довіри γ= 99%. Критична точка розподілу Стьюдента $t(\eta; T-n)$ знаходиться за рівнем значимості η і числом степенів вільності $T-n$. Похибка прогнозу для середнього значення регресанта знаходиться за формулою:

$$\tilde{\sigma}_x = \tilde{\sigma}_u \sqrt{(x^*)^T (X^T X)^{-1} x^*}, \qquad (4.38)$$

де: X – матриця регресорів,

$X^T$ - транспонована матриця,

$(x^*)^T$ - транспонований вектор прогнозних значень регресорів,

$\tilde{\sigma}_u$- вибіркове середньоквадратичне відхилення залишків моделі.

Похибка прогнозу для індивідуального значення регресанта знаходиться за формулою:

$$\tilde{\sigma}_y = \tilde{\sigma}_u\sqrt{1 + (x^*)^T(X^TX)^{-1}x^*}. \quad (4.39)$$

Довірчі прогнозні інтервали наберуть остаточного вигляду: для індивідуального значення регресанта визначаються наступним чином:

$$(\tilde{y}^* - \tilde{\sigma}_y t(\eta; T-n); \tilde{y}^* + \tilde{\sigma}_y t(\eta; T-n)) \quad (4.40)$$

А для середнього значення регресанта:

$$(\tilde{y}^* - \tilde{\sigma}_x t(\eta; T-n); \tilde{y}^* + \tilde{\sigma}_x t(\eta; T-n)) \quad (4.41)$$

Для прогнозних значень головних факторів, які не увійшли у вибірку, в деяких статистичних пакетах (наприклад. *Excel*) побудова прогнозних інтервалів для індивідуального і середнього значень регресанта відсутня. Тому довірчі прогнозні інтервали для прикладу 4.2 потрібно будувати окремо.

Критична точка *t*(0.01;7)=*СТЬЮДРАСПОБР*(0,01;7)=3,499. Вибіркове середньоквадратичне відхилення знаходиться за формулою:

$$\tilde{\sigma}_u = \sqrt{\frac{\sum_{t=1}^{T}\tilde{u}_t^2}{T-n}}. \quad (4.42)$$

Його значення *Стандартна похибка* знаходиться в таблиці Excel *"Регресійна статистика"* (рис. 4.10) $\tilde{\sigma}_u$= 1,767430203 .

Похибка прогнозу $\tilde{\sigma}_x$для середнього значення результуючого показника:

$$(x^*)^T(X^TX)^{-1}x^*=$$

$$\begin{bmatrix}1,110,225\end{bmatrix}\begin{vmatrix}39.3 & -0,115429 & -0,136762\\ -0,115429 & 0,00067 & 0,000297\\ -0,136627 & 0,00029 & 0,00051\end{vmatrix}\begin{vmatrix}1\\110\\225\end{vmatrix}=1,046095$$

$$\tilde{\sigma}_x= 1,767430203,\ \sqrt{1.046095}=1,808.$$

Точність прогнозу для середнього значення регресанта:

$$\delta =3,499\times1,808=6,326.$$

Довірчий прогнозний інтервал для середнього значення:

$$(3,95\text{-}6,33;3,95+6,33=(-2,38;10,28).$$

Похибка прогнозу для індивідуального значення результуючого показника:

$$1+(x^*)^T(X^TX)^{-1}x^*=2,046095,$$

$$\tilde{\sigma}_y=1,767430203\sqrt{2.046095}=2,527.$$

Точність прогнозу для індивідуального значення регресанта:

$$\delta =3,355\times2,527=8,84.$$

Довірчий прогнозний інтервал для індивідуального значення регресанта:

(3,95-8,84;3,95+8,84)=(-4,89;12,79).

Економічний коментар прогнозного інтервалу у випадку часової структури статистичних даних для індивідуального значення результуючого показника: виходячи із статистичних даних на 99% можна гарантувати, що для часових періодів (це може бути одинадцятий період, коли періоди представлені в хронологічному порядку), в яких ціна реалізації 110 грн за 1 кг і середньодушові доходи споживачів 225 грн, торгове підприємство передбачає реалізувати не більше 12,79 кг продукції і може мати не реалізованої продукції (лишитися в запасах) в середньому не більше 4,89 кг.

Економічний коментар прогнозного інтервалу для середнього значення результуючого показника у випадку часової структури статистичних даних: виходячи із статистичних даних на 99% можна гарантувати, що для всіх часових періодів (це можуть бути одинадцять періодів, коли періоди представлені в хронологічному порядку), в яких ціна реалізації прогнозується110 грн за 1 кг і середньодушові доходи споживачів 225 грн, торгове підприємство передбачає по всіх часових періодах реалізувати в середньому не більше 10,28 кг дано\ продукції і може мати не реалізованої продукції (лишитися в запасах) в середньому не більше 2,38 кг.

В Excel і в багатьох статистичних пакетах (наприклад, в таблиці “ВИВЕДЕННЯ ЗАЛИШКІВ” - рис. 4.10) наведено результати передбачення істинних значень реалізації продукції в часових періодах (або в динаміці), в яких прогнозні значення головних факторів (ціни і доходи споживачів) збігаються тільки з їх статистичними даними (цінами і доходами споживачів за статистикою).

| ВИВЕДЕННЯ ЗАЛИШКІВ | | | |
|---|---|---|---|
| Спостереження | Прогнозування Y | Залишки | Стандартні залишки |
| 1 | 24,09333333 | 0,906666667 | 0,581671267 |
| 2 | 30,38666667 | -0,386666667 | -0,248065688 |
| 3 | 22,88 | -2,88 | -1,847661673 |
| 4 | 26,41333333 | -1,413333333 | -0,906722858 |
| 5 | 14,6 | 0,4 | 0,256619677 |
| 6 | 8,306666667 | 1,693333333 | 1,086356632 |
| 7 | 20,89333333 | -0,893333333 | -0,573117278 |
| 8 | 32,70666667 | 2,293333333 | 1,471286147 |
| 9 | 39 | 1 | 0,641549192 |
| 10 | 30,72 | -0,72 | -0,461915418 |

*Рисунок 4.10.* **Точковий прогноз в Excel реалізації продукції для учасників ринку за вибірковими значеннями головних факторів**

*Джерело:* розрахунки авторів

Наприклад, для реалізації продукції за статистикою $x^*_{t2}$ = 25 грн за кг і доходами споживачів $x^*_{t3}$ =200 грн для інших часових періодів (для одинадцятого періоду, коли періоди утворюють динаміку показника) передбачається обсяг реалізації: $\tilde{y}^*$= 24,0933333 кг.

В програмі “Регресія” статистичного пакету “Аналіз даних” відсутнє знаходження індивідуальних прогнозних значень результуючого показника за прогнозованими значеннями факторів, крім фактичних вибіркових значень факторів (в таблиці “ВИВЕДЕННЯ ЗАЛИШКУ” в колонці “Передбачення”). В Excel відсутні також прогнозні довірчі інтервали для індивідуального значення результуючого показника і його середнього значення. В системі SPSS всі зазначені об’єкти присутні. Інформаційна технологія їх пошуку наступна. Після вводу статистичних даних (рис. 4.11) виконується наступна послідовність команд: *Analyze*

*→Regression→Linear* і відкривається діалогове вікно *Linear Regression* (лінійна регресія).

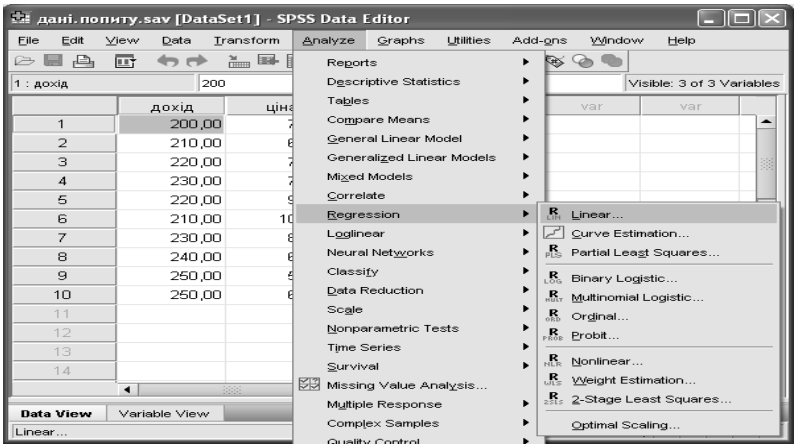



*Рисунок 4.11*. **Введення даних для побудови множинної регресійної моделі в програмному середовищі SPSS**

*Джерело:* розрахунки авторів

Далі, клавіша *Save* (зберегти) призводить до появи діалогового вікна *Linear Regression: Save* (рис. 4.12). У цьому вікні шляхом встановлення прапорця *Predicted Values: Unstandardized* (Прогнозовані значення факторів: нестандартизовані) і *Residuals: Unstandardized* (залишки: нестандартизовані) можна задати розрахунок прогнозованих (теоретичних значень) результуючого показника і залишків за прогнозованими значеннями факторів.

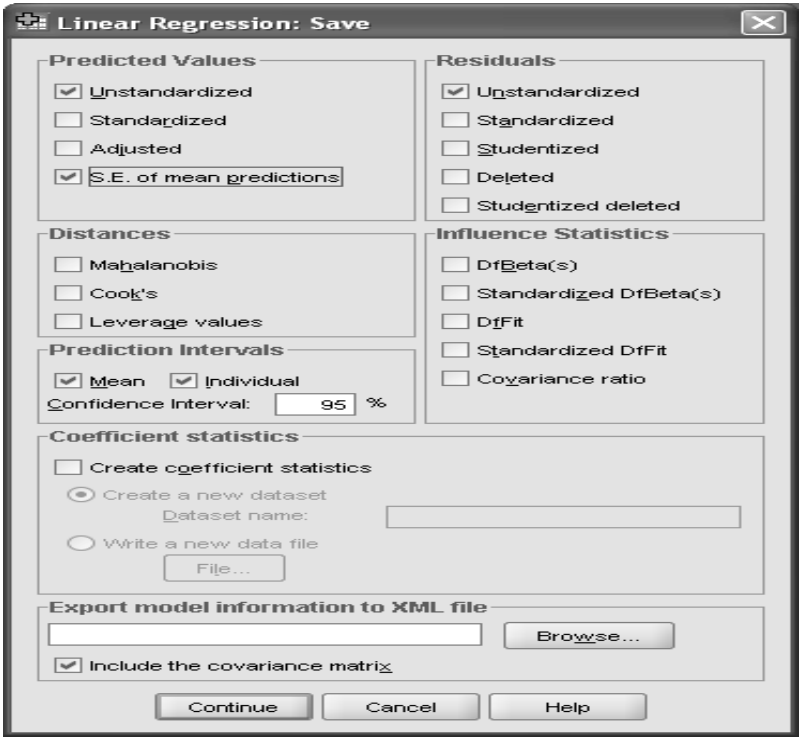



*Рисунок 4.12.* **Результати прогнозованого значення регресанта за очікуваними значеннями головних факторів в програмному середовищі SPSS**

*Джерело:* розрахунки авторів

Крім того, у групі *Productions intervals* (прогнозні інтервали) можна отримати довірчі 95% прогнозні інтервали для середнього значення (*Mean*) у генеральній сукупності, а також прогнозні значення індивідуального (*individual*) значення результуючого показника генеральної сукупності за фактичними вибірковими (не прогнозними!) значеннями факторів.

Класичний лінійний регресійний аналіз проводиться за тією схемою, приведеною вище що і нормальний регресійний аналіз.

Нормальна регресія вимагає виконання положення 3 – збурення є нормально розподіленою випадковою величиною. Невиконання положення 3 в багатьох випадках дає можливість мати закон розподілу, який буде близьким до нормального та дозволить проводити статистичну оцінку самої регресії і її параметрів.

Наступні питання регресії полягають в її узагальненні – порушенні ряду положень крім першого, третього і четвертого. Розглянуті технології регресійного аналізу можуть бути застосовані для розв'язання практичних задач побудови узагальненої регресії і для проведення статистичного аналізу регресивних моделей.

### *Список питань до самоконтролю*:

**1.** Назвіть основні завдання регресії і регресивного аналізу, чим вони зумовлені?

**2.** Охарактеризуйте основні статистичні показники регресії. Поясніть зміст та методику визначення кожного з них.

**3.** Поясніть сутність методу найменших квадратів для побудови моделей регресії.

**4.** Які види статистичних прогнозів за регресією вам відомі? Дайте характеристику кожному з них та поясніть формули обчислення.

**5.** Поясніть сутність інваріантності коефіцієнтів детермінації і множинної кореляції.

**6.** Які статистичні критерії використовуються під час дослідження регресії і? В чому полягає сутність кожного з них?

**7.** За даними Державної служби статистики України отримайте квартальні дані про валовий регіональний продукт у Вашій області за останні декілька років, а також дані трьох важливих фактори впливу на його формування. Проаналізуйте за побудованою регресією залежність продукту від кожного з факторів.

### *Список рекомендованої літератури по темі:*


1. Бахрушин В.Є. Методи аналізу даних : навчальний посібник для студентів / В.Є. Бахрушин. Запоріжжя : КПУ, 2011. 268 с.
2. Єріна А. М. Статистичне моделювання та прогнозування: навч. посібник. К.: КНЕУ, 2001. 170 с.
3. Костюк В. О. Прикладна статистика: навч. посібник. Харків: ХНУМГ ім. О. М. Бекетова, 2015. 191 с.
4. Матковський С.О., Гальків Л.І., Гринькевич О.З., Сорочак О.З. Статистика: Навч. посібник. Львів: «Новий Світ 2000», 2009. 430 с.
5. Руська Р. В. Економетрика: навчальний посібник. Тернопіль : Тайп, 2012. 224 с.
6. Лещинський О.М., Разанцева В.В. ,Юнькова О.О. Економетрія. К.: МАУП. 2003. 208 с.
7. Field, A. (2012). Discovering statistics using IBM SPSS Statistics. 4-th edition. Los Angrles/ London/ New Delphi/ Singapore/ Washingtone DC : Sage. 1467. URL: https://ua1lib.org/book/2823883/3c9a15?dsource=recommend
8. Mayers, A. (2013). Introduction to statistics and SPSS in psychology. Pearson. URL: https://ua1lib.org/book/10985783/e208a5?id=10985783&secret=e

# Розділ 5. МОДЕЛЮВАННЯ РЯДІВ ДИНАМІКИ

## 5.1. Статистичний аналіз показників розвитку явищ у часі

Вивчення суспільно-економічних явищ у розвитку та динаміці виступає одним з найважливіших завдань статистичного моделювання, котре вирішується завдяки формуванню та аналізу динамічних (часових) рядів. *Рядом динаміки називають ряд статистичних показників, які характеризують зміну суспільних явищ у часі.* Це, наприклад, чисельність населення в державі станом на певні дати, урожайність культур певного виду у сільськогосподарських підприємствах регіону за останні декілька років, чисельність працівників на фірмі станом на початок кожного місяця. Ряд динаміки складається з двох складових елементів: періодів часу (t) і рівнів (*y*).

*Рівнем ряду динаміки називають статистичний показник, який характеризує величину суспільного явища на даний момент або за певний період часу.* Рівні динамічного ряду можуть бути представлені абсолютними, відносними або середніми величинами. Вони характеризують величину показника на певний момент часу (певну дату) і за відповідний часовий проміжок - період часу *t* (день, місяць, квартал, рік). В зв'язку з цим розрізняють моментні та інтервальні ряди динаміки. Моментні часові ряди характеризують розмір досліжуваного явища на певний момент часу (наприклад, кількість працівників на підприємстві на перше число кожного місяця тощо), а інтервальні ряди - розмір явища за певний період часу (наприклад, виробництво продукції за рік)[1].

Для статистичного моделювання соціально-економічних явищ у часі, вивчення напряму та інтенсивності їх змін у певних часових періодах використовують систему абсолютних, відносних та

[1] Мармоза А.Т. Практикум з теорії статистики. –К.: Центр учбової літератури, 2013. –484 с.

середніх величин динаміки. До них належать наступні показники: абсолютний приріст, темп (коефіцієнт) зростання, темп приросту, абсолютне значення одного відсотку приросту, середній рівень ряду динаміки, середній абсолютний приріст, середній темп зростання і приросту.

Абсолютні та відносні показники динамічних рядів отримують, порівнюючи між собою вихідні рівні ряду динаміки. При цьому рівень, з яким порівнюють, називають базисним ($y_0$)**,** а порівнюваний - поточним рівнем ($y_i$). У разі порівняння кожного наступного рівня з тим рівнем, який взято за базу порівняння, визначають базисні показники, а при порівнянні кожного наступного рівня з попереднім отримують ланцюгові показники (табл. 5.1).

Абсолютний приріст ($A$) характеризує абсолютну швидкість зростання (зниження) рівнів ряду динаміки та визначається як різниця між двома рівнями, один з яких взято за базу порівняння.

*Таблиця 5.1.* **Система показників динаміки**

| Вид показника динаміки | Базисний показник | Ланцюговий показник |
|---|---|---|
| A | B | C |
| *Абсолютні показники динаміки* | | |
| 1. Абсолютний приріст | $A_{бі}=y_i-y_0$ | $A_{лі}=y_i-y_{i-1}$ |
| *Відносні показники динаміки* | | |
| 1. Темп (коефіцієнт) зростання | $K_{бі}=\frac{y_i}{y_0}$ | $K_{лі}=\frac{y_i}{y_{i-1}}$ |
| 2. Темп приросту | $T_{бі}=\frac{A_{бі}}{y_0}$ | $T_{лі}=\frac{A_{лі}}{y_{i-1}}$ |
| 3. Абсолютне значення одного відсотку приросту | – | $П_{лі}=\frac{A_{лі}}{T_i}*100\%$ |
| 4. Пункти зростання | – | $ПЗР_{лі}=100\%(T_{лі}-T_{лі-1})$ |

Продовження таблиці 5.1

| A | B | C |
|---|---|---|
| *Середні показники рядів динаміки* | | |
| 1. Середній рівень ряду динаміки | Для інтервальних рядів динаміки з рівновіддаленими один від одного рівнями: $\bar{y} = \frac{\sum y}{n}$; Для інтервальних рядів динаміки з нерівновіддаленими рівнями: $\bar{y} = \frac{\sum yt}{\sum t}$; Для моментних рядів динаміки: $\bar{y}_{xp} = \frac{1/2y_1 + y_2 + ... + 1/2y_n}{n-1}$. | |
| 2. Середній абсолютний приріст | $\bar{A} = \frac{\sum A_{ли}}{n}$ *або* $\bar{A} = \frac{y_n - y_0}{n-1}$ | |
| 3. Середній темп (коефіцієнт) зростання | $\bar{K} = \sqrt[n-1]{K_{л1} * K_{л2} * ... * K_{л\,n-1}}$ або $\bar{K} = \sqrt[n-1]{\frac{y_n}{y_0}}$ | |
| 4. Середній темп приросту | $\bar{T} = (\bar{K} * 100\%) - 100\%$. | |

*Джерело*: розробки авторів

Темп (коефіцієнт) зростання (К) характеризує відносну швидкість зміни явища, на практиці визначається як відношення двох рівнів, один з яких взято за базу порівняння. Він визначає, у скільки разів кожен наступний рівень більший або менший за рівень, який було взято за базу порівняння.

Швидкість зміни явища у часі можна також охарактеризувати за допомогою обчислення темпів приросту (*Т*)**,** які є відношенням абсолютного приросту до рівня, взятого за базу порівняння. Темп приросту, як і абсолютний приріст, може бути як додатним у випадку зростання величини явища, так і від'ємним числом, відповідно при зниженні рівня, математично він визначається у вигляді коефіцієнтів або відсотків. На практиці для аналізу динаміки соціально-економічних явищ темпи приросту часто виражають у формі

процентів. В такому випадку вони характеризують, на скільки відсотків збільшився або зменшився досліджуваний рівень порівняно з базисним, який взято за 100%.

Під час аналізу динамічних процесів важлива роль належить обчисленню абсолютного значення одного відсотку приросту (*П*), котрий дає змогу визначити вагомість визначеного відсотка приросту. Його обчислюють як відношення абсолютного приросту до відповідного відсотку приросту. Зауважимо, що розрахунок даного показника має зміст тільки на ланцюговій основі.

Слід відмітити, що в статистичному аналізі динамічних рядах відсотки зростання і приросту інколи порівнюють шляхом визначення різниці рівнів, що називається пунктами зростання (ПЗР). Їх обчислюють як різницю базисних відсотків зростання або приросту двох суміжних періодів.

Для узагальненої характеристики динамічного розвитку суспільно-економічних явищ, обчислюють середні показники рядів динаміки. *Середню з n-рівнів динамічного ряду називають хронологічною середньою або середнім рівнем динаміки* ($\bar{y}$). Методологія обчислення середнього рівня для інтервальних і моментних рядів динаміки істотно відрізняється. Зокрема, у випадку інтервального ряду з рівновіддаленими рівнями середнє значення рівня обчислюють за формулою середньої арифметичної простої. Якщо інтервальний ряд динаміки складається з нерівновіддалених один від одного рівнів, тоді середній рівень визначається за формулою середньої арифметичної зваженої. Для моментного рядах з однаковими проміжками між датами середній рівень обчислюється за формулою середньої хронологічної.

Для формування висновків щодо основної тенденції розвитку явища у часі важливе місце належить визначенню середнього абсолютного приросту та темпу приросту (зниження) величини явища. Середній абсолютний приріст ($\bar{A}$) характеризує середню швидкість) зміни рівня. Середній темп (коефіцієнт) зростання ($\bar{K}$)

характеризує, у скільки разів у середньому кожен даний рівень ряду більший (або менший) від попереднього рівня. Середній темп приросту $(\overline{T})$ визначають за даними середнього темпу зростання, даний показник відображає, на скільки процентів у середньому збільшується (або зменшується величина явища в розрахунку на 1 часовий проміжок.

*Задача 5.1.* За даними про вартість основних засобів на промислових підприємствах регіону у 2015-2020 рр. визначити та проаналізувати систему абсолютних, відносних та середніх показників динаміки (табл. 5.2). За отриманими результатами обчислень зробити висновки. Для розв'язання поставленої задачі складемо розрахункову таблицю для визначення абсолютних (базисних і ланцюгових) та відносних показників динамічного ряду.

*Таблиця 5.2*. **Показники динаміки вартості основних засобів на промислових підприємствах регіону у 2015-2020 рр.**

| Рік | Вартість основних виробничих засобів, млн. грн | Абсолютний приріст | | Темп (коефіцієнт) зростання | | Темп приросту, % | | Абсолютне значення 1% приросту, млн. грн |
|---|---|---|---|---|---|---|---|---|
| | | базисний | ланцюговий | базисний | ланцюговий | базисний | ланцюговий | |
| 2015 | 33228 | – | – | – | – | – | – | – |
| 2016 | 34285 | 1057 | 1057 | 1,0318 | 1,0318 | 3,18 | 3,18 | 332,3 |
| 2017 | 35387 | 2159 | 1102 | 1,0650 | 1,0321 | 6,50 | 3,21 | 342,8 |
| 2018 | 35681 | 2453 | 294 | 1,0738 | 1,0083 | 7,38 | 0,83 | 353,9 |
| 2019 | 35974 | 2746 | 293 | 1,0826 | 1,0082 | 8,26 | 0,82 | 356,8 |
| 2020 | 36922 | 3694 | 948 | 1,1112 | 1,0264 | 11,12 | 2,64 | 359,7 |

*Джерело*: розрахунки авторів

Визначимо абсолютні прирости вартості основних виробничих засобів:

базисні: $А_{б1}=y_1-y_0=34285-33228=1057$ млн. грн;

$А_{б2}=y_2-yo=35387-33228=2159$ млн. грн;

$А_{б3}=у_3-уо=35681-33228=2453$ млн. грн;

ланцюгові: $Ал_1=у_1-у_0=34285-33228=1057$ млн. грн;

$Ал_2=у_2-у_1=35387-34285=1102$ млн. грн;

$Ал_3=у_3-у_2=35681-35387=294$ млн. грн і т.д.

На наступному кроці – визначення низки відносних показників ряду динаміки: коефіцієнтів зростання, темпів приросту, абсолютних значень 1% приросту.

Обчислимо темпи (коефіцієнти) зростання вартості основних засобів:

базисні: $К_{б1}=у_1:у_0=34285:33228=1,0318$;

$К_{б2}=у_2:у_0=35387:33228=1,0650$;

$К_{б3}=у_3:у_0=35681:33228=1,0738$ і т.д.;

ланцюгові: $Кл_1=у_1:у_0=34285:33228=1,0318$;

$Кл_2=у_2:у_1=35387:34285=1,0321$;

$Кл_3=у_3:у_2=35681:35387=1,0083$ і т.д.

Визначимо темпи приросту вартості основних засобів у %:

базисні: $Т_{б1}=(А_{1б}:у_0)\times100\%=(1057:33228)\times100\%=3,18\%$;

$Т_{б2}=(А_{2б}:уо)\times100\%=(2159:33228)\times100\%=6,50\%$;

$Т_{б3}=(А_{3б}:уо)\times100\%=(2453:33228)\times100\%=7,38\%$ і т.д.;

ланцюгові: $Тл_1=(А_{1л}:у_0)\times100\%=(1057:33228)\times100\%=3,18\%$;

$Тл_2=(А_{2л}:у_1)\times100\%=(1102\text{-}34285)\times100\%=3,21\%$;

$Тл_3=(А_{3л}:у_2)\times100\%=(294\text{-}35387)\times100\%=0,83\%$ і т.д.

Знайдемо абсолютне значення одного процента приросту як відношення ланцюгових абсолютних приростів до ланцюгових процентів приросту:

$Пл_1=А_{л1}:Т_{л1}=1057:3,18=332,3$ млн. грн;

$Пл_2=А_{л2}:Т_{л2}=1102:3,21=342,8$ млн. грн;

$Пл_3=А_{л3}:Т_{л3}=294:0,83=353,9$ млн. грн і т. д.

Заключним етапом дослідження є визначення низки середніх показників ряду динаміки вартості основних засобів промислових підприємств регіону - середнього рівня, середнього абсолютного приросту, середнього коефіцієнту зростання та середнього темпу приросту.

Середній рівень динамічного ряду вартості основних засобів слід шукати за формулою середньої арифметичної простої, оскільки даний ряд динаміки є інтервальним з рівновіддаленими проміжками:

$$\bar{y} = \frac{\sum y}{n} = (33228 + 34285 + 35387 + 35681 + 35974 + 36922):6 =$$

$$=211477:6 = 35246 \text{ млн. грн.}$$

Середній абсолютний приріст вартості основних засобів буде таким:

$$\bar{A} = (y_n - y_0)/(n-1) = (36922 - 33228) : 5 = 3594 : 5 = 739 \text{ млн. грн.}$$

Визначимо середній коефіцієнт зростання вартості основних виробничих засобів за 2007-2012 рр.:

$$\bar{K} = \sqrt[n-1]{K_{л1} \cdot K_{л2} \cdot ... \cdot K_{л\,n-1}},$$

де: *Ki* — ланцюгові коефіцієнти зростання; *n* - число рівнів ряду динаміки.

В нашому випадку коефіцієнт зростання буде таким:

$$\bar{K} = \sqrt[5]{1{,}0318 \cdot 1{,}0321 \cdot 1{,}0083 \cdot 1{,}0082 \cdot 1{,}0264} = \sqrt[5]{1{,}1112} = 1{,}021.$$

Отже, середній коефіцієнт зростання вартості основних виробничих засобів становить 1,021або 102,1%.

Визначимо середній темп приросту ($\bar{T}$):

$$\bar{T} = (\bar{K} * 100\%) - 100\% = 1{,}021 * 100\% - 100\% = 2{,}1\%.$$

У середньому кожен рівень порівняно з попереднім збільшується на 2,1%. Таким чином, середньорічна вартість основних виробничих засобів на промислових підприємствах регіону у 2015-2020 рр. становить 35246 млн. грн. Щороку вона зростала в середньому на 739 млн. грн. У цілому даний показник за досліджуваний період (2015-2020 рр.) на даних підприємствах збільшився від 33228 до 36922 млн. грн, тобто на 3694 млн. грн, або на 11,12%. За середнім коефіцієнтом зростання можна встановити, що середній щорічний темп зростання показника становить 2,1% (102,1% - 100%). Із зростанням вартості основних засобів

збільшувалося абсолютне значення одного процента приросту в регіоні з 332,3 млн. грн у 2015 р. до 359,7 млн. грн у 2020 р.

Вивчення показників зміни суспільно-економічного явища у динаміці під час розв'язання прикладних статистичних задач дає змогу дослідити характер цих змін та виявити закономірність їх розвитку.

## 5.2. Прогнозування явищ у часі

*Одним з основних завдань аналізу рядів динаміки є виявлення тенденції розвитку соціально-економічних явищ та прогнозуванню їх на перспективу, що називається вирівнюванням динамічних рядів.*

Часто, щоб виявити тенденцію в рядах динаміки, не досить одного візуального аналізу ряду, якщо його рівні через будь-які об'єктивні або випадкові причини істотно коливаються, зростаючи та знижуючись. В таких випадках потрібно вдаватися до спеціальних прийомів обробки динамічних рядів. До таких прийомів належать укрупнення періодів, вирівнювання ряду динаміки способом ковзної середньої, вирівнювання ряду динаміки за середнім абсолютним приростом, середнім коефіцієнтом зростання і аналітичне вирівнювання за методом найменших квадратів (МНК) (рис. 5.1).

Більшість кількісних методів прогнозування базується на використанні історичної інформації, представленої у вигляді часових рядів, тобто рядів динаміки, які впорядковуються за часовою ознакою. Головна ідея аналізу часових рядів полягає у побудові тренду на основі минулих даних і наступному екстраполюванні цієї лінії у майбутнє. При цьому використовуються складні математичні процедури для отримання точного значення трендової лінії, визначення будь-яких сезонних або циклічних коливань. Для здійснення розрахунків, пов'язаних з аналізом часових рядів, звичайно використовуються спеціальні комп'ютерні програми. Перевага цього методу полягає у тому, що він базується на чомусь іншому, ніж думка експерта, а саме на цифрових даних. Аналіз часових рядів доцільно використовувати тоді, коли в наявності є

достатній обсяг «історичної» інформації, а зовнішнє середовище досить стабільне.

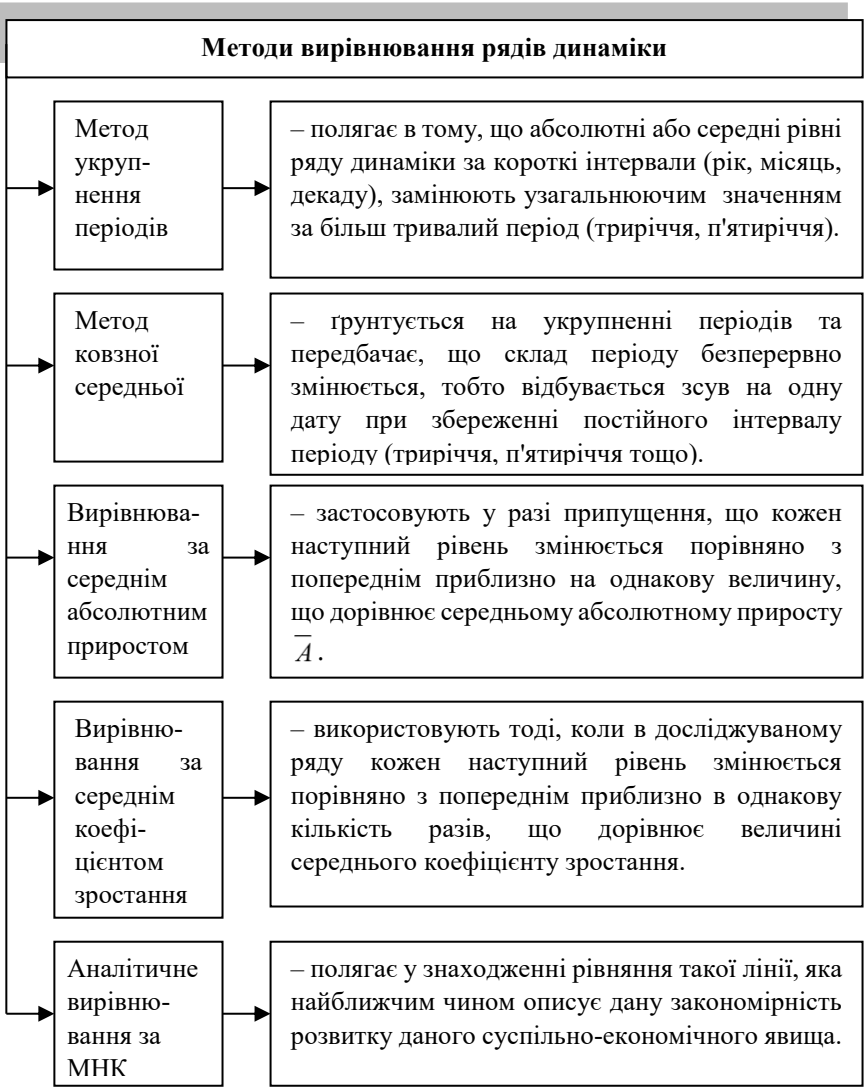



*Рисунок 5.1.* **Методи вирівнювання рядів динаміки**

*Джерело*: розробки авторів

Недоліком екстраполяції можна вважати те, що головне припущення, яке приймається при застосуванні аналізу часових рядів, може бути помилковим - майбутнє насправді може бути несхожим на минуле. До кількісних методів прогнозування належать дві великі підгрупи методів: екстраполяції і моделювання. Методи екстраполяції – це прийоми найменших квадратів, рухомих середніх, експоненційного згладжування. До методів моделювання належать прийоми структурного, сітьового і матричного моделювання.

Під час формування прогнозів з допомогою екстраполяції звичайно спираються на статистично обґрунтовані тенденції зміни тих чи інших кількісних характеристик об'єкта. Екстраполяційні методи є одними з найбільш розповсюджених і розроблених серед усіх способів прогнозування суспільно-економічних явищ.

Як видно з рисунку 5.1, три останні методи передбачають побудову математичного рівняння (функції, що апроксимується), за яким і визначають вирівняні рівні динамічного ряду. *Рівняння, що виражає розрахункові значення рівнів ряду динаміки як деяку функцію часу t, називають трендом.*

Найбільш досконалим і точним прийомом вирівнювання рядів динаміки, який враховує всі рівні вихідного ряду, є аналітичне вирівнювання за методом найменших квадратів**.** Воно ґрунтується на припущенні, що зміни досліджуваного ряду динаміки можна наближено виразити певним математичним рівнянням. Іншими словами, рівні ряду динаміки розглядають як функцію часу, де $\tilde{Y}_t = y\ (t)$ - рівні динамічного ряду, визначені за відповідним рівнянням на момент часу *t*.

Найпростішою та найбільш поширеною формою аналітичного вирівнювання є пряма лінія, рівняння якої визначають за формулою:

$$\tilde{y}_t = a_0 + a_1 t\,, \qquad (5.1)$$

де: $\tilde{y}_t$ - вирівняні значення рівнів динамічного ряду; $t$ – порядкові номери періодів*;* $a_o$ і $a_1$ параметри рівняння шуканої прямої: $a_o$

називають початком відліку, який економічного змісту не має, $a_1$-коефіцієнт регресії або пропорційності, який показує середній щорічний приріст (зниження) явища, яке вивчають.

Параметри $a_0$ і $a_1$ шуканої прямої за методом найменших квадратів знаходять, розв'язуючи таку систему рівнянь:

$$\begin{cases} \sum y = a_0 n + a_1 \sum t \\ \sum yt = a_0 \sum t + a_1 \sum t^2 \end{cases}, \qquad (5.2)$$

де $n$ - число рівнів ряду динаміки.

Поряд з цим при розв'язанні практичних задач статистичного моделювання широко використовуються й нелінійні динамічні моделі: поліноми 2-го, 3-го та вищих порядків, експоненціальні, логарифмічні криві та ін., які більш детально розглядаються в розділі 7.

*Задача 5.2.* За даними про продуктивність праці на підприємстві (табл. 5.3) провести аналітичне вирівнювання показника, використовуючи рівняння прямої лінії.

Складемо таблицю вихідних та розрахункових даних. За результатами обчислень отримаємо:

$$\begin{cases} 2810 = 10a_0 + 55a_1; \\ 16290 = 55a_0 + 385a_1. \end{cases}$$

Розв'язавши дану систему рівнянь, знайдемо:

$$a_1 = 10{,}12 \text{ тис. грн.}$$

Знаючи значення параметра $a_1$ , знайдемо параметр $a_0$, звідки:

$$a_0 = 225{,}3.$$

Тому аналітичне вирівнювання динаміки показника продуктивності праці на підприємстві у рівняння прямої лінії має вигляд:

$$\tilde{y}_t = 225{,}3 + 10{,}12 \cdot t.$$

*Таблиця 5.3.* **Вихідні та розрахункові дані для визначення параметрів лінії тренду показника продуктивності праці**

| Вихідні дані | | | Розрахункові дані | | |
|---|---|---|---|---|---|
| Рік | Продуктивність праці на 1 працівника, тис. грн, *y* | Порядковий номер року, *t* | $t^2$ | *yt* | Очікувані значення $\tilde{y}_t$ |
| 2011 | 230 | 1 | 1 | 230 | 235,42 |
| 2012 | 250 | 2 | 4 | 500 | 245,54 |
| 2013 | 240 | 3 | 9 | 720 | 255,66 |
| 2014 | 280 | 4 | 16 | 1120 | 265,78 |
| 2015 | 300 | 5 | 25 | 1500 | 275,90 |
| 2016 | 290 | 6 | 36 | 1740 | 286,02 |
| 2017 | 280 | 7 | 49 | 1960 | 296,14 |
| 2018 | 290 | 8 | 64 | 2320 | 306,26 |
| 2019 | 300 | 9 | 81 | 2700 | 316,38 |
| 2020 | 350 | 10 | 100 | 3500 | 326,50 |
| *Разом* | 2810 | 55 | 385 | 16290 | 2809,60 |

*Джерело*: розрахунки авторів

Параметр $a_1$ показує, що збільшення показника продуктивності праці на даному підприємстві в середньому збільшується щороку на 10,12 тис. грн.

Підставляючи значення t у рівняння даної лінії, отримаємо повний теоретичний ряд динаміки $\tilde{y}_t$, вирівняний за рівнянням прямої лінії.

Наприклад, для 2011 р., коли t=1:

$$\tilde{y}_{2011} = 225{,}3 + 10{,}12 \cdot 1 = 235{,}42 \ тис.\ грн.$$

Зокрема, для 2012 р. і коли t=2:

$$\tilde{y}_{2012} = 225{,}3 + 10{,}12 \cdot 2 = 245{,}54 \ тис.\ грн.$$

І т. д. для кожного наступного рядка таблиці.

Правильність розрахунків перевіряється сумами:

$$\sum y = \sum \cdot \tilde{y}_t .$$

В теорії статистичного прогнозування виділяють різні класифікації прогнозів, але мета прогнозування залишається спільною – визначення величини показника на перспективу, враховуючи його значення у попередніх часових проміжках [2]. Найбільш популярною є класифікація статистичних прогнозів в залежності від довжини періоду прогнозу, коли розрізняють такі види прогнозів:

- короткострокові (до 5 років);
- середньострокові (від 5 до 10 років);
- довгострокові прогнозування (понад 10 років).

При цьому не тільки будують лінію регресії – так звану зону прогнозу, але й визначають довірчі інтервали для лінії регресії та прогнозного значення, які часто називають оптимістичним і песимістичним прогнозом (сценарієм), визначають еластичність лінії регресії, застосовують низку статистичних критеріїв: Фішера, Стьюдента, Дарбіна-Уотсона, Акаике, Шварца, Хеннана-Куінна та ін.

Обчислені прогнозні значення показника вважають точковим прогнозом, а побудовані довірчі інтервали, де із заданою досить високою ймовірністю має потрапити досліджуваний показник, називають інтервальною оцінкою прогнозу. Зазначимо, що високо достовірним вважається таке прогнозування, де зона прогнозу становить не більше третини попереднього часового проміжку. Для $\tilde{y}_{пр}$, задавшись рівнем значущості $\alpha$, довірчий інтервал, який також в теорії статистики називають оптимістичним та песимістичним прогнозом, формула якого має вигляд:

$$\tilde{y}_{пр} - t_{кр} S < \tilde{y}_{пр} < \tilde{y}_{пр} + t_{кр} S, \qquad (5.3)$$

[2] Єріна А. М. Статистичне моделювання та прогнозування: Навч. посібник. — К.: КНЕУ, 2001. — 170 с.

де: $t_{кр}$- критичне значення $t$ – статистики Стьюдента, яке знаходять за допомогою статистичних таблиць t-розподілу Стьюдента за заданим рівнем ймовірності;

S – виправлена середня квадратична похибка прогнозу, що визначається за формулою:

$$S = \sqrt{1 + \frac{1}{n} + \frac{\left(t_{пр} - \bar{t}\right)^2}{\sum_{i=1}^{n}\left(t_i - \bar{t}\right)^2}}. \qquad (5.4)$$

В практичних задачах часто виникає потреба оцінити варіабельність прогнозованого показника навколо обчисленої лінії тренду. Для цього необхідно оцінити ступінь наближення побудованого лінійного тренду до фактичних даних динамічного ряду визначенням залишкового середнього квадратичного відхилення та коефіцієнта варіації за формулами:

$$\sigma_{зал} = \sqrt{\frac{\sum (y_i - \tilde{y}_i)^2}{n}}; \quad . \; V = \frac{\sigma_{зал}}{\bar{y}} \times 100\%, \qquad (5.5\text{-}5.6)$$

де: n – кількість періодів часу у ряді, в нашому випадку кількість років.

*Задача 5.3.* За даними аналітичного вирівнювання ряду динаміки показника продуктивності праці на підприємстві у рівняння прямої лінії виконаємо прогнозування даного показника на наступні 3 роки. Оцінимо точність відображення тенденції зміни явища за допомогою обчислення коефіцієнта варіації.

В попередній задачі було виконано аналітичне вирівнювання динаміки показника продуктивності праці для підприємства регіону у рівняння прямої лінії:

$$\tilde{y}_t = 225{,}3 + 10{,}12 \cdot t.$$

Підставивши значення t у рівняння лінії, отримаємо прогнозні значення на наступні 3 роки:

на 2021 рік (t=11): $\tilde{y}_t = 225{,}3 + 10{,}12 \cdot 11 = 235{,}42$ тис. грн;

на 2022 рік (t=12): $\tilde{y}_t = 225{,}3 + 10{,}12 \cdot 12 = 245{,}54$ тис. грн;

на 2023 рік (t=13): $\tilde{y}_t = 225{,}3 + 10{,}12 \cdot 13 = 255{,}66$ тис. грн.

Знайдену лінію тренду показано на рис. 5.2.

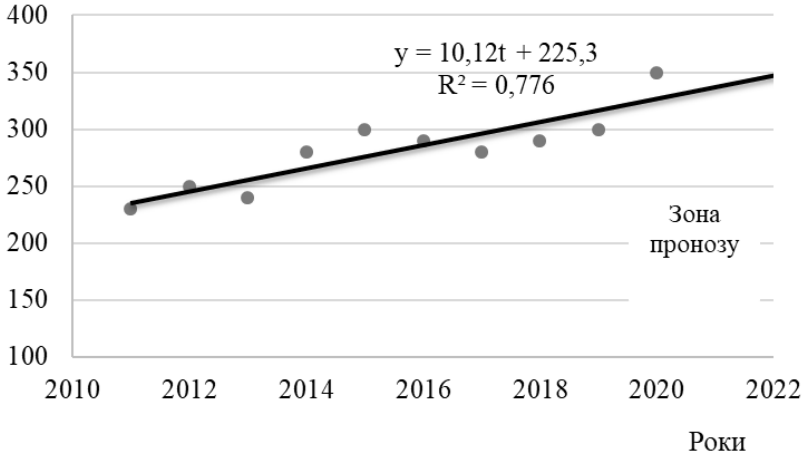



*Рисунок 5.2.* **Прогнозування показника продуктивності праці (тис. грн) на підприємстві за методом НМК**

*Джерело*: розрахунки авторів

З використанням програмних можливостей персонального комп'ютера в Excel можна значно спростити порядок побудови даного графіку. Для цього слід скористатися в меню «Вставка» функцією «Графік», виконати побудову рисунка, а далі – піднести курсор миші до будь-якої точки на графічному полі, натиснути її праву клавішу і вибрати команду: «Добавити лінію тренду». Серед запропонованих типів ліній в нашій задачі доцільно вибрати пряму лінію (хоча може бути і інший тип лінії – парабола, крива лінія n-порядку, логарифмічна тощо). Також потрібно зазначити про виконання комп'ютером прогнозу на 3 наступних часових періоди, визначення математичного рівняння лінії тренду та коефіцієнта

апроксимації $R^2$. Відмітимо, що за графіком та рівнянням лінійного тренду прогноз показника продуктивності праці має збігатися.

Виконаємо характеристику точності відображення тенденції зміни показника продуктивності праці для підприємства регіону, оцінюючи міру наближення побудованого лінійного тренду до фактичних даних динамічного ряду та визначаючи коефіцієнт варіації.

Складемо таблицю 5.4 вихідних та розрахункових даних.

*Таблиця 5.4.* **Вихідні та розрахункові дані для характеристики точності лінії тренду за ступенем варіювання показника**

| Рік | Продуктивність праці на 1 працівника, тис. грн | | Розрахункові дані | |
|---|---|---|---|---|
| | Вихідні дані, y | Вирівняні значення, $\tilde{y}_t$ | $y-\tilde{y}_t$ | $(y-\tilde{y}_t)^2$ |
| 2011 | 230 | 235,42 | -5,42 | 29,38 |
| 2012 | 250 | 245,54 | 4,46 | 19,89 |
| 2013 | 240 | 255,66 | -15,66 | 245,24 |
| 2014 | 280 | 265,78 | 14,22 | 202,21 |
| 2015 | 300 | 275,90 | 24,10 | 580,81 |
| 2016 | 290 | 286,02 | 3,98 | 15,84 |
| 2017 | 280 | 296,14 | -16,14 | 260,50 |
| 2018 | 290 | 306,26 | -16,26 | 264,39 |
| 2019 | 300 | 316,38 | -16,38 | 268,30 |
| 2020 | 350 | 326,50 | 23,50 | 552,25 |
| Разом | 2810 | 2809,6 | - | 2438,80 |

*Джерело*: розрахунки авторів

Залишкове середнє квадратичне відхилення і ступінь варіювання показника продуктивності праці для даного підприємства буде становити:

$$\sigma_{зал} = \sqrt{\frac{2438{,}80}{10}} = 15{,}62 \; тис.грн;$$

Щоб обчислити коефіцієнт варіації даного показника, знайдемо спочатку його середнє значення за формулою середньої арифметичної простої:

$$\bar{y} = \frac{\sum y_i}{n} = \frac{2810}{10} = 281 \, тис. \, грн.$$

$$V = \frac{15{,}62}{281} \times 100\% = 5{,}6\%$$

Таким чином, коливання фактичного показника продуктивності праці навколо знайденої лінії тренду в середньому становить 15,62 тис. грн, або 5,6%, що значно менше критичної величини 33%.

Такий невеликий коефіцієнт варіації показує на те, що рівняння прямої лінії досить точно відображає загальну тенденцію зміни продуктивності праці в часі. Водночас, аналіз динамічного ряду показника продуктивності праці для підприємств регіону засвідчує, що незважаючи на суттєве коливання показника по роках, чітко простежується стійка тенденція його підвищення в останні роки.

*Задача 5.4.* За даними аналітичного вирівнювання ряду динаміки показника продуктивності праці на підприємстві регіону у рівняння прямої лінії та розрахованими прогнозними значеннями даного показника на наступні 3 роки визначимо його оптимістичний та песимістичний прогноз. Оцінимо довірчий інтервал відображення тенденції зміни явища за допомогою побудови графіка.

В попередніх задачах було визначено математичне рівняння вирівнювання ряду динаміки показника продуктивності праці для підприємства регіону:

$$\tilde{y}_t = 225{,}3 + 10{,}12 \cdot t \, .$$

Його прогнозні значення на наступні 3 роки становлять: на 2021 р. – 35,42 тис. грн, на 2022 р. – 45,54 тис. грн, на 2023 р.- 55,66 тис. грн.

Для $\tilde{y}_{пр}$, задавшись рівнем значущості $\alpha$, виконаємо інтервальну оцінку досліджуваного показника, називаючи нижню межу інтервалу оптимістичним, а верхню його межу – відповідно, песимістичним прогнозом.

Для вирішення поставленого завдання складемо таблицю 5.5 для обчислення квадратичних відхилень по кожному часовому проміжку.

Зауважимо, що середнє значення порядкового номеру кожного для ряду динаміки визначимо за формулою середньої арифметичної простої:

$$\bar{t} = \frac{\sum t_i}{n} = \frac{55}{10} = 5{,}5 \, років$$

Таким чином, сума усіх квадратичних відхилень почасових проміжках даного ряду динаміки продуктивності праці на даному підприємстві становить:

$$\sum (t_i - \bar{t})^2 = 82{,}5$$

*Таблиця 5.5.* **Вихідні та розрахункові дані для визначення квадратичних відхилень часових проміжків**

| Рік | Порядковий номер року | Продуктивність праці , тис. грн | Розрахункові дані | |
|---|---|---|---|---|
| | $t$ | $\tilde{y}$ | $t_i - \bar{t}$ | $(t_i - \bar{t})^2$ |
| 2011 | 1 | 230 | -4,5 | 20,25 |
| 2012 | 2 | 250 | -3,5 | 12,25 |
| 2013 | 3 | 240 | -2,5 | 6,25 |
| 2014 | 4 | 280 | -1,5 | 2,25 |
| 2015 | 5 | 300 | -0,5 | 0,25 |
| 2016 | 6 | 290 | 0,5 | 0,25 |
| 2017 | 7 | 280 | 1,5 | 2,25 |
| 2018 | 8 | 290 | 2,5 | 6,25 |
| 2019 | 9 | 300 | 3,5 | 12,25 |
| 2020 | 10 | 350 | 4,5 | 20,25 |
| Разом | 55 | 2810 | - | 82,5 |

*Джерело*: розрахунки авторів

На наступному етапі складемо таблицю 5.6 вихідних і розрахункових даних для визначення виправленої середньої квадратичної похибки S.

Наприклад, для $t_{пр} = 11$:

$$S_{2021р.} = \sqrt{1 + \frac{1}{10} + \frac{30{,}25}{82{,}5}} = \sqrt{1{,}47} = 1{,}21.$$

І т. д. визначимо виправлені середні квадратичні похибки прогнозу для всіх наступних років.

*Таблиця 5.6.* **Вихідні та розрахункові дані для визначення виправленої середньої квадратичної похибки прогнозу**

| Рік | Порядковий номер року | Прогнозні значення продуктивності праці, тис. грн | Розрахункові дані | | |
|---|---|---|---|---|---|
| | $t_{пр}$ | $\tilde{y}_{пр}$ | $t_{пр}-\bar{t}$ | $(t_{пр}-\bar{t})^2$ | S |
| 2021 | 11 | 35,42 | 5,5 | 30,25 | 1,21 |
| 2022 | 12 | 45,54 | 6,5 | 42,25 | 1,27 |
| 2023 | 13 | 55,66 | 7,5 | 56,25 | 1,33 |

*Джерело*: розрахунки авторів

Важлива роль у даному методі відводиться встановленню надійної ймовірності, тобто такого гранично допустимого рівня, коли визначені прогнозні значення будуть вважатися високо достовірними. В більшості економічних завдань приймають надійну ймовірність p=0,95 або 95%. Це означає, що з ймовірністю 95% визначені прогнозні значення даного показника є високо достовірними та очікуваними. Тоді рівень значущості буде:

$$\alpha = 1 - p = 1 - 0{,}95 = 0{,}05 \; або \; 5\%.$$

Це означає, що з ймовірністю 5% очікувані прогнозні значення можуть вийти за межі надійного інтервалу. За таблицею t-статистики Стьюдента знайдемо критичне значення із заданим рівнем значущості 0,05 та при кількості часових проміжків динамічного ряду n=10 років.

$$t_{кр}(\alpha = 0{,}05; n = 10) = 2{,}23.$$

Тепер є всі передумови для визначення оптимістичного і песимістичного прогнозу даного показника на наступні три роки (табл. 5.7). Наприклад,

$\tilde{y}_{пес\ пр}(2021р.) = \tilde{y}_{пр} - t_{кр}S = 35{,}42 - 2{,}23 \times 1{,}21 = 32{,}72 тис.грн;$

$\tilde{y}_{опт\ пр}(2021р.) = \tilde{y}_{пр} + t_{кр}S = 35{,}42 + 2{,}23 \times 1{,}21 = 38{,}12 тис.грн;$

Аналогічно визначають значення оптимістичного і песимістичного прогнозів для даного показника на наступні 2022 та 2023 рр. Межі прогнозування даного показника показано на рис. 5.3.

*Таблиця 5.7.* **Визначення оптимістичного та песимістичного прогнозів показника продуктивності праці**

| Рік | Порядковий номер року | Прогнозні значення продуктивності праці, тис. грн | Розрахункові дані | | |
|---|---|---|---|---|---|
| | $t_{пр}$ | $\tilde{y}_{пр}$ | S | песимістичний прогноз, тис. грн | оптимістичний прогноз, тис. грн |
| 2021 | 11 | 35,42 | 1,21 | 32,72 | 38,12 |
| 2022 | 12 | 45,54 | 1,27 | 42,71 | 48,37 |
| 2023 | 13 | 55,66 | 1,33 | 52,69 | 58,63 |

*Джерело*: розрахунки авторів

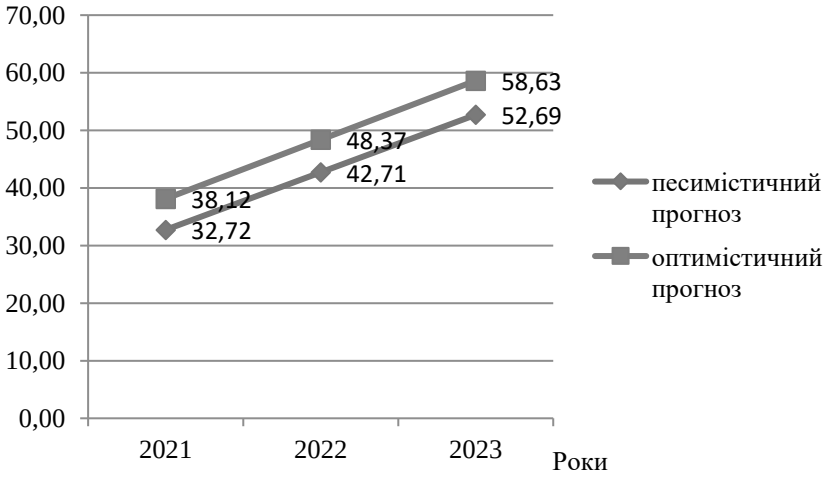



*Рисунок 5.3.* **Інтервальна оцінка прогнозування продуктивності праці (тис. грн) на підприємстві, 2021-2023 рр.**

*Джерело*: розробки авторів

Таким чином, з високою ймовірністю 95% можна стверджувати, що у 2021 р. значення показника продуктивності праці на підприємстві становитиме від 32,72 до 38,12 тис. грн; у 2022 р – від 42,71 до 48,37 тис. грн; у 2023 р. – від 52,69 до 58,63 тис. грн.

Поряд із стандартним програмним середовищем Excel при розв'язанні практичних задач широко застосовуються сучасні прикладні статистичні програми: R, SPSS Statistics, SAS, Mathlab, Gretl та ін. Зокрема, програма SPSS (рис. 5.4) дає можливість отримати коефіцієнти регресії лінії тренду, середні помилки по кожному коефіцієнту, перевірити критерії Стьюдента та Фішера.

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | ,881[a] | ,776 | ,748 | 17,45991 |

a. Predictors: (Constant), VAR00002

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 8451,212 | 1 | 8451,212 | 27,723 | ,001[b] |
| | Residual | 2438,788 | 8 | 304,848 | | |
| | Total | 10890,000 | 9 | | | |

a. Dependent Variable: VAR00003

b. Predictors: (Constant), VAR00002

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | 225,333 | 11,927 | | 18,892 | ,000 |
| | VAR00002 | 10,121 | 1,922 | ,881 | 5,265 | ,001 |

a. Dependent Variable: VAR00003

*Рисунок 5.4.* **Прогнозування показника продуктивності праці (тис. грн) на підприємстві в програмному середовищі SPSS**

*Джерело*: розрахунки авторів

Поряд з цим, з'являється можливість перевірити правило розкладання загального обсягу дисперсії за її розкладанням на міжгрупову та внутрішньогрупову. З цією метою застосовуємо вставку функції «Аналіз»→ «Регресія» → «Лінійна». Як видно з рисунка 5.4, основні результати моделювання суспільно-економічних явищ за допомогою програмного середовища SPSS стверджують про якість виконаних досліджень та високу достовірність побудованого прогнозу.

## 5.3. Особливості перевірки автокореляції у динамічних рядах

Поряд із побудовою лінії тренду під час розв'язання прикладних статистичних задач виникає потреба перевірки відсутності (наявності ) автокореляції у часових рядах. *Автокореляцією* називають наявність тісного взаємозв'язку між послідовними елементами часового чи просторового ряду даних. В економетричних дослідженнях часто виникають такі ситуації, коли дисперсія залишків є сталою, але спостерігається їх коваріація, взаємозв'язок, що називають *автокореляцією залишків.*

Безумовно, наявність автокореляції в рядах динаміки є негативним явищем, що робить неякісною статистичну модель та недостовірними - побудовані статистичні прогнози. В таких випадках застосовують спеціальні методи перевірки відсутності (наявності) автокореляції. Зазначимо, що коли існує кореляція між послідовними значеннями деякої залежної змінної, то буде спостерігатись і кореляція послідовних значень залишків.

Одним з методів виявлення автокореляції відхилень є *перевірка за допомогою вибіркового коефіцієнта кореляції* $r_в$ між двома сусідніми значеннями відхилень $u_i$, що називається автокореляцією залишків. З цією метою використовуються послідовності:

$$u_i = \{u_1, u_2, \ldots, u_{n-1}\} \text{ та } u_{i+1} = \{u_2, u_3, \ldots, u_n\}$$

і обчислюється коефіцієнт кореляції $r_в$ за такою формулою:

$$r_в = \frac{\sum_{i=1}^{n-1} u_i u_{i+1}}{\sum_{i=1}^{n} u_i^2}.$$

Останній вираз можна використовувати для оцінки кореляції між сусідніми відхиленнями $u_i$. Однак, оскільки ці відхилення є випадковими, виникає необхідність перевірки нульової гіпотези $H_0: r_г = 0$ про рівність нулю генерального коефіцієнта кореляції, тобто про відсутність зв'язку між даними рядами $u_i$ та $u_{i+1}$ при альтернативній гіпотезі $H_1: r_г \neq 0$.

За критерій перевірки нульової гіпотези використовують $t$ – статистику Ст'юдента, розрахункове значення якої відповідає виразу:

$$t = \frac{r_в\sqrt{n-1}}{\sqrt{1-r_в^2}} \qquad (5.7)$$

Для необхідного рівня значущості $\alpha$ та кількості ступенів свободи $k=n-1$ знаходять критичне значення $t_{кр}$ цього критерію і порівнюють його з обчисленим значенням. Якщо $|t| > t_{кр}$, то нульову гіпотезу відхиляють, тобто між сусідніми відхиленнями $u_i$ є кореляція і побудовану модель неможливо використовувати для дослідження даного соціально – економічного процесу. Інакше, нульову гіпотезу приймають, тобто між сусідніми відхиленнями тісної автокореляції немає.

Наступним методом перевірки наявності автокореляції між сусідніми відхиленнями $u_i$ є метод Дарбіна-Уотсона. За цим методом також будуються послідовності $u_i$ та $u_{i+1}$ , а потім обчислюють розрахункове значення $d$ – критерію за формулою:

$$d = \frac{\sum_{i=1}^{n-1}(u_{i+1}-u_i)^2}{\sum_{i=1}^{n} u_i^2} \qquad (5.8)$$

Встановимо зв'язок між вибірковим коефіцієнтом кореляції $r_в$ і $d$ – критерієм.

Оскільки $-1 \le r_в \le 1$, то завжди спостерігаємо $0 \le d \le 4$. Якщо між відхиленнями $u_i$ та $u_{i+1}$ є додатна тісна кореляція ($r_в \approx 1$), то сусідні відхилення будуть майже збігатись і $d$ – статистика прямує до нуля. Якщо ж сусідні відхилення некорельовані між собою ($r_в \approx 0$), то $d \to 2$. При значній від'ємній кореляції між $u_i$ та $u_{i+1}$ ($r_в \approx -1$) $d \to 4$.

Випадкова величина $d$ має спеціальний статистичний розподіл. Існують таблиці, що називаються критерієм Дарбіна-Уотсона критичних точок $d_1$ (нижнє значення) та $d_2$ (верхнє значення), величина яких залежить від кількості спостережень $n$, кількості факторів $m$ та величини значущості $\alpha$. Для перевірки наявності кореляції висувається нульова гіпотеза $H_0$ про відсутність кореляції між сусідніми відхиленнями у генеральній сукупності при альтернативній гіпотезі $H_1$: $r_{u_i u_{i+1}} \neq 0$. За вказаною формулою обчислюється величина $d$ – статистики, що спостерігається, і її значення порівнюється з критичними значеннями $d_1$ і $d_2$.

Із наведеного аналізу випливає, що коли $0 \le d < d_1$, то між $u_i$ та $u_{i+1}$ є додатня автокореляція і нульову гіпотезу необхідно відхилити. Якщо $2 > d > d_2$, то між вказаними відхиленнями кореляція відсутня, немає підстав відхиляти нульову гіпотезу, дана часова економетрична модель має зміст. Якщо ж $4 - d_1 < d < 4$, то має місце від'ємна автокореляція залишків і нульову гіпотезу необхідно відхилити. Якщо $2 < d \le 4 - d_2$, то говорять, що тісної від'ємної автокореляції немає, дана часова економетрична модель має зміст.

Якщо між сусідніми відхиленнями є тісна автокореляція, то використовувати дану модель неможливо, оскільки не виконується одна із передумов побудови парної нормальної лінійної регресії. Необхідно виявити причини наявності такої кореляції. Сама статистична модель в таких випадках потребує доопрацювання, а отримані результати та статистичний аналіз не є достовірними та високо якісними.

Автокореляція може бути також наслідком помилкової специфікації статистичної моделі, зокрема наявність автокореляції залишків може означати, що необхідно ввести до моделі нову незалежну змінну, розширити часовий проміжок ряду вихідних даних або розширити великі часові інтервали (роки) на менші проміжки (квартали, місяці).

*Задача 5.4.* На основі застосування програмного середовища SPSS перевірити відсутність автокореляції в ряді динаміки про продуктивність праці на підприємстві (задача 5.2). Виконати перевірку автокореляції часових даних та автокореляції залишків на основі методу Дарбіна -Уотсона.

Для перевірки відсутності автокореляції часових даних використовуємо вставку функції «Аналіз»→ «Прогнозування» →«Автокореляція» (рис. 5.5).

**Autocorrelations**

Series: VAR00003

| Lag | Autocorrelation | Std. Error[a] | Box-Ljung Statistic | | |
|---|---|---|---|---|---|
| | | | Value | df | Sig.[b] |
| 1 | ,414 | ,274 | 2,286 | 1 | ,131 |
| 2 | ,183 | ,258 | 2,791 | 2 | ,248 |
| 3 | -,058 | ,242 | 2,849 | 3 | ,416 |
| 4 | -,021 | ,224 | 2,858 | 4 | ,582 |
| 5 | ,045 | ,204 | 2,907 | 5 | ,714 |
| 6 | -,099 | ,183 | 3,200 | 6 | ,783 |
| 7 | -,356 | ,158 | 8,270 | 7 | ,309 |
| 8 | -,285 | ,129 | 13,157 | 8 | ,107 |

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

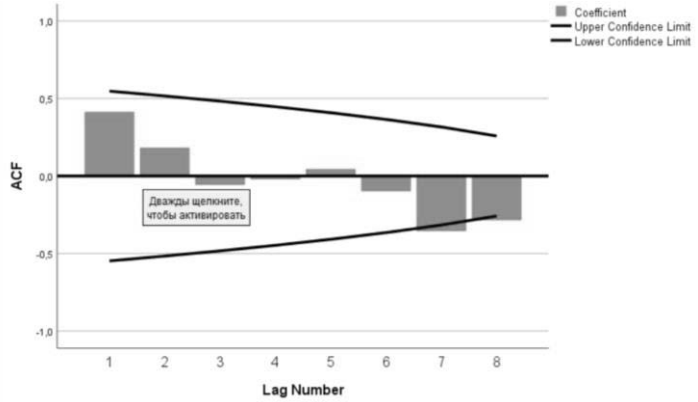



*Рисунок 5.5.* **Перевірка відсутності автокореляції даних продуктивності праці на підприємстві в програмному середовищі SPSS**

*Джерело*: розрахунки авторів

Перевірка на відсутність автокореляції залишків представлена на рис. 5.6.

З рисунків 5.5 та 5.6 слідує, що автокореляція даних та автокореляція залишків знаходиться у допустимих межах. Лише для передостаннього часового проміжку 2019 р. спостерігається незначне перевищення тісноти зв'язку між послідовними даними, що незначним чином виходить за допустимі межі.

**Partial Autocorrelations**

Series: VAR00003

| Lag | Partial Autocorrelation | Std. Error |
|---|---|---|
| 1 | ,414 | ,316 |
| 2 | ,015 | ,316 |
| 3 | -,168 | ,316 |
| 4 | ,070 | ,316 |
| 5 | ,082 | ,316 |
| 6 | -,211 | ,316 |
| 7 | -,347 | ,316 |
| 8 | ,053 | ,316 |

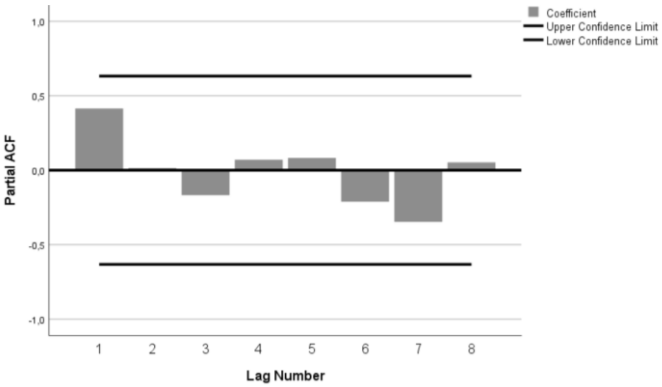



*Рисунок 5.6.* **Перевірка відсутності автокореляції залишків даних продуктивності праці на підприємстві в програмному середовищі SPSS**

*Джерело*: розрахунки авторів

## 5.4. Статистичне моделювання сезонних хвиль

При вивченні змін суспільно-економічних явищ у часі інколи має місце їх *сезонність*, тобто закономірність розвитку в залежності від певного часового проміжку (сезону року, місяця, певного дня на тижні тощо). Наприклад, попит на ринку певного виду (морозиво, газовані напої) продукції влітку може зростати, а взимку зменшуватися. Особливо великий вплив сезонності спостерігається в аграрному виробництві, у сфері туристичних послуг, у рівнях захворюваності на ГРВІ та грип, у обсягах споживання домогосподарствами електроенергії та газу тощо.

Такі показники у динаміці мають вигляд хвилі, яка щоразу то зростає, то спадає. *Сезонними коливаннями (сезонною хвилею) називають циклічні зміни певного суспільно-масового явища, які мають властивість сезонності*[3]. Зауважимо, що для статистичного аналізу таких даних необхідно мати інформацію про значення показника протягом кількох циклів, щоб оцінити справжню закономірність розвитку явища у часі.

Такі коливання спостерігаються в багатьох галузях народного господарства. Особливо вони характерні для сільського господарства, де виробництво продукції значною мірою залежить від природних умов. Продуктивність тварин, використання трудових ресурсів і техніки, переробка сільськогосподарської продукції тощо мають явно виражений сезонний характер.

При вивченні сезонних коливань статистичні дослідження виконуються за такими напрямами та вирішують наступні завдання:

1. вивчається загальна тенденція зміни досліджуваного явища у часі;
2. дається всебічна характеристика ступеня сезонності явища;
3. визначаються фактори, котрі викликають сезонні коливання;

[3] Матковський С.О., Гальків Л.І., Гринькевич О.З., Сорочак О.З. Статистика: Навч. Посібник – Львів: «Новий Світ - 2000», 2009. – 430 с.

4. формуються практичні рекомендації по усуненню або послабленню негативного впливу сезонних явищ на результативні показники ефективності суб'єктів господарювання.

Аналіз сезонних коливань дає змогу дати кількісну оцінку інтенсивності сезонних змін і розробити заходи щодо їх послаблення.

Щоб виявити сезонні коливання, аналізують місячні рівні ряду за один рік або кілька років.

Сезонні коливання в статистиці вимірюють за допомогою розрахунку спеціальних показників. Показники сезонності у вигляді сезонної хвилі можуть бути розраховані різними способами. Способи розрахунку показників сезонності залежать від характеру основної тенденції ряду динаміки.

При розв'язанні практичних задач спочатку вивчають графічне зображення динаміки показника. Якщо на графіку спостерігаються однакові часові періоди (цикли), зміна ознаки має однакову тенденцію в залежності від фактору часу, то це дає підстави для подальшого вивчення її сезонної декомпозиції статистичними методами.

Найбільш поширеним статистичним показником для оцінки сезонності суспільних явищ є *індекс сезонності*, який обчислюють за формулою:

$$I \text{ сез} = \frac{Y_i}{\bar{Y}} \times 100\% \qquad (5.9)$$

Даний показник дає змогу дослідити, наскільки відсотків відхиляється кожне індивідуальне значення ознаки у часі порівняно із його середнім значенням.

При стабільній тенденції у ряду динаміки, в якому внутрішньорічні коливання ознаки відбуваються навколо деякого постійного рівня, показники сезонності визначають як процентне відношення рівнів за кожний місяць до середньомісячного рівня за рік.

Однак окремі часові рівні, наприклад по місяцях року за один рік можуть бути нетиповими через вплив випадкових причин. Тому на практиці індекси сезонності визначають за місячними даними за декілька років (два-три роки і більше). В цьому разі для кожного місяця встановлюють середню величину рівня за кілька років (наприклад, три роки), далі з них розраховують середньомісячний рівень для всього ряду. Після цього кожен середньомісячний рівень порівнюють з середньомісячним річним рівнем за кілька років, а знайдений результат перемножують на сто процентів.

Якщо у часовому ряді має місце певна тенденція розвитку, то спочатку визначають рівняння лінії тренду Y(t), обчислюють вирівняні значення ознаки $\tilde{y}_{ti}$ за даним рівнянням, визначають індекс сезонності, а потім коригують даний показник на сезонність шляхом множення розрахункових значень на пункти індексу. Такий метод моделювання сезонних хвиль дає змогу не тільки провести статистичну оцінку сезонності досліджуваного явища, але й з високою достовірністю виконати прогнозування його розвитку на декілька часових інтервалів вперед.

*Задача 5.5*. За даними показника захворюваності на 10 тис. осіб у регіоні у 2017-2020 рр. визначити лінію тренду та скоригувати її на сезонність. Оцінити на основі залишкової дисперсії показника міру його коливання та ступінь впливу випадкових чинників на тенденцію розвитку явища.

Усі вихідні та розрахункові дані представлено у табл. 5.8. На основі застосування методу найменших квадратів визначено рівняння лінії тренду:

$$\tilde{y}_t = 10{,}925 - 0{,}0294 \cdot t.$$

Підставляючи кожне окреме значення часового проміжку в дану рівняння, отримуємо ряд вирівняних (розрахункових) значень даної ознаки $\tilde{y}_{ti}$.

*Таблиця 5.8.* **Коригування сезонної хвилі показника захворюваності на 10 тис. осіб у регіоні, 2017-2020 рр.**

| Рік | Квартал | №з/п | Показник захворюваності | Вирівняні значення | Індекс сезонності | Коригування показника на сезонність | Залишки для оцінки залишкової дисперсії | |
|---|---|---|---|---|---|---|---|---|
| | | $t_i$ | $y_i$ | $\tilde{y}_{ti}$ | $I_{сез}$ | $\frac{\tilde{y}_t \times I_{сез}}{100}$ | $u_i$ | $u_i^2$ |
| A | B | C | D | E | F | G | H | I |
| 2017 | 1 | 1 | 11,2 | 10,9 | 104,9 | 11,4 | -0,2 | 0,0536 |
| | 2 | 2 | 12,0 | 10,9 | 112,4 | 12,2 | -0,2 | 0,0462 |
| | 3 | 3 | 9,1 | 10,8 | 85,2 | 9,2 | -0,1 | 0,0190 |
| | 4 | 4 | 10,2 | 10,8 | 95,6 | 10,3 | -0,1 | 0,0160 |
| 2018 | 1 | 5 | 12,9 | 10,8 | 120,8 | 13,0 | -0,1 | 0,0155 |
| | 2 | 6 | 11,5 | 10,7 | 107,7 | 11,6 | -0,1 | 0,0063 |
| | 3 | 7 | 8,3 | 10,7 | 77,8 | 8,3 | 0,0 | 0,0012 |
| | 4 | 8 | 10,4 | 10,7 | 97,4 | 10,4 | 0,0 | 0,0002 |
| 2019 | 1 | 9 | 11,3 | 10,7 | 105,9 | 11,3 | 0,0 | 0,0002 |
| | 2 | 10 | 11,4 | 10,6 | 106,8 | 11,4 | 0,0 | 0,0022 |
| | 3 | 11 | 8,1 | 10,6 | 75,9 | 8,0 | 0,1 | 0,0031 |
| | 4 | 12 | 10,6 | 10,6 | 99,3 | 10,5 | 0,1 | 0,0104 |
| 2020 | 1 | 13 | 11,0 | 10,5 | 103,0 | 10,9 | 0,1 | 0,0186 |
| | 2 | 14 | 12,3 | 10,5 | 115,2 | 12,1 | 0,2 | 0,0347 |
| | 3 | 15 | 11,5 | 10,5 | 107,7 | 11,3 | 0,2 | 0,0423 |
| | 4 | 16 | 9,0 | 10,5 | 84,3 | 8,8 | 0,2 | 0,0345 |
| Сума | - | - | 170,8 | 170,8 | - | 170,8 | 0,0 | 0,3040 |
| Середнє | - | - | 10,7 | 10,7 | - | - | - | - |

*Джерело*: розрахунки авторів

Індекс сезонності обчислюємо шляхом співставлення фактичного значення показника захворюваності із його середнім значенням за даний період. Наприклад, для першого проміжку:

$$Iсез1 = \frac{11,2}{10,7} * 100\% = 104,9\%$$ і т. д.

Скоригуємо значення показника на індекс сезонності множенням розрахункових показників на їх пункти сезонності. Наприклад, для першого кварталу 2017 р. коригування показника захворюваності на індекс сезонності буде виконуватися таким чином:

$$Yкор\ 1 = \frac{10,9*104,9}{100} = 11,4\ .$$

Графічне зображення показника захворюваності у регіоні на 10 тис. осіб населення та його коригування на сезонність представлено на рисунку 5.7. Як видно з рисунка, фактичні та розрахункові дані є досить близькими. Тобто, побудована лінія тренду та її коригування на сезонність досить точно описують тенденцію розвитку явища.

Як бачимо із табл. 5.8, існують незначні відхилення розрахункових значень показника від його фактичних рівнів, що зумовлене його випадковими коливаннями у часі. Обчислимо абсолютні та відносні показники таких випадкових коливань. Найбільш поширеною абсолютною мірою цих коливань є залишкова дисперсія, яка визначається за формулою:

$$S_u^2 = \frac{\sum_{i=1}^{n} u_i^2}{n-m} \qquad (5.10)$$

де: $u_i^2$ – квадрати відхилень розрахункових значень ознаки від її фактичних величин;

n – кількість експериментів (в нашому випадку n=16);

m – число факторних ознак (в нашому випадку m=2).


*Рисунок 5.7*. **Графічне зображення показника захворюваності та його коригування на сезонність у регіоні на 10 тис. осіб**

*Джерело*: розрахунки авторів

Тоді,

$$S_u^2 = \frac{0{,}3040}{16-2} = 0{,}0217.$$

За даними залишкової дисперсії визначимо середнє квадратичне відхилення випадкових залишків:

$$S_u = \sqrt{S_u^2} = \sqrt{0{,}0217} = 0{,}147.$$

Таким чином, кожне індивідуальне значення випадкових коливань ознаки відхиляється відносно їх середнього значення на ±0,147. Відносною мірою випадкових сезонних коливань є відносний коефіцієнт варіації сезонних хвиль:

$$V_u = \frac{S_u}{\bar{Y}} * 100\%, \qquad (5.11)$$

де: $\bar{Y}$ – середній рівень динамічного ряду.

Отже,

$$V_u = \frac{0{,}147}{10{,}7} * 100\% = 1{,}4\%.$$

Таке значення коефіцієнта варіації випадкових сезонних коливань свідчить про незначний вплив випадкових чинників на динаміку рівня захворюваності на ГРВІ у регіоні.

Наступним методом моделювання сезонних хвиль виступає метод ковзної середньої, коли відбувається укрупнення періодів і кожен наступний рівень ряду зсувається на попередній період. Програмне середовище SPSS дає змогу побудувати таку модель на основі застосування функції «Аналіз»→ «Прогнозування» →«Сезонна декомпозиція» (рис. 5.8). Як видно з рисунку, лінія вирівнювання показника з урахуванням фактору сезонності є досить близькою до моделювання сезонної хвилі за МНК.

Такі статистичні моделі дають змогу не тільки оцінити вплив випадкових факторів на сезонну хвилю, але й змоделювати розвиток явища на перспективу. Зауважимо, що математичне формулювання лінії тренду може бути те тільки прямою лінією, але й кривою - параболою, гіперболою, показниковою, логарифмічною, степеневою та ін. функціями, що детально описано в наступних розділах посібника.

Таким чином, метод розрахунку індексів сезонності на основі осереднених значень досліджуваного показника заснований на властивості явищ мати однакові тенденції впродовж кожного окремого сезону та відкидає можливість еволюції сезонного фактору. Крім того, більш достовірними є обчислення, виконані для економік, які стабільно розвиваються, аніж для країн з перехідною економікою. Більшість досліджуваних явищ мають певну тенденцію до зміни (зростання або зменшення), тому для визначення показників сезонності необхідна нейтралізація еволюції тренду, що дає змогу здійснювати прогноз на перспективу з урахуванням сезонних коливань.

## Seasonal Decomposition

**Model Description**

| | | |
|---|---|---|
| Model Name | | MOD_3 |
| Model Type | | Multiplicative |
| Series Name | 1 | VAR00002 |
| Length of Seasonal Period | | 4 |
| Computing Method of Moving Averages | | Span equal to the periodicity and all points weighted equally |

Applying the model specifications from MOD_3

**Seasonal Decomposition**

Series Name: VAR00002

| DATE_ | Original Series | Moving Average Series | Ratio of Original Series to Moving Average Series (%) | Seasonal Factor (%) | Seasonally Adjusted Series | Smoothed Trend-Cycle Series | Irregular (Error) Component |
|---|---|---|---|---|---|---|---|
| Q1 2017 | 11,200 | . | . | 109,2 | 10,252 | 10,773 | ,952 |
| Q2 2017 | 12,000 | . | . | 108,4 | 11,066 | 10,805 | 1,024 |
| Q3 2017 | 9,100 | 10,6250 | 85,6 | 82,0 | 11,097 | 10,869 | 1,021 |
| Q4 2017 | 10,200 | 11,0500 | 92,3 | 100,3 | 10,169 | 10,887 | ,934 |
| Q1 2018 | 12,900 | 10,9250 | 118,1 | 109,2 | 11,808 | 10,910 | 1,082 |
| Q2 2018 | 11,500 | 10,7250 | 107,2 | 108,4 | 10,605 | 10,690 | ,992 |
| Q3 2018 | 8,300 | 10,7750 | 77,0 | 82,0 | 10,121 | 10,496 | ,964 |
| Q4 2018 | 10,400 | 10,3750 | 100,2 | 100,3 | 10,368 | 10,350 | 1,002 |
| Q1 2019 | 11,300 | 10,3500 | 109,2 | 109,2 | 10,343 | 10,310 | 1,003 |
| Q2 2019 | 11,400 | 10,3000 | 110,7 | 108,4 | 10,513 | 10,324 | 1,018 |
| Q3 2019 | 8,100 | 10,3500 | [illegible] | 82,0 | 9,877 | 10,245 | ,964 |
| Q4 2019 | 10,600 | 10,2750 | [illegible] | 100,3 | 10,568 | 10,383 | 1,018 |
| Q1 2020 | 11,000 | 10,5000 | 104,8 | 109,2 | 10,069 | 10,881 | ,925 |
| Q2 2020 | 12,300 | 11,3500 | 108,4 | 108,4 | 11,343 | 11,306 | 1,003 |
| Q3 2020 | 11,500 | 10,9500 | 105,0 | 82,0 | 14,023 | 11,446 | 1,225 |
| Q4 2020 | 9,000 | . | . | 100,3 | 8,973 | 11,516 | ,779 |

Дважды щелкните, чтобы активировать

*Рисунок 5.8.* **Моделювання сезонної хвилі показника захворюваності на ГРВІ (у розрахунку на 10 тис. осіб) у регіоні в програмному середовищі SPSS**

*Джерело*: розрахунки авторів

***Список питань до самоконтролю*:**

**1.** Назвіть основні завдання дослідження рядів динаміки, чим вони зумовлені?

**2.** Охарактеризуйте основні статистичні показники для дослідження суспільно-економічних явищ у часі?

**3.** Які статистичні методи дослідження трендів у динаміці вам відомі? Поясніть зміст та методику визначення кожного з них.

**4.** Поясніть сутність методу найменших квадратів для побудови лінії тренду явища у динаміці.

**5.** Які види статистичних прогнозів вам відомі? Дайте характеристику кожному з них та поясніть формули обчислення.

**6.** Дайте визначення автокореляції даних та поясніть сутність методів щодо перевірки її наявності (відсутності). У разі наявності сформулюйте рекомендації по її усуненню.

**7.** Які статистичні критерії використовуються під час дослідження явищ у динаміці? В чому полягає сутність кожного з них?

**8.** Яким чином моделюють сезонні хвилі у динамічних процесах? Поясніть сутність та методику статистичних обчислень у кожному випадку.

**9.** Розробіть і за даними Державної служби статистики України заповніть статистичну таблицю, яка б характеризувала кількість юридичних осіб за організаційно-економічними формами

господарювання Вашої області упродовж останніх 5 років. Проаналізуйте наведені дані. Визначте абсолютні, відносні та середні показники динаміки даного явища. Зобразіть результати обчислень графічно та проаналізуйте одержані результати.

**10.** Розробіть макет статистичної таблиці, яка б характеризувала динаміку середньомісячної заробітної плати (номінальної та реальної) у Вашому регіоні за останні 10 років та заповніть її за даними Державної служби статистики України. Побудуйте лінію тренду для даного показника, виконайте прогнозування на наступні 3 роки. Одержані дані зобразіть графічно та проаналізуйте основні результати.

**11.** За даними Державної служби статистики України отримайте дані про валовий регіональний продукт у Вашій області за останні 5 років заповніть їх у статистичну таблицю. Проаналізуйте наведені дані. Перевірте наявність (відсутність) автокореляції за методом Дарбіна-Уотсона. Зобразіть результати обчислень графічно в програмному середовищі Excel чи SPSS та проаналізуйте одержані результати.

**12.** Розробіть макет статистичної таблиці, яка б характеризувала сезонність обсягів споживання електроенергії у Вашому регіоні та заповніть її за даними Державної служби статистики України. Побудуйте сезонну хвилю даного показника, виконайте коригування даних на індекс сезонності. Одержані дані зобразіть графічно та проаналізуйте основні результати.

### *Список рекомендованої літератури по темі:*


1. Дербенцев В.Д., Сердюк О.А., Соловйов В.М., Шарапов О.Д. Синергетичні та економфізичні методи дослідження динамічних та структурних характеристик економічних систем. Монографія. Черкаси: Брама-Україна, 2010. 287 с.
2. Єріна А. М. Статистичне моделювання та прогнозування: Навч. посібник. К.: КНЕУ, 2001. 170 с.
3. Зражевський О.Г. Методи побудови моделей для довгострокового прогнозування фінансових часових рядів. Системні дослідження та інформаційні технології, 2010, №1. С. 123-142.
4. Ковтун Н. В. Теорія статистики: підручник. Київ : Знання, 2012. 400 с.
5. Костюк В. О. Прикладна статистика: навч. Посібник. Харк. нац. ун-т міськ. госп-ва ім. О. М. Бекетова. Харків : ХНУМГ ім. О. М. Бекетова, 2015. 191 с.
6. Мармоза А.Т. Практикум з теорії статистики. К.: Центр учбової літератури, 2013. 484 с.
7. Матковський С.О., Гальків Л.І., Гринькевич О.З., Сорочак О.З. Статистика: Навч. Посібник Львів: «Новий Світ 2000», 2009. 430 с.
8. Руська Р. В. Економетрика : навчальний посібник. Тернопіль : Тайп, 2012. 224 с.
9. Field, A. (2012) Discovering statistics using IBM SPSS Statistics. 4-th edition. Los Angrles/ London/ New Delphi/ Singapore/ Washingtone DC : Sage. 1467 p. URL: https://ua1lib.org/book/2823883/3c9a15?dsource=recommend
10. Mayers, A. (2013) Introduction to statistics and SPSS in psychology. Pearson. 626 p. URL: https://ua1lib.org/book/10985783/e208a5?id=10985783&secret=e208a5.

# Розділ 6. ВИКОРИСТАННЯ ЛОГІТ ТА ПРОБІТ РЕГРЕСІЙНИХ МОДЕЛЕЙ В АНАЛІЗІ БІНАРНОЇ КЛАСИФІКАЦІЇ

## 6.1. Сутність методу логіт- та пробіт-регресії та умови його застосування

Логістична регресія (logit-регресійна модель) – це різновид множинної регресії, загальне призначення якої полягає в аналізі зв'язку між декількома незалежними змінними (регресорами або предикторами) і залежною змінною. Ідея логістичної регресії зароджувалася в роботах різних авторів ще у 1950-х, проте в нинішньому вигляді остаточно була сформульована в середині 1960х (D.R. Cox *Some procedures associated with the logistic qualitative response curve*).

Моделі логістичної регресії застосовуються у різних сферах, а саме:

• Медицині (для визначення ймовірності успішного лікування тощо)

• Соціології.

• Маркетингових дослідженнях (для передбачення схильності до покупки).

• При вирішенні завдань класифікації (скоринг в банках, маркетинг тощо).

Бінарна логістична регресія, як випливає з назви, застосовується у випадку, коли залежна змінна є бінарною (тобто може приймати тільки два значення). Іншими словами, за допомогою логістичної регресії можна оцінювати вірогідність того, що подія настане для конкретного спостереження.

Як відомо, все регресійні моделі можуть бути представлені у загальному вигляді за формулою:

$$y = F(x_1, x_2, \ldots, x_n) \tag{6.1}$$

Наприклад, в множинній лінійній регресії передбачається, що залежна змінна є лінійною функцією незалежних змінних, тобто:

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 \ldots + b_n \cdot x_n. \tag{6.2}$$

Даний вид регресії можна використовувати для оцінки ймовірності результату події, проте множинна регресія не «знає», що змінна відгуку бінарна за своєю природою. Це неминуче призведе до моделі, що розраховує значення більші за 1 і менші 0. Але такі значення взагалі не припустимі для початкової задачі. Таким чином, множинна регресія просто ігнорує обмеження на діапазон значень для у.

Для вирішення проблеми завдання регресії може бути сформульоване інакше: замість передбачення бінарної змінної, ми передбачаємо безперервну змінну зі значеннями на відрізку [0,1] при будь-яких значеннях незалежних змінних. Це досягається застосуванням наступного регресійного рівняння (логіт-еквіваленту):

$$p = \frac{1}{1+e^{-y}} \tag{6.3}$$

де: p – ймовірність того, що відбудеться подія;
e – основа натуральних логарифмів;
y – стандартне рівняння регресії.

Залежність, що зв'язує ймовірність події і величину у, показана на рис. 6.1:

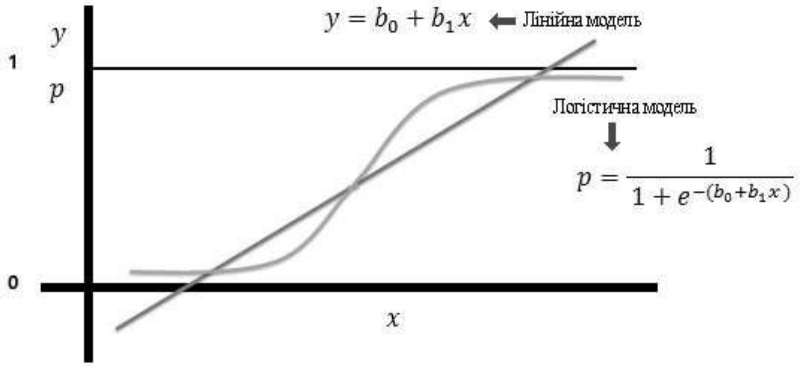



*Рисунок. 6.1.* **Порівняння логістичної кривої та лінійної регресії**

Пояснимо необхідність перетворення. Припустимо, що ми розмірковуємо про нашу залежну змінну в термінах основної ймовірності *P*, що лежить між 0 і 1. Тоді перетворимо цю ймовірність *P*:

$$P' = ln\left(\frac{P}{1-P}\right) \tag{6.4}$$

Це перетворення зазвичай називають логістичним або логіт-перетворенням. Теоретично P' може приймати будь-яке значення. Оскільки логістичне перетворення вирішує проблему про обмеження на 0-1 інтервалі для початкової залежної змінної (ймовірності), то ці перетворені значення можна використовувати в звичайному лінійному регресійному рівнянні. А саме, якщо провести логістичне перетворення обох частин описаного вище рівняння, ми отримаємо стандартну модель лінійної регресії.

Зі всього випливає, що логіт-модель може бути застосована лише за умов відповідності факторних змінних логнормальному розподілу, а отже ці змінні мають бути неперервними.

Існує кілька способів знаходження коефіцієнтів логістичної регресії. На практиці часто використовують метод максимальної правдоподібності. Він застосовується в статистиці для отримання оцінок параметрів генеральної сукупності за даними вибірки. Основу методу складає функція правдоподібності (likehood function), яка виражає щільність ймовірності спільної появи результатів вибірки $Y_1$, $Y_2$, ..., $Y_k$:

$$L = (Y_1, Y_2, \ldots, Y_k; \theta) = p(Y_1; \theta) \cdot \ldots \cdot p(Y_k; \theta) \tag{6.5}$$

Згідно з методом максимальної правдоподібності в якості оцінки невідомого параметра приймається таке значення $\Theta = \Theta(Y_1, \ldots, Y_k)$, яке максимізує функцію L.

Знаходження оцінки спрощується, якщо максимізувати не саму функцію L, а натуральний логарифм ln(L), оскільки максимум обох функцій досягається при одному і тому ж значенні θ:

$$L \cdot (Y; \theta) = ln(L(Y; \theta)) \rightarrow max \tag{6.6}$$

У разі бінарної незалежної змінної, яку ми маємо в логістичній регресії, розрахунок можна продовжити наступним чином. Позначимо через $P_i$ ймовірність появи одиниці: $P_i$ = Prob ($Y_i$ = 1). Ця ймовірність буде виглядати таким чином:

$$P_i = F(X_i W), \qquad F(z) = \frac{1}{1+e^{-z}} \tag{6.7}$$

де $X_i$ – рядок матриці регресорів;
W – вектор коефіцієнтів регресії.

Логарифмічна функція правдоподібності дорівнює:

$$L^* = \sum_{i \in I1} lnP_i(W) + \sum_{i \in I0} \ln(1 - P_i(W)) =$$

$$= \sum_{i=1}^{k} [Y_i ln P_i(W) + (1 - Y_i) ln(1 - P_i(W))], \quad (6.8)$$

де $I_0$, $I_1$ – безлічі спостережень, для яких $Y_i = 0$ і $Y_i = 1$ відповідно.

Інтерпретація коефіцієнтів логіт-моделі відрізняється від моделі лінійної регресії. У моделі бінарного вибору коефіцієнти показують наскільки зміниться ймовірність отримання значення Z = 1 при зміні величини незалежної змінної на одиницю і при незмінних значеннях інших змінних. Негативний знак при коефіцієнті регресії говорить про зменшення ймовірності при збільшенні відповідних змінних, позитивний – про збільшення.

Для полегшення інтерпретації коефіцієнтів регресії необхідно знайти граничний ефект, процедура розрахунку якого реалізується наступним чином:

1. Розраховується z значення, як лінійна комбінація наступного виду :

$$z = b_0 + b_1 \bar{x}_1 + \ldots + b_n \bar{x}_n \quad (6.9)$$

2. Обчислюється f(x) за формулою:

$$f(x) = \frac{1}{1+e^{-z}} \quad (6.10)$$

3. Граничний ефект для кожного коефіцієнта логістичної регресії розраховується за формулою:

$$b'_i = b_i \times f(x) \quad (6.11)$$

Таким чином, логіт-регресійна модель вирішуючи задачу класифікації, і на відміну від звичайної лінійної регресії, передбачає безперервну змінну зі значеннями на відрізку [0,1] при будь-яких значеннях незалежних змінних, що здійснюється шляхом застосування логіт-перетворення. Логіт-регресійна модель вимагає відповідності розподілу неперервних змінних лог-нормальному розподілу випадкових величин. При оцінці коефіцієнтів моделі використовується метод максимальної правдоподібності, а для їх інтерпретації розраховуються граничні ефекти.

Оцінка адекватності побудованої моделі заснована на аналізі тестових характеристик і статистичній перевірці гіпотез. Для оцінки статистичної надійності оцінок параметрів застосовується p-value; для аналізу рівняння в цілому перевірка нульової гіпотези про значимість коефіцієнту проводиться за допомогою відношення правдоподібності (LM). В якості критичної статистики тесту береться різниця максимумів логарифмічних функцій правдоподібності.

У всіх цьому тесті нульова гіпотеза формується наступним чином:

$$H_0: Q\beta = r\,, \tag{6.12}$$

де: Q – відома матриця обмежень;

β – вектор параметрів, що тестуються;

r – вектор констант.

Сутність нульової гіпотези полягає в тому, що коефіцієнти при всіх включених в модель змінних одночасно дорівнюють нулю. Якщо нульова гіпотеза відхиляється, то, значить, в моделі присутні фактори, що статистично значуще впливають на ендогенну змінну.

Таким чином, логістична регресія дає можливість передбачити наявність або відсутність ознаки на підставі значень набору змінних-предикторів. Вона подібна до моделі лінійної регресії, але може бути застосована для моделей, де залежна змінна має тільки два значення. Основною перевагою логістичної регресії є можливість її застосування до більш широкого діапазону ситуацій, ніж дискримінантний аналіз.

Не дивлячись на те, що в регресійному аналізі економічних процесів частіше використовуються інші види регресій (через неперервний характер залежної величини), логістична регресія широко використовується для управління кредитними ризиками та у маркетингових дослідженнях. Найбільшого поширення даний вид

регресії набув саме у медичній сфері. Аналіз може проводитись як для визначення імовірності появи хвороби, так і на стадії перебігу захворювання (виникнення ускладнень).

Логістична регресія також є одним з методів статистичного контролю якості продукції – встановлення відповідності продукції та процесів вимогам нормативно-технічної документації.

Пробіт-модель – альтернативна модель двоїчного вибору. Ідея пробіт-аналізу вперше була опублікована Честером Бліссом в 1934 р в статті про вплив пестицидів на відсоток вбитих шкідників. Блісс запропонував для обліку відсотка убитих шкідників використовувати імовірнісний блок – probability unit (або probit). Дане їм визначення трохи відрізнялося від того, що використовується в наш час в статистичному аналізі. Остаточне визначення пробістичного аналізу дав Джон Фінні у своїй праці «Пробіт-аналіз», що була опублікована у 1971 році при Кембріджському університеті.

Для пробіт-аналізу використовується стандартний нормальний розподіл для моделювання залежності F(Z). Функція ймовірності залежить від змінної Z, яка своєю чергою залежить від обраних факторів:

$$p_i = F(Z_i)\ . \tag{6.13}$$

Наприклад, в множинній лінійній регресії передбачається, що залежна змінна є лінійною функцією незалежних змінних, тобто:

$$z = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 \ldots + b_n \cdot x_n. \tag{6.14}$$

Основною умовою застосування пробіт-моделей – це відповідність стандартному нормальному розподілу, а тому можливості пробіт-моделей значно ширше ніж логіт-моделі, оскільки вона може бути застосована не тільки для неперервних

змінних, що підпорядковуються нормальному розподілу, а й для категоріальних змінних, які мають біноміальний розподіл, який може бути апроксимований стандартним нормальним розподілом.

Для оцінювання параметрів, як і в логіт-моделі, використовується метод максимальної правдоподібності. Граничний ефект змінної $x_i$ – дорівнює похідної функції ймовірності за цією змінною.

Оскільки $f(Z)$ – похідна функції (функція щільності) стандартного нормального розподілу $F(Z_i)$, то вона виглядає наступним чином:

$$f(x) = p = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \tag{6.15}$$

Враховуючи цю особливість, у даній моделі можуть використовуватись змінні, які мають біноміальний розподіл, який можна апроксимувати нормальним. В цілому біноміальний розподіл не може в точності співпадати з нормальним в силу двох причин. По-перше, будь-який нормальний розподіл може давати результати у вигляді чисел із дробом (дробною частиною), у той час як біноміально розподілена величина може приймати лише цілі значення. По-друге, біноміальний розподіл при значенні імовірності, що відмінне від 0,5, завжди буде асиметричним, у той час, як нормальний розподіл є ідеально симетричним навколо середнього значення випадкової величини. Але біноміальний розподіл можна добре апроксимувати нормальним у випадку, коли n достатньо велике, а імовірність настання події не занадто близька до 0 або 1. Для цього необхідно розширити межі біноміально розподіленої величини на 0,5 (в меншу сторону для нижньої межі, в більшу – для верхньої) та здійснити нормування на основі середнього значення та квадратичного відхилення за формулою:

$$Z = \frac{x - \overline{x}}{\sigma}, \tag{6.16}$$

де x – верхня або нижня межа значень, збільшена або зменшена на 0,5;

$\overline{x}$ – середнє значення;

σ – середньоквадратичне відхилення.

Отримані нормовані значення знаходять у таблиці нормального розподілу і визначають відповідну імовірність настання події.

Для знаходження коефіцієнтів пробіт-моделі, за аналогією до логіт-моделі, використовується метод максимальної правдоподібності (6.5-6.8).

Для інтерпретації коефіцієнтів пробіт-регресійної моделі необхідно розрахувати граничні ефекти для кожного коефіцієнта. Дана процедура реалізується наступним чином:

Процедура оцінювання пробіт-моделі аналогічна логіт-моделі і включає наступні етапи:

1) Визначення залежної змінної (за необхідності представленняїї у формі бінарної змінної) і факторів, що її визначають.

2) Визначення Z, як лінійної комбінації незалежних змінних.

3) Побудова рівняння для ймовірності події та знаходження похідних (для оцінювання кумулятивного і граничного впливу факторів).

4) Оцінювання параметрів пробіт-моделі (використовується метод максимальної правдоподібності).

5) Визначення кумулятивного і граничного ефектів – з цією метою необхідно провести подальші розрахунки, підставивши отримані коефіцієнти у формулу:

$$\beta_i = b_i \cdot F(Z_i). \tag{6.16}$$

На практиці зустрічаються ситуації, коли необхідно дослідити не дві альтернативи, а декілька. Якщо ці альтернативи є невпорядкованими, то застосовують множинну (multinomial) пробіт-регресійну модель, у випадку упорядкованих альтернатив говорять про впорядковану (ordered) пробіт-модель.

Таким чином, аналогічно до логіт-моделі, пробіт-модель дозволяє передбачати значення бінарної результуючої змінної на основі відомих значень незалежних ознак. Але застосування probit-аналізу має певні переваги над логіт-аналізом, адже допускає використання дихтономічних та категоріальних змінних, що підпорядковуються біноміальному розподілу, адже вони можуть бути апроксимовані нормальним розподілом.

Для дослідження прогностичної здатності моделі можна використовувати класифікаційну таблицю, яка представляє собою розподіл правильних і помилкових класифікацій досліджуваної вибірки об'єктів. Побудова таблиці засновується на використанні порогового значення С і на обчисленні очікуваних значень залежної змінної. В результаті, чим більше отриманих правильних класифікацій і чим менше значення оцінок ймовірності помилок, тим вище прогностична якість побудованої моделі.

## 6.2. ROC-аналіз оцінювання прогностичної здатності моделі

ROC-аналіз (Receiver Operator Characteristic) – аналіз, який найбільш часто використовується для представлення результатів бінарної класифікації. ROC-аналіз використовується для дослідження прогностичної здатності моделі. На його основні розраховуються показники чутливості і специфічності моделі (частіше виявлення позитивних чи негативних наслідків).

Оскільки класів два, один з них називається класом з позитивними наслідками, другий – з негативними наслідками. ROC-аналіз показує залежність кількості вірно класифікованих позитивних прикладів від кількості невірно класифікованих негативних прикладів. У термінології ROC-аналізу перші називаються істинно позитивними, другі – хибно негативною множиною. При цьому передбачається, що у класифікатора є деякий параметр, що варіює і ми будемо отримувати те чи інше розбиття на два класи. Цей параметр часто називають порогом, або точкою відсікання (cut-off value). Залежно від нього будуть виходити різні величини помилок I і II роду. У логістичної регресії поріг відсікання змінюється від 0 до 1 – це і є розрахункове значення рівняння регресії.

Для розуміння суті помилок I і II роду розглянемо чотири клітинну таблицю, яка будується на основі результатів класифікації моделі і фактичної (об'єктивної) приналежності прикладів до класів.

*Таблиця 6.1.* **Матриця оцінювання прогностичної здатності моделі**

| Теоретичні значення | Фактичні значення | |
|---|---|---|
| | Позитивні | Негативні |
| Позитивні | TP | FP |
| Негативні | FN | TN |

Нижче представлене пояснення щодо складових табл. 6.1:

– TP (True Positives) – кількість вірно класифікованих позитивних прикладів (так звані істинно позитивні випадки);

– TN (True Negatives) – кількість вірно класифікованих негативних прикладів (істинно негативні випадки);

– FN (False Negatives) – кількість позитивних прикладів, класифікованих як негативні (помилка I роду). Це так званий

«помилковий пропуск», коли подія, яка нас цікавить, помилково не виявляється (хибно негативні випадки);

– FP (False Positives) – негативні зразки, класифіковані як позитивні. Це помилка 2-го роду (хибно позитивні випадки).

Що є позитивною подією, а що – негативною залежить від конкретного завдання. При аналізі частіше оперують не абсолютними показниками, а відносними.

Частка істинно позитивних прикладів (True Positives Rate):

$$TPR = \frac{TP}{TP+FN} \cdot 100\% \tag{6.17}$$

Частка хибно позитивних прикладів (False Positives Rate):

$$FPR = \frac{FP}{TN+FP} \cdot 100\% \tag{6.18}$$

Для аналізу результатів класифікації необхідно також розрізняти поняття чутливість і специфічність моделі. Ними визначається об'єктивна цінність будь-якого бінарного класифікатора.

Чутливість (Sensitivity) – це і є частка істинно позитивних випадків:

$$S_e = TPR \tag{6.19}$$

Специфічність (Specificity) – частка істинно негативних випадків, які були правильно ідентифіковані моделлю:

$$S_p = \frac{TN}{TN+FP} \cdot 100\% \tag{6.20}$$

Модель з високою чутливістю часто дає істинний результат за наявності позитивного результату (виявляє позитивні приклади).

Навпаки, модель з високою специфічністю частіше дає істинний результат при наявності негативного результату (виявляє негативні приклади).

В параграфі 6.4 представлено практику використання логіт- та пробіт-моделей в середовищах STATISTICA та SPSS.

## 6.3. Сутність тесту Мантеля-Ханзела та умови його використання для аналізу прихованих факторів впливу

У рамках логістичного та пробістичного регресійного аналізу при вивченні залежності між вихідною змінною та вхідними незалежними змінними важливими моментом постає вивчення та аналіз прихованих факторів впливу. Вплив даних факторів не коректно вимірювати шляхом простого включення їх у модель в якості незалежних змінних. Доречніше застосувати стратифікацію сукупності за даною ознакою та вивчити елементи, що належать до однієї страти, окремо, а потім порівняти отримані результати для різних страт і в цілому.

Для вимірювання рівня впливу стратифікованої змінної на результуючу (з урахуванням впливу факторної ознаки) спершу необхідно подати дані для кожної страти окремо та в цілому по сукупності у вигляді чотириклітинкової таблиці. Приклад таблиці для кожної страти наведений у табл. 6.2.

Для сукупності в цілому значення a,b,c,d дорівнюватимуть сумі відповідних значень для усіх страт.

*Таблиця 6.2.* **Чотири клітинна таблиця**

| | | Наявність результуючої ознаки | | Загалом |
|---|---|---|---|---|
| | | Так (1) | Ні (0) | |
| Наявність факторної ознаки | Так (1) | $a_i$ | $b_i$ | $a_i + b_i$ |
| | Ні (0) | $c_i$ | $d_i$ | $c_i + d_i$ |
| | Загалом | $a_i + c_i$ | $b_i + d_i$ | $n_i$ |

Далі доцільно розрахувати Risk Ratio (RR) і Odds Ratio (OR) окремо для кожної страти і для вибірки в цілому. Дані коефіцієнти характеризують щільність зв'язку незалежної та залежної ознаки (OR) та ризик появи певної події (вихідної ознаки) у групі із наявним впливом факторної ознаки у порівнянні з групою, де дана ознака відсутня (RR) [15, с. 598-600].

Дані коефіцієнти розраховують для кожної страти розраховуються за формулами:

$$OR = \frac{a_i d_i}{b_i c_i} \tag{6.21}$$

$$RR = \frac{a_i / (a_i + b_i)}{ci / (ci + di)} \tag{6.22}$$

За аналогією до кожної страти, дані показники розраховуються і в цілому для усієї сукупності, при цьому значення a, b, c, d в цілому для сукупності дорівнюватимуть сумі відповідних значень для усіх страт.

Показником наявності прихованого впливу різних страт є перевищення показників в цілому для сукупності над показниками окремо по стратах.

Завдання тесту Мантеля-Ханзела полягає у перевірці істотності прихованого впливу виділених страт на результуючу ознаку з урахуванням впливу факторної ознаки. При цьому нульова гіпотеза

($H_0$) формується як відсутність розбіжності між впливом окремих страт. Перевірка істотності даного впливу здійснюється за допомогою $\chi^2$, практичне значення якого розраховується за формулою:

$$\chi^2_{MH} = \frac{(|O-E|-0,5)^2}{V} \tag{6.23}$$

де O – фактичні частоти;
E – теоретичні частоти;
V – варіація.

Фактичні частоти дорівнюють значенню $a$ для усієї сукупності, а теоретичні частоти та варіація розраховуються за формулами:

$$E = \sum_{i=1}^{k} E_i = \sum_{i=1}^{k} \frac{(a_i+b_i)(a_i+c_i)}{n_i} \tag{6.24}$$

$$V = \sum_{i=1}^{k} V_i = \sum_{i=1}^{k} \frac{(a_i+b_i)(a_i+c_i)(b_i+c_i)(b_i+d_i)}{n_i^2(n_i-1)} \tag{6.25}$$

Розрахункове значення $\chi^2_{MH}$ порівнюється з теоретичним значенням $\chi^2$ при сталій кількості ступенів свободи (df=1) та заданому рівні істотності (1-$\alpha$). У випадку, якщо $\chi^2_{MH}<\chi^2$, то $H_0$ приймається, а істотність впливу виділених страт визнається випадковою, у протилежному випадку – нульова гіпотеза відхиляється, і визнається істотність впливу стратифікованої змінної на результат.

Таким чином, за допомогою розрахунку показників Odds Ratio та Risk Ratio, можна зробити загальний висновок про наявність чи відсутність прихованого впливу страт, а тест Мантеля-Ханзеля дозволяє перевірити істотність впливу виділених страт та оцінити вплив бінарної незалежної змінної на результуючу з урахуванням стратифікації сукупності.

## 6.4. Практика використання logit- та probit- регресійних моделей

### 6.4.1. Використання logit-регресійної моделі при дослідженні економічних процесів

Усі статистичні методи мають особливості використання залежно від предметної області дослідження. Логістична регресія не є винятком, тому було розглянуто її застосування у різних сферах діяльності, зокрема – економічній. На макрорівні застосування цієї моделі є практично неможливим, оскільки залежна змінна має бути бінарною, а економічний результат зазвичай це неперервна величина у вартісному вираженні. Проте логістична регресія широко використовується у маркетингу для сегментації ринку, виявлення цільової аудиторії, тощо. Також широкого застосування логістична регресія набула у банківській діяльності для оцінки кредитних ризиків.

Скорингова модель дозволяє банку на основі класифікації та визначення характерних ознак надійних, ненадійних та безнадійних клієнтів щодо погашення кредитної заборгованості, які отримані за допомого аналізу кредитних історій попередніх позичальників та визначити якої величини є ймовірність того, що окремо взятий позичальник поверне кредит у визначений термін. Це реалізовується за допомогою розрахунку інтегрального показника, на основі значення якого, здійснюється розподіл позичальників відносно бар'єру надійності – клієнти з показниками оцінки вище за бар'єр відносяться до надійних та отримують кредит, ті ж, що мають оцінки нижче за бар'єр потрапляють до списку неблагонадійних.

Завданням logit-регресійної моделі є виявлення благонадійних та неблагонадійних потенційних позичальників. Оскільки у скорингу загальноприйнятим є те, що чим вище рейтинг клієнта, тим вище його кредитоспроможність, то будемо вважати позитивним

результатом успішне погашення позики, а негативним – дефолт по кредиту.

Тоді проектуючи при цих умовах визначення чутливості і специфічності на скоринг, можна зробити висновок, що скорингова модель з високою специфічністю відповідає консервативній кредитній політиці (частіше відбувається відмова у видачі кредиту), а з високою чутливістю – політиці ризикованих кредитів. У першому випадку мінімізується кредитний ризик, пов'язаний з втратами позики і відсотків, і додатковими витратами на повернення кредиту, а в другому – комерційний ризик, пов'язаний з упущеною вигодою.

До моделі було включено наступні показники:

- вік клієнта;
- стать (ч/ж);
- чи перебуває клієнт у шлюбі? (так / ні);
- кількість осіб на утриманні, осіб;
- перевірений сукупний наявний дохід, грош. Од.;
- досвід роботи, років;
- термін проживання в регіоні, років;
- ринкова вартість нерухомості у власності, тис. дол .;
- щомісячний платіж по кредиту, грош. Од .;
- Залежна змінна (1 – благонадійний, 0 – неблагонадійний позичальник).

Розподіл залежної змінної наступний: 492 благонадійних позичальника з 999.

Модель була побудована у ПП SPSS Statistic. Програма надає результат у вигляді двох блоків: Block 0: Beginning Block та Block 1: Method = Enter. Перший блок передбачає попередній аналіз сукупності, а другий – результати застосування методу.

У таблиці Omnibus Tests of Model Coefficients відображаються результати оцінки якості наближення статистичної моделі (рис.6.2). Загальна значимість всієї моделі висока (Sig. <0,001). Тому побудовану модель слід визнати значущою і практично придатною.

**Omnibus Tests of Model Coefficients**

| | | Chi-square | df | Sig. |
|---|---|---|---|---|
| Step 1 | Step | 855,942 | 9 | ,000 |
| | Block | 855,942 | 9 | ,000 |
| | Model | 855,942 | 9 | ,000 |

*Рисунок. 6.2.* **Результати Омнімбус тесту для скорингової моделі**

Наступна таблиця Model Summary (рис.6.3) дозволяє оцінити частку сукупної дисперсії, що описується побудованої моделлю (величина R Square). Рекомендується використовувати величину Nagelkerke. Позитивним результатом можна вважати величину Nagelkerke R Square, що перевищує 0,50.

**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 528,741[a] | ,575 | ,767 |

*Рисунок. 6.3.* **Критерії адекватності для скорингової моделі**

Тобто, 76,7% варіації благонадійності позичальника пояснюється варіацією включених в модель факторів, а 23,3 – невідомими.

Наступним є результати класифікації (рис. 6.4), в якій реально спостережувані показники приналежності до тієї чи іншої з двох досліджуваних груп зіставляються з передбаченими на основі логістичної регресійної моделі. У нашому випадку з рядка Overall Percentage видно, що побудована модель дозволяє коректно

класифікувати 88,7% клієнтів. Також можна зробити відповідні висновки про коректність класифікації для кожної з двох розглянутих груп.

**Classification Table**[a]

| Observed | | | Predicted | | |
|---|---|---|---|---|---|
| | | | Благонадійність | | Percentage Correct |
| | | | Да | He| | |
| Step 1 | Благонадійність | Да | 428 | 64 | 87,0 |
| | | He| | 49 | 458 | 90,3 |
| | Overall Percentage | | | | 88,7 |

a. The cut value is ,500

*Рисунок. 6.4.* **Класифікаційна таблиця для скорингової моделі**

Для аналізу результатів класифікації необхідно також розрізняти поняття чутливість і специфічність моделі за формулами 6.9-6.12.

*Таблиця 6.2.* **Розрахункові значення істинно/хибно позитивних прикладів, чутливості та специфічності скорингової моделі**

| Показник | Значення, у % |
|---|---|
| TRP | 89,7 |
| FRP | 28,8 |
| Se | 89,7 |
| Sp | 71,2 |

Тобто розрахована скорингова модель з високою чутливістю (Se>Sp) відповідає політиці ризикованих кредитів, де мінімізується комерційний ризик, пов'язаний з упущеною вигодою.

З наступної рис. 6.5 можна з'ясувати статистичну значущість включених незалежних змінних, а також коефіцієнти регресійної функції. На підставі цих коефіцієнтів можна спрогнозувати

приналежність до певної групи кожного конкретного респондента у вибірці.

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 1[a] | Платежі | -,001 | ,000 | 109,367 | 1 | ,000 | 1,001 |
| | Стать(1) | ,669 | ,234 | 8,196 | 1 | ,004 | 1,953 |
| | Шлюб(1) | ,241 | ,227 | 1,128 | 1 | ,288 | 1,272 |
| | Забезпечення | -1,874 | ,177 | 112,247 | 1 | ,000 | 6,515 |
| | Дохід | ,001 | ,000 | 186,782 | 1 | ,000 | ,999 |
| | Досвід | -,003 | ,028 | ,014 | 1 | ,905 | ,997 |
| | Термін | -,009 | ,011 | ,785 | 1 | ,376 | ,991 |
| | Нерухомість | ,011 | ,006 | 3,858 | 1 | ,049 | ,989 |
| | Вік | ,003 | ,021 | ,016 | 1 | ,899 | 1,003 |
| | Constant | 2,858 | ,693 | 17,031 | 1 | ,000 | 17,431 |

a. Variable(s) entered on step 1: Платежі, Стать, Шлюб, Забезпечення, Дохід, Досвід, Термін, Нерухомість, Вік.

*Рисунок. 6.5.* **Коефіцієнти логістичної регресії для скорингової моделі**

З включених змінних статистично значимими (p-value<0,05) виявились платежі по кредиту, стать, кількість людей на забезпеченні, дохід, нерухомість та вільний член рівняння. Рівняння матиме наступний вигляд:

$$Z = 2{,}858 - 0{,}01 \cdot b_1 + 0{,}669 \cdot b_2 - 1{,}874 \cdot b_4 + 0{,}01 \cdot b_5 + \\ + 0{,}11 \cdot b_8$$

Можна зробити наступні висновки:

– чим менші місячні платежі, тим шанси отримати кредит більші в 1,001 рази

– за умови того, що клієнт чоловік, шанси отримати кредит більші в 1,953 рази

– чим менша кількість осіб на забезпечення, тим шанси отримати кредит більші в 6,515 разів

– чим вище дохід, тим шанси отримати кредит більше в 0,999 разів

– чим вища вартість нерухомості, тим шанси отримати кредит більші в 0,989 рази

Отже, побудовану логістичну регресійну модель можна використовувати на практиці для визначення благонадійності клієнтів у банку зібравши основні дані про кредитну історію позичальника.

### 6.4.2. Використання logit-регресійної моделі у медичних дослідженнях

Логістична регресія має найширше застосування саме у медичній статистиці. За допомогою цієї моделі розраховують імовірності результату лікування, імовірність виникнення ускладнень, побічних ефектів від ліків та ефективність медичного препарату.

Якщо міркувати в термінах медицини – завдання діагностики захворювання, де модель класифікації пацієнтів на хворих і здорових називається діагностичним тестом, то вийде наступне:

– чутливий діагностичний тест проявляється в гіпердіагностиці – максимальному запобіганні пропуску хворих;

– специфічний діагностичний тест діагностує лише достеменно хворих. Це важливо в разі, коли, наприклад, лікування хворого пов'язане з серйозними побічними ефектами і гіпердіагностика пацієнтів не бажана.

Часто логічним продовженням використання даної моделі є побудова логлінійних моделей – таблиць виживання.

У цій роботі було розглянуто результати ефективності препарату. До моделі були включені наступні показники:

– вік пацієнта, років

- стать пацієнта (ч/ж)
- рівень здоров'я (високий, середній, низький)
- надані ліки (новий препарат/існуючий препарат)
- доза (велика/мала)
- час прийняття ліків, год.
- залежна змінна (1 – є ефект, 0 – немає ефекту).

Розподіл залежної змінної наступний: 153 позитивних результати з 200.

У таблиці Omnibus Tests of Model Coefficients відображаються результати оцінки якості наближення статистичної моделі (рис.2.4). Загальна значимість всієї моделі висока (Sig. <0,001). Тому побудовану модель слід визнати значущою і практично придатною.

**Omnibus Tests of Model Coefficients**

| | | Chi-square | df | Sig. |
|---|---|---|---|---|
| Step 1 | Step | 54,728 | 6 | ,000 |
| | Block | 54,728 | 6 | ,000 |
| | Model | 54,728 | 6 | ,000 |

*Рисунок. 6.6.* **Результати Омнімбус тесту для моделі ефективності препарату**

Результати наступних розрахунків Model Summary (рис.6.7) дозволяють оцінити частку сукупної дисперсії, що описується побудованої моделлю (величина R Square). Позитивним результатом можна вважати величину Nagelkerke R Square, що перевищує 0,50.

**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 163,372[a] | ,239 | ,361 |

*Рисунок. 6.7.* **Критерії адекватності для моделі ефективності препарату**

Тобто, 36,1% варіації ефективності лікування пояснюється варіацією включених в модель факторів, а 63,9 – невідомими.

Результати класифікації, які представлені на рис. 6.8, наочно демонструють зіставлення показників приналежності до тієї чи іншої з двох досліджуваних груп, що спостерігаються, з передбаченими на основі логістичної регресійної моделі. У нашому випадку з рядка Overall Percentage видно, що побудована модель дозволяє коректно класифікувати 80,5% пацієнтів. Проте у медичних дослідження такий рівень класифікації є недостатньо високим, тому необхідно підвищити точність класифікації за допомогою збільшення кількості тих, кому ліки не допомогли чи включити інші предиктори.

**Classification Table[a]**

| Observed | | | Predicted | | |
|---|---|---|---|---|---|
| | | | Effect status | | Percentage Correct |
| | | | Censored | Taken effect | |
| Step 1 | Effect status | Censored | 19 | 28 | 40,4 |
| | | Taken effect | 11 | 142 | 92,8 |
| | Overall Percentage | | | | 80,5 |

a. The cut value is ,500

*Рисунок. 6.8.* **Класифікаційна таблиця для моделі ефективності препарату**

Для аналізу результатів класифікації необхідно також розрізняти поняття чутливість і специфічність моделі. Результати розрахунку яких подано у табл. 6.3

*Таблиця 6.3.* **Розрахункові значення істинно/хибно позитивних прикладів, чутливості та специфічності моделі ефективності препарату**

| Показник | Значення, у % |
|---|---|
| TRP | 63,3 |
| FRP | 16,5 |
| Se | 63,3 |
| Sp | 83,5 |

Тобто розрахована модель з високою специфічністю (Sp>Se) відповідає діагностуванню лише тих, для кого препарат є ефективним.

Рис. 6.9 наочно демонструє статистичну значущість включених незалежних змінних, а також коефіцієнти регресії. На підставі цих коефіцієнтів можна спрогнозувати приналежність до певної групи кожного конкретного респондента у вибірці.

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) | 95% C.I.for EXP(B) | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Lower | Upper |
| Step 1[a] | dosage | ,014 | ,400 | ,001 | 1 | ,973 | 1,014 | ,463 | 2,219 |
| | health | ,855 | ,294 | 8,433 | 1 | ,004 | 2,351 | 1,320 | 4,185 |
| | treatment | ,352 | ,400 | ,778 | 1 | ,378 | 1,423 | ,650 | 3,114 |
| | gender | ,092 | ,399 | ,053 | 1 | ,818 | 1,096 | ,501 | 2,398 |
| | age | ,011 | ,020 | ,302 | 1 | ,583 | 1,011 | ,973 | 1,051 |
| | time | -,519 | ,086 | 36,611 | 1 | ,000 | ,595 | ,503 | ,704 |
| | Constant | 1,077 | 1,307 | ,679 | 1 | ,410 | 2,936 | | |

a. Variable(s) entered on step 1: dosage, health, treatment, gender, age, time.

*Рисунок. 6.9.* **Коефіцієнти логістичної регресії для моделі ефективності препарату**

З включених змінних статистично значимими (p-value<0,05) виявились рівень здоров'я і час . Рівняння матиме наступний вигляд:

$$Z = 1{,}077 + 0{,}855 \cdot b_2 - 0{,}516 \cdot b_6$$

Можна зробити наступні висновки:

- чим кращий початковий стан здоров'я у пацієнта, тим більші в 2,351 разів шанси ефективної дії ліків;
- чим пізніший час прийняття препарату, тим менші в 0,599 разів шанси ефективної дії ліків

Отже, побудована логістична регресія потребує доопрацювання для підвищення ідентифікаційної здатності. Проте було виявлено важливі аспекти дії препарату – його можна приймати як дорослим, так і дітям, і чоловікам і жінкам, ефективність не залежить від величини дози, проте час прийняття є важливим.

### 6.4.3. Використання logit-регресійної моделі у контролі якості

Ще однією з можливих сфер застосування логістичної регресії є контроль якості на виробництві. Для даної роботи було обрано дані за результатами перевірки яблучного соку на ріст бактерій Аліциклобациллус.

Бактерії Аліциклобациллус це аеробні термофільні ацидофільні грампозитивні палички. Їх спори витримують вплив кислого середовища і високих температур, в тому числі, при пастеризації. Вперше аліциклобацілли виявили в гарячих джерелах, а пізніше в деревині і в сушених приквітниках гібіскуса (каркаде) і в соках з фруктів і овочів: груш, яблук, персиків, манго, апельсинів, малини, томатів. За деякими даними, одним з джерел Аліциклобациллус можуть бути харчові ароматизатори. Для людини аліциклобацілли не є небезпечними, проте вони викликають псування продукції. Багато

з цих бактерій, наприклад, Аліциклобациллус ацидотеррестріс і Аліциклобациллус ацидусалдаріус виробляють гваякол – речовину, яка додає фруктовим і овочевим сокам специфічний смак і запах, схожий з запахом диму або копчених продуктів. Відповідно до Технічного Регламенту Митного Союзу сокову продукцію з фруктів і овочів, в соках з абрикосів, персиків і груш необхідно перевіряти на вміст спороутворюючих аеробних і факультативно-анаеробних мікроорганізмів.

Головна мета моделі ідентифікувати зростання бактерій даної групи, для цього були включені наступні параметри:

– рівень pH;

– концентрація низину;

– температура;

– градус Brix (Брікс) – сахароза в рідині;

– залежна змінна (1 – є зростання Аліциклобациллус , 0 – немає зростання).

Розподіл залежної змінної наступний: 26 випадків зростання з 74.

Результати Omnibus Tests of Model Coefficients відображають оцінки якості наближення статистичної моделі (рис. 6.10). Загальна значимість всієї моделі висока (Sig. <0,001). Тому побудовану модель слід визнати значущою і практично придатною.

**Omnibus Tests of Model Coefficients**

| | | Chi-square | df | Sig. |
|---|---|---|---|---|
| Step 1 | Step | 43,615 | 4 | ,000 |
| | Block | 43,615 | 4 | ,000 |
| | Model | 43,615 | 4 | ,000 |

*Рисунок. 6.10.* **Результати Омнімбус тесту для моделі зростання бактерій**

Наступна таблиця Model Summary (рис. 6.11) дозволяє оцінити частку сукупної дисперсії, яка описується побудованої моделлю

(величина R Square). Позитивним результатом можна вважати величину Nagelkerke R Square, що перевищує 0,50.

**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 52,331[a] | ,445 | ,613 |

*Рисунок. 6.11.* **Критерії адекватності для моделі зростання бактерій**

Тобто, 61,3% варіації ефективності лікування пояснюється варіацією включених в модель факторів, а 38,7 – невідомими.

Наступними є результати класифікації (рис. 6.12), в якій реально спостережувані показники приналежності до тієї чи іншої з двох досліджуваних груп зіставляються з передбаченими на основі логістичної регресійної моделі.

**Classification Table[a]**

| Observed | | | Predicted | | |
|---|---|---|---|---|---|
| | | | VAR00005 | | Percentage Correct |
| | | | ,00 | 1,00 | |
| Step 1 | Зріст | ,00 | 42 | 6 | 87,5 |
| | | 1,00 | 6 | 20 | 76,9 |
| | Overall Percentage | | | | 83,8 |

a. The cut value is ,500

*Рисунок. 6.12.* **Класифікаційна таблиця для моделі зростання бактерій**

У нашому випадку з рядка Overall Percentage видно, що побудована модель дозволяє коректно класифікувати 83,8% партій

соку. Тобто розрахована модель з високою чутливістю (Se>Sp) дає істинний результат при наявності позитивного результату (виявляє позитивні приклади), що у нашому випадку – наявність зростання бактерій (табл. 6.4).

*Таблиця 6.4.* **Розрахункові значення істинно/хибно позитивних прикладів, чутливості та специфічності моделі зростання бактерій**

| Показник | Значення, у % |
|---|---|
| TRP | 87,5 |
| FRP | 23,1 |
| Se | 87,5 |
| Sp | 76,9 |

З наступної таблиці (рис. 6.13) можна з'ясувати статистичну значущість включених незалежних змінних, а також коефіцієнти регресійної функції. На підставі цих коефіцієнтів можна спрогнозувати приналежність до певної групи кожного конкретного респондента у вибірці.

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 1[a] | Ph | 1,886 | ,541 | 12,142 | 1 | ,000 | 6,593 |
| | Nisin | -,066 | ,019 | 12,105 | 1 | ,001 | ,936 |
| | Temprature | ,110 | ,048 | 5,362 | 1 | ,021 | 1,117 |
| | Brix | -,312 | ,143 | 4,741 | 1 | ,029 | ,732 |
| | Constant | -7,246 | 3,219 | 5,069 | 1 | ,024 | ,001 |

a. Variable(s) entered on step 1: Ph, Nisin, Temprature, Brix.

*Рисунок. 6.13.* **Коефіцієнти логістичної регресії для моделі зростання бактерій**

З включених змінних всі виявилися статистично значимими (p-value<0,05). Рівняння матиме наступний вигляд:

$$Z = -7{,}246 + 1{,}886 \cdot b_1 - 0{,}66 \cdot b_2 + 0{,}110 \cdot b_3 - 0{,}312 \cdot b_4$$

Можна зробити наступні висновки:

– чим вищий рівень pH, тим в 6,593 разів більше шансів зростання бактерій Аліциклобациллус;

– чим більший рівень низину, тим в 0,936 разів менше шансів зростання зростання бактерій Аліциклобациллус;

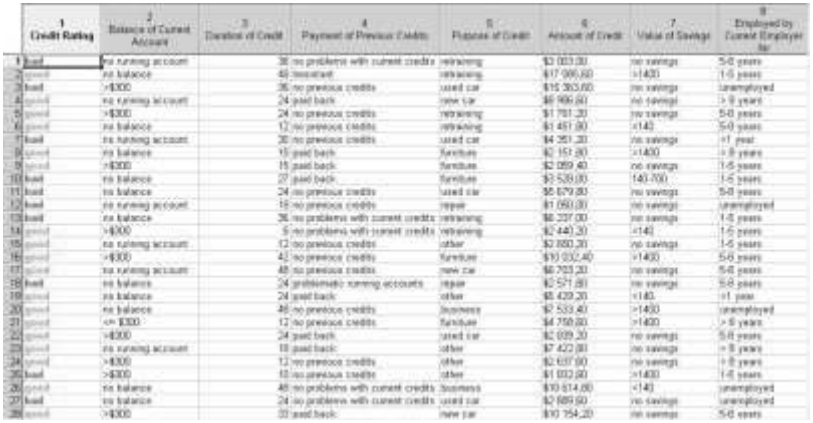
– чим вища температура, тим в 1,117 разів більше шансів зростання бактерій Аліциклобациллус;

– чим вищий градус Брікса, тим в 0,732 разів менше шансів зростання бактерій Аліциклобациллус;

Отже, отримана модель дозволяє здійснювати контроль за якістю харчової продукції, а саме якістю яблучного соку та регулювати рівень вищезазначених факторів щоб запобігти виникненню бактерій та псуванню продукції.

### 6.4.4. Використання probit-регресійної моделі в аналізі впливу факторів на кредитний рейтинг позичальників

Для аналізу були взяті дані, що наведені у файлі CreditScoring.sta (Statistica/Examples/) – фрагмент якого представлений на рис. 6.14.



| | 1 Credit Rating | 2 Balance of Current Account | 3 Duration of Credit | 4 Payment of Previous Credits | 5 Purpose of Credit | 6 Amount of Credit | 7 Value of Savings | 8 Employed by Current Employer for |
|---|---|---|---|---|---|---|---|---|
| 1 | bad | no running account | 36 | no problems with current credits | retraining | $2 003,00 | no savings | 5-8 years |
| 2 | good | no balance | 48 | insolvent | retraining | $17 999,60 | <1400 | 1-5 years |
| 3 | bad | >4000 | 36 | no previous credits | used car | $15 383,60 | no savings | unemployed |
| 4 | good | no running account | 24 | paid back | new car | $8 998,60 | no savings | > 8 years |
| 5 | good | >4000 | 24 | no previous credits | retraining | $1 781,20 | no savings | 5-8 years |
| 6 | good | no balance | 12 | no previous credits | retraining | $1 481,80 | <140 | 5-8 years |
| 7 | bad | no running account | 30 | no previous credits | used car | $4 351,20 | no savings | <1 year |
| 8 | good | no balance | 15 | paid back | furniture | $2 157,80 | >1400 | > 8 years |
| 9 | good | >4000 | 15 | paid back | furniture | $2 099,40 | no savings | 1-5 years |
| 10 | bad | no balance | 27 | paid back | furniture | $3 529,00 | 140-700 | 1-5 years |
| 11 | bad | no balance | 24 | no previous credits | used car | $5 679,80 | no savings | 5-8 years |
| 12 | bad | no running account | 18 | no previous credits | repair | $1 050,20 | no savings | unemployed |
| 13 | bad | no balance | 36 | no problems with current credits | retraining | $6 207,00 | no savings | 1-5 years |
| 14 | good | >4000 | 9 | no problems with current credits | retraining | $2 440,20 | <140 | 1-5 years |
| 15 | good | no running account | 12 | no previous credits | other | $2 850,20 | no savings | 1-5 years |
| 16 | good | >4000 | 42 | no previous credits | furniture | $10 032,40 | >1400 | 5-8 years |
| 17 | good | no running account | 48 | no previous credits | new car | $6 703,20 | no savings | 5-8 years |
| 18 | bad | no balance | 24 | problematic running accounts | repair | $2 571,80 | no savings | 5-8 years |
| 19 | good | no balance | 24 | paid back | other | $5 429,20 | <140 | <1 year |
| 20 | good | no balance | 48 | no previous credits | business | $7 533,40 | >1400 | unemployed |
| 21 | good | <= 1000 | 12 | no previous credits | furniture | $4 758,80 | >1400 | > 8 years |
| 22 | good | >4000 | 24 | paid back | used car | $2 039,20 | no savings | 5-8 years |
| 23 | good | no running account | 18 | paid back | other | $7 422,80 | no savings | > 8 years |
| 24 | good | >4000 | 12 | no previous credits | other | $2 937,60 | no savings | > 8 years |
| 25 | bad | >4000 | 18 | no previous credits | other | $1 002,60 | >1400 | 1-5 years |
| 26 | good | no balance | 48 | no problems with current credits | business | $10 874,00 | <140 | unemployed |
| 27 | bad | no balance | 24 | no problems with current credits | used car | $2 689,60 | no savings | unemployed |
| 28 | good | >4000 | 33 | paid back | new car | $10 764,20 | no savings | 5-8 years |

*Рисунок. 6.14.* **Фрагмент бази даних, що характеризує кредитний рейтинг позичальників CreditScoring.sta**

База даних містить інформацію про 1000 позичальників, які характеризуються 19 ознаками (рис.6.15).

| | Name | Type | MD code | Length | Long Name (label or formula) |
|---|---|---|---|---|---|
| 1 | Credit Rating | Double | -9999 | | |
| 2 | Balance of Current Account | Double | -9999 | | |
| 3 | Duration of Credit | Double | -9999 | | 'Duration of Credit in Months |
| 4 | Payment of Previous Credits | Double | -9999 | | |
| 5 | Purpose of Credit | Double | -9999 | | |
| 6 | Amount of Credit | Double | -9999 | | 'Amount of Credit in Dollars |
| 7 | Value of Savings | Double | -9999 | | 'Value of Savings in Dollars |
| 8 | Employed by Current Employer for | Double | -9999 | | |
| 9 | Installment in % of Available Income | Double | -9999 | | |
| 10 | Marital Status | Double | -9999 | | |
| 11 | Gender | Double | -9999 | | |
| 12 | Living in Current Household for | Double | -9999 | | |
| 13 | Most Valuable Assets | Double | -9999 | | |
| 14 | Age | Double | -9999 | | |
| 15 | Further running credits | Double | -9999 | | |
| 16 | Type of Apartment | Double | -9999 | | |
| 17 | Number of previous credits at this bank | Double | -9999 | | |
| 18 | Occupation | Double | -9999 | | |
| 19 | TrainTest | Double | -9999 | | =rnd(1)<=0,67 |

*Рисунок. 6.15.* **Характеристика кредитного рейтингу позичальників бази даних CreditScoring.sta**

Результативна змінна – Credit Rating, являє собою якісну бінарну змінну, яку з метою подальшого аналізу було перетворено і представлено через 0 та 1:

| Кредитний рейтинг | Кількісне представлення змінної |
|---|---|
| bad | 0 |
| good | 1 |

База даних містить лише 3 кількісних змінних – строк кредиту, сума кредиту та вік позичальника. Для цих змінних розрахуємо описові статистики, перевіримо на нормальність розподілу, за допомогою статистики Колмогорова-Смірнова, та наявність екстремальних викидів.

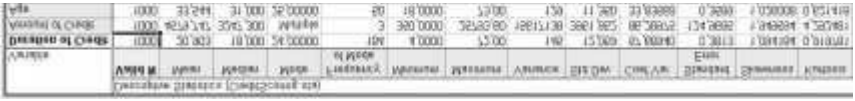


*Рисунок. 6.16.* **Статистична характеристика неперервних змінних**

За результатами розрахунків, представлених на рис. 6.16, можна констатувати наступне:

- Середня тривалість кредиту становить 20 місяців. Найкоротший строк на який було взято кредит складав 4 місяці, а найдовший – 72. Розподіл має правобічну асиметрію. Вибірка не є однорідною.
- Сума кредиту коливається від 350 до, майже, 26 000 дол. Середгє значення становить 4 579,7 дол. Також, розподілу притаманна правобічна асиметрія. Вибірка не є однорідною.
- Вік дебіторів коливається від 18 до 73 років. Середній вік становить 34 роки. Характерна правобічна асиметрія. Вибірка не є однорідною.

Нижче наведені графіки розподілу числових ознак (рис. 6.17):

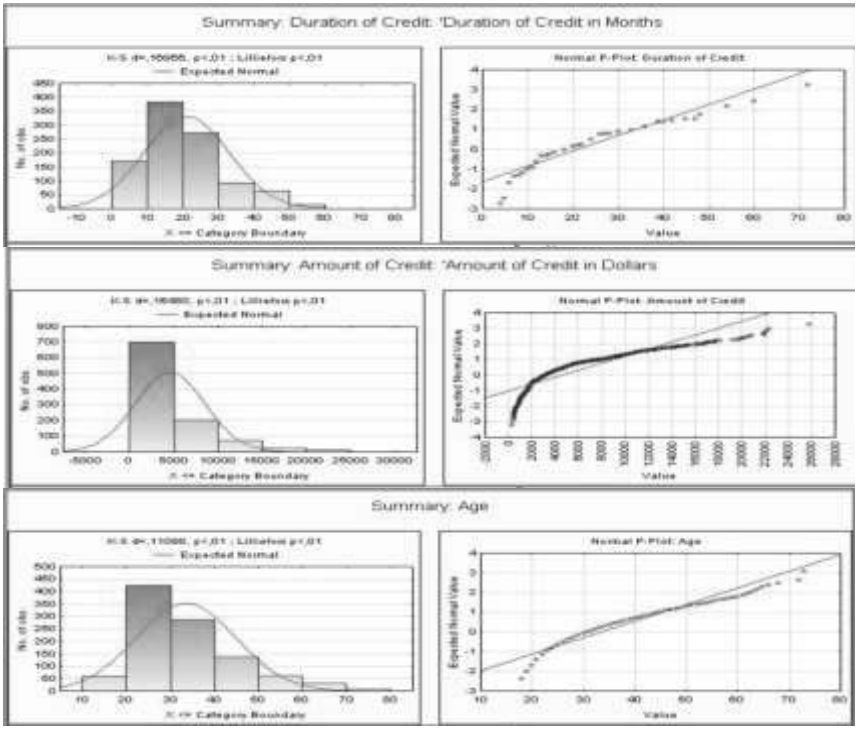



*Рисунок. 6.17.* **Графічне зображення розподілів неперервних змінних**

Кількість ступенів свободи: N=1000-1=999, з рівнем значущості 0,05:

$d_t = \frac{1{,}36}{\sqrt{999}} = 0{,}043 > d_p = 0{,}16866$, тобто гіпотеза про нормальність розподілу строку кредиту не відхиляється.

$d_t = 0{,}043 > d_p = 0{,}16480$ - гіпотеза про нормальність розподілу суми кредиту не відхиляється.

$d_t = 0{,}043 > d_p = 0{,}11098$ – Вік, також можна вважати нормально розподіленим.

Для аналізу викидів було побудовано коробку з вусами (рис. 6.18). Як видно з рисунку, для тривалості кредиту та віку дебітора характерні помірні викиди, а от для суми кредиту – екстремальні викиди.

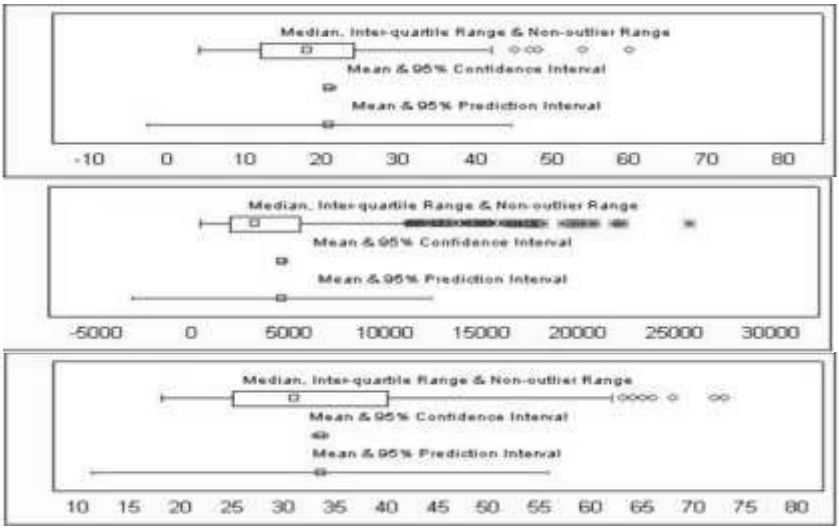



*Рисунок. 6.18.* **Коробки з вусами для неперервних змінних**

Побудувавши матрицю кореляції для вищезазначених неперервних змінних (рис. 6.19), робимо висновок про наявність прямого помірного зв'язку між тривалістю кредиту та його розміром:

| Variable | Correlations (CreditScoring.sta)<br>Marked correlations are significant at p < ,05000<br>N=1000 (Casewise deletion of missing data)<br>**Means** | Std.Dev. | Age | Duration of Credit | Amount of Credit |
|---|---|---|---|---|---|
| **Age** | 33,544 | 11,350 | 1,000000 | -0,037711 | 0,032169 |
| Duration of Credit | 20,903 | 12,059 | -0,037711 | 1,000000 | 0,624988 |
| Amount of Credit | 4579,747 | 3951,852 | 0,032169 | 0,624988 | 1,000000 |

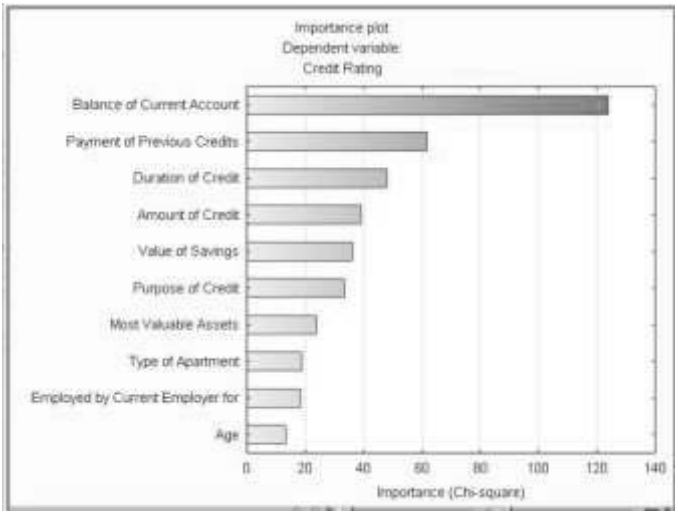
*Рисунок. 6.19.* **Матриця кореляції для неперервних змінних**

Наступним кроком виділимо найбільш значущі предиктори впливу на кредитний рейтинг, для подальшого аналізу, на основі розрахунку статистики $\chi 2$. З цією метою використаємо модуль (Feature Selection and Variable Screening).

Як показують результати аналізу (рис. 6.20), фактори, що представлені змінними: «баланс на поточному рахунку», «виплати по кредитах», «строк кредиту», «сума кредиту», «ціль кредиту», «найбільш цінні активи», «тип житла» та «працевлаштування на поточній роботі» виділяються як найбільш важливі, з рівнем істотності 0,05.

В пробіт-модель імовірності підвищення кредитного рейтингу в якості залежної змінної виступає фактор кредитного рейтингу, який приймає значення 0 та 1, де 0 – негативний, 1 – позитивний.

На основі попереднього розвідувального аналізу було визначено, що у модель необхідно включити змінні «баланс на поточному рахунку», «виплати по кредитах», «строк кредиту», «сума кредиту», «ціль кредиту», «найбільш цінні активи», «тип житла» та «працевлаштування на поточній роботі». При спробі включення у модель змінних «сума кредиту», «ціль кредиту», «найбільш цінні активи», «тип житла» та «працевлаштування на поточній роботі» програмою Statistica було зазначено, що данні змінні є надлишковим, а оцінки зміщеними, тому включення цих факторів у модель не є доцільними.




| | Best predictors for catec | |
|---|---|---|
| | **Chi-square** | p-value |
| **Balance of Current Account** | 123,7209 | 0,000000 |
| Payment of Previous Credits | 61,6914 | 0,000000 |
| Duration of Credit | 47,9200 | 0,000000 |
| Amount of Credit | 39,0862 | 0,000001 |
| Value of Savings | 36,0989 | 0,000000 |
| Purpose of Credit | 33,3564 | 0,000116 |
| Most Valuable Assets | 23,7196 | 0,000029 |
| Type of Apartment | 18,6740 | 0,000088 |
| Employed by Current Employer for | 18,3683 | 0,001045 |
| Age | 13,2465 | 0,066325 |

*Рисунок. 6.20.* **Результати розрахунків визначення найбільш значущих предикторів впливу на кредитний рейтинг**

За результатами моделювання були отримані оцінки, що представлені на рис. 6.21.

| | Model: Probit regression N of 0's: 300 1's: 700 (CreditScoring.sta)<br>Dep. var: Credit Rating Loss: Max likelihood<br>Final loss: 513,22193924 Chi?( 3)=195,28 p=0,0000 | | | |
|---|---|---|---|---|
| N=1000 | **Const.B0** | Balance of Current Account | Payment of Previous Credits | Duration of Credit |
| **Estimate** | -0,434334 | 0,368401 | 0,220596 | -0,021638 |

*Рисунок. 6.21.* **Результати розрахунків пробіт-моделі**

Оскільки *p-value*<0,05, то це свідчить про адекватність побудованої моделі. Наступним кроком для інтерпретації отриманих коефіцієнтів необхідно розрахувати граничні ефекти (рис. 6.22, табл. 6.5).

| | Means and Standard Deviations (CreditScoring.sta) | | | |
|---|---|---|---|---|
| Variable | **Mean** | Std.Dev. | Minimum | Maximum |
| **Balance of Current Account** | 2,57700 | 1,25764 | 1,000000 | 4,00000 |
| Payment of Previous Credits | 2,54500 | 1,08312 | 0,000000 | 4,00000 |
| Duration of Credit | 20,90300 | 12,05881 | 4,000000 | 72,00000 |
| Credit Rating | 0,70000 | 0,45849 | 0,000000 | 1,00000 |

*Рисунок. 6.22.* **Результати розрахунків описових статистик**

*Таблиця 6.5.* **Розрахунок граничних ефектів: 1 етап**

| Змінна | b | середнє | b × середнє |
|---|---|---|---|
| Баланс на рахунку | 0,36840 | 2,57700 | 0,94937 |
| Виплати по попередніх кредитах | 0,22060 | 2,54500 | 0,56142 |
| Тривалість кредиту | -0,02164 | 20,90300 | -0,45230 |
| Константа | -0,43433 | 1,00000 | -0,43433 |
| Всього | x | x | **0,62415** |

Отримане значення z підставляємо у формулу нормального розподілу:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-0,5z^2} = 0,39894 * 2,71828^{-0,5*0,39} = 0,39894 * 0,823 = 0,3283.$$

*Таблиця 6.6.* **Розрахунок граничних ефектів: 2 етап**

| Змінна | b | f(x) | b` |
|---|---|---|---|
| Баланс на рахунку | 0,36840 | 0,32830 | 0,12095 |
| Виплати по попередніх кредитах | 0,22060 | 0,32830 | 0,07242 |
| Тривалість кредиту | -0,02164 | 0,32830 | -0,00710 |
| Константа | -0,43433 | 0,32830 | -0,14259 |

Таким чином, зростання балансу на рахунку дебітора сприяє збільшенню імовірності присвоєння позитивного кредитного рейтингу на 12,09 %. За умов відсутності проблем з виплатами по попередніх кредитах – ймовірність позитивного рейтингу зростає на 7,24%.

*ROC-аналіз чутливості та специфічності*

Для побудованої вище пробіт-моделі проведемо ROC-аналіз. Таблиця класифікації має вигляд:

*Таблиця 6.7.* **Розрахунок граничних ефектів**

| Classification of Cases (CreditScoring.sta)<br>Odds ratio: 5,6274 Perc. correct: 74,60% | | | |
|---|---|---|---|
| | Pred. - Bad | Pred. - Good | Percent - Correct |
| Bad | 122 (TN) | 178(FP) | 40,66667 |
| Good | 76 (FN) | 624(TP) | 89,14286 |

Чутливість моделі $= TRP = \frac{TP}{TP+FN} = \frac{624}{624+76} = 89{,}14\%$.

Специфічність моделі $= 1 - FPR = \frac{TN}{TN+FP} = \frac{122}{122+178} =$ 40,67%.

Чутливість більша за специфічність, що свідчить про те, що модель правильно ідентифікує позитивні приклади, тобто наявна лояльна політика ідентифікації дебіторів.

Загальний показник розпізнавальних можливостей моделей:

$DS = \frac{TP+TN}{TP+FN+TN+FP} = \frac{624+122}{624+76+122+178} = 74{,}6\%$, що свідчить про високі розпізнавальні здібності пробіт-моделі.

Відношення шансів: $W = \frac{TP \cdot TN}{FN \cdot FP} = \frac{624 \cdot 122}{76 \cdot 178} = 5{,}63$ – індикатор кількісної міри тісноти зв'язку двох ознак – таким чином, можемо стверджувати, що ймовірність правильно присвоїти кредитний рейтинг у 5,63 рази більша, ніж зробити це хибно.

### *Список завдань до самоконтролю:*

**1.** За даними CreditScoring.sta про кредитний рейтинг позичальників банку провести аналіз впливу факторів на кредитний рейтинг позичальників, використовуючи:

- розвідувальний аналіз даних.
- логіт-модель імовірності підвищення кредитного рейтингу.
- ROC-аналіз чутливості та специфічності.

**2.** За даними завдання 1 провести аналіз впливу факторів на кредитний рейтинг позичальників, використовуючи пробіт-модель імовірності підвищення кредитного рейтингу, здійснити ROC-аналіз чутливості та специфічності, зробити висновки та порівняти отримані результати.

**3.** За даними CreditScoring2.sta про тип ризику позичальників банку (змінна «Type of Risk»: «Bad» або «Good») провести аналіз впливу факторів на тип ризику позичальників, використовуючи:

1) розвідувальний аналіз даних.
2) логіт-модель імовірності підвищення кредитного рейтингу.
3) ROC-аналіз чутливості та специфічності.

**4.** За даними завдання 3 провести аналіз впливу факторів на на тип ризику позичальників, використовуючи пробіт-модель імовірності підвищення кредитного рейтингу, здійснити ROC-аналіз чутливості та специфічності, зробити висновки та порівняти отримані результати.

**5.** За даними про службовців адміністративного відділу (див. табл. 6.1, дод. 6) провести аналіз впливу факторів на підвищення заробітної плати, використовуючи наступні методи:

1) Побудувати логіт-модель і пробіт-модель імовірності підвищення заробітної плати. Порівняти отримані результати.

2) Провести ROC-аналіз чутливості та специфічності.

*Таблиця 6.1.* **Дані про службовців адміністративного відділу**

| № | Заробітна плата, дол. | Стать | Вік | Стаж | Рівень підготовки* | Підвищення заробітної плати** | Підвищення посади** |
|---|---|---|---|---|---|---|---|
| 1 | 37360 | Ж | 42 | 3 | B | 0 | 0 |
| 2 | 53174 | М | 54 | 10 | B | 1 | 0 |
| 3 | 52722 | М | 47 | 10 | A | 0 | 0 |
| 4 | 53423 | М | 47 | 1 | B | 0 | 0 |
| 5 | 50602 | М | 44 | 5 | B | 1 | 0 |
| 6 | 49033 | М | 42 | 10 | A | 1 | 1 |
| 7 | 24395 | М | 30 | 5 | A | 0 | 0 |
| 8 | 24395 | Ж | 52 | 6 | A | 0 | 0 |
| 9 | 43124 | М | 48 | 8 | A | 1 | 0 |
| 10 | 23975 | Ж | 58 | 4 | A | 0 | 0 |
| 11 | 53174 | М | 46 | 4 | C | 1 | 1 |

| № | Заробітна плата, дол. | Стать | Вік | Стаж | Рівень підготовки* | Підвищення заробітної плати** | Підвищення посади** |
|---|---|---|---|---|---|---|---|
| 12 | 58515 | М | 36 | 8 | С | 1 | 0 |
| 13 | 56194 | М | 49 | 10 | В | 1 | 1 |
| 14 | 49033 | Ж | 55 | 10 | В | 1 | 1 |
| 15 | 44884 | М | 41 | 1 | А | 0 | 0 |
| 16 | 53479 | Ж | 52 | 5 | В | 1 | 1 |
| 17 | 46574 | М | 57 | 8 | А | 0 | 0 |
| 18 | 58968 | Ж | 61 | 10 | В | 1 | 1 |
| 19 | 53174 | М | 50 | 5 | А | 0 | 0 |
| 20 | 53627 | М | 47 | 10 | В | 1 | 1 |
| 21 | 49033 | М | 54 | 5 | В | 1 | 0 |
| 22 | 54981 | М | 47 | 7 | А | 0 | 0 |
| 23 | 62530 | М | 50 | 10 | В | 1 | 1 |
| 24 | 27525 | Ж | 38 | 3 | А | 0 | 0 |
| 25 | 24395 | М | 31 | 5 | А | 0 | 0 |
| 26 | 56884 | М | 47 | 10 | А | 1 | 1 |
| 27 | 52111 | М | 56 | 5 | А | 0 | 0 |
| 28 | 44183 | Ж | 38 | 5 | В | 0 | 0 |
| 29 | 24967 | Ж | 55 | 6 | А | 0 | 0 |
| 30 | 35423 | Ж | 47 | 4 | А | 0 | 0 |
| 31 | 41188 | Ж | 35 | 2 | В | 0 | 0 |
| 32 | 27525 | Ж | 35 | 3 | А | 0 | 0 |
| 33 | 35018 | М | 39 | 1 | А | 0 | 0 |
| 34 | 44183 | М | 41 | 2 | А | 0 | 0 |
| 35 | 35423 | М | 44 | 1 | А | 0 | 0 |
| 36 | 43033 | М | 53 | 8 | А | 0 | 0 |

| № | Заробітна плата, дол. | Стать | Вік | Стаж | Рівень підготовки* | Підвищення заробітної плати** | Підвищення посади** |
|---|---|---|---|---|---|---|---|
| 37 | 40741 | М | 47 | 2 | A | 0 | 0 |
| 38 | 49033 | М | 42 | 10 | A | 1 | 1 |
| 39 | 56294 | Ж | 44 | 6 | C | 1 | 0 |
| 40 | 47180 | Ж | 45 | 5 | C | 1 | 0 |
| 41 | 46574 | М | 56 | 8 | A | 1 | 1 |
| 42 | 52722 | М | 33 | 8 | C | 1 | 1 |
| 43 | 51237 | М | 53 | 2 | B | 0 | 0 |
| 44 | 53627 | М | 52 | 8 | A | 1 | 1 |
| 45 | 53174 | М | 54 | 10 | A | 1 | 1 |
| 46 | 56234 | М | 49 | 10 | B | 1 | 1 |
| 47 | 49033 | Ж | 53 | 10 | B | 1 | 1 |
| 48 | 40033 | М | 43 | 9 | A | 1 | 1 |
| 49 | 55549 | М | 35 | 8 | C | 1 | 1 |
| 50 | 51237 | М | 56 | 1 | C | 1 | 0 |
| 51 | 35200 | Ж | 38 | 1 | B | 0 | 0 |
| 52 | 50174 | Ж | 42 | 5 | A | 0 | 0 |
| 53 | 24352 | Ж | 35 | 1 | A | 0 | 0 |
| 54 | 27525 | Ж | 40 | 3 | A | 0 | 0 |
| 55 | 29606 | Ж | 34 | 4 | B | 1 | 0 |
| 56 | 24352 | Ж | 35 | 1 | A | 0 | 0 |
| 57 | 47180 | Ж | 45 | 5 | B | 0 | 0 |
| 58 | 49333 | М | 54 | 10 | A | 1 | 1 |
| 59 | 53174 | М | 47 | 10 | A | 1 | 1 |
| 60 | 53429 | Ж | 45 | 7 | B | 1 | 0 |
| 61 | 53627 | М | 47 | 10 | A | 1 | 1 |

| № | Заробітна плата, дол. | Стать | Вік | Стаж | Рівень підготовки* | Підвищення заробітної плати** | Підвищення посади** |
|---|---|---|---|---|---|---|---|
| 62 | 26491 | Ж | 46 | 7 | A | 1 | 0 |
| 63 | 42961 | М | 36 | 3 | B | 1 | 0 |
| 64 | 53174 | М | 45 | 5 | A | 0 | 0 |
| 65 | 36292 | М | 46 | 0 | A | 0 | 0 |
| 66 | 37292 | М | 47 | 1 | A | 0 | 0 |
| 67 | 41188 | Ж | 34 | 3 | B | 0 | 0 |
| 68 | 57242 | Ж | 45 | 7 | C | 1 | 0 |
| 69 | 53429 | Ж | 44 | 6 | C | 1 | 0 |
| 70 | 53174 | М | 50 | 10 | B | 1 | 1 |
| 71 | 44138 | Ж | 38 | 2 | B | 0 | 0 |

*) періодично добровільно службовцю пропонують пройти курс перепідготовки: службовці, що не пройшли перепідготовку, отримують кваліфікацію «А», після проходження одного курсу – кваліфікацію «В», а після другого заключного курсу – «С»;
**) 0 – не підвищено; 1 – підвищено.

### *Список рекомендованої літератури по темі:*

1. Davis J., Goadrich M. The Relationship Between Precision-Recall and ROC Curves // Proc. Of 23 International Conference on Machine Learning, Pittsburgh, PA, 2006
2. Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers // 2004 Kluwer Academic Publishers.

3. Pain medicine testing [Електронний ресурс] // Center for Machine Learning and Intelligent Systems – Режим доступу: http://cml.ics.uci.edu/
4. Paul O' Dea and Josephine Griffith and Colm O' Riordan. Combining Feature Selection and Neural Networks for Solving Classification Problems.
5. Spermann A. The Probit Model [Електронний ресурс] / A. Spermann – Режим доступу: https://www.empiwifo.uni-freiburg.de/lehre-teaching-1/summer-term-09/materials-microeconometrics/probit_7-5-09.pdf – Назва з екрана.
6. W.E.L. Pena, P.R. De Massaguer, A.D.G. Zuniga, and S.H. Saraiva (2011). "Modeling the Growth Limit of Alicyclobacillus Acidoterrestris CRA7152 in Apple Juice: Effect of pH, Brix, Temperature, and Nisin Concentration," Journal of Food Processing and Preservation, Vol. 35, pp. 509–517.
7. Zaghdoudi T. Bank Failure Prediction with Logistic Regression International Journal of Economics and Financial Issues, Vol. 3, No. 2, 2013, pp.537–543
8. Zweig M.H., Campbell G. ROC Plots: A Fundamental Evaluation Tool in Clinical Medicine // Clinical Chemistry, Vol. 39, No. 4, 1993.
9. Барсегян А.А., Купріянов М.С., Степаненко В.В., Холод І.І. Технології аналізу даних: Data Mining, Visual Mining, Text Mining, OLAP. – СПб .: БХВ-Петербург, 2007. – 384 с.
10. Бююль А., Цёфель П. SPSS: Мистецтво обробки інформації. Аналіз статистичних даних і відновлення прихованих закономірностей: Пер. з нім. – СПб .: ДиаСофтЮП, 2005. – 608 с.
11. Модуль Regression [Електронний ресурс] // IBM Knowledge Center – Режим доступу: https://www.ibm.com/support/knowledgecenter/

ru/SSLVMB_22.0.0/kc_gen/com.ibm.spss.statistics.help_statistics_mainhelpgen13.html

12. Паклин Н.Б., Орєшков В.І. Бізнес-аналітика: від даних до знань: Учеб. допомога. 2-е изд., Перераб. і доп. – СПб .: Пітер, 2010. – 704 с.
13. Помазанов М. Кількісний аналіз кредитного ризику / М. Помазанов // Банківські технології. – 2004. – № 2. – С. 22–28.
14. Циплаков А.А. Деякі економетричні методи. Метод максимальної правдоподібності в економетрії. Навчальний посібник.
15. Чубукова І.А. Data mining: навчальний посібник – М .: Інтернет-університет інформаційних технологій: БИНОМ: Лабораторія знань, 2006. – 382 с.

# Розділ 7. НЕЛІНІЙНІ МОДЕЛІ У СТАТИСТИЧНОМУ АНАЛІЗІ

## 7.1 Структурна модель статистичної залежності

Дослідження нелінійних (реальних) процесів і явищ в системах різної природи завжди пов'язано з виникненням задачі вивчення та оцінювання взаємозалежності їх параметрів, характеристик, показників, інформаційних потоків з метою отримання нових достовірних даних для потреб моделювання, прогнозування, прийняття адекватних рішень, нових планів, стратегій, знань тощо.

Для розглядання такої задачі можна застосувати відомий принцип «Чорного ящика», запропонований англійським кібернетиком У.Р. Ешбі. На рис. 7.1 показана загальна структурна модель статистичної залежності: рис. 7.1а – детермінованої, рис. 7.1б – стохастичної. Змінні *X, Y, E* визначають відповідно вхідний вплив (*X*), результат впливу (*Y*) та випадковий вплив (*E*) і подаються у вигляді векторних змінних різної розмірності:

$$X = (x_1, x_2, \cdots, x_n)^T,\ Y = (y_1, y_2, ..., y_m)^T,\ E = (e_1, e_2, ..., e_k)^T.$$

Крім того, компоненти векторів *X, Y, E* можуть бути функціями від часу, тобто представляти собою часові процеси.

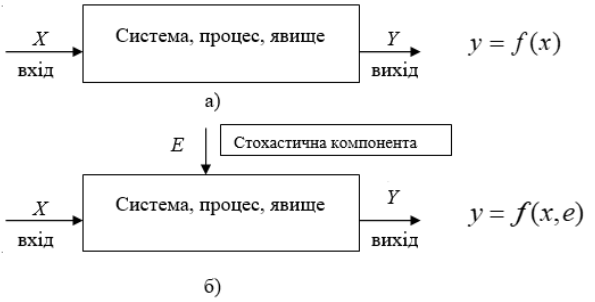



*Рисунок 7.1.* **Загальна структурна модель статистичної залежності: а – детермінованої, б – стохастичної**

В моделі, рис. 7.1, змінна *X* описує умови функціонування об'єкта і в різних задачах має назву: незалежна, екзогенна змінна; факторна ознака; предиктор або регресор.

Вихідна змінна *Y* характеризує поведінку, результат функціонування системи і в статистичних моделях її називають: залежна, ендогенна, результуюча змінна; відгук; регресанд.

Вектор *E* (зовнішній, випадковий вплив) складається з латентних (прихованих) стохастичних компонентів, які відображають вплив неврахованих (невизначених, невідомих) факторів, а також випадкові помилки при вимірюванні показників, що аналізуються.

В задачах статистичного аналізу компоненти векторів *X, Y, E* можуть бути різних типів, що суттєво впливає на вибір методів і моделей: *кількісні*, які можна вимірювати за визначеною шкалою; *порядкові*, що дозволяють упорядковувати дані за якісною ознакою; *класифікаційні*, що поділяють сукупність об'єктів (даних) на однорідні за певною властивістю.

В процесі проведення математико–статистичного моделювання необхідно також враховувати типи взаємозв'язку змінних: *рівноправний* (двосторонній) або *нерівноправний* (причина – наслідок), а також види даних: *перехресні вибірки* даних та *часові ряди*. Перехресними є дані за якимось показником, що отримані для різних однотипних об'єктів (підприємств, областей, регіонів) в один момент часу (тобто часова приналежність несуттєва). Часовий ряд характеризує один об'єкт в різні моменти часу. Послідовні значення часових рядів можуть бути певним чином пов'язані між собою, що визначаються закономірними відхиленнями від загальної тенденції розвитку, або виявляються часові лаги. Тому методи обробки перехресних і часових даних дещо відрізняються.

База статистичних даних основується на спостереженнях. Формуючи спостереження, слід забезпечити порівнянність даних у просторі та часі. Тому початкові дані (спостереження, факторні

ознаки) повинні підпорядковуватись вимогам однаковості за: ступенем агрегування та однорідністю структури одиниць вибірки, методами розрахунку показників у часі та просторі, періодичністю обліку окремих змінних та зовнішніми умови. Кожний із факторів $X_i$ має бути значущим та обґрунтованим теоретично та їх кількість не повинна перевищувати однієї третини числа спостережень у вибірці (довжини часового ряду), що суттєво впливає на досліджувані показники *Y*. Крім того, потрібно враховувати та оцінювати можливість виникнення таких явищ, як мультиколінеарність та автокореляція.

Отже, в процесі проведення статистичного аналізу та математико-статистичного моделювання слід враховувати різні типи (компоненти векторів *X, Y*) та види (перехресні, часові) даних, типи взаємозв'язку (рівноправний, нерівноправний), види залежності, зв'язку між явищами та процесами (функціональний, стохастичний, кореляційний), а також, тип задачі дослідження: однофакторна $y = f(x)$ та багатофакторна $y = f(x_1, x_2, \ldots, x_n)$.

Загальну задачу статистичного аналізу можна сформулювати так (рис. 7.1а): на основі спостережень (вимірювань) змінних *X* та *Y* побудувати таку функцію $f(x)$, що найкраще буде відновлювати значення регресанду *Y*, відповідно заданим значенням регресора *X*.

Розв'язок даної задачі передбачає вибір математичного виразу для опису залежності $y = f(x)$ та критерію якості апроксимації, відповідно до якого буде визначатись найкращий спосіб відновлення значень *Y*.

Однак, перш ніж приступати до розв'язування такої задачі з використанням статистичних методів та будувати відповідні математичні моделі, потрібно визначити прикладну мету (цілі) проведення статистичного аналізу і моделювання.

## 7.2 Апроксимація даних нелінійними функціями

Ясність прикладної цілі визначає послідовність виконання різних етапів проведення статистичного аналізу, вибір загальної структури функції $f(x)$ та моделі, інтерпретацію отриманих результатів моделювання. До типових кінцевих цілей проведення аналізу та математико–статистичного моделювання можна віднести:

1 - встановлення факту наявності (відсутності) статистично значимого зв'язку між змінними *Y* та *X*;

2 - відновлення (прогноз) значень регресанду *Y* відповідно заданим значенням регресора *X*;

3 - виявлення причинного зв'язку між змінними *X* та *Y*.

Мета 1 не потребує розв'язування задачі вибору виду функції $f(x)$ та її побудови, а забезпечується методами кореляційного аналізу, які залежать від природи змінних (кількісні, порядкові, класифікаційні), вибраного показника статистичної залежності (коефіцієнт або індекс кореляції , ранговий коефіцієнт кореляції тощо), постановки конкретної задачі: оцінювання ступеню залежності двох та більше явищ; відбір факторів, що суттєво впливають на результативну ознаку; знаходження невідомого причинного зв'язку, встановлення структури зв'язку між компонентами змінних *X* та *Y*.

Для мети 2 важливим є отримання значень функції апроксимації $f(x)$, а не її структура, тобто функція $f$ повинна показати числову залежність змінних *Y* та *X*, а не їх змістовний зв'язок.

Мета 3 передбачає проникнення у «фізичний механізм» статистичного зв'язку, тобто у механізм перетворення вхідних впливів *X* та *E* в результуючі показники *Y*. Тому, головним завданням тут є визначення структури функції $f(x)$, що буде покладена в основу математичної моделі, яка повинна адекватно описувати

об'єкт дослідження, надавати достовірні результати моделювання та можливість їх змістовної інтерпретації.

Найбільш розповсюджений клас функцій апроксимації, що використовуються у нелінійних моделях при статистичному аналізі, складають узагальнені багаточлени в базисі функцій $\varphi_0(x), \varphi_1(x), ..., \varphi_{m-1}(x)$ у вигляді психологічно[1]:

$$P(x) = \sum_{k=0}^{m-1} a_k \varphi_k(x), \tag{7.1}$$

де $a_k$ – числові коефіцієнти.

Наприклад:

алгебраїчні багаточлени, що породжені базисом $1, x^1, x^2, x^3, \dots, x^{m-1}$, подаються у такому вигляді (операторній (8.2) та розгорнутій (8.3) формах):

$$P(x) = \sum_{K=0}^{m-1} a_K x^K, (m = 2, 3, 4, \dots), \tag{7.2}$$

$$P(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{m-1} x^{m-1}; \tag{7.3}$$

експоненціальні функції (7.4а) та тригонометричні багаточлени (7.4б), що породжені базисом на основі комплексних чисел (гармонійного ряду):

$$y = a_0 e^{a_1 x}, \tag{7.4а}$$

$$P(x) = \sum_{k=-m}^{m} a_k e^{jk\frac{\pi}{L}x}, (m = 0,1,2,\dots, L = 1,2,3,\dots); \tag{7.4б}$$

дрібнораціональні функції

[1] Бродський Ю. Б., Малютіна В. П. Економіко – математичне моделювання. Конспект лекцій // Житомир: ЖНАЕУ, 2010. – 116 с.

$$R(x) = \frac{a_0 + a_1 x + \cdots + a_m x^m}{b_0 + b_1 x + \cdots + b_n x^n}; \tag{7.5}$$

сплайни – кусочно–поліноміальні функції невисокої степені, частіше третьої (кубічні сплайни).

Важливою особливістю алгоритмічних багаточленів (7.2) – їх лінійність відносно невідомих коефіцієнтів $a_k$, які визначають параметри нелінійної, як правило, моделі. Ця особливість надає можливість не тільки швидко і досить просто розраховувати робочі параметри моделей, а й дозволяє будувати ефективні алгоритми апроксимації.

Один із методів математичної обробки статистичних даних – є метод найменших квадратів (МНК), результатом застосування якого є отримання числових коефіцієнтів емпіричної формули (вибраного багаточлена *P(x)*). Тому, побудова оптимального апроксимуючого багаточлена *P(x)* зводиться до знаходження коефіцієнтів, які мінімізують функцію помилок *S* (квадрати відхилень вибраного багаточлена $P(x_i)$ та значень статистичної вибірки $y_i$)

$$S = \sum_{i=0}^{n-1} \left(P(x_i) - y_i\right)^2 \Rightarrow \min, \tag{7.6}$$

$$a_k = \arg\min S(a_k). \tag{7.7}$$

Таким чином, алгоритм обробки статистичних даних за допомогою МНК складається з етапів[2]:

1. Обирають апроксимуючу функцію $\varphi(x)$.

2. Визначають $\min S$, як $\frac{\partial S}{\partial a_k} = 0;$ тобто знаходять систему частинних похідних і прирівнюють їх до нуля.

[2] Бродський Ю. Б., Малютіна В. П. Економіко – математичне моделювання. Конспект лекцій // Житомир: ЖНАЕУ, 2010. – 116 с.

3. Розв'язують систему рівнянь відносно коефіцієнтів $a_k$.

4. Записують апроксимуючий поліном $P(x)$ з урахуванням числових значень коефіцієнтів $a_k$.

Кількість рівнянь повинна бути не меншою за кількість незалежних коефіцієнтів. При цьому чим більше вимірювань, тобто чим більшою мірою система перевизначена (надлишок інформації), тим краще, бо тоді випадкові похибки окремих вимірювань вилучають одна одну і рішення стає більш достовірним, тобто багаточлен *P(x)* більш адекватно описує систему або процес, що досліджується[3].

Приклад апроксимації даних нелінійними функціями поданий на рис. 7.2, де представлено число загиблих ***N*** при сильних землетрусах за 1999 – 2011 роки.

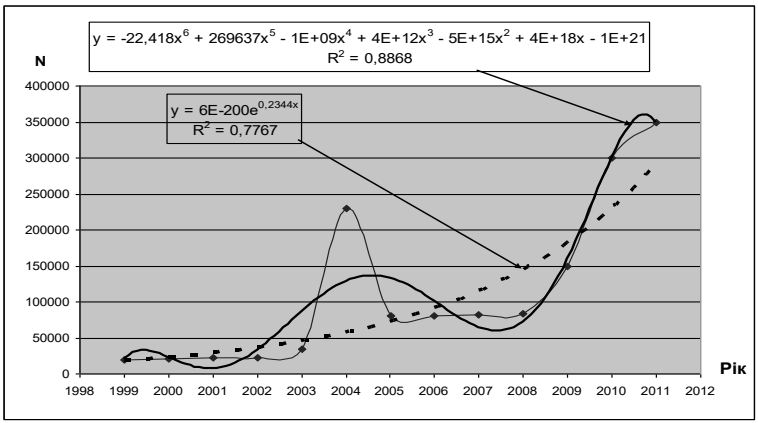



*Рисунок 7.2.* **Статистика загиблих при сильних землетрусах протягом 1999 – 2011 років**

*Джерело*: власні розробки автора

[3] Бродський Ю. Б., Малютіна В. П. Економіко – математичне моделювання. Конспект лекцій // Житомир: ЖНАЕУ, 2010. – 116 с.

Побудовані нелінійні моделі вказують на тенденцію циклічності (поліноміальна) та різкого підвищення числа загиблих (експоненціальна) за досліджений період, що свідчить про вступ людства в епоху природних катаклізмів, до якої людство поки що не готове технологічно, економічно, юридично і психологічно[4].

## 7.3 Оцінювання результатів дослідження

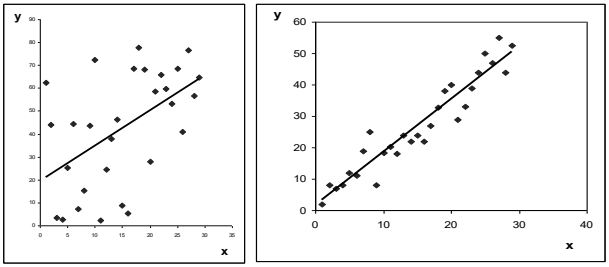
Достовірність результатів математико-статистичного моделювання і, відповідно побудованих математичних моделей, оцінюється відомими статистичними показниками. Нагадаємо, що основними задачами кореляційного та регресійного аналізу є виявлення та оцінювання щільності зв'язку між випадковими змінними, установлення математичної форми та вивчення причинно–наслідкової залежності. Методи, що використовуються для розв'язування указаних задач, залежать від: природи досліджуваних випадкових змінних (кількісні, порядкові, класифікаційні); вибраного статистичного показника (коефіцієнт або індекс кореляції); конкретно поставленої задачі (оцінка показника статистичної залежності, перевірки гіпотези про його значення, встановлення структури зв'язку між компонентами змінних).

Оцінювання щільності зв'язку виконують за допомогою таких статистичних показників, як: коефіцієнт кореляції $r$, кореляційне відношення $\eta$, індекс кореляції $R$ та коефіцієнт детермінації $R^2$. Визначимо їх зміст та способи розрахунку.

Якщо розглянути найбільш важливий для практики і теорії випадок лінійної залежності $y = a_0 + a_1 x$, то на перший погляд оцінити щільність зв'язку $y$ від $x$ можна коефіцієнтом регресії $a_1$,

[4] Бродський Ю. Б., Ганношин В. П., Пінкін А. А. Аномальна електрична складова електромагнітного поля землі як передвісник виникнення землетрусу. *Вісник ЖНАЕУ*. 2012. № 2 (31), т. 1. С. 280–286.

оскільки, він показує, на скільки одиниць в середньому змінюється $y$, коли $x$ збільшується на одиницю. Однак, по-перше, дві кореляційні залежності можуть мати однакові значення $a_1$, тобто однакові кутові коефіцієнти рівняння регресії, але різну щільність зв'язку – різні кореляційні поля (рис. 7.3а,б); по-друге, коефіцієнт регресії $a_1$ залежить від одиниць вимірювання змінних (показника розмірності фізичної величини). Тому, показником щільності зв'язку є коефіцієнт кореляції $r$, який залежить від коефіцієнта регресії $a_1$ та середньо квадратичних відхилень змінних $s_x$ та $s_y$.




а) б)

*Рисунок 7.3.* **Приклад кореляційного поля**

Для залежності $y(x) = a_{0_{yx}} + a_{1_{yx}} x$ величина

$$r = a_{1_{yx}} \frac{s_x}{s_y} \tag{7.8}$$

показує на скільки зміниться $s_y$ в середньому $y$, коли $x$ збільшиться на одне $s_x$.

Навпаки, для залежності $x(y) = a_{0_{xy}} + a_{1_{xy}} y$ величина

$$r = a_{1_{xy}} \frac{s_y}{s_x}. \tag{7.9}$$

Якщо знайти добуток (8.8) і (8.9), отримаємо

$$r^2 = a_{1_{yx}} a_{1_{xy}}, \tag{7.10}$$

або

$$r = \pm\sqrt{a_{1_{yx}} a_{1_{xy}}}, \tag{7.11}$$

тобто коефіцієнт кореляції $r$ змінних $x$ та $y$ є середнє геометричне коефіцієнтів регресії із врахуванням їхнього знаку.

Приведемо іншу формулу для розрахунку кореляції та розглянемо його основні властивості:

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{s_x s_y}, \tag{7.12}$$

де: $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$;

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$; $s_x = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}}$; $s_y = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n}}$.

Основні властивості коефіцієнта кореляції:

- знак коефіцієнта кореляції $r$ співпадає зі знаком коефіцієнта регресії $a_{1_{yx}}\left(a_{1_{xy}}\right)$: якщо $a_1 \rangle 0$, $r \rangle 0$, то кореляційний зв'язок між змінними називають прямим, якщо $a_1 \langle 0$, $r \langle 0$ - зворотним;
- діапазон значень коефіцієнта кореляції: $-1 \le r \le 1$. В залежності від наближення $|r|$ до 1 розрізнюють зв'язок: слабкий

$|r| \approx 0{,}1....0{,}3$; помірний $|r| \approx 0{,}3...0{,}5$; помітний $|r| \approx 0{,}5...0{,}7$; достатньо тісний $|r| \approx 0{,}7...0{,}9$; тісний або сильний $0{,}9 \le |r| < 1$;

- якщо всі значення змінних збільшити (або зменшити) пропорційно (на одне, або в одне і те ж число разів), то величина коефіцієнта кореляції не зменшиться;
- при $r = \pm 1$ кореляційний зв'язок представляє лінійну функціональну залежність;
- при $r = 0$ – лінійний кореляційний зв'язок відсутній. Однак, це не значить, що не існує взагалі кореляційної (нелінійної), а тим більш статистичної залежності.

Отже, коефіцієнт кореляції є показником щільності зв'язку лише у випадку *лінійної* залежності між двома змінними. Для нелінійної залежності застосовують інші показники оцінювання інтенсивності зв'язку: кореляційне відношення (індекс кореляції *R*) та коефіцієнт детермінації $R^2$.

Указані показники використовуються в моделі $y(x) = f(x) + \varepsilon$, де $\varepsilon$ – випадкова змінна, а змінна $x$ може бути вектором.

Позначаємо через $s_y^2$ загальну дисперсію випадкової величини $y$, через $s_f^2$ – дисперсію функції *f(x)*, а через $s_\varepsilon^2$ – залишкову дисперсію, що визначається випадковою величиною $\varepsilon$:

$$s_y^2 = \frac{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2}{n}, \tag{7.13}$$

$$s_f^2 = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \overline{y}\right)^2}{n}, \tag{7.14}$$

$$s_\varepsilon^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}, \tag{7.15}$$

де $\hat{y}_i$ – значення функції регресії $f(x)$ в точках $x_1, x_2, ... x_i, ... x_n$ ($f(x)$ побудована на основі експериментальних даних $(x_1, y_1)$, $(x_2, y_2), \ldots (x_i, y_i), \ldots (x_n, y_n)$ і залежить від $k$ параметрів $a_k$): $\hat{y}_i = f(x_i), i = \overline{1,n}$;

$k$ – число степенів свободи.

Дисперсії (7.13), (7.14) і (7.15) зв'язані рівнянням:

$$s_y^2 = s_f^2 + s_a^2. \tag{7.16}$$

Індекс кореляції $y$ по $x$ (теоретичне кореляційне відношення) $R_{yx}$ характеризує розсіювання точок кореляційного поля відносно побудованої лінії регресії $f(x)$ і визначається співвідношенням

$$R_{yx} = \sqrt{\frac{s_f^2}{s_y^2}} = \sqrt{1 - \frac{s_\varepsilon^2}{s_y^2}}. \tag{7.17}$$

Емпіричне кореляційне відношення $\eta$ є показником степені розсіювання точок кореляційного поля відносно емпіричної лінії регресії (ламаної, що з'єднує значення $\tilde{y}_i = y_i + \varepsilon_i$, де $\varepsilon_i$ – випадкові, не враховані фактори) і визначається:

$$\eta_{yx} = \sqrt{\frac{s_{\tilde{y}_i}^2}{s_y^2}}, \tag{7.18}$$

де:

$$s^2_{\tilde{y}_i} = \frac{\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2}{n}. \qquad (7.19)$$

Чим тісніший зв'язок, тим більший вплив на варіацію змінної $y$ оказують зміни $x$ порівняно з неврахованими факторами, тим вище значення $\eta_{yx}$.

Емпіричне кореляційне відношення $\eta_{yx}$ перебільшує тісноту зв'язку, тому частіше використовують індекс кореляції $R$. Хоча, для розрахунку $\eta$ рівняння регресії знати не потрібно, на відміну від $R$. Кореляційне відношення $\eta$ та індекс кореляції $R$ зв'язані з коефіцієнтом кореляції $r$ наступним чином:

$$0 \le |r| \le R \le \eta \le 1. \qquad (7.20)$$

На відміну від коефіцієнта кореляції $r$ (для якого $r = r_{yx} = r_{xy}$) $\eta_{yx} \neq \eta_{xy}$. У випадку лінійної залежності між змінними $R_{yx} = R_{xy} = |r|$. Різниця між кореляційним відношенням $\eta$ (індексом кореляції $R$) та коефіцієнтом кореляції $r$ дозволяє перевірити *лінійність* кореляційної залежності.

Оцінимо граничні значення індексу кореляції $0 \le R_{yx} \le 1$:

- якщо $R_{yx} = 0$, тоді з (8.17) $s^2_f = 0$, або $s^2_y = s^2_\varepsilon$, що означає повну відсутність будь-якого впливу змінної $x$ на змінну $y$, тобто відсутність кореляційного зв'язку між $x$ та $y$;
- якщо $R_{yx} = 1$, тоді з (8.17) $s^2_\varepsilon = 0$, що означає наявність функціональної залежності між змінними $x$ та $y$.

Квадрат індексу кореляції $R^2$ часто називають коефіцієнтом детермінації. Він показує, яка доля дисперсії результуючої величини $y$ визначається варіацією (дисперсією) функції $f(x)$, що залежить від впливу змінної $x$. Коефіцієнт детермінації використовують як міру адекватності підбору функції регресії для апроксимації початкових даних і визначається із (7.17):

$$R_{yx}^2 = \frac{s_f^2}{s_y^2}. \tag{7.21}$$

Діапазон змінювання коефіцієнта детермінації відповідає граничній області індексу кореляції:

$$0 \le R^2 \le 1.$$

У випадку лінійної регресії коефіцієнти детермінації та кореляції співпадають $R^2 = r^2$.

Таким чином, наближеність коефіцієнта детермінації до одиниці показує рівень відповідності (точності) вибраної функції регресії експериментальним даним.

Крім коефіцієнта детермінації та індексу кореляції (кореляційних відношень) для оцінки адекватності функції регресії початковим даним часто користуються показником середньої відносної похибки апроксимації:

$$\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|. \tag{7.22}$$

Чим менше $\bar{\varepsilon}$, тим краще функція регресії апроксимує експериментальні дані.

В практичних дослідженнях характеристики (параметри) випадкової величини, як правило, невідомі, тому виникає задача їх оцінювання, яка вирішується методом вибіркового аналізу. Тобто, параметри генеральної сукупності оцінюється за результатами аналізу вибірки, рис. 7.4, що приводить до відповідної похибки.

Тому, потрібно так витягти вибірку і провести аналіз (обробити дані), щоб похибка оцінювання була мінімальною.

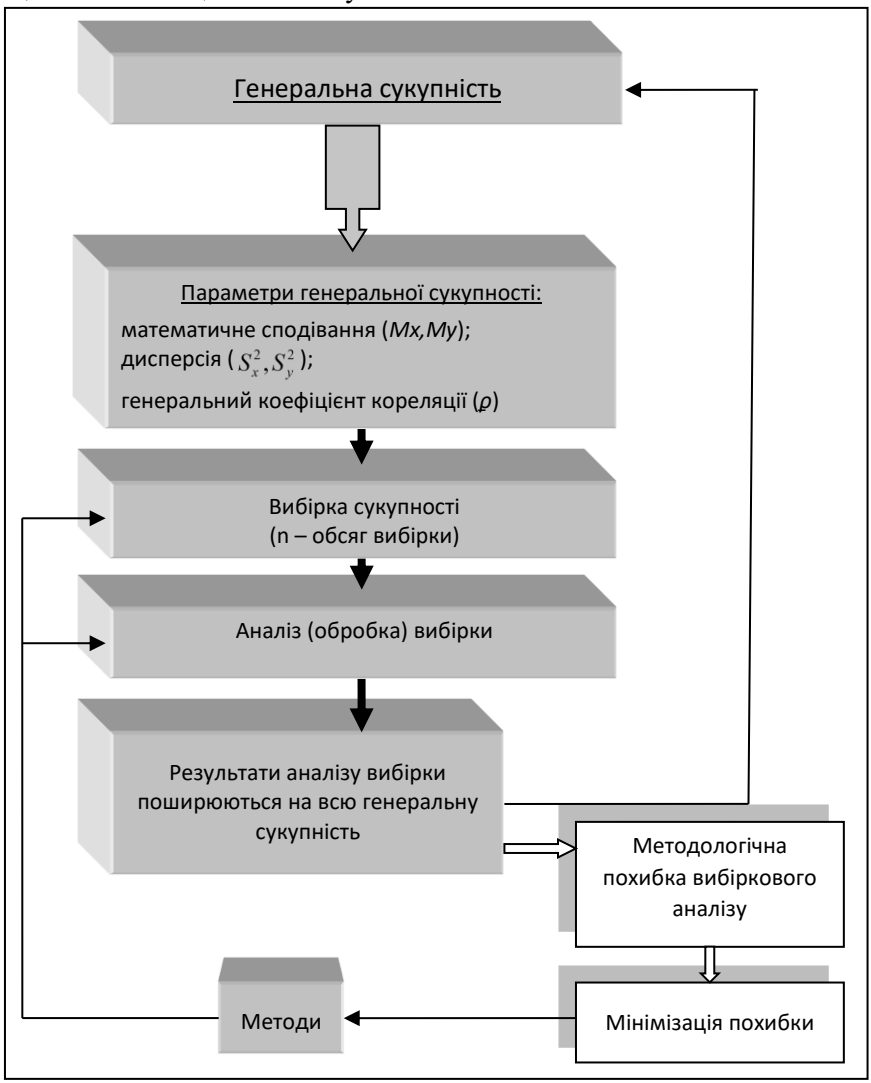



*Рисунок 7.4.* **Алгоритм аналізу вибірки**
*Джерело*: власні розробки автора

Отже, елементи вибірки – це випадкові величини. Функції від вибіркових значень – статистики (середнє, дисперсія, тощо) – використовується для побудови оцінок. Для цього розподіл вибраної статистики повинен бути зосереджений достатньо близько біля невідомого значення параметра Θ (імовірність відхилень невелика), а точність оцінювання збільшувалась при збільшенні об'єму вибірки.

Оцінки невідомих параметрів Θ бувають двох типів: точкові та інтервальні.

Точкова оцінка $\hat{\Theta}_n$ – це оцінка, що має конкретне числове значення. Інтервальна оцінка подається довірчим інтервалом значень $\Theta_{1n} \le \Theta \le \Theta_{2n}$, всередині якого знаходиться істинне значення параметра Θ із заданою імовірністю $p$ (на практиці: 0,95; 0,98; 0,99). Величину $p$ ще називають надійністю, $\alpha = (1 - p)$ – рівнем значимості відхилення оцінки.

Будь–яка статистика має бути перевірена на значущість за допомогою спеціальних критеріїв, які допомагають розрізнити похибку вимірювання від значущої інформації, що закладена у значення функції від елементів отриманої вибірки. Неперевірений статистичний результат є лише деякою гіпотезою, яка може бути прийнята або відхилена.

Для кожної конкретної ситуації можна сформулювати множину різних гіпотез, із яких слід вибирати ті, що чітко сформульовані, однозначні, прості та краще відповідають цілям дослідження.

Критерій перевірки статистичної гіпотези – це процедура підготовки та прийняття рішення про підтвердження або відхилення даної гіпотези.

### *Список питань до самоконтролю:*

**1.** Поясніть компоненти загальної структурної моделі статистичної залежності.

**2.** Що слід враховувати в процесі проведення статистичного аналізу та моделювання?

**3.** Сформулювати загальну задачу статистичного аналізу.

**4.** Визначити типові цілі статистичного аналізу і моделювання. Вирішення яких завдань вони передбачають?

**5.** Назвіть найбільш розповсюджений клас функцій, що використовується у нелінійних моделях.

**6.** Запишіть математичний вираз поліному 4-го степеню в операторній та розгорнутій формах.

**7.** Яка важлива особливість алгоритмічних багаточленів?

**8.** Поясніть алгоритм обробки статистичних даних за методом найменших квадратів.

**9.** Назвіть статистичні показники оцінювання достовірності результатів математико-статистичного моделювання.

**10.** Чим визначається коефіцієнт кореляції? Проведіть аналіз його властивостей?

**11.** Що характеризує і чим визначається індекс кореляції (кореляційне відношення)?

**12.** Що дозволяє перевірити порівняння кореляційного відношення та коефіцієнта кореляції?

**13.** Проведіть аналіз граничних значень індексу кореляції.

**14.** Що визначає коефіцієнт детермінації?

**15.** Поясніть алгоритм аналізу вибірки.

### *Список рекомендованої літератури:*

1. Бережна Л.В., Снитюк О.І. Економіко – математичні методи та моделі в фінансах. – К: Кондор, - 2009. – 301 с.
2. Бродський Ю. Б., Ганношин В. П., Пінкін А. А. Аномальна електрична складова електромагнітного поля землі як передвісник виникнення землетрусу. Вісник ЖНАЕУ. 2012. № 2 (31), т. 1. С. 280–286.
3. Бродський Ю. Б., Малютіна В. П. Економіко – математичне моделювання. Конспект лекцій // Житомир: ЖНАЕУ, 2010. – 116 с.
4. Лещинський О.Л. Економетрія: Навч. посіб. для студ. вищ. навч. закл. /О.Л. Лещинський, В.В. Рязанцева, О.О. Юнькова. – К.: МАУП, 2003. – 208 с.
5. Томашевський В. М. Моделювання систем [під ред. М. З. Згуровського] / Томашевський В. М. – К. : Видавнича група BHV, 2005. – 352 с.

# Розділ 8. МОДЕЛЮВАННЯ ІНДЕКСНИХ СИСТЕМ

## 8.1. Введення в теорію індексів

### 8.1.1. Поняття індексів. Індивідуальні та зведені індекси

*Індекс* – відносна величина, що характеризує зміну рівня певного явища в часі, просторі або порівняно з планом (нормою, стандартом). Наприклад, посадовий оклад доцента зріс у січні порівняно з груднем минулого року в 1,48 рази.

Залежно від мети порівняння індекси поділяються на три групи:

1. Індекси динаміки – характеризують зміну явища в часі.
2. Територіальні індекси – характеризують зміну явища у просторі.
3. Індекси норми (стандарту) – характеризують зміну явища порівняно з нормативним (стандартизованим, еталонним) рівнем.

Наприклад, індекс динаміки ціни одиниці товару розраховується таким чином: $i_p = \frac{p_1}{p_0}$;

територіальний індекс цін – $i_p = \frac{p_A}{p_B}$;

індекс норми можна показати на прикладі виконання плану – $i_q = \frac{q_1}{q_{пл}}$.

Залежно від охоплення одиниць сукупності розрізняють індивідуальні та зведені індекси.

*Індивідуальні* індекси характеризують зміну одного явища (видобутку вугілля на шахті, ціни на товар).

*Зведені* індекси характеризують зміну рівня показника, що належить до сукупності. При цьому сукупність може складатись з однорідних або неоднорідних елементів.

Наприклад, у першому випадку ми маємо дані про видобуток вугілля кількома шахтами, врожайність кількох зернових (тобто однорідних у деякому розумінні) культур, про ціни на картоплю в кількох продавців. У другому випадку – обсяг виробництва різнорідної продукції одним або декількома підприємствами, ціни на різні продукти в місті. У першому випадку можна обчислити середній для сукупності рівень. Зміну цих середніх рівнів характеризують *зведеним індексом середніх величин*, у другому випадку користуються *зведеним агрегатним індексом*.

Якщо зміна явища вивчається не за два, а за більше періодів, то кожен з них позначається відповідно "0", "1", "2", "3" тощо. Якщо за базу порівняння береться рівень попереднього періоду, то індекси динаміки називають *ланцюговими*, якщо один і той же початковий рівень – *базисними*.

Наприклад, обсяг виробництва підприємством у 2010 р. позначимо через $q_0$, у 2011 – $q_1$, ..., у 20016 р. – $q_6$.

Тоді $i_{6/0} = \frac{q_1}{q_0} \times \frac{q_2}{q_1} \times \frac{q_3}{q_2} \times \frac{q_4}{q_3} \times \frac{q_5}{q_4} \times \frac{q_6}{q_5} = \frac{q_6}{q_0}$.

Існує загальне правило для оцінювання взаємозв'язку для індексів: індекси пов'язані між собою так само, як їх індексовані величини. Так, якщо треба визначити, як зміниться валовий збір, за умови що урожайність збільшилася у 1,2 рази, а розмір посівної площі зменшився на 10 %. Відомо, що валовий збір обчислюється як добуток урожайності та посівної площі, тобто $ВЗ = У \times П$. Звідси: $i_{ВЗ} = i_У \times i_П = 1{,}2 \times 0{,}9 = 1{,}08$. Таким чином, валовий збір збільшиться на 8 %.

### 8.1.2. Агрегатні індекси

Зведені агрегатні індекси використовуються у випадках неоднорідної сукупності. Так, якщо реалізується порівнянний товар (овочі та фрукти), то індекс фізичного обсягу реалізації може мати вигляд

$$I_q = \frac{\Sigma q_1}{\Sigma q_0}.$$

Якщо ж реалізуються різні, непорівнянні товари, фізичний обсяг яких може вимірюватися як за допомогою різних одиниць виміру (кг, шт., л), так і однакових, тоді зіставлення загальних фізичних обсягів реалізованого товару немає сенсу, наприклад обсяг реалізації продовольчих і непродовольчих товарів. Таким чином, загальний індекс фізичного обсягу не може виглядати, як $I_q = \frac{\Sigma q_1}{\Sigma q_0}$

.

Виникає необхідність у приведенні різнорідних товарів до порівнянного виду. У даному випадку використовується такий сумірник, як ціна. Тоді зведений індекс фізичного обсягу набуває вигляду

$$I_q = \frac{\Sigma pq_1}{\Sigma pq_0}$$

Однак виникає питання, на якому рівні слід фіксувати ціни: на поточному чи базисному? Обсяг реалізації (товарооборот або виручка від реалізації) у поточному періоді порівняно з базисним може змінюватися під впливом двох факторів: зміни фізичного обсягу реалізованих товарів (одного виду або кількох) та зміни цін. Зрозуміло, що на цю зміну можуть впливати як один із факторів, так і обидва. Крім того, діяти вони можуть в одному чи в протилежних напрямах. У таких випадках для оцінки впливу всіх факторів, а також кожного з них окремо використовують систему індексів. Якщо ми

зафіксуємо ціни на базисному рівні, тоді отримана величина $\sum p_0 q_1$ у чисельнику буде характеризувати виручку від реалізації поточного періоду у порівнянних цінах, тобто такий показник має економічний зміст і може інтерпретуватися. Така система зважування на базисному рівні називається *системою Ласпереса*, а індекс фізичного обсягу набуває вигляду

$$I_q = \frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \tag{8.1}$$

і називається *зведеним індексом фізичного обсягу товарообороту*.

При розрахунку зведеного індексу цін змінюється ціна, а фізичний обсяг в даній індексній системі фіксується на поточному рівні згідно із системою Пааше.

Тоді індекс цін має такий вигляд:

$$I_p = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \tag{8.2}$$

Між індексом фізичного обсягу та індексом ціни існує взаємозв'язок:

$$I_{pq} = I_q \times I_p = \frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} \tag{8.3}$$

Індекси такого типу отримали назву агрегатних, оскільки їх чисельники і знаменники є агрегатами, тобто величинами, які мають економічний зміст. Так, $\sum p_0 q_1$ – є обсягом реалізації поточного періоду в цінах базисного періоду, або, як кажуть, у фіксованих цінах, $\sum p_0 q_0$ – обсяг реалізації базисного періоду, $\sum p_1 q_1$ – обсяг реалізації поточного періоду.

На основі агрегатних індексів можна розрахувати абсолютні зміни виручки від реалізації в цілому і під впливом окремих факторів: ціни і фізичного обсягу, як різницю між чисельником і знаменником відповідної формули.

Абсолютна зміна за рахунок фактора ціни розраховується за формулою

$$\Delta_p = \sum p_1 q_1 - \sum p_0 q_1 \tag{8.4}$$

Абсолютна зміна за рахунок фактора фізичного обсягу:

$$\Delta_q = \sum p_0 q_1 - \sum p_0 q_0 \tag{8.5}$$

Загальна абсолютна зміна під впливом двох факторів:

$$\Delta_{pq} = \sum p_1 q_1 - \sum p_0 q_0 = \left(\sum p_0 q_1 - \sum p_0 q_0\right) + \left(\sum p_1 q_1 - \sum p_0 q_1\right)$$

(8.6) Таким чином, між абсолютними змінами також існує взаємозв'язок, а це дає можливість оцінити частку впливу кожного з факторів на загальну зміну виручки від реалізації. Частка впливу фактору ціни можна визначити за формулою

$$d_{\Delta_p} = \frac{\left|\Delta p\right|}{\left|\Delta p\right| + \left|\Delta q\right|} \times 100\ \% \tag{8.7}$$

На другий фактор припадає $100 - d_{\Delta_p}$. Абсолютні зміни слід брати за модулем, оскільки вплив факторів може бути різноспрямований.

Розглянемо порядок розрахунку агрегатних індексів за даними про продаж продовольчих товарів в торгівельній мережі. Визначимо зведений індекс ціни, фізичного обсягу та товарообороту, а також абсолютний приріст товарообороту за рахунок цих факторів.

*Таблиця 8.1.* **Обсяги реалізації та цін в торгівельній мережі**

| **Товар** | **Продано, т** | | **Ціна 1 кг, грн** | |
|---|---|---|---|---|
| | базисний ( $q_0$ ) | звітний ( $q_1$ ) | базисний ( $p_0$ ) | звітний ( $p_1$ ) |
| А | 15 | 10 | 10 | 12 |
| Б | 50 | 45 | 25 | 35 |
| Разом | 65 | 55 | x | x |

Побудуємо додаткову розрахункову таблицю 8.2:

*Таблиця 8.2.* **До розрахунку агрегатних індексів**

| **Товар** | **Виручка від реалізації, тис. грн** | | |
|---|---|---|---|
| | $p_0q_0$ | $p_1q_1$ | $p_0q_1$ |
| А | 10×15=150 | 12×10=120 | 10×10=100 |
| Б | 25×50=1250 | 35×45=1575 | 25×45=1125 |
| Разом | 1400 | 1695 | 1225 |

Зведений індекс товарообороту

$$I_{pq} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} = \frac{1695}{1400} = 1{,}2107 \,(+21{,}07\ \%),$$

абсолютна зміна виручки від реалізації в цілому становить

$$\Delta_{pq} = \sum p_1 q_1 - \sum p_0 q_0 = 1695 - 1400 = +295 \,(\text{тис. грн}).$$

Таким чином, виручка від реалізації зросла на 21,07 %, що становить 295 тис. грн.

Проаналізуємо вплив кожного з факторів:

а) фізичного обсягу:

$$I_q = \frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} = \frac{1225}{1400} = 0{,}875 \,(-12{,}5\ \%),$$

абсолютна зміна виручки від реалізації за рахунок фізичного обсягу становить $\Delta_q = \sum p_0 q_1 - \sum p_0 q_0 = 1225 - 1400 = -175$ (тис. грн), фізичний обсяг продажу товарів у середньому знизився на 12,5 %, що призвело до втрат виручки від реалізації на 175 тис. грн;

б) цін:

$$I_p = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} = \frac{1695}{1225} = 1{,}3837 \,(+38{,}4\ \%),$$

абсолютна зміна виручки від реалізації за рахунок фактора ціни становить $\Delta_p = \sum p_1 q_1 - \sum p_0 q_1 = 1695 - 1225 = +470$ (тис. грн), ціни товарів в середньому зросли на 38,4 %, що призвело до приросту виручки від реалізації на 470 тис. грн.

Перевіримо взаємозв'язок індексів і абсолютних змін:

$$I_{pq} = I_q \times I_p = 0{,}875 \times 1{,}3837 = 1{,}2107,$$

$$\Delta_{pq} = \Delta_p + \Delta_q = 470 + (-175) = +295 \text{ (тис. грн)}$$

Оцінимо частку впливу кожного з факторів:

$$d_{\Delta_p} = \frac{|\Delta_p|}{|\Delta_p| + |\Delta_q|} \times 100\% = \frac{470}{470 + |-175|} \times 100\% = 73\%,$$

$$d_{\Delta_q} = 100 - d\Delta_p = 100 - 73 = 27\%.$$

Таким чином, більш ніж дві третини загальної зміни виручки від реалізації відбулося за рахунок зростання цін, що призвело до зниження попиту, у результаті чого втрати виручки від реалізації становили 175 тис. грн, однак це було повністю компенсовано приростом виручки за рахунок підвищення цін, у результаті чого виручка зросла на 295 тис. грн.

### 8.1.3. Середньозважений арифметичний і гармонічний індекси

Агрегатні індекси можуть бути обчислені за допомогою індивідуальних індексів. Наприклад, необхідно обчислити індекс фізичного обсягу за умов відсутності індивідуальних даних про ціни та фізичний обсяг і наявності індивідуальних індексів фізичного обсягу. Тоді можемо скористатись підстановкою, тобто замінити $\sum p_0 q_1$ на $\Sigma i_q p_0 q_0$, оскільки $q_1 = i_q \times q_0$. Формула індексу фізичного обсягу набуває вигляду

$$I_q = \frac{\Sigma i_q p_0 q_0}{\Sigma p_0 q_0} \qquad (8.8)$$

такий індекс називається *середньозваженим арифметичним індексом фізичного обсягу*.

Таким чином, *середньозважений індекс* – це середній з індивідуальних індексів, зважених на агрегати, що мають відповідну розмірність і однакову фіксацію рівнів. Агрегатами можуть виступати вартісні $\sum p_i q_i$ (товарооборот), $\sum c_i q_i$ (витрати на виробництво) або трудові $\sum t_i q_i$ (витрати праці) показники.

Аналогічну заміну можна здійснити в агрегатному індексі ціни, якщо у знаменнику замінити $\sum p_0 q_1$ на $\Sigma \frac{p_1 q_1}{i_p}$, оскільки

$$p_0 = \frac{p_1}{i_p}.$$

Тоді формула індексу ціни набуває вигляду:

$$I_p = \frac{\Sigma p_1 q_1}{\Sigma \frac{p_1 q_1}{i_p}} \qquad (8.9)$$

такий індекс називається *середньозваженим гармонічним індексом цін.*

Розглянемо розрахунок індексу рівня споживання продуктів в розрахунку на одну особу на прикладі (табл.8.3).

*Таблиця 8.3.* **До розрахунку середньозважених індексів**

| **Товарні групи** | **Обсяг споживання в поточних цінах, грн** | | **Індекси цін $i_p$** | $\frac{p_1 q_1}{i_p}$ |
|---|---|---|---|---|
| | I квартал $p_0 q_0$ | II квартал $p_1 q_1$ | | |
| М'ясопродукти | 320000 | 315000 | 0,90 | 350 |
| Молокопродукти | 28000 | 26530 | 0,95 | 27926 |
| Хлібопродукти | 32000 | 32817 | 0,98 | 33487 |
| Разом | 380000 | 374527 | х | 411413 |

Додатковою умовою є зростання чисельності населення на 4 %.

Рівень споживання на одну особу розраховується як відношення зведеного обсягу споживання на чисельність населення: $\partial = \frac{q}{T}$. Таким чином $I_\partial = \frac{I_q}{I_T}$.

Відповідно до умови $I_T = 1{,}04$. Залишається обчислити $I_q$.

$$I_q = \frac{\Sigma(p_1 q_1)/i_p}{\Sigma p_0 q_0} = \frac{411413}{380000} = 1{,}081 \qquad (108{,}1\%),$$

тоді $I_\partial = \frac{1{,}081}{1{,}040} = 1{,}038$ (103,8 %).

Отже, рівень споживання продуктів на одну особу в середньому зріс на 3,8 %, що збільшило загальні витрати на 31413 грн (411413–380000).

### 8.1.4. Індекси середніх величин

Поряд з необхідністю оцінювання динаміки рівня показника (наприклад, рівня цін) для окремого підприємства, перед статистикою постає завдання необхідності вивчення динаміки показника в сукупності (наприклад, рівня цін на окремий товар по групі торгових підприємств). Тобто потрібно обчислити середній рівень показника й оцінити його в динаміці. Якщо вивчається динаміка середніх рівнів показників, то використовують систему зведених індексів середніх величин.

Як уже відомо, на величину середнього рівня впливають два фактори:

- зміна рівня значення самого показника;
- зміна структури сукупності, тобто розподілу одиниці сукупності за рівнем показника.

Таким чином, система індексів середніх величин буде включати три індекси:

1. Індекс змінного складу, який характеризує динаміку середнього рівня під впливом двох факторів: як зміни показника в середньому, так і зміни структури сукупності. Індекс змінного складу можна записати таким чином:

$$I_{\bar{x}}^{зс} = \frac{\bar{x}_1}{\bar{x}_0}.$$

Якщо взяти до уваги, що $\bar{x} = \frac{\sum xf}{\sum f}$, то формула індексу змінного складу набуває такого вигляду:

$$I_{\bar{x}}^{зс} = \frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_0}{\sum f_0} \qquad (8.10)$$

Різниця між двома дробами характеризує абсолютну зміну середнього рівня показника:

$$\Delta\bar{x} = \frac{\sum x_1 f_1}{\sum f_1} - \frac{\sum x_0 f_0}{\sum f_0} \qquad (8.11)$$

Якщо ми згадаємо, що $\frac{f}{\sum f}$ характеризує структуру, то, позначивши це співвідношення через $d$, можемо записати індекс змінного складу у такому вигляді:

$$I_{\bar{x}}^{зс} = \frac{\sum x_1 d_1}{\sum x_0 d_0},$$

а абсолютну зміну можна записати, як

$$\Delta\bar{x} = \sum x_1 d_1 - \sum x_0 d_0.$$

2. Індекс фіксованого складу дає можливість оцінити динаміку рівня в середньому за рахунок зміни тільки значень самого показника

(*x*) при зафіксованій структурі на поточному рівні згідно із системою Пааше. Тоді індекс фіксованого складу має вигляд

$$I_{\bar{x}}^{фс} = \frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_1}{\sum f_1} = \frac{\sum x_1 f_1}{\sum x_0 f_1} \qquad (8.12)$$

Таким чином, індекс фіксованого складу не що інше, як агрегатний індекс.

Різниця між чисельником і знаменником цього індексу характеризує абсолютну зміну агрегату під впливом зміни рівня показника в середньому:

$$\Delta_x = \sum x_1 f_1 - \sum x_0 f_1 \qquad (8.13)$$

Аналогічно можемо записати індекс фіксованого складу через частки в такому вигляді:

$$I_{\bar{x}}^{фс} = \frac{\sum x_1 d_1}{\sum x_0 d_1},$$

а абсолютну зміну можна записати як

$$\Delta_x = \sum x_1 d_1 - \sum x_0 d_1 .$$

3. Індекс структурних зрушень характеризує зміну середнього рівня під впливом зрушень у структурі сукупності $\frac{f}{\sum f}$ при зафіксованому значенні показника (*x*) на базовому рівні згідно із системою Ласпейреса. Тоді індекс структурних зрушень набуває вигляду

$$I_{\bar{x}}^{сз} = \frac{\sum x_0 f_1}{\sum f_1} \div \frac{\sum x_0 f_0}{\sum f_0} \qquad (8.14)$$

Таким чином, індекс структурних зрушень можна записати й інакше, а саме:

$$I_{\bar{x}}^{сз} = \frac{\sum x_0 f_1}{\sum x_0 f_0} \div \frac{\sum f_1}{\sum f_0},$$

а це дозволяє стверджувати, що структурні зрушення формуються також під впливом двох індексів: агрегатного індексу фізичного обсягу – $I_f$ і загального індексу фізичного обсягу – $I^f$ .

Різниця між двома дробами характеризує абсолютну зміну середнього рівня показника під впливом структурної складової:

$$\Delta_d = \frac{\sum x_0 f_1}{\sum f_1} - \frac{\sum x_0 f_0}{\sum f_0} \qquad (8.15)$$

Аналогічно можемо записати індекс структурних зрушень через частки в такому вигляді:

$$I_{\bar{x}}^{сз} = \frac{\sum x_0 d_1}{\sum x_0 d_0},$$

а абсолютну зміну можна записати як

$$\Delta_d = \sum x_0 d_1 - \sum x_0 d_0 .$$

Отже, індексна модель, що характеризує взаємозв'язок між розглянутими індексами, може бути дво- і трифакторною. Двофакторна індексна модель включає індекс фіксованого складу та індекс структурних зрушень і записується у вигляді добутку цих індексів:

$$I_{\bar{x}}^{зс} = I_{\bar{x}}^{фс} \times I_{\bar{x}}^{сз} \qquad (8.16)$$

Трифакторна індексна модель включає два агрегатних індекси та один загальний індекс фізичного обсягу і записується в такому вигляді:

$$I_{\bar{x}}^{зс} = I_x \times I_f \div I^f \qquad (8.17)$$

Розглянемо цю індексну систему на прикладі індексу середньої урожайності пшениці: озимої та ярової. Як відомо, урожайність (*У*) – це валовий збір культури (*ВЗ*) зібраний з одиниці площі (*П*). Тоді, урожайність окремої культури розраховується за формулою:

$$У = \frac{ВЗ}{П}, \quad \text{а середня урожайність } \bar{У} = \frac{\sum ВЗ}{\sum П}.$$

Якщо валовий збір представити у вигляді добутку урожайності і посівних площ, то формула середньої урожайності набуває такого вигляду:

$$\bar{У} = \frac{\sum У \times П}{\sum П}.$$

Тоді маємо право записати систему індексів середніх величин змінного, фіксованого складу і структурних зрушень. Розглянемо застосування індексної системи на прикладі середньої врожайності двох сортів пшениці: озимої і ярової ( табл.8.4).

*Таблиця 8.4.* **Динаміка врожайності озимої та ярової пшениці**

| **Сорт пшениці** | **Урожайність, ц/га** | | **Посівна площа, тис. га** | |
|---|---|---|---|---|
| | **1-й рік ($У_0$)** | **2-й рік ($У_1$)** | **1-й рік ($П_0$)** | **2-й рік ($П_1$)** |
| Озима | 38,0 | 31,0 | 7000 | 5205 |
| Ярова | 27,5 | 25,0 | 20 | 55 |
| Разом | х | х | 7020 | 5260 |

Уведемо умовні позначення (див. табл.). Тоді індекс змінного складу середньої урожайності має такий вигляд:

$$I_{\bar{У}}^{зс} = \frac{\sum У_1 П_1}{\sum П_1} \div \frac{\sum У_0 П_0}{\sum П_0}.$$

Для його розрахунку побудуємо додаткову таблицю 8.5:

*Таблиця 8.5.* **До розрахунку індексів середніх величин**

| Вид пшениці | Валовий збір, тис. ц | | |
|---|---|---|---|
| | $У_0 П_0$ | $У_1 П_1$ | $У_0 П_1$ |
| Озима | 38×7000=266000 | 31×5205=161355 | 38×5205=197790,0 |
| Ярова | 27,5×20=550 | 25×55=1375 | 27,5×55=1512,5 |
| Разом | 266550 | 162730 | 199302,5 |

Підставимо отримані результати у формулу індексу змінного складу:

$$I_{\bar{y}}^{зс} = \frac{162730}{5260} \div \frac{266550}{7020} = \frac{30,94}{37,97} = 0,815 \text{ (-18,5 \%)}.$$

Урожайність знизилась на 18,5 %, що становить практично 7 ц/га (30,94-37,97).

Визначимо, як на зниження середньої урожайності вплинув фактор урожайності озимої та ярової пшениці. Для цього розрахуємо індекс фіксованого складу:

$$I_{\bar{y}}^{фс} = \frac{\sum y_1 п_1}{\sum y_0 п_1} = \frac{162730}{199302,5} = 0,816 \text{ (-18,4 \%)}.$$

Абсолютна зміна становить

$$\Delta_y = \sum y_1 п_1 - \sum y_0 п_1 = 162730 - 199302,5 = -36572,5 \text{ (тис. ц)}.$$

Таким чином, за рахунок зниження урожайності як озимої, так і ярової культур у середньому на 18,4 % втрати валового збору склали 3 млн 657 тис. 250 тон.

Для оцінювання впливу структурної складової побудуємо індекс структурних зрушень середньої урожайності:

$$I_{\bar{y}}^{сз} = \frac{\sum y_0 п_1}{\sum п_1} \div \frac{\sum y_0 п_1}{\sum п_1}.$$

Підставимо дані з розрахункової таблиці :

$$I_{\bar{y}}^{сз} = \frac{199302,5}{5260} \div 37,97 = 0,9979 \text{ (-0,21 \%)}.$$

Значних структурних зрушень у посівних площах не відбулося. Під впливом структурної складової середня урожайність знизилася на 0,21 %.

Усі розраховані індекси взаємопов'язані: $0,816 \times 0,9979 = 0,814$.

Розглянемо дані про роботу двох шахт, які утворюють трест. Обчислимо індивідуальні та зведені індекси продуктивності праці,

загальний приріст видобутку вугілля в цілому і за рахунок окремих факторів (табл. 8.6).

Добудуємо таблицю (стовпчики 5–8) та обчислимо підсумковий рядок.

Продуктивність праці – це видобуток вугілля за одиницю часу. Для шахти № 1: $w_1 = 88:40 = 2,2$ (т/люд.-днів);

$$w_0 = 40:20 = 2,0 \text{ (т/л-днів)}.$$

Отже, індивідуальний індекс продуктивності праці для неї становить 1,1 (2,2/2,0), а для шахти № 2 – 1,02 (1,53/1,5).

*Таблиця 8.6.* **До розрахунку індексів середніх величин**

| **№ шахти** | **Видобуто вугілля, тис. т** | | **Відпрацьовано тис. люд.-днів** | | **Продуктивність праці, т/люд.-днів** | | **Частка у витратах праці** | |
|---|---|---|---|---|---|---|---|---|
| | 1-й рік $q_0$ | 2-й рік $q_1$ | 1-й рік $T_0$ | 2-й рік $T_1$ | 1-й рік $w_0$ | 2-й рік $w_1$ | 1-й рік $d_0$ | 2-й рік $d_1$ |
| **А** | **1** | **2** | **3** | **4** | **5=1/3** | **6=2/4** | **7** | **8** |
| 1 | 40 | 88,0 | 20 | 40 | 2,0 | 2,2 | 20/50 =0,4 | 0,67 |
| 2 | 45 | 30,6 | 30 | 20 | 1,5 | 1,53 | 30/50 =0,6 | 0,33 |
| Разом | 85 | 118,6 | 50 | 60 | 85/50 =1,7 | 118,6/60 =1,977 | 1,0 | 1,00 |

Таким чином, продуктивність праці на першій шахті зросла в 1,1 рази, або на 10 %. Зазначимо, що цей індекс ми називаємо *індивідуальним*, оскільки він належить до *одиниці сукупності* – шахти № 1.

Однак, з іншого боку, він характеризує зміну *середньої* продуктивності праці всіх шахтарів шахти № 1, тобто *сукупності*. Аналогічно для другої шахти. Але якщо продуктивність праці на одній шахті зросла на 10 %, а на другій – на 2 %, то це зовсім не обов'язково, що в цілому по тресту вона зросте в таких межах.

Обчислимо зведений індекс середньої продуктивності:

$$\boldsymbol{I_{\bar{w}}^{зс} = \frac{\bar{w}_1}{\bar{w}_0} = \frac{\sum q_1}{\sum T_1} \div \frac{\sum q_0}{\sum T_0} = \frac{118{,}6}{60} \div \frac{85}{50} = \frac{1{,}977}{1{,}7}} = 1{,}163 \text{ (+16,3 \%)}.$$

Виявляється, що продуктивність праці по тресту зросла на 16,3 %, тобто більше, ніж на кращій шахті № 1.

Такі випадки не поодинокі. Наприклад, середня заробітна плата збільшилась, незважаючи на те, що водночас одна частина колективу у травні заробила такі ж гроші, що й у квітні, а друга частина отримала менші гроші. Таким чином, індекс змінного складу може виходити за межі індивідуальних індексів. У нашому прикладі він показує зміну середньої продуктивності за рахунок зміни продуктивності по кожній шахті, а також за рахунок змін у структурі витрат часу.

Для того щоб визначити вплив першого фактору, обчислимо індекс фіксованого складу:

$$I_{\bar{w}}^{фс} = \frac{\sum w_1 T_1}{\sum w_0 T_1} = \frac{118{,}6}{2{,}0 \times 40 + 1{,}5 \times 20} = \frac{118{,}6}{110} = 1{,}078 \text{ (+7,8 \%)}.$$

На відміну від індексу змінного складу, індекс фіксованого складу ніколи не виходить за межі індивідуальних індексів. То ж, у середньому продуктивність праці зросла на 7,8 %.

Поглянемо на таблицю з обчисленими даними. Легко помітити, що:

1) продуктивність праці на першій шахті була вищою в кожному році;

2) частка відпрацьованого часу на першій шахті (а, як бачимо, використовується він більш ефективно, ніж на другій) істотно зросла (67 % проти 33 %).

Ці зміни у структурі відпрацьованого часу позитивно позначились на зміні середньої продуктивності:

$$I_{\bar{w}}^{сс} = \frac{\sum w_0 T_1}{\sum T_1} \div \frac{\sum w_0 T_0}{\sum T_0} = \frac{I_{\bar{w}}^{зс}}{I_{\bar{w}}^{фс}} = \frac{1{,}163}{1{,}078} = 1{,}079 \text{ (+7,9 \%)}.$$

Тепер визначимо зміну видобутку вугілля по тресту – загальну і за рахунок окремих факторів:

$$\Delta_q = \sum q_1 - \sum q_0 = 118{,}6 - 85 = 33{,}6 \text{ (тис. т).}$$

Фактори, які впливають на зміну загального видобутку вугілля, можуть бути різними, але врешті-решт їх можна звести до двох: продуктивності праці та загальних витрат часу:

$$\Delta q_w = (\bar{w}_1 - \bar{w}_0)\sum T_1 = (1{,}977 - 1{,}7) \times 60 = 16{,}6 \text{ (т с. т).}$$

$$\Delta q_T = (\sum T_1 - \sum T_0)\bar{w}_1 = (60 - 50) \times 1{,}7 = 17 \text{ (тис. т).}$$

Перевіримо:

$$\Delta q = \Delta q_w + \Delta q_T = 16{,}6 + 17 = 33{,}6 \text{ (тис. т).}$$

### 8.1.5. Територіальні індекси

Одним із різновидів індексів середніх величин є територіальні індекси. Територіальний індекс дозволяє порівняти середні рівні показників за окремими об’єктами, регіонами або країнами. Побудова системи територіальних індексів має певні особливості:

- необхідно обґрунтувати регіон (об’єкт), що буде обраний за базу порівняння;
- важливе значення має визначення порядку фіксації значень показників ($x_j$) і структурних складових ($d_j$).

Щодо бази порівняння, то вона може обратися довільно (залежно від мети порівняння) або базою порівняння повинен виступати однаковий рівень – це або середній рівень, або стандарт (еталонне значення). Середнє значення ознаки обчислюється за вертикальною структурою як середня арифметична зважена, а середнє значення для частки обчислюється за горизонтальним розподілом також як середня арифметична зважена за двома об’єктами (територіями).

Розглянемо правило та особливості побудови територіальних індексів змінного складу, фіксованого складу і структурних зрушень.

Індекс змінного складу має класичну формулу розрахунку і показує, у скільки разів середній рівень показника об'єкта А більше за середнє значення об'єкта Б:

$$I_{\bar{x}} = \frac{\sum x_A f_A}{\sum f_A} \div \frac{\sum x_Б f_Б}{\sum f_Б} = \frac{\sum x_A d_A}{\sum x_Б d_Б} \quad (8.18)$$

Таким чином, для зручності інтерпретації цього індексу краще об'єкт з меншим рівнем показників обирати базою порівняння. Однак це не є правилом. Можна інтерпретувати й інакше. Наприклад, якщо індекс змінного складу становив 0,9, то можна сказати, що середній рівень у регіоні А становить 90 % від середнього рівня в регіоні Б.

Індекс фіксованого складу будується залежно від того, чи порівнянні структури об'єктів, оскільки треба визначитися із фіксацією структури. Якщо структури порівнянні, то для фіксації може бути обраний або об'єкт А, або об'єкт Б. Наприклад, має місце такий розподіл часток (табл.8.7):

*Таблиця 8.7.* **До розрахунку територіальних індексів**

| Елементи | Структура об'єкта А, % | Структура об'єкта Б, % |
|---|---|---|
| 1 | 25 | 20 |
| 2 | 75 | 80 |
| Разом | 100 | 100 |

Це порівняння структура. Тоді формула індексу фіксованого складу може мати такий вигляд:

$$I_{\bar{x}} = \begin{cases} \frac{\sum x_A f_A}{\sum f_A} \div \frac{\sum x_Б f_A}{\sum f_A} = \frac{\sum x_A d_A}{\sum x_Б d_A} \\ \frac{\sum x_A f_Б}{\sum f_Б} \div \frac{\sum x_Б f_Б}{\sum f_Б} = \frac{\sum x_A d_Б}{\sum x_Б d_Б} \end{cases} \quad (8.19)$$

При розподілі часток як у таблиці 8.7 – маємо приклад, коли структури не порівнянні**.**

*Таблиця 8.8.* **До розрахунку територіальних індексів**

| Елементи | Структура об'єкта А, % | Структура об'єкта Б, % |
|---|---|---|
| 1 | 25 | 50 |
| 2 | 75 | 50 |
| Разом | 100 | 100 |

Тоді формула індексу фіксованого складу може мати такий вигляд:

$$I_{\bar{x}} = \begin{bmatrix} \frac{\sum x_A f_{AБ}}{\sum f_{AБ}} \div \frac{\sum x_Б f_{AБ}}{\sum f_{AБ}} = \frac{\sum x_A d_{AБ}}{\sum x_Б d_{AБ}} \\ \frac{\sum x_A d_{st}}{\sum x_Б d_{st}} \end{bmatrix}, \qquad (8.20)$$

де: $f_{AБ}$ – сума частот по двох об'єктах (територіях); $d_{AБ}$ – структура по двох регіонах, яка обчислюється за формулою

$$d_{\text{АБ}} = \frac{f_A + f_Б}{\sum(f_A + f_Б)}.$$

Індекс структурних зрушень будується на фіксації середнього рівня для двох об'єктів (територій) у розрізі окремих елементів і характеризує співвідношення середніх рівнів при зафіксованій структурі елементів:

$$I_{\bar{x}} = \frac{\sum \bar{x}_{AБ} f_A}{\sum f_A} \div \frac{\sum \bar{x}_{AБ} f_Б}{\sum f_Б} = \frac{\sum \bar{x}_{AБ} d_A}{\sum \bar{x}_{AБ} d_Б} \qquad (8.21)$$

де: $\bar{\text{x}}_{\text{АБ}} = \frac{\text{x}_\text{А} d_A + \text{x}_\text{Б} d_\text{Б}}{d_A + d_\text{Б}}$.

Важливо зауважити, що між територіальними індексами змінного, фіксованого складу та структурних зрушень не існує взаємозв'язку.

Розглянемо індексну систему на прикладі ефективності виробництва овочів в двох регіонах А і В (табл. 8.9).

*Таблиця* 8.9. **Ефективність виробництва овочів в регіонах**

| **Технологія виробництва** | **Валовий збір, тис. т** | | **Собівартість 1 т, гр. од.** | | |
|---|---|---|---|---|---|
| | **Регіон А ($f_A$)** | **Регіон Б ($f_Б$)** | **Регіон А ($x_A$)** | **Регіон Б ($x_Б$)** | **У середньому по двох регіонах ($\bar{x}_{АБ}$)** |
| Інтенсивна | 90 | 210 | 130 | 110 | (130×90+110×210) /(210+90)=116 |
| Традиційна | 210 | 140 | 150 | 160 | (150×210+160×140)/ (210+140)=154 |
| Разом | 300 | 350 | x | x | x |

За наведеними даним визначимо територіальні індекси середньої собівартості змінного, фіксованого складу і структурних зрушень. Побудуємо додаткову розрахункову таблицю 8.10.

*Таблиця* 8.10. **До розрахунку територіальних індексів**

| $d_A$ | $d_Б$ | $d_{АБ}$ | $x_А d_A$ | $x_Б d_Б$ | $x_А d_{АБ}$ | $x_Б d_{АБ}$ | $\bar{x}_{АБ} d_A$ | $\bar{x}_{АБ} d_Б$ |
|---|---|---|---|---|---|---|---|---|
| 90/300 =0,3 | 0,6 | (90+210)/ (300+350) =0,46 | 130× 0,3 | 110× 0,6 | 130× 0,46 | 110× 0,46 | 116× 0,3 | 116×0, 6 |
| 210/300 =0,7 | 0,4 | (210+140)/ (300+350) =0,54 | 150× 0,7 | 160× 0,4 | 150× 0,54 | 160× 0,54 | 154× 0,7 | 154×0, 4 |
| 1,0 | 1,0 | 1,00 | 144 | 130 | 140,8 | 137 | 142,6 | 131,2 |

Як бачимо з таблиці, структури валового збору за інтенсивною і традиційною технологіями непорівнянні: значна питома вага (60 %) всього виробництва в регіоні Б здійснюється за інтенсивною технологією, водночас в регіоні А за інтенсивною технологією виробляється тільки 30 % овочів. Це дозволяє припустити, що в регіоні Б собівартість усього виробництва менша і його доцільно обрати за базу порівняння. Тоді індекс змінного складу розраховується за формулою:

$$I_{\bar{x}} = \frac{\sum x_A d_A}{\sum x_Б d_Б} = \frac{144}{130} = 1{,}108 \text{ (більше на 10,8 %).}$$

Таким чином, середня собівартість виробництва овочів у регіоні А на 10,8 % вища, ніж у регіоні Б, що становить 14 гр. од. (144-130).

За непорівнянності структур індекс фіксованого складу середньої собівартості обчислюється за формулою

$$I_{\bar{x}} = \frac{\sum x_A d_{АБ}}{\sum x_Б d_{АБ}} = \frac{140{,}8}{137} = 1{,}028 \text{ (більше на 2,8 %).}$$

Це свідчить про те, що в середньому собівартість у регіоні А вище на 2,8 % за рахунок значно високого рівня собівартості за інтенсивною технологією.

Індекс структурних зрушень показує, що собівартість у регіоні А на 8,7 % вище, ніж у регіоні Б за рахунок орієнтації на виробництво овочів за традиційною технологією:

$$I_{\bar{x}} = \frac{\sum \bar{x}_{АБ} d_A}{\sum \bar{x}_{АБ} d_Б} = \frac{142{,}6}{131{,}2} = 1{,}087 \text{ (більше на 8,7 %).}$$

Таким чином, на високий рівень собівартості в регіоні А більшою мірою впливає фактор структури виробництва, а саме недооцінювання переваг виробництва за інтенсивною технологією.

## 8.2. Багатофакторні індексні системи

У навколишньому світі всі явища взаємопов'язані між собою, будь яке з них є наслідком дії певної множини причин і водночас причиною інших явищ. При статистичному моделюванні враховується характер зв'язку та особливості наявної інформації. За своїм характером зв'язки поділяють на *стохастичні* (різновидом стохастичних зв'язків є кореляційні), та жорстко детерміновані – *функціональні*. Перші відображають стохастичний характер причинно-наслідкових відносин, другі – адитивні чи мультиплікативні зв'язки між елементами розрахункових формул показників. Відповідно вибирається функціональна форма моделі: кореляційні зв'язки переважно описуються регресійними моделями, функціональні – балансовими або індексними. Отже, статистичний аналіз складних явищ, що змінюються в часі та просторі, їх загального зв'язку та взаємозумовленості, пов'язаний із застосуванням багатофакторних індексних систем.[1]

Основою індексної моделі є мультиплікативний зв'язок між певною множиною показників. Багатофакторний індексний аналіз дозволяє кількісно виміряти вплив декількох факторів $x_i$ на зміну того чи іншого економічного показника $y$, який розглядається як результативний:

$$y = x_1\, x_2 x_3 \ldots x_n \tag{8.22}$$

При вивченні функціональних зав'язків побудова багатофакторних індексних моделей, що відображають результативний показник як добуток взаємодії складових його факторів, має ґрунтуватися на знанні певних принципів, що

1 Єріна А. М. Статистичне моделювання та прогнозування [Текст]: підручник / А. М. Єріна, Д. Л. Єрін; Держ. ВНЗ "Київ. нац. екон. ун-т ім. Вадима Гетьмана". Київ: КНЕУ, 2014. 348 с.

випливають з об'єктивних особливостей взаємозв'язку між явищами. При цьому індексований складний показник, умовно приймається за результат, являє собою добуток безпосередньо пов'язаних між собою ознак - факторів, кількість яких може бути різною. Можливо побудова з п'яти, шести і більше факторів-множників.[2]

Послідовність факторів в індексній моделі не може бути довільною, вона визначається економічним змістом показників та методикою їх розрахунку. Кожний наступний фактор-множник розраховується на одиницю попереднього, і відповідно, добуток будь-якої кількості факторів є економічно змістовною величиною. Багатофакторні індексні моделі використовують при проведенні аналізу фінансової діяльності компанії та економічної ефективності (ефективності капітальних вкладень, аналізу динаміки оплати праці, собівартості, прибутку, рівня рентабельності тощо), для аналізу ефективності сфери послуг, соціально-економічного розвитку країни.

Наприклад, прибутковість активів компанії $y$ є функцією прибутковості продажу продукції $x_1$ та оборотності мобільних активів $z_1$, тобто $y = x_1 z_1$. Оборотності мобільних активів $z_1$, в свою чергу, є функцією оборотності матеріальних запасів $x_2$ та частки матеріальних запасів у мобільних активах $z_2$, отже, $y = x_1 x_2 z_2$. [3]

Схематично послідовність розширення моделі можна представити так:

$$y = x_1 z_1 = x_1 x_2 z_2 = x_1 x_2 z_m x_m \qquad (8.23)$$

2 Андрієнко В. Ю. Статистичні індекси в економічних дослідженнях / В. Ю. Андрієнко. К. : Академперіодика, 2004. 118с.

3 Єріна А. М. Статистичне моделювання та прогнозування [Текст]: підручник / А. М. Єріна, Д. Л. Єрін; Держ. ВНЗ "Київ. нац. екон. ун-т ім. Вадима Гетьмана". Київ: КНЕУ, 2014. 348 с.

При побудові багатофакторної індексної моделі наведена функція розглядається для двох періодів – базисного (взятий за базу порівняння попередній рік або квартал) і поточного:

$$y_0 = x_{10}\, x_{20}\, x_{30} \ldots x_{m0} \tag{8.24}$$

$$y_1 = x_{11}\, x_{21}\, x_{31} \ldots x_{m1} \tag{8.25}$$

У практиці статистики обчислення багатофакторної індексної моделі може проводиться щомісячно, щоквартально та за рік. Період, з яким проводиться порівняння, вважається базисним.

Характерною рисою мультиплікативної індексної моделі є взаємозв'язок факторів: чисельник розрахункової формули одного фактору є знаменником розрахункової формули наступного. Введення в ланцюгову схему нового фактору означає деталізацію функціонального зв'язку і не змінює його сутності. А ступінь деталізації залежить від мети дослідження. Отже, при дослідженні взаємопов'язаних дії багатьох чинників на загальний результативний показник використовується прийом виявлення відокремленого впливу кожного фактора шляхом послідовної зміни факторів. З цією метою вплив всіх інших включених у модель факторів елімінують (ізолюють).

За визначенням А.М. Єріної «абсолютну і відносну зміну показника-функції *y* можна розкласти за факторами-множниками $x_i$»[4].

В рамках індексної моделі, в якій відтворюються взаємозв'язки між показниками, здійснюється оцінювання ступеня та абсолютного розміру впливу кожного з них на динаміку функції:

4 Єріна А. М. Статистичне моделювання та прогнозування [Текст]: підручник / А. М. Єріна, Д. Л. Єрін; Держ. ВНЗ "Київ. нац. екон. ун-т ім. Вадима Гетьмана". Київ: КНЕУ, 2014. 348 с.

$$I_y = I_{x_1}\, I_{x_2} I_{x_3}\, \dots I_{x_m} \tag{8.26}$$

Для виявлення відокремленого впливу кожного фактору при розрахунку частинного індексу $I_{x_i}$ всі фактори-множники, окрім $x_i$ , фіксуються на постійному рівні. Найчастіше фактори, розміщені в ланцюгу перед $x_i$ , фіксуються на рівні поточного періоду, а розміщені після $x_i$ — на рівні базисного періоду. Принцип послідовно-ланцюгового елімінування впливу (відносна зміна) фактору $x_2$ в моделі $y = x_1\, x_2\, x_3$ реалізується таким чином:

$$I_{x_2} = \frac{x_{11}\, x_{21}\, x_{30}}{x_{11}\, x_{20}\, x_{30}} \tag{8.27}$$

Абсолютний вплив фактору $x_2$ на $y$ визначається за такою ж схемою:

$$\mathrm{A}_{x_2} = x_{11}\, (x_{21} - x_{20})\, x_{30} \tag{8.28}$$

Абсолютний вплив факторів можна визначити з використанням $I_{x_i}$ – відповідних частинних індексів. При послідовному множенні (за ланцюговою схемою) базисного рівня показника-функції на індекси факторів визначаються розрахункові рівні, тобто такі рівні, які мав би показник $y$ під впливом *i*-го фактора та при незмінному рівні решти факторів.

Методику побудови багатофакторної індексної моделі розглянемо на прикладі аналізу взаємозв'язків економічних, демографічних та інших аспектів функціонування системи соціального страхування, результати якого впроваджені у практичну роботу Фонду соціального страхування України.[5]

5 Мазуренко О. К., Горна М.О. Методичні підходи до формування стратегії розвитку у системі послуг соціального захисту / О. К. Мазуренко, М. О. Горна // Економічний аналіз. 2015. Т. 19, № 1. С. 69-75.

Сформуємо систему первинних абсолютних показників, що характеризують економічні, демографічні та інші аспекти функціонування системи соціального страхування як елемента національної економіки України. Так як система показників, що використовується для аналізу, має характеризувати попит і пропозицію по даному виду соціальних послуг, їх ресурсне забезпечення, в тому числі оцінку трудових, матеріальних та фінансових ресурсів, то включаємо у модель:

валовий внутрішній продукт (ВВП);

чисельність населення (Н);

чисельність населення працездатного віку (ПН);

чисельність зайнятого населення (ЗН);

чисельність застрахованих осіб у системі соціального страхування (ЗО);

страхові виплати (пенсії за віком, допомога по безробіттю, оплата лікарняних, допомога постраждалим на виробництві тощо) загальнообов'язкового державного страхування (СВ);

фонд оплати праці (ФОП);

доходи фондів соціального страхування (Д);

витрати фондів соціального страхування (В).[6]

Наступним кроком аналізу є оцінювання в динаміці впливу факторів на якість виконання обов'язків страховика щодо надання послуг соціального страхування. Методику побудови багатофакторної індексної моделі розглянуто на прикладі взаємозв'язку показника середнього страхового внеску *у* з індикаторами соціально-демографічного стану. Для кожного фонду соціального страхування середній страховий внесок розраховуються відношенням одержаних надходжень до застрахованих осіб.

6 Мазуренко О. К., Горна М.О. Методичні підходи до формування стратегії розвитку у системі послуг соціального захисту / О. К. Мазуренко, М. О. Горна // Економічний аналіз. 2015. Т. 19, № 1. - С. 69-75.

Динаміку цього показника можна розкласти за такою множиною факторів:

$x_1$- частка страхових внесків у фонді оплати праці;

$x_2$- середня заробітна плата;

$x_3$- рівень зайнятості застрахованих осіб.

Взаємозв'язок між цими показниками має вигляд:

$$\frac{\text{доходи фондів}}{\text{чисельність застрахованих осіб}} = \frac{\text{доходи фондів}}{\text{фонд оплати праці}} \cdot \frac{\text{фонд оплати праці}}{\text{чисельність зайнятих застрахованих осіб}} \cdot \frac{\text{чисельність зайнятих застрахованих осіб}}{\text{чисельність застрахованих осіб}} \quad (8.29)$$

А отже, мультиплікативний зв'язок між відповідними індексами виражений формулами:

$$I_y = I_{x_1} I_{x_2} I_{x_3} \quad (8.30)$$

Відповідно:

$$\frac{x_{11}\, x_{21}\, x_{31}}{x_{10}\, x_{20}\, x_{30}} = \frac{x_{11}\, x_{20}\, x_{30}}{x_{10}\, x_{20}\, x_{30}} \cdot \frac{x_{11}\, x_{21}\, x_{30}}{x_{11}\, x_{20}\, x_{30}} \cdot \frac{x_{11}\, x_{21}\, x_{31}}{x_{11}\, x_{21}\, x_{30}} \quad (8.31)$$

Ваги в індексах-співмножниках фіксуються за схемою: в індексі першого фактора (частка страхових внесків у фонді оплати праці) – на рівні базисного періоду, в індексі другого фактора (середня заробітна плата) – ті, що праворуч від індексованої величини, на рівні базисного періоду, ті що ліворуч, – на рівні поточного періоду, а в індексі третього фактору (рівень зайнятості застрахованих осіб) – всі ваги фіксуються на рівні поточного періоду (вони розміщені ліворуч від індексованої величини).

Порядок розрахунку зміни будь-якого фактору на динаміку середнього страхового внеску схематично можна представити, якщо базисний його рівень позначити $y_0$, розрахунковий рівень для першого фактору – $y'$, для другого — $y''$, для другого — $y'''$, то тоді:

$$y_0 \longrightarrow \underset{\llcorner \quad A_{x_1} \quad \lrcorner}{\left(y' = I_{x_1} y_0\right)} \longrightarrow \underset{\llcorner \quad A_{x_2} \quad \lrcorner}{\left(y'' = I_{x_2} y'\right)} \longrightarrow \underset{\llcorner \quad A_{x_3} \quad \lrcorner}{\left(y''' = I_{x_3} y''\right)}$$ [7]

Так, у базисному році середній страховий внесок всіх видів загальнообов'язкового державного соціального страхування становив 107,26 умовних одиниць, а в поточному році – 113,39 умовних одиниць, тобто зріс на 6,13 умовних одиниці (у.о.), а індекс середнього страхового внеску становить – 1,057. Індекси введених у модель факторів-множників і розрахунок вкладу кожного з них в абсолютний приріст середнього страхового внеску подано в табл.8.11.

*Таблиця 8.11.* **До розрахунку багатофакторної індексної моделі середнього страхового внеску**

| Фактор | Індекс *i*-го фактора | Абсолютний внесок *i*-го фактора в приріст середнього страхового внеску, умовних одиниць (у.о.) |
|---|---|---|
| частка страхових внесків у фонді оплати праці | 0,927 | – 1,22 |
| середня заробітна плата | 1,134 | + 6,5 |
| рівень зайнятості застрахованих осіб | 1,005 | + 0,85 |

*Джерело*: розраховано за даними Державної служби статистики України. URL: http://www.ukrstat.gov.ua/

7 Єріна А. М. Статистичне моделювання та прогнозування [Текст]: підручник / А. М. Єріна, Д. Л. Єрін; Держ. ВНЗ "Київ. нац. екон. ун-т ім. Вадима Гетьмана". Київ: КНЕУ, 2014. 348 с.

За даними таблиці 8.11, абсолютний приріст середнього страхового внеску в розмірі 6,13 у.о. розкладено за факторами. Два фактори, окрім частки страхових внесків у фонді оплати праці, мали позитивний вплив на динаміку середнього страхового внеску. Серед них найвагоміший вплив фактора $x_2$ – середня заробітна плата, на наступному місці фактор – рівень зайнятості застрахованих осіб. Зауважимо, що негативний вплив фактора $x_1$– частки страхових внесків у фонді оплати праці, пояснюється зростанням рівня тіньової економіки.[8]

## 8.3. Соціально-економічна нормаль

Систему взаємозв'язаних показників можна представити у матричному вигляді. На головній діагоналі матриці за певною стратегією розміщуються *m* абсолютних величин $q_i$, на основі яких можна визначити *m(m – 1)* відносних величин $x_i = \frac{q_i}{q_j}$, де $i \neq j$. Очевидно, що недіагональні елементи, симетрично розташовані щодо головної діагоналі, є оберненими одна до одної величинами, тобто $x_{ij} = \frac{1}{x_{ij}}$ . Система взаємозв'язаних абсолютних і відносних величин утворює квадратну матрицю. Аналогічно складається матриця індексів.[9]

8 Мазуренко О.К., Горна М.О. Теоретико-методологічні засади аналізу діяльності фондів соціального страхування /Соціальне забезпечення в контексті вступу в ЄС: монографія / За загальною редакцією Дерій Ж.В. К.: Видавничий дім «Кондор», 2017. С.6-28.

9 Єріна А. М. Статистичне моделювання та прогнозування [Текст]: підручник / А. М. Єріна, Д. Л. Єрін; Держ. ВНЗ "Київ. нац. екон. ун-т ім. Вадима Гетьмана". - Київ: КНЕУ, 2014. 348 с.

Для оцінки динамічності реакції системи загальнообов'язкового державного соціального страхування України, базуючись на взаємозв'язку динаміки окремих показників, використаємо метод побудови соціально-економічної нормалі. У статистиці соціально-економічна нормаль являє собою теоретично обґрунтоване оптимальне співвідношення темпів зміни показників системи обслуговування населення послугами державного соціального страхування, яке забезпечує гармонійний розвиток і стабільну якість роботи соціальної сфери. Порядок взаємозв'язку факторів, які можуть бути включені в соціально-економічну нормаль, залежить від мети дослідження й аналізу.[10]

Вишикуємо темпи зростання сформованої вище системи первинних абсолютних показників у соціально-економічної нормалі. Розташовані в симетричних відносно одиничної діагоналі клітинах темпи зростання побудовані по взаємообернених показниках, і тому змінюються у протилежних напрямках при вдосконаленні якості послуг соціального страхування. Якщо $K_{\text{ВВП/ФОП}} < 1$, то $K_{\text{ВВП/ФОП}} \geq 1$. Проведено оцінку кожної пари темпів зростання по взаємообернених відносних показниках та одержана матриця, в якій елементи, що опинились під головною діагоналлю, відповідають умовам гармонійного розвитку і стабільності діяльності фондів соціального страхування (табл. 8. 12):

В такий спосіб визначено соціальні, демографічні та економічні фактори впливу на систему загальнообов'язкового державного соціального страхування України, та базуючись на їх взаємозв'язку теоретично обґрунтовано, що для своєчасного та якісного обслуговування населення послугами державного соціального страхування оптимальними співвідношеннями є:

10 Мазуренко О. К., Горна М.О. Методичні підходи до формування стратегії розвитку у системі послуг соціального захисту / О. К. Мазуренко, М. О. Горна // Економічний аналіз. 2015. Т. 19, № 1. С. 69-75.

- темп зростання ВВП на одну особу $K_{ВВП}$ не менший, ніж темп зростання відношення фонду оплати праці до ВВП $K_{ФОП}$;
- темп зростання відношення фонду оплати праці до ВВП $K_{ФОП}$ не менший, ніж темп зростання середніх страхових виплат $K_{СВ}$;
- темп зростання середніх страхових виплат $K_{СВ}$ не менший, ніж темп зміни частки застрахованих осіб серед зайнятого населення $K_{ЗО}$;
- темп зміни частки застрахованих осіб серед зайнятого населення $K_{ЗО}$ не менший, ніж темп зростання частки зайнятого серед працездатного населення $K_{ЗН}$;
- темп зростання частки зайнятого до працездатного населення $K_{ЗН}$ не менший, ніж темп зміни частки чисельності працездатного до всього населення $K_{ПН}$ , де $K$ – темпи зростання (зменшення) показників[11].

*Таблиця 8.12.* **До розрахунку показників соціально-економічної нормалі**

| | *ВВП* | *ФОП* | *СВ* | *ЗО* | *ЗН* | *ПН* | *Н* |
|---|---|---|---|---|---|---|---|
| *ВВП* | *1* | | | | | | |
| *ФОП* | *К' ВВП/ФОП* | *1* | | | | | |
| *СВ* | *К'ВВП/СВ* | *К'ФОП/СВ* | *1* | | | | |
| *ЗО* | *К'ВВП/ЗО* | *К'ФОП/ЗО* | *К'СВ/ЗО* | *1* | | | |
| *ЗН* | *К'ВВП/ЗН* | *К'ФОП/ЗН* | *К'СВ/ЗН* | *К'ЗО/ЗН* | *1* | | |
| *ПН* | *К'ВВП/ПН* | *К'ФОП/ПН* | *К'СВ/ПН* | *К'ЗО/ПН* | *К' ЗН/ПН* | *1* | |
| *Н* | *К'ВВП/Н* | *К'ФОП/Н* | *К'СВ/Н* | *К'ЗО/Н* | *К'ЗН/Н* | *К'ПН/Н* | *1* |

11 Мазуренко О.К., Горна М.О. Теоретико-методологічні засади аналізу діяльності фондів соціального страхування /Соціальне забезпечення в контексті вступу в ЄС: монографія / За загальною редакцією Дерій Ж.В. К.: Видавничий дім «Кондор», 2017. С.6-28

Одержана соціально-економічна нормаль має вигляд:

$$K_{\text{ВВП}} > K_{\text{ФОП}} > K_{\text{СВ}} > K_{\text{ЗО}} > K_{\text{ЗН}} > K_{\text{ПН}} \qquad (8.31)$$

де: $K$ – темпи зростання (зменшення) показників,.

У таблиці 8.13 наведено результати аналізу умов діяльності системи соціального страхування та соціально-демографічного стану в Україні на основі одержаної релевантної нормалі.

Послідовність розміщення показників у моделі відповідає економічній нормалі, тобто стратегії розвитку економічної системи соціального страхування, згідно з якою темпи зростання кінцевих результатів мають бути вищими за темпи зростання витрат і ресурсів. Заливка даних означає, що значення показника не відповідає умовам нормалі.

Аналіз динаміки наведених показників дозволяє зробити наступний висновок. Отримані нерівності свідчать про те, що фактичні співвідношення показників – індикаторів умов діяльності фондів соціального страхування та соціально-демографічного стану в Україні були наближені до оптимальних і відповідали вище зазначеним припущенням лише у перші роки досліджуваного періоду та частково у 4-7 роках. Довготривалі посткризові процеси спричинили невідповідність нерівності у 8-10 роках, темпи зростання фонду оплати праці до валового внутрішнього продукту менші за темпи зростання середніх страхових виплат.

*Таблиця 8.13.* **Динаміка співвідношень показників соціально-економічної нормалі**

| Порядковий номер періоду | *Темп зростання* ВВПу розрахунку на одну особу | *Темп зростання* Фонду оплати праці (ФОП) до ВВП | *Темп зростання* середніх страхових виплати (СВ) | *Темп зростання* частки застрахованих осіб серед зайнятих (ЗО) | *Темп зростання* частки зайнятого населення серед працездатного (ЗН) | *Темп зростання* частки працездатного населення (ПН) |
|---|---|---|---|---|---|---|
| 1 | 1,332 | 1,021 | 1,085 | 1,016 | 1,013 | 1,001 |
| 2 | 1,322 | 1,302 | 1,283 | 1,022 | 1,005 | 1,004 |
| 3 | 0,967 | 0,984 | 1,121 | 0,893 | 0,963 | 1,005 |
| 4 | 1,248 | 1,202 | 1,204 | 1,007 | 1,005 | 1,002 |
| 5 | 1,209 | 1,046 | 1,095 | 1,088 | 1,003 | 1,002 |
| 6 | 1,083 | 1,027 | 1,054 | 1,008 | 1,006 | 0,998 |
| 7 | 1,046 | 1,058 | 1,025 | 1,002 | 1,007 | 0,998 |
| 8 | 1,075 | 1,135 | 1,044 | 0,949 | 0,913 | 1,050 |
| 9 | 1,266 | 0,970 | 1,032 | 1,007 | 0,953 | 0,939 |
| 10 | 1,209 | 0,903 | 1,021 | 1,002 | 0,980 | 0,999 |

Протягом кризових для України років, коли відбувається заглиблення економічної та соціально-демографічної кризи, не виконується майже жодна з умов соціально-економічної нормалі. Старіння населення України спричиняє зменшення частки працездатного населення на протязі останніх років, а ситуація на ринку праці (безробіття та високий рівень тіньових зарплат) призводить до зниження темпів зростання частки застрахованих осіб серед зайнятих в порівнянні з часткою зайнятого населення серед працездатного. Щодо зростання ВВП у розрахунку на одну особу необхідно зауважити, що цей показник дійсно зростає, але тільки у гривневому еквіваленті, незважаючи на зростання номінального валового внутрішнього продукту, населення України дедалі біднішає. Сучасний стан характеризується низькими показниками якості життя, несприятливою демографічною ситуацією, поганим

станом здоров'я населення, соціальною нерівністю, значним поширенням бідності та безробіття. Все це ускладнює виконання фондами соціального страхування обов'язків по забезпеченню застрахованих осіб своєчасно і якісно належними послугами та виплатами.

Основними напрямками використання одержаної нормалі є:

• аналіз по обласних відділеннях фондів соціального страхування, групування областей по групах з повною або частковою невідповідністю нормалі;

• побудова індексної системи, що характеризуватиме вплив факторів на якість надання послуг соціального страхування.

## 8.4. Особливості використання індексного методу

Індекс будь-якого показника, що характеризує результативність інвестиційної діяльності, містить фактор нерівномірності розвитку[12] окремих функціональних елементів інвестиційної системи (підприємств, галузей, регіонів, ринків) і, таким чином, визначає структурну неоднорідність. Нерівномірність розвитку окремих складових інвестиційного процесу неминуче веде до зміни видової структури і технічного рівня виробничого капіталу. У зв'язку з цим виникає потреба у вивченні впливу факторів нерівномірності на динаміку результативного показника. Інакше кажучи, існує необхідність проведення аналізу структурних зрушень та визначення їх впливу на динаміку результатів інвестиційної діяльності на всіх рівнях.

Структурні зрушення можна оцінити і за допомогою абсолютних змін питомих ваг, виражених у відсоткових пунктах.

[12] Нерівномірність розвитку інвестиційної діяльності характеризується якісною та кількісною неоднаковістю процесів, що відбуваються в інвестиційній сфері.

Однак при аналізі результативності інвестиційної діяльності важливо не тільки оцінити інтенсивність структурних зрушень, але й виміряти ступінь їх впливу на динаміку середнього рівня досліджуваного показника.

Це завдання вирішують за допомогою індексів структурних зрушень. Як відомо з теорії статистики, у найпростішому варіанті ця система включає три індекси: індекс динаміки середніх рівнів (індекс змінного складу), індекс середніх рівнів за умови незмінної структури та індекс структурних зрушень. Особливо корисно використовувати цю систему при оцінюванні впливу структурного фактора на зміну дохідності інвестиційного портфеля або портфеля цінних паперів. Як відомо, структура портфелю інвестицій визначає тип інвестора та основну мету інвестиційної діяльності: (переважання в портфелі низькодоходних цінних паперів характеризує консервативність інвестора і його несхильність до ризику; і навпаки, віддання переваги цінним паперам з високим рівнем дохідності дає підстави для висновку про агресивний характер інвестора і його схильність до ризику).

Крім того, важливою складовою портфельного аналізу є динаміка структури портфеля вкладень. Якщо структура портфеля досить постійна, це свідчить про пасивну політику управління портфелем вкладень. А про існування усталеної тенденції до зміни структури портфеля свідчитимуть істотні структурні зрушення в портфелі вкладень, що дозволяє вважати відповідну політику управління інвестиційним портфелем активною. За допомогою системи індексів середньої дохідності інвестиційного портфеля можна відстежувати динаміку політики управління портфелем вкладень. Можна застосовувати як просту, так і складну систему взаємопов'язаних індексів середньої дохідності.

У простій схемі індекс змінного складу середньої дохідності будують для портфеля цінних паперів, і він характеризує зміну середнього рівня дохідності як у результаті динаміки рівня

дохідності окремих цінних паперів (*i*), так і в результаті структурних зрушень у портфелі цінних паперів (*d*):

$$I_{\bar{i}} = \frac{\sum i_1 d_1}{\sum i_0 d_0} \tag{8.32}$$

Індекс фіксованого складу елімінує вплив структурних зрушень у портфелі на зміну середньої дохідності. При його побудові ваги, відповідно до прийнятої системи зважування, приймають на рівні звітного періоду, але для потреб аналізу дохідності портфеля цінних паперів доцільно обрати за вагу базовий період, оскільки в чисельнику ми можемо отримати таку важливу характеристику, як середній рівень дохідності поточного періоду за умови незмінної структури:

$$I_i = \frac{\sum i_1 d_0}{\sum i_0 d_0} \tag{8.33}$$

Використання класичного принципу фіксації на поточному періоді для цього індексу недоцільне, оскільки отримана в знаменнику величина буде характеризувати середню дохідність минулого періоду за умови поточної структури. Але для розв'язання цієї проблеми у статистичній літературі пропонують скористатися принципом, який уперше запропонував Фішер, оскільки це сприятиме більш об'єктивному аналізу. Основна ідея індексу Фішера – врахування як базового, так і поточного періодів. У нашому випадку індекс буде розраховано як середню геометричну з індексів середньої дохідності, побудованих на основі поточної та базової структури:

$$I_i^F = \left( \frac{\sum i_1 d_1}{\sum i_0 d_1} \times \frac{\sum i_1 d_0}{\sum i_0 d_0} \right)^{1/2} \tag{8.34}$$

В індексі структурних зрушень класично незмінними (на рівні базисного періоду) приймаються значення дохідності окремих цінних паперів:

$$I_d = \frac{\sum i_0 d_1}{\sum i_0 d_0}$$

Однак, враховуючи попередні зауваження щодо побудови індексу фіксованого складу, необхідно використовувати або поточну фіксацію:

$$I_d = \frac{\sum i_1 d_1}{\sum i_1 d_0},$$

або застосувати індекс Фішера:

$$I_d^F = \left( \frac{\sum i_1 d_1}{\sum i_1 d_0} \times \frac{\sum i_0 d_1}{\sum i_0 d_0} \right)^{1/2} \qquad (8.35)$$

У межах цієї індексної системи, яка характеризує динаміку середнього рівня дохідності портфеля цінних паперів, тобто основного показника ефективності фінансових інвестицій, індекс структурних зрушень вимірює вплив структурної зміни того інвестиційного ресурсу, ефективність якого він відображає, на динаміку середньої дохідності.

При дослідженні впливу структурних зрушень на результативність інвестиційної діяльності використання одного індексу структурних зрушень не достатньо. Це пов'язано насамперед з тим, що важливо не стільки кількісно оцінити цей вплив, скільки визначити, які саме структурні зрушення відбулися і які з них істотно вплинули на динаміку інвестиційної діяльності.

Тому, аналізуючи структурні зрушення в інвестиційному портфелі, який може містити різні інвестиційні інструменти, застосовують складнішу систему взаємопов'язаних індексів. Але тоді важливо врахувати наявність вкладеної структури: за першою ознакою (за напрямами вкладень – *D*) і за другою ознакою (за інвестиційними інструментами – *d*). При побудові такої індексної системи використовують групові середні відповідно до базового і поточного періодів:

$$\bar{i}_0 = \sum \left( \sum i_0 d_0 \right) D_0 \qquad (8.36)$$

$$\bar{i}_1 = \sum\left(\sum i_1 d_1\right) D_1 \tag{8.37}$$

Індекс змінного складу цієї системи характеризує динаміку середньої дохідності інвестиційного портфеля за умови існування розподілу всередині системи за двома групувальними ознаками:

$$I_{\bar{i}} = \frac{\sum\left(\sum i_1 d_1\right) D_1}{\sum\left(\sum i_0 d_0\right) D_0} \tag{8.38}$$

Індекс фіксованого складу, що визначає зміну середньої дохідності за умови постійного складу інвестиційного портфеля, який доцільно фіксувати або на базовому рівні, як і у попередній індексній системі, відступаючи від класичного представлення індексу:

$$I_{id} = \frac{\sum\left(\sum i_1 d_1\right) D_0}{\sum\left(\sum i_0 d_0\right) D_0} \tag{8.39}$$

або скористатися формулою Фішера:

$$I_{id}^F = \left( \frac{\sum\left(\sum i_1 d_1\right) D_0}{\sum\left(\sum i_0 d_0\right) D_0} \times \frac{\sum\left(\sum i_1 d_1\right) D_1}{\sum\left(\sum i_0 d_0\right) D_1} \right)^{1/2} \tag{8.40}$$

Індекс структурних зрушень, який вимірює вплив змін у розподілі між напрямами інвестування на динаміку середньої дохідності, але без урахування структури інвестиційних інструментів, доцільно розраховувати за умови поточного зважування:

$$I_D = \frac{\sum\left(\sum i_1 d_1\right) D_1}{\sum\left(\sum i_1 d_1\right) D_0} \tag{8.41}$$

або за формулою Фішера

$$I_D^F = \left( \frac{\sum\left(\sum i_1 d_1\right) D_1}{\sum\left(\sum i_1 d_1\right) D_0} \times \frac{\sum\left(\sum i_0 d_0\right) D_1}{\sum\left(\sum i_0 d_0\right) D_0} \right)^{1/2} \tag{8.42}$$

Крім того, у межах цієї індексної системи можна оцінити вплив структурних зрушень усередині груп:

$$I_d = \frac{\sum\left(\sum i_1 d_1\right) D_0}{\sum\left(\sum i_1 d_0\right) D_0} \tag{8.43}$$

а також загальні структурні зрушення:

$$I_{dD} = \frac{\sum\left(\sum i_1 d_1\right) D_1}{\sum\left(\sum i_1 d_0\right) D_0} \tag{8.44}$$

Між частковими індексами структурних зрушень можна побудувати взаємозв'язок:

$$I_{dD} = I_d \times I_D .$$

Загальний взаємозв'язок матиме такий вигляд:

$$I_{\bar{i}} = I_i \times I_d \times I_D$$

Методичною особливістю побудови системи взаємозалежних індексів середніх величин є порівнянність структурних частин досліджуваної сукупності в часі. У процесі зміни структури інвестиційного портфеля деякі інструменти можуть продаватися або зникати, а інші – купуватися або з'являтися. У такому разі непорівнянність структурних частин портфеля унеможливлює розкладання індексу середнього рівня за розглянутою схемою. Виникає необхідність коригування індексної системи. Така індексна система включає, поряд з індексами фіксованого складу і структурних зрушень, розрахованих для сукупності однотипних інструментів, індекси, що характеризують вплив нових структурних частин і зниклих елементів портфеля. Тоді індексна система набуває вигляду:

$$I_{\bar{i}} = I_i \times I_{d^0} \times I_{d^+} \times I_{d^-}$$

Усі індекси розраховуються співвідношенням середніх рівнів дохідності за два періоди.

Індекс змінного складу характеризує динаміку середньої дохідності портфеля цінних паперів під впливом усіх факторів:

$$I_{\bar{i}} = \frac{\sum i_1 d_1}{\sum i_0 d_0}.$$

Індекс фіксованого складу показує, як у середньому змінилася дохідність портфеля з однотипним складом цінних паперів:

$$I_i = \frac{\sum i_1 d_1^0}{\sum i_0 d_1^0} \tag{8.45}$$

Вплив перерозподілу цінних паперів у портфелі з однотипним складом цінних паперів характеризує індекс структурних зрушень:

$$I_{d^0} = \frac{\sum i_1 d_1^0}{\sum i_1 d_0^0} \tag{8.46}$$

Індекси структурних зрушень $I_{d^+}, I_{d^-}$ показують, як змінилася середня дохідність портфеля, відповідно, унаслідок купівлі нових і продажу (погашення) старих цінних паперів:

$$I_{d^+} = \frac{\sum i_1 d_1}{\sum i_1 d_1^0}, \tag{8.47}$$

$$I_{d^-} = \frac{\sum i_0 d_0^0}{\sum i_0 d_0}. \tag{8.48}$$

Аналогічно можна розкласти абсолютну зміну середньої дохідності портфеля цінних паперів залежно від зміни середнього рівня індексованого показника.

## 8.5. Узагальнюючі характеристики ринку цінних паперів. Фондові індекси

Для характеристики рівня та динаміки цін на акції та інші цінні папери, що перебувають в обігу, одержання загальної уяви про стан і розвиток фондового ринку, тобто для поточного та перспективного аналізу ринку, використовують фондові індекси. На першому етапі визначають середню ціну акції за формулою середньої арифметичної:

$$\overline{P} = \frac{\sum_{i=1}^{n} P_{i,t}}{n},$$

де $\overline{P}$ – середня ціна акції; $P_{i,t}$ – ціна кожної *i*-ї акції в момент часу *t*; *n* – кількість акцій.

Зміни середньої ціни акції в часі характеризують зростання або зниження курсу акцій. У разі значного коливання цін для розрахунку використовують формулу середньої геометричної

$$\overline{P} = \sqrt[n]{P_1 \cdot P_2 \cdot P_3 \cdot ... \cdot P_n}.$$

На фондовому ринку з кожним днем складається дедалі більше різноманітних індексів. Для аналізу та макроекономічних зіставлень, об'єктивної оцінки динаміки цінової ситуації на фондовому ринку, оцінки якості цінних паперів, на практиці використовують такі показники:

- інтегральні індекси, що відображають динаміку фондового ринку загалом або важливого його сегмента;
- галузеві (субсекторні) індекси, на їхній основі формують узагальнені, інтегральні індекси;
- рейтингові оцінки, на основі яких дається якісна та кількісна оцінка окремих видів цінних паперів.

Для обчислення фондових індексів використовують такі характеристики:

1. Список індексу (набір акцій – представників). Під час складання списку індексу критеріями відбору акцій корпорацій є забезпечення їхньої репрезентативності та надійності самої корпорації, що випускає акцію.

2. Метод усереднення. Вибір методу розрахунку: середньої арифметичної або середньої геометричної величини.

3. Види ваг, які використовують під час зважування акцій, що входять до списку індексу. Це можуть бути такі вагові показники: курсова вартість акції корпорації (індекс із ціновим зважуванням) або капіталізація корпорації-емітента (індекс з ринковим зважуванням).

4. Базове значення індексу – величина індексу в періоді, що прийнятий за базу. Окрім індексів з ціновим зважуванням, усім індексам притаманна така характеристика. Базове значення індексу для зручності зазвичай округлюється до 10, 50, 100 чи 1000.

5. Статистична база, на основі якої виконуються розрахунки індексу. Це можуть бути показники, які характеризують результати торгів на біржовому або позабіржовому ринку цінних паперів, чи на їх сукупності.

З огляду на методику обчислення фондових індексів, їх можна поділити на дві групи:

– індекси стану, в основу яких покладено ціновий фактор і вони мають вартісну розмірність;

– індекси динаміки, в основу яких покладено не саму вартість, а величину її змін за відповідний період часу.

Більшість індексів належить до першої групи. Вони розраховуються як середня арифметична з визначених до розгляду акцій (зважених або незважених за кількістю їх продажу). В основу другого типу індексів покладено добуток темпів зростання цін акцій, які ввійшли в лістинг з котирування за поточний день порівняно з

попереднім днем. Загальне значення індексу обчислюється за формулою середньої геометричної з цієї величини.

У світовій практиці використовують методичні прийоми обчислення фондових індексів, які базуються на розрахунку:

темпів зростання (зниження) середньоарифметичної ціни акцій обмеженої кількості обраних корпорацій (*індекс Dow Jones*);

- темпів зростання (зниження) середньозваженої (за кількістю акцій, що перебувають в обігу) ціни всієї множини акцій корпорацій (*індекс Standard & Poor's );*
- середньогеометричного значення темпів зростання (зниження) цін акцій (*Value line composite index*).

Розрахунок середньої ціни акцій та її індексування надає фондовому індексу зручну та зіставну форму. За всіх рівних умов за основу розрахунку обирають метод середньої геометричної величини:

$$I = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{P_{i,t}}{P_{i,0}} \quad , \tag{8.50}$$

або $$I = \sqrt[n]{\prod_{i=1}^{n} \frac{P_{i,t}}{P_{i,0}}} \tag{8.51}$$

Для оцінювання стану та динаміки фондового ринку в науковій літературі пропонуються такі варіанти розрахунку фондових індексів :

1. Індекс поточного стану фондового ринку

$$I_t = \frac{\sum_{i=1}^{n} P_{i,t}}{\sum_{i=1}^{n} P_{i,t\text{-}1}} \quad , \tag{8.52}$$

де $P_{i,t}$ – ціна продажу *i*-ї акції в *t*-й біржовий день; якщо в цей день акція не продавалась, то – ціна попиту; $P_{i,t-1}$ – ціна продажу *i*-ї акції в попередній біржовий день.

2. Базовий індекс фондового ринку:

$$I_0 = \frac{\sum_{i=1}^{n} P_{i,t}}{\sum_{i=1}^{n} P_{i,0}}, \tag{8.53}$$

де $P_{i,0}$ – номінальна ціна *i*-ї акції.

3. Індекс, зважений на обсяг ринкової капіталізації**:**

3.1. $$I_t = \frac{\sum P_{i,t} \cdot q_{i,0}}{\sum P_{i,0} \cdot q_{i,0}} I_0, \tag{8.54}$$

де: $P_{i,t}$ – ціна акції *i*-го виду в поточний момент часу; $P_{i,0}$ – ціна акції *i*-го виду в базовий момент часу; $q_{i,0}$ – кількість акцій *i*-го виду в обігу в базовий момент часу; $P_{i,t} \cdot q_{i,0}$ – ринкова вартість, або капіталізація акцій *i*-го виду, розрахована на основі поточної ціни і кількості акцій в обігу у базовий момент часу; $P_{i,0} \cdot q_{i,0}$ – показник капіталізації базового періоду; $I_0$ – базове значення індексу.

Для безпосереднього урахування властивості транзитивності в часі, використовується модифікована формула:

3.2. $$I_t = \frac{\sum P_{i,t} \cdot q_{i,t}}{\sum P_{i,t-1} \cdot q_{i,t}} I_{t-1}, \tag{8.55}$$

де: $q_{i,t}$ – кількість акцій *i*-го виду в обігу за новим складом сукупності; $I_{t-1}$ – значення індексу на момент, що передував зміні сукупності.

4. Індекс, розрахований як середня геометрична з темпів зростання (зниження) цін акцій:

$$I_t = \sqrt[n]{\prod_{i=1}^{n} \frac{P_{i,t}}{P_{i,0}}} = \sqrt[n]{\frac{P_{1,t}}{P_{1,0}} \cdot \frac{P_{2,t}}{P_{2,0}} \cdot \frac{P_{3,t}}{P_{3,0}} \cdot \ldots \cdot \frac{P_{n,t}}{P_{n,0}}}, \qquad (8.56)$$

де: $P_{i,t}$ – ціна акції *i*-ї компанії в поточному періоді; $P_{i,0}$ – ціна акції *i*-ї компанії в базовому періоді; *n* – кількість компаній в переліку.

Методику розрахунку вищезазначених індексів ррозглянемо за даними табл. 8.14.

*Таблиця 8.14.* **Результати торгів акціями**

| Емітент | Кількість акцій в обігу, тис. шт. | Курс акції (ринкова ціна), грн, у періоді | | Темп зростання, % | Ринкова капіталізація, млн грн, у періоді | |
|---|---|---|---|---|---|---|
| | | базисний | поточний | | базисний | поточний |
| A | 30 | 25,00 | 30,00 | 120,0 | 750 | 900 |
| B | 40 | 10,00 | 11,00 | 110,0 | 400 | 440 |
| C | 15 | 100,00 | 90,00 | 90,0 | 1500 | 1350 |
| D | 10 | 50,00 | 55,00 | 110,0 | 500 | 550 |
| Разом | 70 | 46,25 | 46,50 | x | 3150 | 3240 |

Середній курс акцій розрахований за формулою середньої арифметичної простої.

Розрахунок індексу здійснено на основі:

– темпів зростання (зниження) середньоарифметичної ціни акцій обмеженої кількості обраних корпорацій

$$I_t = \frac{46{,}50}{46{,}25} = 1{,}005, \text{ або } 100{,}5\ \%;$$

– темпів зростання (зниження) середньозваженої (за кількістю акцій, що перебувають в обігу) ціни всієї множини акцій корпорацій

$$I_t = \frac{3240}{3150} = 1{,}029 \text{, або } 102{,}9\ \%;$$

– середньогеометричного значення темпів зростання (зниження) цін акцій

$$I_t = \sqrt[4]{1{,}2 \cdot 1{,}1 \cdot 0{,}9 \cdot 1{,}1} = \sqrt[4]{1{,}3068} = 1{,}143\text{, або } 114{,}3\ \%.$$

Усі обчислені індекси характеризують загальну тенденцію змін (зростання) курсу акцій.

В Україні однією з найбільш відомих статистичних характеристик фондового ринку є ПФТС-індекс. Він розраховується одним із найбільших організаторів торгівлі на українському ринку – асоціацією ПФТС і є офіційним показником Першої фондової торговельної системи (до вересня 1999 р. – Позабіржової фондової торговельної системи).

Індекс ПФТС — ціновий індекс, зважений за обсягом емісії (free float), що реально доступна для широкого кола інвесторів. Перелік акцій для розрахунку Індексу ПФТС формується з цінних паперів, що входять до Біржового списку ПФТС, на підставі даних про ринкову капіталізацію, обсяг торгів, кількість угод та інших факторів, що впливають на ліквідність акцій.

ПФТС-індекс розраховується на основі простих акцій підприємств, що пройшли лістинг ПФТС. Період для розрахунку індексу – поточний, щоденний та щотижневий. Щоденний ПФТС-індекс розраховується кожного робочого дня наприкінці торгової сесії, щотижневий ПФТС-індекс – наприкінці кожного робочого тижня. Розраховується ПФТС-індекс на основі так званого Переліку акцій – ПФТС-переліку, який переглядається кожного місяця.

Під час відбору акцій підприємств у список індексу включаються підприємства-емітенти, що пройшли лістинг ПФТС і належать до першого або другого рівнів списку ПФТС. Водночас до Переліку акцій підприємств, що входять до індексу, відбираються акції, за якими в ПФТС було зареєстровано найбільшу кількість

двосторонніх угод. Як показує практика, це майже всі компанії з першого рівня та кілька компаній з другого рівня Списку ПФТС.

ПФТС-індекс почали розрахувати з 01.10.1997 р. Його базове значення становить 100. Під час розрахунку індексу враховуються всі угоди, що були зареєстровані ПФТС за період, і задовольняють наведеним нижче умовам. Розрахунок індексу здійснюється тільки на основі офіційних результатів торгів в ПФТС.

ПФТС-індекс розраховується за формулою

$$I_{pfts} = I_{pfts_O} \frac{\sum MC_{i,t}}{\sum MC_{i,0}}, \tag{8.57}$$

де: $I_{pfts_O}$ – базове значення індексу; $\sum MC_{i,t}$ – сума ринкової капіталізації всіх акцій ПФТС-переліку в поточному періоді; $\sum MC_{i,0}$ – сума ринкової капіталізації всіх акцій ПФТС-переліку в базовому періоді.

Капіталізація розраховується за такою формулою:

$$MC_{i,t} = q_i \cdot Pl_{i,t},$$

де $q_i$ – кількість простих акцій, випущених певним емітентом (методика не розподіляє державну і недержавну частки акцій); $Pl_{i,t}$ – ціна останньої угоди $i$-ї акції в поточному періоді, якщо вона задовольняє таку умову:

$$B_{i,t} \le Pl_{i,t} \le A_{i,t},$$

де: $B_{i,t}$ – значення найкращої (найвищої) ціни купівлі; $A_{i,t}$ – значення найкращої (найнижчої) ціни продажу.

Якщо в поточному періоді ціна останньої угоди не відповідає цій умові, то для щотижневого ПФТС-індексу за основу береться ціна попередньої угоди за поточний період, яка відповідає наведеній вище умові.

Для щоденного ПФТС-індексу в разі відсутності зареєстрованої угоди за цією акцією за поточний період береться ціна, що розраховується за формулою

$$Pl_{i,t} = \frac{B_{i,t} + A_{i,t}}{2},$$

де: $B_{i,t}$ – значення найкращої (найвищої) ціни купівлі на момент закриття торгової сесії ПФТС; $A_{i,t}$ – значення найкращої (найнижчої) ціни продажу на момент закриття торгової сесії ПФТС.

У разі поновлення ПФТС-переліку з метою уникнення різкої зміни значення індексу в поточному періоді, розрахунок ПФТС-індексу здійснюється за новим переліком за формулою

$$I^{*}_{pfts} = I_{pfts(t\text{-}1)} \frac{\sum MC_{i,t}}{\sum MC_{i,t-1}},$$

де: $I_{pftst\text{-}1}$ – базове значення індексу, розраховане на ($t$–1) період з новим Переліком акцій індексу; $\sum MC_{i,t}$ – сума ринкової капіталізації всіх акцій з нового Переліку акцій індексу в поточному періоді; $\sum MC_{i,t-1}$ – сума ринкової капіталізації всіх акцій з нового Переліку акцій індексу в базовому ($t$–1) періоді.

ПФТС-індекс є об'єктивним індикатором ринкових процесів, що забезпечується завдяки чіткій методиці, відкритості та рівнодоступності первинної інформації. Перша фондова торговельна система – неприбуткова організація, яка зацікавлена в довірі до себе інвесторів, відповідно, у наданні їм повної та об'єктивної інформації. Положення про ПФТС-індекс наведено в дод. 8.

Серед інших індексів українського ринку цінних паперів слід зазначити індекс Wood 15, запропонований компанією Wood Company; індекс ProU-50, розроблений спеціалістами компанії "Проспект Інвестментс" та індекси КАС-20 (простий і зважений), розроблені аналітиками компаній "Альфа-Капітал". Серед перерахованих індексів найбільш репрезентативним є ProU-50, який

охоплює акції 50 емітентів, Wood 15, відповідно, – 15, а КАС – по 20. "Проспект Інвестментс" обраховує також індекс ProU-10 за 10 найпривабливішими емітентами з основної вибірки, однак індекс не є основним.

Під час формування бази всіх перерахованих індексів розробники орієнтуються насамперед на показник капіталізації, причому для Wood 15 ринкова капіталізація претендента для включення до цього індексу має на 5 % перевищувати аналогічний показник останньої у списку компанії.

Індекс Wood 15 розраховується як відношення поточної загальної ринкової капіталізації бази індексу до початкової ринкової капіталізації бази індексу і помножується на 1000. Початкове значення індексу в момент його першої публікації (13.06.97 р.) – 1000.

Індекс ProU-50 базується на стандартній методиці капіталізованих індексів, в основу розрахунку яких покладено зміни капіталізації компаній, що входять до вибірки. Базове значення індексу (01.01.97р.) – 100. Індекс ProU-50 на поточну дату розраховується за формулою

$$I_t = I_0 \cdot \frac{\sum MC_1}{\sum MC_0},$$

де: $MC_1$ – сумарна капіталізація 50 емітентів на поточну дату; $MC_0$ – сумарна капіталізація 50 емітентів на 01.01.97; $I_0$ – базове значення індексу.

Розрахунок капіталізації кожного емітента здійснюється за формулою

$$MC = q \cdot P,$$

де: $MC$ – капіталізація емітента; $q$ – кількість акцій; $P$ – ціна акції.

Індекс КАС-20 офіційно розраховується з 01.01.97 р. Простий базовий індекс КАС-20 (s-simple) призначеній для портфельних інвесторів, ойго визначають за формулою

$$\text{КАС-20(s)}=\frac{\sum_{i=1}^{20}\left(P_{\text{bid}\,i,t}+P_{\text{ask}\,i,t}\right)}{\sum_{i=1}^{20}\left(P_{\text{bid}\,i,0}+P_{\text{ask}\,i,0}\right)},$$

де $P_{\text{bid}\,i,t}$; $P_{\text{bid}\,i,0}$ – котирування на купівлю акції *i*-го підприємства в поточному та базисному періодах, відповідно; $P_{\text{ask}\,i,t}$; $P_{\text{ask}\,i,0}$ – котирування на купівлю акцій *i*-го підприємства в поточному та базисному періодах, відповідно.

Зважений індекс КАС-20 (w-weight) офіційно розраховується за формулою

$$\text{КАС-20(w)}=\frac{\sum_{i=1}^{20}M_{\text{cap}_{i,t}}}{\sum_{i=1}^{20}M_{\text{cap}_{i,0}}},$$

де $M_{\text{cap}_{i,t}}$ – ринкова капіталізація *i*-го підприємства в поточному періоді; $M_{\text{cap}_{i,0}}$ – ринкова капіталізація *i*-го підприємства в базовому періоді.

Під час відбору емітентів для розрахунку ProU та КАС ураховується ліквідність цінних паперів та рівномірне представлення компаній важливих галузей економіки (залежно від частки тієї чи іншої галузі у ВВП). Щодо визначення ціни акції, то "Проспект Інвестментс" використовує для свого індексу найкращу ціну купівлі, а "Альфа Капітал" – середньоарифметичну між найкращою ціною купівлі та продажу за котируванням ПФТС. Wood Company враховують додатково до котирувань в ПФТС тенденції біржового та "телефонного" ринків.

### *Список питань до самоконтролю:*

**1.** Що являє собою індекс у статистиці? Як класифікуються індекси?

**2.** Що характеризує індивідуальний індекс?

**3.** У чому полягає суть і методика побудови зведених індексів?

**4.** Як обчислюють агрегатний індекс цін? Вкажіть, який елемент агрегату – індексована величина, а який – "вага", на якому рівні вона фіксується.

**5.** Як обчислюють агрегатний індекс фізичного обсягу? Вкажіть, який елемент агрегату – індексована величина, а який – сумірник, на якому рівні він фіксується.

**6.** Наведіть два приклади індивідуальних та зведених індексів.

**7.** Покажіть взаємозв'язок індексів фізичного обсягу виробництва, собівартості та витрат на виробництво.

**8.** Чим відрізняється факторний аналіз, який проводиться при вивченні статистичного зв'язку від індексного факторного аналізу?

**9.** На які складові частини можна поділити абсолютний приріст витрат на виробництво?

**10.** Припустимо, урожайність жита збільшилась в 1,4 рази, а урожайність гречки зросла на 20 %. В яких межах може знаходитися значення індексів фіксованого та змінного складу?

**11.** Доведіть, що продуктивність праці підвищиться на 25 %, якщо трудомісткість зменшиться на 20 %.

**12.** Побудуйте зведені індекси продуктивності праці, виходячи з того, що продуктивність – це обсяг продукції, яка виробляється за одиницю часу: $W = q/T$.

**13.** Побудуйте зведені індекси динаміки цін та фізичного обсягу для однорідних та різнорідних товарів. Що буде братися за сумірник, а що за вагу?

**14.** Яка економічна суть різниці між чисельником та знаменником агрегатного індексу продуктивності праці?

**15.** Що являє собою середньозважений арифметичний індекс фізичного обсягу? Доведіть, що він тотожний цьому ж індексу агрегатної форми.

**16.** Що являє собою середньозважений гармонічний індекс ціни? Доведіть, що він тотожний цьому ж індексу агрегатної форми.

**17.** Як обчислити суму економії (перевитрат) за рахунок зміни цін?

**18.** Які індекси називаються взаємопов’язаними? Наведіть приклади.

**19.** Який зв’язок між індексом трудомісткості і продуктивності праці; індексом цін і індексом купівельної спроможності грошової одиниці?

**20.** Що характеризує собою індекс середнього рівня показника фіксованого складу, як він обчислюється?

**21.** Що характеризує собою індекс структурних зрушень, як він обчислюється?

**22.** Який зв'язок існує між індексами середньої ціни на товар "А" змінного складу, фіксованого складу і структурних зрушень?

**23.** Як розкладається на основі багатофакторної індексної моделі абсолютний приріст показника, що вивчається, за факторами?

**24.** Маємо дані про продаж товарів:

| Товар | Продано, т | | Ціна 1 кг, гр. од. | |
|---|---|---|---|---|
| | базисний період | звітний період | базисний період | звітний період |
| Груші | 15,0 | 16,2 | 2,5 | 3,0 |
| Яблука | 50,0 | 51,0 | 4,5 | 7,0 |
| Разом | 65,0 | 67,2 | x | x |

Обчисліть індивідуальні та зведені індекси ціни та фізичного обсягу реалізованого товару, покажіть зв'язок між ними. Визначте приріст товарообігу в цілому і за рахунок окремих факторів. Зробіть висновки.

**25.** Маємо дані про продаж товарів у магазині:

| Товарні групи | Реалізовано в минулому році, млн гр. од. | Темпи приросту (зниження) фізичного обсягу проданого товару порівняно з минулим роком, % |
|---|---|---|
| Продовольчі товари | 150 | -2 |
| Побутова техніка | 200 | +5 |
| Одяг і взуття | 30 | +20 |
| Разом | 380 | x |

Визначте зведений індекс фізичного обсягу та абсолютну зміну виручки від реалізації за рахунок фізичного обсягу. Результати проаналізуйте.

**26.** Маємо дані про продаж товарів за два квартали поточного року:

| Товари | Товарооборот у діючих цінах, гр. од. | | Зміна середніх цін у II кв. порівняно з I кв., % |
|---|---|---|---|
| | I кв. | II кв. | |
| Солодощі | 60 | 64 | -20 |
| Напої | 42 | 44 | +10 |
| Галантерея | 35 | 38 | без зміни |
| Разом | 137 | 146 | x |

Визначте зведені індекси та абсолютну зміну товарообороту в цілому і за рахунок окремих факторів. Результати проаналізуйте.

**27.** Маємо дані про виробництво озимих зернових культур :

| Вид культури | Урожайність, ц/га | | Посівні площі, тис. га | |
|---|---|---|---|---|
| | 1-й рік | 2-й рік | 1-й рік | 2-й рік |
| Пшениця | 38 | 31 | 7000 | 5200 |
| Жито | 24 | 23 | 5700 | 4500 |
| Разом | x | x | 11270 | 9700 |

Визначте: 1) зведений індекс урожайності і посівних площ; 2) абсолютну зміну валового збору за рахунок зміни урожайності окремих культур і за рахунок зміни розміру посівних площ, а також під впливом двох факторів; 3) перевірте взаємозв'язок розрахованих індексів. Розрахунки представте в таблиці. Результати проаналізуйте.

**28.** Маємо дані про обсяги виробництва на підприємствах регіону у січні та лютому:

| Підприємство | Січень | | Лютий | |
|---|---|---|---|---|
| | Випуск тис. шт. | Витрати на виробництво, млн грн | Випуск тис. шт. | Витрати на виробництво, млн грн |
| А | 60 | 24 | 80 | 20 |
| Б | 60 | 20 | 120 | 18 |
| Разом | 120 | 44 | 200 | 38 |

Розрахуйте зведені індекси витрат на виробництво, а також абсолютну зміну витрат на виробництво в цілому і за рахунок окремих факторів: собівартості і фізичного обсягу випущеної продукції. Перевірте взаємозв'язок розрахованих індексів. Зробіть висновки.

**29.** Є дані про середньорічний надій молока від однієї корови у фермерських господарствах двох регіонів**:**

| Регіон | Продуктивність, кг | | Поголів'я, тис. гол. | |
|---|---|---|---|---|
| | Попередній період | Поточний період | Попередній період | Поточний період |
| А | 2060 | 2120 | 20 | 22 |
| Б | 2040 | 1950 | 40 | 45 |
| Разом | х | х | 60 | 67 |

Визначте: 1) зведені індекси обсягу виробництва, середньорічного надою і поголів'я; 2) абсолютну зміну обсягу виробництва молока в поточному періоді порівняно з попереднім, за рахунок зміни надою і поголів'я. Поясніть, який із двох факторів більше вплинув на обсяг виробництва молока. Розрахунки оформіть в таблиці. Результати проаналізуйте.

**30.** Є дані про реальну заробітну плату і чисельність робітників та службовців по двох видах діяльності:

| Вид діяльності | Середня заробітна плата одного працівника, гр. од. | | Чисельність працівників, тис. | |
|---|---|---|---|---|
| | Базис | Факт | Базис | Факт |
| Промисловість | 1600 | 2200 | 150 | 140 |
| Торгівля | 1800 | 2400 | 20 | 50 |
| Разом | х | х | 170 | 190 |

Визначте: 1) зведені індекси заробітної плати і чисельності працівників; 2) зведений індекс витрат на оплату праці, використовуючи взаємозв'язок індексів; 3) абсолютну зміну витрат на оплату праці за рахунок зміни середньої заробітної плати і чисельності працівників. Поясніть, який із двох факторів більше вплинув на зміну загального фонду оплати праці. Розрахунки оформіть в таблиці. Зробіть висновки.

**31.** Маємо дані про товарооборот продовольчих і непродовольчих товарів за два періоди:

| Вид товарів | Товарооборот базисного періоду, млн грн | Товарооборот звітного періоду, млн грн | Індекс цін |
|---|---|---|---|
| Продовольчі | 26,1 | 210 | 5,3 |
| Непродовольчі | 17,5 | 130 | 4,8 |
| Разом | 43,6 | 340 | х |

Визначте: 1) відносну зміну товарообороту у фактичних цінах і середній відсоток зростання цін; 2) абсолютну зміну товарообороту, що отримана за рахунок зростання цін. Зробіть висновки.

**32.** Маємо дані про товарооборот підприємства торгівлі:

| Найменування товару | Товарооборот у грудні, тис. грн | Фізичний обсяг продажу, тис. кг | |
|---|---|---|---|
| | | грудень | січень |
| Цукерки | 200 | 40 | 40 |
| Печиво | 100 | 50 | 55 |
| Халва | 25 | 5 | 6 |
| Разом | 325 | 95 | 101 |

Визначте абсолютну та відносну зміну товарообороту за рахунок впливу фізичного обсягу проданих товарів. Розрахунки оформіть в таблиці. Результати проаналізуйте.

**33.** Маємо дані про товарооборот підприємства торгівлі:

| Найменування товару | Товарооборот у січні, тис. грн | Ціна за 1 кг, грн | |
|---|---|---|---|
| | | грудень | січень |
| Сир | 150 | 6,5 | 6,8 |
| Олія | 120 | 7,0 | 7,5 |
| Разом | 270 | x | x |

Визначте абсолютну та відносну зміну товарообороту за рахунок зміни цін на товари. Розрахунки оформіть в таблиці. Результати проаналізуйте.

**34.** Є дані про продуктивність праці і структуру чисельності працівників по двох регіонах:

| Галузь | Продуктивність праці, тис. грн | | Структура чисельності працюючих, % | | |
|---|---|---|---|---|---|
| | Регіон А | Регіон Б | Регіон А | Регіон Б | У середньому по двох регіонах |
| Видобувна | 30 | 35 | 40 | 20 | 30 |
| Переробна | 60 | 72 | 60 | 80 | 70 |

Визначте територіальні індекси змінного і фіксованого складу середньої продуктивності праці, а також індекс структурних зрушень, прийнявши за базу порівняння регіон А. Розрахунки оформіть в таблиці. Результати проаналізуйте.

**35.** Маємо дані про вклади населення в ощадні і комерційні банки по двох регіонах:

| Тип банків | Середній розмір вкладу, гр. од. | | Сума вкладів, млн гр. од. | |
|---|---|---|---|---|
| | Регіон А | Регіон Б | Регіон А | Регіон Б |
| Ощадний | 1350 | 1240 | 5000 | 6000 |
| Комерційний | 5850 | 4700 | 1450 | 4000 |
| Разом | x | x | 6450 | 10000 |

Визначте територіальний індекс середнього розміру вкладу змінного та фіксованого складу, прийнявши за базу порівняння регіон Б, а також індекс структурних зрушень. Результати проаналізуйте.

**36.** Для окремої компанії (фірми, корпорації) прибутковість капіталу розраховується відношенням чистого прибутку до власного капіталу. Динаміку цього показника можна розкласти за такою множиною факторів:

a — чистий прибуток на одиницю валового обороту (реалізації продукції, послуг);

b — оборотність поточних активів;

c — поточна ліквідність;

d — частка поточних пасивів у залучених коштах, (коефіцієнт заборгованості);

f — співвідношення залучених і власних коштів.

Побудуйте багатофакторну індексну модель на прикладі взаємозв'язку показника прибутковості капіталу з індикаторами фінансового стану та платоспроможності підприємства.

**37.** Динаміка матеріальних витрат на виробництво продукції залежить від матеріаломісткості продукції, оборотності та розміру оборотного капіталу. За поточний квартал матеріальні витрати зросли з 200 до 221 млн грн. Визначте абсолютний вплив металомісткості та оборотності капіталу на динаміку матеріальних витрат.

| Показник | Індекс |
|---|---|
| Матеріальні витрати | 1,105 |
| Оборотний капітал | 1,12 |
| Матеріаломісткість продукції | 1,05 |
| Оборотність капіталу | 0,94 |

**38.** На підставі залежності побудованої багатофакторної індексної моделі проаналізуйте динаміку результативної ознаки (середніх страхових виплат) за рахунок зміни виплат постраждалим, зміни вірогідності настання страхового випадку, зміни рівня зайнятості. Індексна система описує вплив кожного з факторів, включених у модель наступним чином:

$$Iy = Id \times Ie \times If.$$

Індекси введених у модель факторів-множників і розрахунок вкладу кожного з них в абсолютний приріст середніх страхових виплат наведено в таблиці:

| Фактор | Індекс *i*- го фактора | Абсолютний внесок *i*- го фактора в приріст середніх страхових виплат, грн |
|---|---|---|
| d. Середні виплати постраждалим | 1,168 | + 201,2 |
| e. Вірогідність настання страхового випадку | 1,053 | + 65,5 |
| f. Рівень зайнятості | 0,907 | – 100,5 |
| Разом | x | + 166,2 |

**39.** За наведеними даними (в млн умовних одиниць) побудуйте 4-факторну індексну модель ефективності комерційної діяльності фірми, вимірником якої є балансова рентабельність виробничого капіталу. Оцініть абсолютний вплив на динаміку цього показника кожного фактора.

| Показник | Базисний період | Поточний період |
|---|---|---|
| Балансовий прибуток | 8,0 | 7,6 |
| Виручка від реалізації продукції | 45,2 | 48,0 |
| Витрати на виробництво продукції | 38,0 | 41,0 |
| Виробничий капітал | 87,4 | 91,8 |
| У т. ч. оборотний капітал | 11,6 | 10,6 |

**40.** За наведеними даними побудуйте індексно-матричну модель розвитку промисловості регіону і зробіть висновки щодо збалансованості динаміки показників інтенсивності та ефективності промислового виробництва.

| Показник | Індекс |
|---|---|
| Товарна продукція промисловості | 0,90 |
| Основні виробничі фонди | 1,02 |
| Матеріальні витрати | 0,95 |
| Споживання електроенергії | 1,07 |
| Витрати праці, людино-годин | 0,98 |

**41.** За минулий рік темпи приросту макропоказників становили: ВВП — 0,7%, матеріальних витрат — 1,5%, енерговитрат — 2,3%, кількості робочих місць — 0,2%. Проведіть діагностику збалансованості економічного розвитку за умови енергозберігаючої економічної стратегії.

**42.** У таблиці наведено індексно-матричну модель економічного розвитку умовної країни за певний період. На головній діагоналі розміщено індекси макропоказників (D — національний дохід, M — матеріальні витрати, F — виробничі фонди, T — чисельність зайнятих працівників).

| Показник нормалі | D | M | F | T |
|---|---|---|---|---|
| D | 1,142 | | | |
| M | *Im* = 1,005 | 1,136 | | |
| F | *If* = 0,935 | *In* = 0,930 | 1,222 | |
| T | *Iq* = 1,171 | *Il* = 1,165 | *Ir* = 1,253 | 0,975 |

Запишіть економічну нормаль, згідно з якою темпи зростання кінцевих результатів мають бути вищими за темпи зростання витрат і ресурсів Проаналізуйте співвідношення індексів, якщо *Iq* — продуктивності праці, *If* — фондовіддачі, *Im* — матеріаловіддачі, *Ir* — фондоозброєності праці, *In* — співвідношення матеріальних витрат і вартості основних фондів та виявіть диспропорції у використанні живої та уречевленої праці.

**43.** Індекси введених у модель факторів-множників і розрахунок вкладу кожного з них в абсолютний приріст середніх страхових виплат наведено в таблиці:

| Фактор | Індекс *i*-го фактора | Абсолютний внесок *i*-го фактору в приріст середнього страхового внеску, умовних одиниць (у.о.) |
|---|---|---|
| частка страхових внесків у фонді оплати праці | 0,927 | – 1,22 |
| середня заробітна плата | 1,134 | + 6,5 |
| рівень зайнятості застрахованих осіб | 1,005 | + 0,85 |

На підставі залежності побудованої багатофакторної індексної моделі проаналізуйте динаміку середнього страхового внеску з індикаторами соціально-демографічного стану:

$x_1$- частка страхових внесків у фонді оплати праці;

$x_2$- середня заробітна плата;

$x_3$- рівень зайнятості застрахованих осіб.

**44.** Прибутковість капіталу умовної фірми становила: в базисному періоді — 115,1%, у поточному — 129,0%, тобто прибутковість зросла на 13,9 відсоткових пунктів, індекс

прибутковості — 1,121. Проведіть аналіз прибутковості капіталу за рахунок факторів. Індекси включених у модель факторів-множників наведено в таблиці:

| Фактор | Індекс фактора |
|---|---|
| a. Чистий прибуток на одиницю валового обороту (реалізації продукції, послуг) | 1,057 |
| b. Оборотність поточних активів | 0,986 |
| c. Поточна ліквідність | 1,012 |
| d. Частка поточних пасивів у залучених коштах, (коефіцієнт заборгованості) | 1,025 |
| f. Співвідношення залучених і власних коштів | 1,037 |

**45.** Проведіть аналіз прибутковості капіталу за рахунок факторів. Розрахунок внеску кожного з них в абсолютний приріст прибутковості капіталу наведено в таблиці:

| Фактор | Абсолютний внесок фактора в приріст прибутковості, в.п. |
|---|---|
| a. Чистий прибуток на одиницю валового обороту (реалізації продукції, послуг) | +6,6 |
| b. Оборотність поточних активів | –1,7 |
| c. Поточна ліквідність | +1,4 |
| d. Частка поточних пасивів у залучених коштах, (коефіцієнт заборгованості) | +3,0 |
| f. Співвідношення залучених і власних коштів | +4,3 |

**46.** За даними про ціни та обсяги цінних паперів розрахувати фондові індекси за формулою середньої арифметичної простої та зваженої та зробити висновки щодо динаміки цін на акції представлених компаній. Результати порівняйте.

| Емітент | Номінал дол. | I період | | II період | | III період | |
|---|---|---|---|---|---|---|---|
| | | Ціна дол. | Кількість, тис. шт. | Ціна, дол. | Кількість, тис. шт. | Ціна, дол. | Кількість, тис. шт. |
| A | 150 | 155 | 10 | 80 | 20 | 85 | 20 |
| B | 200 | 210 | 5 | 220 | 5 | 120 | 10 |
| C | 160 | 180 | 10 | 95 | 20 | 85 | 20 |
| D | 120 | 150 | 5 | 145 | 15 | 150 | 15 |

**47.** За даними про ціни та обсяги цінних паперів розрахувати фондові індекси за формулою середньої геометричної простої та зваженої та зробити висновки щодо динаміки цін на акції представлених компаній. Результати порівняйте.

| Емітент | Номінал, дол. | I період | | II період | | III період | |
|---|---|---|---|---|---|---|---|
| | | Ціна дол. | Кількість, тис. шт. | Ціна, дол. | Кількість, тис. шт. | Ціна, дол. | Кількість, тис. шт. |
| A | 50 | 55 | 10 | 53 | 20 | 60 | 20 |
| B | 200 | 210 | 5 | 110 | 10 | 120 | 10 |
| C | 150 | 180 | 10 | 185 | 10 | 85 | 20 |
| D | 100 | 150 | 5 | 155 | 5 | 150 | 10 |
| E | 100 | 110 | 30 | 105 | 60 | 110 | 60 |

### *Список рекомендованої літератури по темі:*

1. Андрієнко В. Ю. Статистичні індекси в економічних дослідженнях / В. Ю. Андрієнко. - К. : Академперіодика, 2004. 118 с. URL: https://www.myslenedrevo.com.ua/uk/Sci/Economics/StatIndices/Preface.html
2. Галицька Е.В., Ковтун Н.В. Фінансова статистика: навчальний посібник для студ. вищих закл. освіти. К.: Кондор, 2008. 440 с.

3. Горна М. О. Статистичний аналіз загальнообов'язкового державного страхування: соціально-економічна нормаль / М. О. Горна // Вісник Академії праці і соціальних відносин Федерації профспілок України. 2014. № 1. С. 30-34. URL: http://nbuv.gov.ua/UJRN/VAPSV_2014_1_6.
4. Економічна статистика: підручник: у 2 ч. – Ч.1. Макроекономічна статистика/ [І.Г. Манцуров, А.М. Єріна, О.К. Мазуренко та ін.]; за наук.ред.чл.-кор. НАНУ І.Г. Манцурова. К.: КНЕУ, 2013. 325 с.
5. Єріна А. М. Статистичне моделювання та прогнозування [Текст]: підручник / А. М. Єріна, Д. Л. Єрін; Держ. ВНЗ "Київ. нац. екон. ун-т ім. Вадима Гетьмана". Київ: КНЕУ, 2014. 348 с.
6. Ковтун Н.В. Теорія статистики: підручник / Н.В.Ковтун. К .: Знання, 2012. 399 с.
7. Ковтун Н. В. Фінансова статистика / Н. В. Ковтун, Е. В. Галицька, О.К. Примєрова. К. : Видавничо-поліграфічний центр "Київський університет", 2017. 623 с.
8. Мазуренко О. К., Горна М.О. Методичні підходи до формування стратегії розвитку у системі послуг соціального захисту / О. К. Мазуренко, М. О. Горна // Економічний аналіз. 2015. Т. 19, № 1. С. 69-75.
9. Мазуренко О.К., Горна М.О. Теоретико-методологічні засади аналізу діяльності фондів соціального страхування /Соціальне забезпечення в контексті вступу в ЄС: монографія / За загальною редакцією Дерій Ж.В. К.: Видавничий дім «Кондор», 2017. С.6-28.
10. Цінні папери : підручник / В. Д. Базилевич, В. М. Шелудько, Н. В. Ковтун та ін. К., 2011.
11. Tsatsulin A.N. Multivariable Modeling in the Analysis of Current Assets in the Format of the Hybrid Model t/v-models. Administrative Consulting. 2017;(4):73-96. (In Russ.) https://doi.org/10.22394/1726-1139-2017-4-73-96

12. Офіційний сайт Державної служби статистики України. URL: http ://www.ukrstat.gov.ua/
13. Сайт ПФТС фондова біржа : Підсумки торгів. URL: http://www.pfts.com/uk/trade-results/
14. Сайт Агентства з розвитку інфраструктури фондового ринку України : Інформація емітентів. Річна звітність емітентів цінних паперів. URL: http://www.smida.gov.ua/reestr/smreestr.php?info=at/
15. Сайт Національного банку України : Статистика : URL: http://www.bank.gov.ua/control/uk/publish/article?art_id=65162&cat_id=36674
16. Сайт Державної служби статистики України : Статистична інформація. URL : http://www.ukrstat.gov.ua/

# Розділ 9. СТАТИСТИЧНЕ МОДЕЛЮВАННЯ ВИПАДКОВИХ ПРОЦЕСІВ

## 9.1. Оцінка генераторів випадкових чисел

Для моделювання та аналізу випадкових явищ (випадкових подій, випадкових величин, випадкових процесів тощо) потрібні набори випадкових чисел, які мають певний необхідний для моделі розподіл. У багатьох випадках такий розподіл легко отримується із рівномірного розподілу, який у свою чергу можна отримати за допомогою генераторів випадкових чисел (ГВЧ). Оскільки таких генераторів існує багато і вони продукують набори рівномірно розподілених чисел різної «якості», то нижче ми розглянемо нескладний критерій, за яким можна самостійно оцінити рівномірність розподілу отриманого набору чисел.

Псевдовипадковими числами як правило називають послідовність чисел, властивості яких певною мірою близькі до властивостей рівномірно розподілених чисел на деякому відрізку. Міру близькості розподілу псевдовипадкових чисел до рівномірного можна визначати по різному, зокрема, з використанням критерію Колмогорова-Смирнова.

Нехай $\xi_1, \xi_2, \ldots, \xi_n$ – вибірка із розподілу $F(x)$, нехай $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{\xi_i \le x\}$. Позначимо через

$$D_n = \sup_x |F_n(x) - F(x)|.$$

Тоді $\sqrt{n}D_n \xrightarrow{D} \sup_{t\in[0,1]} |B(t)|$, $n \to \infty$, де $B(t) = W(t) - tW(1)$, $t \in [0,1]$ – броунівський міст ($W(t)$ – процес Вінера), а саме

$$P\left(\lim_{n\to\infty} \sqrt{n}D_n \le x\right) = K(x) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2x^2}.$$

Математичне сподівання розподілу Колмогорова

$$m = \ln 2 \sqrt{\frac{\pi}{2}} \approx 0.868731.$$

Дисперсія розподілу Колмогорова

$$\sigma^2 = \frac{\pi^2}{12} - m^2 \approx 0.067773,$$

і, отже, середньоквадратичне відхилення дорівнює $\sigma \approx 0.260333$.

Оскільки розподіл $K(x)$ є одномодальний, то, використовуючи нерівність Височанського-Петуніна, маємо

$$P\left(\left|\lim_{n\to\infty} \sqrt{n} D_n - m\right| \geq \lambda\sigma\right) \leq \frac{4}{9\lambda^2}, \qquad \lambda > \sqrt{\frac{8}{3}}.$$

Для $\lambda = 2$, тобто $\lambda\sigma = 2 \cdot 0.260333 = 0{,}520666$, маємо

$$P\left(\left|\lim_{n\to\infty} \sqrt{n} D_n - 0.868731\right| \geq 0{,}520666\right) \leq \frac{1}{9}$$

або

$$P\left(0{,}348065 \leq \lim_{n\to\infty} \sqrt{n} D_n \leq 1{,}389397\right) \geq \frac{8}{9}.$$

У якості прикладів, перевіримо кілька варіантів ГВЧ Лемера.

*Задача 9.1.* Спочатку візьмемо константи: $n = 30$, $x_0 = 3$, $a = 106$, $c = 1283$, $m = 100$

Підставляючи у формулу Лемера: $x_{k+1} = (ax_k + c)\, mod(m)$, маємо послідовність: $x_0 = 3, x_1 = (106 \cdot 3 + 1283)(mod 100) = 1$, $x_2 = 89$, і т. д. Отримали таку вибірку:

3, 1, 89, 17, 85, 93, 41, 29, 57, 25, 33, 81, 69, 97, 65, 73, 21, 9, 37, 5, 13, 61, 49, 77, 45, 53, 1, 89, 17, 85.

Легко бачити, що поділивши кожне число цієї послідовності на 100, ми отримаємо числа $\xi_0 = \frac{x_0}{100} = \frac{3}{100}$, $\xi_1 = \frac{x_1}{100} = \frac{1}{100}$ і так далі, які мають бути рівномірно розподіленими на відрізку $[0,1]$.

Складемо варіаційний ряд вибірки $\{\xi_k\}$:

$\frac{1}{100}, \frac{1}{100}, \frac{3}{100}, \frac{5}{100}, \frac{9}{100}, \frac{13}{100}, \frac{17}{100}, \frac{17}{100}, \frac{21}{100}, \frac{25}{100}, \frac{29}{100}, \frac{33}{100}, \frac{37}{100}, \frac{41}{100}, \frac{45}{100}, \frac{49}{100}, \frac{53}{100}, \frac{57}{100}, \frac{61}{100}, \frac{65}{100}, \frac{69}{100}, \frac{73}{100}, \frac{77}{100}, \frac{81}{100}, \frac{85}{100}, \frac{85}{100}, \frac{89}{100}, \frac{89}{100}, \frac{93}{100}, \frac{97}{100}$.

Емпірична функція розподілу $F_n^*(x)$ приймає такі значення:
$0, \frac{2}{30}, \frac{3}{30}, \frac{4}{30}, \frac{5}{30}, \frac{6}{30}, \frac{8}{30}, \frac{8}{30}, \frac{9}{30}, \frac{10}{30}, \frac{11}{30}, \frac{12}{30}, \frac{13}{30}, \frac{14}{30}, \frac{15}{30}, \frac{16}{30}, \frac{17}{30}, \frac{18}{30},$
$\frac{19}{30}, \frac{20}{30}, \frac{21}{30}, \frac{22}{30}, \frac{23}{30}, \frac{24}{30}, \frac{26}{30}, \frac{26}{30}, \frac{28}{30}, \frac{28}{30}, \frac{29}{30}, 1.$

Через відносно великий розмір вибірки ми не будемо будувати графік емпіричної функції $F_n^*(x)$, але неважко переконатись, що

$$D_n = \sup_{x \geq 0} |F_n^*(x) - U(x)| = \frac{29}{300},$$

де $U(x)$ – функція розподілу рівномірного на $[0,1]$ розподілу, тобто

$$\frac{dU(x)}{dx} = I_{[0,1]}(x) = \begin{cases} 1, x \in [0,1], \\ 0, x \notin [0,1]. \end{cases}$$

Отже, $\sqrt{n} D_n = \sqrt{30} \cdot \frac{29}{300} \approx 0.529465 \in (0{,}348065; 1{,}389397)$ і гіпотеза, що $\{\xi_k\}$ (чи $\{x_k\}$) рівномірно розподілені приймається.

*Задача 9.2*. Тепер розглянемо ГВЧ Лемера у випадку, коли $n = 50$, $x_0 = 3$, $a = 106$, $c = 1283$, $m = 50$.

Підставляючи у формулу Лемера: $x_{k+1} = (ax_k + c)\ mod(m)$, ми отримаємо таку послідовність псевдовипадкових чисел:
3, 1, 39, 17, 35, 43, 41, 29, 7, 25, 33, 31, 19, 47, 15, 23, 21, 9, 37, 5, 13, 11, 49, 27, 45, 3, 1, 39, 17, 35, 43, 41, 29, 7, 25, 33, 31, 19, 47, 15, 23, 21, 9, 37, 5, 13, 11, 49, 27, 45.

Аналогічно прикладу 1 із послідовності $\{x_k\}$ отримуємо послідовність чисел $\xi_0 = \frac{x_0}{50} = \frac{3}{50}$, $\xi_1 = \frac{x_1}{50} = \frac{1}{50}$ і так далі, які мають бути рівномірно розподіленими на відрізку $[0,1]$.

Утворивши варіаційний ряд послідовності $\{\xi_k\}$, маємо
$\frac{1}{50}, \frac{1}{50}, \frac{3}{50}, \frac{3}{50}, \frac{5}{50}, \frac{5}{50}, \frac{7}{50}, \frac{7}{50}, \frac{9}{50}, \frac{9}{50}, \frac{11}{50}, \frac{11}{50}, \frac{13}{50}, \frac{13}{50}, \frac{15}{50}, \frac{15}{50}, \frac{17}{50},$
$\frac{17}{50}, \frac{19}{50}, \frac{19}{50}, \frac{21}{50}, \frac{21}{50}, \frac{23}{50}, \frac{23}{50}, \frac{25}{50}, \frac{25}{50}, \frac{27}{50}, \frac{27}{50}, \frac{29}{50}, \frac{29}{50}, \frac{31}{50}, \frac{31}{50}, \frac{33}{50}, \frac{33}{50},$
$\frac{35}{50}, \frac{35}{50}, \frac{37}{50}, \frac{37}{50}, \frac{39}{50}, \frac{39}{50}, \frac{41}{50}, \frac{41}{50}, \frac{43}{50}, \frac{43}{50}, \frac{45}{50}, \frac{45}{50}, \frac{47}{50}, \frac{47}{50}, \frac{49}{50}, \frac{49}{50}.$

Емпірична функція розподілу $F_n^*(x)$ випадкової послідовності $\{\xi_k\}$ приймає значення:

$$0, \frac{1}{50}, \frac{2}{50}, \frac{3}{50}, \frac{4}{50}, \frac{5}{50}, \frac{6}{50}, \frac{7}{50}, \frac{8}{50}, \frac{9}{50}, \frac{10}{50}, \frac{11}{50}, \frac{12}{50}, \frac{13}{50}, \frac{14}{50}, \frac{15}{50}, \frac{16}{50},$$
$$\frac{17}{50}, \frac{18}{50}, \frac{19}{50}, \frac{20}{50}, \frac{21}{50}, \frac{22}{50}, \frac{23}{50}, \frac{24}{50}, \frac{25}{50}, \frac{26}{50}, \frac{27}{50}, \frac{28}{50}, \frac{29}{50}, \frac{30}{50}, \frac{31}{50}, \frac{32}{50}, \frac{33}{50},$$
$$\frac{34}{50}, \frac{35}{50}, \frac{36}{50}, \frac{37}{50}, \frac{38}{50}, \frac{39}{50}, \frac{40}{50}, \frac{41}{50}, \frac{42}{50}, \frac{43}{50}, \frac{44}{50}, \frac{45}{50}, \frac{46}{50}, \frac{47}{50}, \frac{48}{50}, \frac{49}{50}, 1.$$

Неважко переконатись, що $D_n = \sup_{x\geq 0}|F_n^*(x) - U(x)| = \frac{1}{50}$ і, отже, $\sqrt{n}D_n = \sqrt{50}\cdot\frac{1}{50} \approx 0.141421$. Оскільки значення величини $\sqrt{n}D_n$ у цьому випадку не належить інтервалу $(0{,}348065; 1{,}389397)$ псевдовипадкова послідовність чисел $\{\xi_k\}$ (чи $\{x_k\}$) не може розглядатись як рівномірно розподілена за нашим критерієм.

*Задача 9.3*. У цьому прикладі параметри ГВЧ Лемера такі: $n = 20$, $x_0 = 8321$, $a = 8253729$, $c = 2396403$, $m = 32768$.

Із рекурсії $x_{k+1} = (8253729x_k + 2396403)\ mod(32768)$ отримуємо такий набір псевдовипадкових чисел:
8321, 12948, 27143, 8321, 12948, 27143, 21210, 30989, 26784, 11667, 21478, 26521, 29868, 1823, 10994, 27685, 22200, 30379, 22526, 1713, 3780, 3799, 23306.

Із цієї послідовності отримуємо послідовність псевдовипадкових чисел $\xi_k = \frac{x_k}{32768}$, $k = 0{,}1, \ldots, 19$, яка є рівномірно розподіленою на $[0{,}1]$: $\xi_0 = \frac{x_0}{32768} = \frac{8321}{32768}$, $\xi_1 = \frac{x_1}{32768} = \frac{12948}{32768}$ і так далі.

Обчислюючи статистику $D_n$, маємо

$$D_n = \sup_{x\geq 0}|F_n^*(x) - U(x)| = \frac{20257}{81920}.$$

Оскільки $\sqrt{n}D_n = \sqrt{15}\cdot\frac{20257}{81920} \approx 0.957703 \in (0.348065; 1.389397)$ гіпотеза про рівномірний розподіл $\{\xi_k\}$ (чи $\{x_k\}$) приймається.

Отже, використовуючи $8/9$-довірчий інтервал, ми доходимо висновку, що ГВЧ Лемера у задачах 9.1 та 9.3 дає рівномірно розподілені набори псевдовипадкових чисел.

Однак, оскільки у прикладі 1 маємо $\left|\sqrt{n}D_n - \ln 2\sqrt{\frac{\pi}{2}}\right| \approx |0.529465 - 0.868731| = 0.339266$, а у задачі 9.3 відповідно $|0.957703 - 0.868731| = 0.088972$, то ГВЧ Лемера при параметрах прикладу 1.3 дає набагато кращий рівномірно розподілений набір чисел, ніж при параметрах задачі 9.1.

Використовуючи такий підхід можна перевіряти якість різних генераторів випадкових чисел у тому числі і створених самостійно.

## 9.2. Моделювання випадкових подій

За допомогою ГВЧ легко моделювати несумісні події, наприклад, настання події $A$ чи $\bar{A}$, за умови, що відома ймовірність $P(A)$. Це частинний випадок такої загальної ситуації: нехай $A_1, A_2, \dots, A_n$ – повна група подій з ймовірностями $P(A_1), P(A_2), \dots, P(A_n)$. Для моделювання настання подій досить на відрізку $[0,1]$ відкласти точки з координатами $P(A_1), P(A_1) + P(A_2), \dots, P(A_1) + P(A_2) + \dots + P(A_{n-1})$. Тоді будемо вважати, шо подія $A_k$ настає, якщо випадкове рівномірно розподілене на $[0,1]$ число $u$, яке згенерував ГВЧ, задовольняє нерівність

$$\sum_{i=1}^{k-1} P(A_i) < u \le \sum_{i=1}^{k} P(A_i), k = 1, \dots, n,$$

тут вважаємо, що $\sum_{i=1}^{0} P(A_i) = 0$.

Для моделювання умовної події $A$ за умови $B$, яка відбувається з ймовірністю $P(A/B) = \frac{P(A \cap B)}{P(B)}$ спочатку моделюємо настання події $B$, використовуючи повну групу подій $B$ та $\bar{B}$. Вважаємо, що подія $B$ настає, якщо рівномірно розподілене на $[0,1]$ число $u$ задовольняє нерівність $u \le P(B)$. Якщо подія $B$ не відбулась, то вважаємо, що умовна подія $A$ за умови $B$ не відбулась. Якщо ж $B$ відбулась, то

моделюється настання події $A \cap B$, яка відбувається, якщо інше згенероване рівномірно розподілене на $[0,1]$ число $v$ задовольняє $v \leq P(A \cap B)$, то умовна подія $A$ за умови $B$ відбулась, у противному разі (якщо $v > P(A \cap B)$) подія $A$ за умови події $B$ не відбулась.

У випадку, коли $A$ та $B$ сумісні події для моделювання їх настання необхідно знати ймовірності чотирьох несумісних подій:

$$H_1 = A \cap B, H_2 = \bar{A} \cap B, H_3 = A \cap \bar{B}, H_4 = \bar{A} \cap \bar{B}.$$

*Задача 9.4. Довести, що набір* $H_i$, $i = 1, \dots, 4$ *утворює повну групу випадкових подій.*

Далі моделюємо настання подій $H_i$, $i = 1, \dots, 4$, як набір, що утворює повну групу подій. Оскільки $A = H_1 \cup H_3$, то подія $A$ наступає, коли наступає подія $H_1$ або подія $H_3$. Аналогічно, подія $B = H_1 \cup H_2$ і вона наступає, коли наступає подія $H_1$ або подія $H_2$,

## 9.3. Моделювання випадкових величин

Спочатку розглянемо методи моделювання неперервних випадкових величин. Нехай $\eta$ – рівномірно розподілена випадкова величина на $[0,1]$.

Припустимо, що випадкова величина $\xi$ має функцію розподілу $F_\xi$, для якої існує обернена функція $F_\xi^{-1}$. Якщо випадкова величина $\eta$ – рівномірно розподілена на $[0,1]$, то величина $F_\xi^{-1}(\eta)$ має такий же розподіл, як і $\xi$.

Дійсно,

$$P\left(F_\xi^{-1}(\eta) \leq x\right) = P\left(\eta \leq F_\xi(x)\right) = F_\xi(x).$$

Описаний метод називається методом *оберненої функції*.

*Приклад*. Нехай $\xi$ має експоненціальний розподіл з параметром $\lambda > 0$:

$$F_\xi(x) = 1 - e^{-\lambda x}, \qquad x \geq 0.$$

Існує обернена функція

$$F_\xi^{-1}(x) = -\frac{1}{\lambda}\ln(1-x).$$

Отже, для того, щоб отримати вибірку об'єму $n$ із експоненціального розподілу досить згенерувати вибірку із рівномірного на $[0,1]$ розподілу $\eta_i$, $i = 1,2,\dots,n$, а потім отримати набір чисел $\xi_i = -\frac{1}{\lambda}\ln(1-\eta_i)$, $i = 1,2,\dots,n$, який і вибіркою із експоненціального розподілу.

*Зауваження 9.3.1*. Легко бачити, що якщо $\eta_i$ рівномірно розподілена на $[0,1]$ випадкова величина, то $1-\eta_i$ також рівномірно розподілена на $[0,1]$, тому при моделюванні вибірки із експоненціального розподілу можна розглядати набір чисел $\xi_i = -\frac{1}{\lambda}\ln(\eta_i)$, $i = 1,2,\dots,n$, який також є вибіркою із експоненціального розподілу.

*Задача 9.5*. Для отримання вибірки із розподілу Ерланга-$m$ слід взяти до уваги, що випадкова величина з таким розподілом є сумою $m$ незалежних експоненційно розподілених випадкових величин. Отже, для того, щоб отримати вибірку об'єму $n$ із розподілу Ерланга-$m$ потрібно згенерувати $nm$ випадкових величин $\eta_i$, $i = 1,2,\dots,nm$, рівномірно розподілених на $[0,1]$, і отримати із них вибірку об'єму $nm$ із експоненціального розподілу $-\frac{1}{\lambda}\ln(\eta_i)$, $i = 1,2,\dots,nm$. Розділити цей набір на $n$ множин по $m$ чисел у кожній $\xi_k^i = -\frac{1}{\lambda}\ln(\eta_i)$, $i = 1,\dots,n$, $k = (i-1)m+1,(i-1)m+2,\dots,im$.

Набір чисел

$$\zeta_i = \sum_{k=(i-1)m+1}^{im} \xi_k^i, i = 1,2,\dots,n$$

і є вибіркою об'єму $n$ із розподілу Ерланга-$m$.

*Задача 9.6.* Часто буває так, що навіть коли нам відомо про існування оберненої функції до функції розподілу, знайти її вираз у замкнутій формі непросто, або й неможливо. Це стосується,

наприклад, стандартного нормального розподілу, у якого функція розподілу $N(x)$ має вигляд $N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$.

Для того, щоб змоделювати вибірку із стандартного нормального розподілу з функцією розподілу користуються центральною граничною теоремо (ЦГТ), яка стверджує, що послідовність незалежних однаково розподілених випадкових величин $\{\xi_k, k \in \mathbb{N}\}$ з скінченними математичним сподіванням $E\xi_k = a$ та дисперсією $D\xi_k = \sigma^2$ задовольняє ЦГТ

$$\lim_{n\to\infty} P\left\{\frac{\sum_{k=1}^{m} \xi_k - ma}{\sigma\sqrt{m}} \le x\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt.$$

Відомо, що якщо випадкова величина $\eta_k$ рівномірно розподілена на $[0,1]$, то $E\eta_k = \frac{1}{2}$, $D\eta_k = \sigma^2 = \frac{1}{12}$. Покладемо

$$\xi = \sum_{k=1}^{12} \eta_k - 6.$$

Тоді $E\xi = 0$, $D\xi = \sigma^2 12 = 1$. Оскільки $\xi = \frac{\sum_{k=1}^{n} \eta_k - na}{\sigma\sqrt{n}}$ при $m = 12$, то за ЦГТ $\xi$ з певним припущенням можна вважати стандартно нормально розподіленою випадковою величиною.

Отже, щоб отримати вибірку об'єму $n$ із стандартного нормального розподілу генеруємо $n$ серій рівномірно розподілених на $[0,1]$ псевдовипадкових чисел по 12 чисел у кожній серії: $\eta_k^i$, $k = 1, \dots, 12$, $i = 1,2, \dots, n$. Далі отримуємо набір випадкових чисел $\{\xi_i, i = 1, \dots, n\}$ виду

$$\xi_i = \sum_{k=1}^{12} \eta_k^i - 6, \qquad i = 1,2, \dots, n,$$

який можна вважати вибіркою об'єму $n$ із стандартного нормального розподілу.

Моделювання дискретних випадкових величин за допомогою рівномірного розподілу часто буває складнішим, ніж моделювання неперервних розподілів. Розглянемо кілька прикладів моделювання таких величин.

*Задача 9.7.* Розглянемо як за допомогою рівномірно розподіленої на $[0,1]$ випадкової величини $\eta$ моделювати геометрично розподілену випадкову величину $\xi$ з розподілом

$$P(\xi = k) = pq^k, k = 0,1, \dots, q = 1 - p.$$

**Твердження 9.3.1.** *Випадкова величина* $\left[\frac{\ln \eta}{\ln q}\right]$ *має геометричний розподіл.*

*Доведення*. Дійсно

$$\begin{aligned} P\left(\left[\frac{\ln \eta}{\ln q}\right] = k\right) &= P\left(k \le \frac{\ln \eta}{\ln q} < k + 1\right) \\ &= P\big((k+1)\ln q \le \ln \eta < k \ln q\big) \\ &= P\big(q^{(k+1)} \le \eta < q^k\big) = q^k(1-q). \end{aligned}$$

Нехай маємо $n$ рівномірно розподілених на $[0,1]$ випадкових величин $\eta_i,\ i = 1,2, \dots, n$. Тоді $\xi_i = \left[\frac{\ln \eta_i}{\ln q}\right],\ i = 1,2, \dots, n,$ – вибірка об'єму $n$ із геометричного розподілу.

*Задача 9.8.* Для моделювання випадкової величини $\xi$, що має розподіл Пуассона

$$P(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0,1, \dots$$

користуються методом, подібним до методу оберненої функції для неперервних розподілів. Функція розподілу Пуассона має вигляд

$$F(x) = P(\xi \le x) = \sum_{i=0}^{[x]} \frac{\lambda^i}{i!} e^{-\lambda},$$

де $\sum_{k=0}^{[x]} = 0$, якщо $x < 0$.

**Твердження 9.3.2.** *Нехай* $\eta \in [0,1]$ *– число отримане ГВЧ. Нехай* $\xi \in \mathbb{N} \cup \{0\}$ *мінімальне таке, що* $\eta < \sum_{i=0}^{\xi} \frac{\lambda^i}{i!} e^{-\lambda}$. *Тоді* $\xi$ *має розподіл Пуассона.*

*Доведення.* Дійсно, для $k \in \mathbb{N} \cup \{0\}$, припускаючи, що $\sum_{i=0}^{-1} \frac{\lambda^i}{i!} e^{-\lambda} = 0$, маємо:

$$P(\xi = k) = P\left(\sum_{i=0}^{k-1} \frac{\lambda^i}{i!} e^{-\lambda} \le \eta < \sum_{i=0}^{k} \frac{\lambda^i}{i!} e^{-\lambda}\right) = \sum_{i=0}^{k} \frac{\lambda^i}{i!} e^{-\lambda} - \sum_{i=0}^{k-1} \frac{\lambda^i}{i!} e^{-\lambda}$$
$$= \frac{\lambda^k}{k!} e^{-\lambda}.$$

***Список питань до самоконтролю:***

**1.** Який процес називається «броунівським мостом»?

**2.** Сформулюйте критерій Колмогорова-Смирнова.

**3.** Що таке нерівність Височанського-Петуніна і які її переваги перед нерівністю Чебишова?

**4.** Що таке генератор випадкових чисел Лемера?

**5.** Чи завжди генератор випадкових чисел Лемера дає задовільні набори рівномірно розподілених випадкових чисел?

**6.** Як за допомогою вибірки із рівномірно розподілених випадкових чисел утворити вибірку:

a) із геометричного розподілу;
b) із нормального розподілу;
c) експоненціального розподілу;
d) із розподілу Ерланга?

### *Список рекомендованої літератури по темі:*

1. Marsaglia G., Tsang W., Wang J. (2003) Evaluating Kolmogorov's distribution. J. Stat. Softw. 8, 1–4.
2. Justel A., Peña D., Zamar R. (1997) A multivariate Kolmogorov–Smirnov test of goodness of fit. Statistics & Probability Letters. 35 (3): 251–259.
3. Высочанский Д. Ф., Петунин Ю. И. (1979) Обоснование правила 3-sigma для одномодальных распределений. Теорія ймовірностей і мат. статистика, вип. 21, 23-35.

# Розділ 10. ЗАДАЧІ ОПТИМІЗАЦІЇ У СТАТИСТИЧНОМУ МОДЕЛЮВАННІ

## 10.1. Поняття про оптимізаційну задачу

Оптимізація будь-якого об'єкта, процесу або явища пов'язана із вибором серед багатьох альтернативних шляхів вирішення проблеми єдиного і найкращого. Такий вибір пов'язаний із процесом прийняття управлінського рішення. Важливо, щоб прийняте рішення забезпечувало досягнення максимально можливих результатів у рамках визначеної цілі. Задачі вибору найкращого варіанта з-поміж сукупності альтернатив із урахуванням цілі належать до класу задач оптимізації, для яких характерна низка особливостей. По-перше, необхідність в оптимізації виникає у разі існування певної проблеми, осмислення якої дає можливість сформулювати конкретну ціль. Наприклад, проблемою підприємства може бути низький рівень рентабельності діяльності, спричинений високими витратами. Тоді ціллю оптимізації діяльності підприємства буде мінімізація собівартості виготовлення продукції. З іншого боку, низька ефективність може бути спричинена недостатніми доходами. У такому випадку ціллю буде забезпечення максимальної виручки від реалізації виготовленої продукції.

По-друге, особливістю оптимізаційних задач є те, що їх розв'язання здійснюється з метою пошуку найкращих шляхів (методів, алгоритмів, технологій) досягнення поставленої цілі. Невідомим у таких задачах можуть бути: обсяг виготовлення кожного окремого виду продукції, який би забезпечив максимальний прибуток або мінімальні витрати; пошук видів та кількості техніки, необхідної для максимізації рівня виробництва; перелік інвестиційних проєктів, що забезпечать максимальну ефективність через певний період тощо.

По-третє, використання тих чи інших інструментів у задачах оптимізації завжди обмежене низкою чинників. Зазвичай це ресурсні обмеження та обмеження щодо максимально та/або мінімально можливих значень невідомих змінних. Наприклад, максимально вигідним для сільськогосподарського підприємства є вирощування та експорт зерна, тоді як обсяг його виробництва обмежений наявною у нього площею сільськогосподарських угідь та грошовими ресурсами. Іншим прикладом є обмеження обсягів виробництва продукції наявним попитом та виробничими потужностями: недоцільно виробляти продукцію в обсязі більшому, ніж потребує ринок, або може виготовити підприємство. Отже, задача оптимізації передбачає пошук екстремального (максимального або мінімального) значення функціоналу, що у формальному вигляді відображає цільову установку в рамках визначених обмежень.

Існує кілька класифікаційних ознак задач оптимізації. Першою і найбільш вживаною ознакою є рівень детермінованості (від англ. Determinate – детермінований, визначений). За цією ознакою виділяють детерміновані та стохастичні задачі і відповідні моделі, всі параметри (показники) яких – детерміновані величини, що набувають лише конкретних, наперед відомих значень. Прикладом детермінованого параметра задачі оптимізації є ціна продукції, визначена заздалегідь оформленим контрактом із покупцем. У цьому випадку виробник наперед знає ціну і вона не залежить від дії випадкових ринкових чинників.

Натомість, значення випадкових величин можна передбачити лише з певною ймовірністю. Наприклад, у задачах планування обсягів виробництва продукції зазвичай практично неможливо точно передбачити, якою буде виробнича собівартість продукції. Тому цей показник є випадковою величиною. Переважна більшість реальних економічних об’єктів характеризується невизначеністю та випадковістю і тому задачі оптимізації їх діяльності мали б бути стохастичними. Однак, стохастичні задачі характеризуються

складністю формалізації та розв'язання. Тому часто їх намагаються замінити детермінованими аналогами.

Іншим критерієм класифікації задач оптимізації є врахування або неврахування фактору часу. За цією ознакою виділяють статичні (від слова «статика» – відсутність руху) та динамічні (від слова «динаміка» – рух) задачі. У статичних задачах фактор часу не враховується. Тобто жодний із параметрів статичної задачі оптимізації не є функцією від часу. Класичними прикладами таких задач є задачі оптимізації маршруту перевезень у конкретний момент часу, визначення оптимальних обсягів виробництва продукції на певний період (наприклад, на наступний квартал, півріччя, рік тощо).

Натомість, у динамічних задачах хоча б один параметр залежить від часу. Така залежність відображається у вигляді функції параметра задачі від фактору часу: $p = f(t)$, де $p$ – параметр задачі оптимізації (наприклад, ціна, обсяг реалізації, собівартість, наявна площа сільськогосподарських угідь тощо); $t$ – фактор часу (місяць, квартал, рік тощо).

Третьою класифікаційною ознакою задач оптимізації є вид функції. За цією ознакою задачі поділяються на лінійні та нелінійні. У лінійних задачах всі функції-компоненти є лінійними. У спрощеному вигляді лінійна функція від однієї змінної має вигляд:

$$f(x) = a + bx, \tag{10.1}$$

де $a, b$ – коефіцієнти (параметри) лінійної функції;

$x$ – змінна лінійної функції.

Функції цілі та обмежень у задачах оптимізації в зазвичай мають вигляд функцій від кількох змінних. Лінійні функції від багатьох змінних у формальному вигляді відображають так:

$$f(x_1, x_2, ..., x_n) = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + ... + a_nx_n, \tag{10.2}$$

де: – коефіцієнти (параметри) лінійної функції;

$n$ – кількість змінних $x$.

Якщо хоча б один компонент задачі оптимізації є нелінійним, то вона класифікується як нелінійна оптимізаційна задачі. Найбільш розповсюдженими типами нелінійних функцій, які використовуються у задачах економіко-математичного моделювання є квадратичні ( $y = a + bx + cx^2$ ), логарифмічні ( $y = a + b\ln(x)$ ), степеневі ( $y = ax^b$ ), показникові ( $y = ab^x$ ) та інші функції. Розподіл задач оптимізації за типом функцій потрібен для того, щоб обрати метод її розв'язання, оскільки пошук рішення у нелінійних задачах потребує використання специфічного методичного інструментарію.

За областю допустимих значень задачі оптимізації поділяються на неперервні та дискретні. Компоненти вектору невідомих змінних у задачах неперервного моделювання можуть набувати будь-яких значень з інтервалу дійсних чисел. Прикладом неперервної величини є обсяг виробництва продукції, що вимірюється в одиницях маси (г, кг, т тощо). У такій задачі невідомі змінні можуть набувати абсолютно всіх значень (0,01; 1,3; 12; 1020 тощо).

У дискретних задачах компоненти вектору невідомих змінних (тобто змінні *x*) є дискретними величинами. Дискретна величина – це величина, можливі значення якої утворюють скінчену (злічену) послідовність. Дискретна величина може набувати лише конкретних значень з інтервалу дійсних чисел. Наприклад, у досліджуваному цеху один станок за зміну може виготовити або 5, або 10, або 15 одиниць продукції, а виготовлення продукції в інших обсягах технологічно неможливе. Задача оптимізації, у якій невідомим буде обсяг виробництва продукції за зміну, є дискретною, оскільки у ній значення невідомої величини може набувати лише наперед встановлених значень: або 5, або 10, або 15 од.

Специфічними задачами дискретного моделювання є цілочислові задачі, у яких компоненти вектору невідомих змінних представлені цілими числами. Цілочислові оптимізаційні задачі

використовуються для пошуку невідомих змінних, значення яких є неподільними числами (наприклад, кількість одиниць техніки, чисельність персоналу, поголів'я тварин тощо).

До частинного випадку задачі цілочислового моделювання належать задачі двійкового програмування. У таких задачах невідомі змінні – це булеві величини, які набувають тільки значень 0 або 1. Найчастіше задачі двійкового програмування використовують в проєктному менеджменті, коли необхідно визначити, які проєкти варто фінансувати: якщо $x=1$, то інвестиційний проєкт доцільно впроваджувати; якщо $x=0$ – фінансування проєкту економічно невигідне. Іншим прикладом застосування задач із булевими змінними є оптимізація асортименту продукції. У цьому випадку, якщо $x=1$, то відповідну продукцію слід виготовляти, а якщо $x=0$ – її виробництво недоцільне.

Ще однією класифікаційною ознакою задач оптимізації є кількість критеріїв оптимальності, показників якості, цільових функцій. За цією ознакою оптимізаційні задачі поділяються на однокритеріальні (наприклад, максимум прибутку) та багатокритеріальні (наприклад забезпечення одночасно і мінімуму собівартості виготовленої продукції і її максимально високої якості).

## 10.2. Формалізація задачі оптимізації на прикладі лінійної моделі

Процес розв'язання задачі оптимізації починається з етапу формалізації, тобто перетворення вербального подання задачі в логіко-математичну конструкцію з метою формування відповідної математичної моделі з урахуванням всіх компонентів: критерію, інструментальних змінних, умов-обмежень. Досліджуючи моделі, можна зробити висновки відносно законів і закономірностей функціонування реальних соціально-економічних систем.

Критерій оптимальності – це показник якості, за яким запропоноване дослідником рішення визнається оптимальним, тобто

найкращим серед всіх можливих альтернативних варіантів з урахуванням заданих умов-обмежень. Критерій оптимальності напряму пов'язаний із ціллю, якої прагне досягти суб'єкт оптимізації. Наприклад, забезпечення максимуму ефективності функціонування підприємства. Для кожної окремої системи у різні періоди часу та за різних зовнішніх обставин показники ефективності різні. Для одного підприємства визначальною є максимізація продуктивності праці, а для іншого – максимальне підвищення якості продукції. У наведених прикладах в якості показників ефективності слід використовувати продуктивність та якість, відповідно. Саме вони є критеріями оптимальності.

У формальному вигляді критерій оптимальності набуває вигляду функції, яка прямує до екстремуму, тобто найбільшого або найменшого значення на заданій множині:

$$F \to \begin{Bmatrix} \max \\ \min \end{Bmatrix}, \tag{10.3}$$

де: $F$ – показник ефективності.

Другим ключовим поняттям задачі оптимізації є інструментальні змінні, під якими розуміються інструменти досягнення визначених критерієм оптимізації цілей. Найчастіше в якості інструментальних змінних на підприємствах використовують обсяги виробництва продукції, витрати, обсяги перевезень тощо. Наприклад, у задачах максимізації прибутку підприємства інструментальними змінними, значення яких потрібно визначити, можуть бути оптимальні обсяги виробництва. Під оптимальними, у даному випадку, маються на увазі такі обсяги виготовлення кожного виду продукції, які забезпечуватимуть отримання максимально можливого прибутку.

Інструментальні змінні в задачі оптимізації мають вигляд вектору:

$$X = (x_1, x_2, x_3, ..., x_n)^T, \tag{10.4}$$

де: $x_1, x_2, x_3, ..., x_n$ – невідомі значення 1-ої, 2-ої, 3-ої, …, *n*-ої змінної.

Значення інструментальних змінних відображають, що необхідно зробити, щоб досягти цілі: скільки виготовити продукції кожного окремого виду, щоб прибуток був максимально можливим; які інвестиційні проєкти впроваджувати, щоб забезпечити максимальний інвестиційний дохід у довгостроковій перспективі; які площі сільськогосподарських угідь відвести під кожну окрему сільськогосподарську культуру, щоб забезпечити максимальний прибуток від виробництва продукції рослинництва тощо. Всі функції оптимізаційної моделі – це функції, які відображають залежності фінансово-економічних показників від інструментальних змінних. Зокрема, критерій оптимальності має вигляд залежності показника ефективності економічної системи від інструментальних змінних.

Третім поняттям задач оптимізації є умови, які обмежують можливі варіанти розв'язку задачі. Такі умови називаються обмеженнями оптимізаційних задач, що мають вигляд рівнянь або, як правило, нерівностей. Наприклад, для соціально-економічної системи природа обмежень полягає у тому, що альтернативні варіанти розв'язку задач оптимізації залежать від низки чинників, основними серед яких є наявні ресурси, потужності, ринкова кон'юнктура, домовленості із контрагентами тощо. Наприклад, прибуток підприємства – показник ефективності – залежить від обсягів виробництва (чим більше максимально рентабельної продукції виготовляється, тим більший розмір отриманого прибутку). Але підприємство може виготовити лише стільки продукції, на скільки у нього вистачить ресурсів. Крім того, якщо раніше було укладено контракти з покупцями на певний обсяг продукції, то незважаючи на її ефективність, підприємство змушене

виготовити той обсяг, що вказано у контракті. Не може суб'єкт бізнесу виробити товару більше, ніж дозволяють його виробничі потужності.

Отже, процес розв'язування задачі оптимізації передбачає пошук найкращого (оптимального) способу досягнення поставленої цілі за конкретних умов. Задача оптимізації включає три компоненти[1]:

1. Цільова функція, яка має вигляд функції показника ефективності від інструментальних змінних, що прямує до екстремуму:

$$F(x) = c_1x_1 + c_2x_2 + ... + c_nx_n = C \cdot X \to \left\{ \begin{matrix} \max \\ \min \end{matrix} \right\} \qquad (10.5)$$

де $c_1, c_2, ..., c_n$ – коефіцієнти цільової функції при першій, другій, …, *n*-ій змінній; $C$ – вектор констант, які відображають коефіцієнти цільової функції $C = (c_1, c_2, ..., c_n)^T$; $X$ – вектор інструментальних змінних $X = (x_1, x_2, ..., x_n)^T$; $x_1, x_2, ..., x_n$ – значення невідомої першої, другої, …, *n*-ої змінних; $n$ – кількість змінних $x$.

2. Ресурсні обмеження у вигляді лінійних нерівностей:

$$\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n \le b_1 \Rightarrow A_1 \cdot X \le b_1 \\
&a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n \le b_2 \Rightarrow A_2 \cdot X \le b_2 \\
&\text{.....} \\
&a_{i1}x_1 + a_{i2}x_2 + ... + a_{in}x_n \le b_i \Rightarrow A_i \cdot X \le b_i \\
&\text{.....} \\
&a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n \le b_m \Rightarrow A_m \cdot X \le b_m
\end{aligned} \qquad (10.6)$$

[1] Бродський Ю. Б., Малютіна В. П. Економіко – математичне моделювання. Конспект лекцій // Житомир: ЖНАЕУ, 2010. – 116 с.

де $a_{ij}$ – константа $i$-го обмеження при $j$-тій інструментальній змінній; $A_1$, $A_2$, …, $A_i$, …, $A_m$ – вектор констант 1-го, 2-го, …, $i$-го, …, $m$-го обмежень, що відповідають вектору інструментальних змінних; $X$ – вектор інструментальних змінних; $b_1$, $b_2$, …, $b_i$, …, $b_m$ – константа 1-го, 2-го, …, $m$-го обмеження;

3. Граничні умови, що обмежують область можливих значень інструментальних змінних. Граничні умови – це мінімальні та/або максимальні значення невідомих інструментальних змінних:

$$\begin{aligned} & X_1^{\min} \le x_1 \le X_1^{\max} \\ & X_2^{\min} \le x_2 \le X_2^{\max} \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ & X_j^{\min} \le x_j \le X_j^{\max} \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ & X_n^{\min} \le x_n \le X_n^{\max}, \end{aligned} \qquad (10.7)$$

де $X_1^{\min}$, $X_2^{\min}$, $X_j^{\min}$, $X_n^{\min}$ – нижня межа (тобто мінімальне значення) 1-ої, 2-ої, …, $j$-ої, …, $n$-ої змінних; $X_1^{\max}$, $X_2^{\max}$, $X_j^{\max}$, $X_n^{\max}$ – верхня межа (тобто максимальне значення) 1-ої, 2-ої, …, $j$-ої, …, $n$-ої змінних.

У більшості задач оптимізації обов'язковою граничною умовою є умова невід'ємності інструментальних змінних:

$$X \ge 0. \qquad (10.8)$$

В операторній формі задача оптимізації має вигляд:

$$F(x)=\sum_{j=1}^{n}c_j x_j \to \begin{Bmatrix}\max\\ \min\end{Bmatrix},$$

$$\sum_{j=1}^{n}a_{ij}x_j \le b_i, \text{ або } a_{i1}x_1+a_{i2}x_2+...+a_{jn}x_n \le b_i, \qquad (10.9)$$

$$x_j \ge 0,\ j=\overline{1,n};\ i=\overline{1,m}.$$

Найбільш розповсюдженою в економічних дослідженнях є задача оптимізації плану виробництва, яка у формалізованому вигляді не відрізняється від задачі, представленій у (10.9). Відмінною рисою задачі оптимізації плану виробництва є конкретизація констант у цільовій функції та ресурсних обмеженнях:

$$F(x)=\sum_{j=1}^{n}c_j x_j \to \begin{Bmatrix}\max\\ \min\end{Bmatrix},$$

$$\sum_{j=1}^{n}a_{ij}x_j \le b_i,\ x_j \ge 0,\ j=\overline{1,n};\ i=\overline{1,m}, \qquad (10.10)$$

де: $c_j$ – показник ефективності (прибуток, виручка від реалізації, витрати тощо), що припадає на одиницю продукції $j$-го виду; $x_j$ – невідомий обсяг виробництва продукції $j$-го виду; $a_{ij}$ – норма витрат ресурсів, необхідна для виготовлення одиниці продукції $j$-го виду; $b_i$ – наявний обсяг ресурсу $i$-го виду.

У задачі оптимізації плану виробництва ліва частина ресурсного обмеження $\sum_{j=1}^{n}a_{ij}x_j$ відображає використані ресурси, а права $b_i$ – наявний обсяг цього ресурсу. Економічний зміст ресурсного обмеження полягає у тому, що підприємство має використати ресурсів не більше, ніж у нього є в наявності. Ресурсні умови обмежують значення невідомих змінних. Наприклад, відомо, що підприємство може вирощувати соєві боби, пшеницю та

цукровий буряк. Виробництво соєвих бобів є найбільш рентабельним і, чим більше буде обсяг виробництва бобів, тим більші прибутки матиме підприємство. Однак, обсяг виготовлення цього виду продукції обмежується наявними ресурсами. Зокрема, маючи у розпорядженні 1000 га, сільськогосподарське підприємство не зможе засіяти сою на більший площі.

Частинним випадком задачі лінійної оптимізації є задачі цілочислового моделювання, у яких значення вектору невідомих змінних $X = (x_1, x_2, x_3, ..., x_n)^T$ є цілими числами. У формальному вигляді задачі цілочислового моделювання відображаються так:

$$F(x) = \sum_{j=1}^{n} c_j x_j \to \begin{Bmatrix} \max \\ \min \end{Bmatrix},$$

$$\sum_{j=1}^{n} a_{ij} x_j \begin{Bmatrix} \le \\ = \\ \ge \end{Bmatrix} b_i, \qquad (10.11)$$

$x_j \ge 0$, $j = \overline{1,n}$; $i = \overline{1,m}$, $x_j$ – цілі числі.

До окремого типу задач цілочислового програмування належать задачі із булевими змінними.

$$F(x) = \sum_{j=1}^{n} c_j x_j \to \begin{Bmatrix} \max \\ \min \end{Bmatrix},$$

$$\sum_{j=1}^{n} a_{ij} x_j \le b_i, \; x_j \ge 0, \qquad (10.12)$$

$$j = \overline{1,n}; \; i = \overline{1,m}, \; x_j \in \{0;1\}.$$

Отже, у формалізованому вигляді задачі оптимізації включають три компоненти: цільову функцію, ресурсні обмеження та граничні умови. Цільова функція задачі лінійної оптимізації – це

функція, що відображає лінійну залежність показника ефективності (наприклад, прибутку, доходу, собівартості, продуктивності тощо) від інструментальних змінних, що прямує до екстремального значення (максимуму або мінімуму). Ресурсні умови обмежують невідомі значення інструментальних змінних і мають вигляд рівнянь або нерівностей.

### 10.3. Задача стохастичного моделювання

Часто прийняття управлінських рішень здійснюється в умовах відсутності повної та точної інформації про середовище економічних систем, що спричиняє невизначеність майбутніх результатів управління. Тому, будь-який вид господарської діяльності здійснюється в умовах невизначеності, врахування якої є необхідною умовою успішного ведення бізнесу. Невизначеність – це характеристика зовнішнього та внутрішнього середовища суб'єкта прийняття рішення, існування якої зумовлене неможливістю отримати повні та точні знання щодо проблеми.

Детерміновані лінійні моделі, описані вище, не враховують дію невизначеності на суб'єкти підприємництва. Зокрема всі параметри таких моделей (наприклад, коефіцієнти цільових функцій, наявні обсяги ресурсів, норми витрат тощо) є сталими величинами. Однак, для вирішення багатьох фінансово-економічних проблем необхідно враховувати невизначеність. У таких випадках використовуються стохастичні моделі, які передбачають використання випадкових, а не детермінованих величин. Значення випадкових величин можна передбачити лише з певною ймовірністю, оскільки вони залежать від випадкових чинників. У формальному вигляді випадкові чинники відображаються у вигляді вектора: $\overline{\omega} = (\omega_1, \omega_2, ..., \omega_k)^T$, де $\omega_1, \omega_2, ..., \omega_k$ – перший, другий, …, *k*-тий випадковий чинник. Залежність показника $a$ від випадкових чинників відображається як функція $a = f(\omega)$.

У задачах оптимізації випадковими можуть бути:

– вектор коефіцієнтів цільової функції $\bar{c}(\omega)$;

– вектор коефіцієнтів обмежень $\bar{a}(\omega)$;

– вектор коефіцієнтів констант обмежень (наприклад, обсягів наявних ресурсів) $\bar{b}(\omega)$.

Варіанти цільової функції задачі стохастичного програмування (табл. 10.1) [2]:

*Таблиця 10.1.* **Формалізація цільової функції задачі стохастичного моделювання**

| Формалізований вигляд | Умовні позначення | Інтерпретація |
|---|---|---|
| $P\left\{\sum_{j=1}^{n} c_j(\omega)x_j \begin{Bmatrix}\geq\\ \leq\end{Bmatrix}\xi\right\} \to \{\max\}$. | $c_j(\omega)$ – коефіцієнт цільової функції для *j*-тої змінної, який є випадковою величиною; $x_j$ – шукане значення *j*-тої змінної; $n$ – кількість змінних; $\xi$ – верхня (нижня) межа показника цільової функції | Максимум ймовірності того, що показник цілі буде більшим (меншим) за пороговий рівень |
| $F = M\left\{\sum_{j=1}^{n} c_j(\omega)x_j\right\} \to \begin{Bmatrix}\max\\ \min\end{Bmatrix}$. | $M\left\{\sum_{j=1}^{n} c_j(\omega)x_j\right\}$ – математичне сподівання показника цільової функції | Максимум (мінімум) математичного сподівання показника цілі |
| $F = D\left\{\sum_{j=1}^{n} c_j(\omega)x_j\right\} \to \min$. | $D\left\{\sum_{j=1}^{n} c_j(\omega)x_j\right\}$ – дисперсія показника цільової функції | Мінімум дисперсії показника цілі |

[2] Николюк О. М., Бродський Ю. Б., Молодецька К. В. Оптимізаційні методи і моделі: підручник // Житомир: вид-во «Житомирський національний агроекологічний університет», 2019. – 144 с.

– максимізація ймовірності потрапляння показника цілі в інтервал;

– максимізація (мінімізація) математичного сподівання показника цілі;

– мінімізація дисперсії показника цільової функції.

Вибір цільової функції має враховувати особливості об'єкту моделювання. Відображені у табл. 10.1 функції цілі є узагальненими та можуть корегуватись залежно від умов кожної конкретної задачі.

Аналогічно багатоваріантності стохастичних цільових функцій, існує кілька видів стохастичних обмежень:

– умова перевищення ймовірності того, що показник обмеження буде більшим (меншим) за граничний рівень;

– умова перевищення (неперевищення) математичного очікування показника певного граничного рівня;

– умова неперевищення середньоквадратичного відхилення показника обмеження певного граничного рівня.

В основу стохастичних обмежень можуть бути покладені як ймовірності виконання певної умови, так і математичне сподівання та дисперсія (табл. 10.2).

Специфіка стохастичних моделей визначає необхідність їх адаптації до надбудови «Пошук рішення» *MS Excel*. Зазначена надбудова призначена для розв'язування детермінованих задач лінійного та нелінійного моделювання, а стохастичні моделі із ймовірністю не належать до жодного із цих типів задач економіко-математичного моделювання. Тому першим етапом розв'язування задач стохастичного моделювання є приведення їх до детермінованого вигляду. Для цього спочатку слід розглянути типи стохастичних задач.

*Таблиця 10.2.* **Формалізація обмежень задачі стохастичного моделювання**

| Формалізований вигляд | Умовні позначення | Інтерпретація |
|---|---|---|
| $P\left\{\sum_{j=1}^{n} a_{ij}(\omega)x_j \begin{Bmatrix}\le\\ \ge\end{Bmatrix} b_i\right\} \ge p_i$;<br>$P\left\{\sum_{j=1}^{n} a_{ij}(\omega)x_j \begin{Bmatrix}\le\\ \ge\end{Bmatrix} b_i(\omega)\right\} \ge p_i$ | $x_j$ – шукана змінна *j*-го виду; $a_{ij}(\omega)$ – випадкова величина норми витрат *i*-го виду ресурсу, що припадає на одиницю змінної *j*-го виду; $b_i$ – граничний рівень *i*-го обмеження; $b_i(\omega)$ – граничний рівень *i*-го обмеження, який є випадковою величиною; $p_i$ – мінімальна ймовірність того, що показник *i*-го обмеження більший (менший) за граничний рівень $b_i$ або $b_i(\omega)$ | Ймовірність того, що показник обмеження $\left(\sum_{j=1}^{n} a_{ij}(\omega)x_j\right)$ перевищує (не перевищує) граничний рівень $b_i$ $(b_i(\omega))$, має бути не меншою за $p_i$. Для ресурсних обмежень, це означає, що ймовірність того, що ресурсів вистачить, має бути не меншою за $p_i$ |
| $M\left\{\sum_{j=1}^{n} a_{ij}(\omega)x_j \right\} \begin{Bmatrix}\le\\ \ge\end{Bmatrix} b_i$;<br>$\sum_{j=1}^{n} a_{ij}x_j \begin{Bmatrix}\le\\ \ge\end{Bmatrix} M\{b_i(\omega)\}$;<br>$M\left\{\sum_{j=1}^{n} a_{ij}(\omega)x_j \right\} \begin{Bmatrix}\le\\ \ge\end{Bmatrix} M\{b_i(\omega)\}$ | $M\left\{\sum_{j=1}^{n} a_{ij}(\omega)x_j \right\}$ – математичне сподівання показника *i*-го обмеження; $M\{b_i(\omega)\}$ – – математичне сподівання граничного рівня *i*-го обмеження | Умова, що показник *i*-го обмеження $\left(\sum_{j=1}^{n} a_{ij}(\omega)x_j\right)$ або його математичне сподівання не має перевищувати (бути меншим) граничного рівня $b_i$ або математичного сподівання $b_i(\omega)$ |

Продовження таблиці 10.2

| Формалізований вигляд | Умовні позначення | Інтерпретація |
|---|---|---|
| $D\left\{\sum_{j=1}^{n} a_{ij}(\omega)x_j\right\} \le \lambda_i$. | $D\left\{\sum_{j=1}^{n} a_{ij}(\omega)x_j\right\}$ – дисперсія показника *i*-го обмеження; $\lambda_i$ – граничний рівень показника *i*-го обмеження | Умова, що дисперсія показника *i*-го обмеження $\left(\sum_{j=1}^{n} a_{ij}(\omega)x_j\right)$ не має перевищувати граничний рівень $\lambda_i$ |

Залежно від типу цільової функції задачі стохастичного моделювання поділяються на *M*-задачі та *P*-задачі. Критерієм оптимальності *M*-задачі є максимум (мінімум) математичного сподівання показника цілі, а *P*-задачі – максимум ймовірності того, що показник цілі потрапить у визначену область:

1. *M-задача*: знайти оптимальні значення змінних $x_j$, які б забезпечили отримання максимуму математичного сподівання випадкової величини показника цілі $F(\omega) = \sum_{j=1}^{n} c_j(\omega) \cdot x_j$ (із математичним сподіванням $\sum_{j=1}^{n} \overline{c_j} x_j$) за умови непервищення випадкової величини обмеження $A(\omega) = \sum_{j=1}^{n} a_{ij}(\omega)x_j$ (із математичним сподіванням $\sum_{j=1}^{n} \overline{a_{ij}} x_j$ та дисперсією $\sum_{j=1}^{n} \sigma_{aij}{}^2 x_j^2$) заздалегідь визначеного числа випадкової величини $b_i(\omega)$ (із математичним сподіванням $\overline{b_i}$ та дисперсією $\sigma_{b_i}^2$) із ймовірністю не менше $p_i$.

Використовуючи положення теорії ймовірності М-задачу можна привести до детермінованого вигляду (табл. 10.3).

*Таблиця 10.3.* **Формалізація *M*-задачі стохастичного моделювання**

| Задача стохастичного програмування | Детермінована задача стохастичного програмування |
|---|---|
| $F = \sum_{j=1}^{n} c_j(\omega) \cdot x_j \to \max$ | $F = \sum_{j=1}^{n} \overline{c_j} \cdot x_j \to \max$ |
| *за обмежень:* | |
| $P\left\{\sum_{j=1}^{n} a_{ij}(\omega) x_j \le b_i(\omega)\right\} \ge p_i$, $x_j \ge 0$ | $\Phi^{-1}(p_i) \cdot \sqrt{\sum_{j=1}^{n} {\sigma_{aij}}^2 x_j^2 + {\sigma_{bi}}^2} \le \bar{b} - \sum_{j=1}^{n} \overline{a_{ij}} x_j$, $x_j \ge 0$ |
| $P\left\{\sum_{j=1}^{n} a_{ij}(\omega) x_j \le b_i\right\} \ge p_i$, $x_j \ge 0$ | $\Phi^{-1}(p_i) \cdot \sqrt{\sum_{j=1}^{n} {\sigma_{aij}}^2 x_j^2} \le b_i - \sum_{j=1}^{n} \overline{a_{ij}} x_j$, $x_j \ge 0$ |

Примітка: $\Phi^{-1}(k) = \frac{1}{\Phi(k)}$; $\Phi(k)$ – функція Лапласа у точці $k$.

Задачі стохастичного моделювання у детермінованому вигляді належать до класу нелінійних задач оптимізації. Розглянемо приклад створення комп'ютерної форми стохастичної *M*-задачі та *P*-задачі з трьома невідомими змінними $x_1$, $x_2$, $x_3$.

*M-задача* звучить таким чином: знайти оптимальні значення змінних $x_1$, $x_2$, $x_3$, які б забезпечили отримання максимуму математичного сподівання показника цілі. Показник цілі $F$ – це випадкова величина, яка має вигляд: $F(\omega) = \sum_{j=1}^{3} c_j(\omega) \cdot x_j$. Математичне сподівання показника цілі $F$ відображено в такий спосіб: $\sum_{j=1}^{3} \overline{c_j} x_j$. У задачі необхідно врахувати наступну умову: ймовірність того, що випадкова величина обмеження $A(\omega) = \sum_{j=1}^{3} a_j x_j$ (використаний обсяг конкретного ресурсу) не буде перевищувати

наперед встановленого значення $\xi$, має бути не меншою ніж $p$. Математичне сподівання та дисперсія показника обмеження $A(\omega) = \sum_{j=1}^{3} a_j x_j$ розраховуються за формулами $\sum_{j=1}^{3} \overline{a_{ij}} x_j$ й $\sum_{j=1}^{3} {\sigma_{a_{ij}}}^2 x_j^2$, відповідно. Формалізовану модель такої задачі, приведену до детермінованого вигляду відображено у табл. 10.3.

2. Р-задача: знайти оптимальні значення змінних $x_j$, які б забезпечили отримання максимуму ймовірності досягнення показником цілі, що є випадковою величиною $F(\omega) = \sum_{j=1}^{n} c_j(\omega) \cdot x_j$ (із математичним сподіванням $\sum_{j=1}^{n} \overline{c_j} x_j$ та дисперсією $\sum_{j=1}^{n} {\sigma_{c_j}}^2 x_j^2$), наперед встановленого значення $\xi$. При цьому, із ймовірністю не менше $p_i$ має бути виконано умову непервищення випадкової величини обмеження $A(\omega) = \sum_{j=1}^{n} a_{ij}(\omega) x_j$ (із математичним сподіванням $\sum_{j=1}^{n} \overline{a_{ij}} x_j$ та дисперсією $\sum_{j=1}^{n} {\sigma_{a_{ij}}}^2 x_j^2$) заздалегідь визначеного числа випадкової величини $b(\omega)$ (із математичним сподіванням $\overline{b_i}$ та дисперсією ${\sigma_{bi}}^2$). Результати приведення елементів Р-задачі до детермінованого вигляду відображено у табл. 10.4.

*Р-задача* звучить таким чином: знайти оптимальні значення змінних $x_1$, $x_2, x_3$, які б забезпечили отримання максимуму ймовірності того, що показник цілі досягне мінімально прийнятного рівня. Показник цілі $F$ – це випадкова величина, яка має вигляд: $F(\omega) = \sum_{j=1}^{3} c_j(\omega) \cdot x_j$. Математичне сподівання показника цілі $F$ має вигляд: $\sum_{j=1}^{3} \overline{c_j} x_j$, а його дисперсія – $\sum_{j=1}^{3} {\sigma_{c_j}}^2 x_j^2$.

*Таблиця 10.4.* **Формалізація *P*-задачі стохастичного моделювання**

| Задача стохастичного програмування | Детермінована задача стохастичного програмування |
|---|---|
| $F = P\left\{\sum_{j=1}^{n} c_j(\omega)x_j \geq \xi\right\} \to \max$<br>де $\xi = const$ | $F = \dfrac{\sum_{j=1}^{n} \bar{c}_j x_j - \xi}{\sqrt{\sum_{j=1}^{n} \sigma_{c_j}{}^2 x_j^2}} \to \max$ |
| *за обмежень:* | |
| $P\left\{\sum_{j=1}^{n} a_{ij}(\omega)x_j \leq b_i(\omega)\right\} \geq p_i,$<br>$x_j \geq 0$ | $\Phi^{-1}(p_i)\cdot\sqrt{\sum_{j=1}^{n} \sigma_{aij}{}^2 x_j^2 + \sigma_{bi}{}^2} \leq \bar{b} - \sum_{j=1}^{n} \overline{a_{ij}} x_j,$<br>$x_j \geq 0$ |
| $P\left\{\sum_{j=1}^{n} a_{ij}(\omega)x_j \leq b_i\right\} \geq p_i,$<br>$x_j \geq 0$ | $\Phi^{-1}(p_i)\cdot\sqrt{\sum_{j=1}^{n} \sigma_{aij}{}^2 x_j^2} \leq b_i - \sum_{j=1}^{n} \overline{a_{ij}} x_j,$<br>$x_j \geq 0$ |

У задачі необхідно врахувати таку умову: ймовірність того, що випадкова величина обмеження $A(\omega) = \sum_{j=1}^{3} a_j x_j$ (використаний обсяг конкретного ресурсу) не буде перевищувати наперед встановленого значення $\xi$, має бути не меншою ніж $p$. Математичне сподівання та дисперсія показника обмеження $A(\omega) = \sum_{j=1}^{3} a_j x_j$ розраховуються за формулами $\sum_{j=1}^{3} \overline{a_{ij}} x_j$ й $\sum_{j=1}^{3} \sigma_{aij}{}^2 x_j^2$, відповідно. Формалізовану модель такої задачі, приведену до детермінованого вигляду відображено у табл. 10.4.

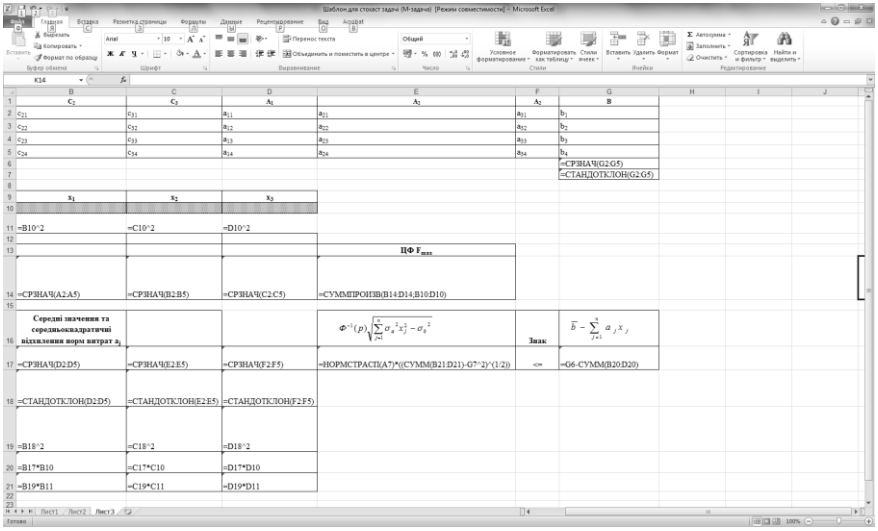
За умовою задачі відомі ряди розподілу випадкових величин $C_1$ (наприклад, ціна або прибуток від першого виду продукції), $C_2$

(наприклад, ціна або прибуток від другого виду продукції), $C_3$ (наприклад, ціна або прибуток від третього виду продукції), $A_1$ (норма витрат ресурсу на виготовлення одиниці першого виду продукції), $A_2$ (норма витрат ресурсу на виготовлення одиниці другого виду продукції), $A_3$ (норма витрат ресурсу на виготовлення одиниці третього виду продукції), $B$ – наявний обсяг ресурсу:

$$C_1 = \begin{pmatrix} c_{11} \\ c_{12} \\ c_{13} \\ c_{14} \end{pmatrix},\ C_2 = \begin{pmatrix} c_{21} \\ c_{22} \\ c_{23} \\ c_{24} \end{pmatrix},\ C_3 = \begin{pmatrix} c_{31} \\ c_{32} \\ c_{33} \\ c_{34} \end{pmatrix},$$

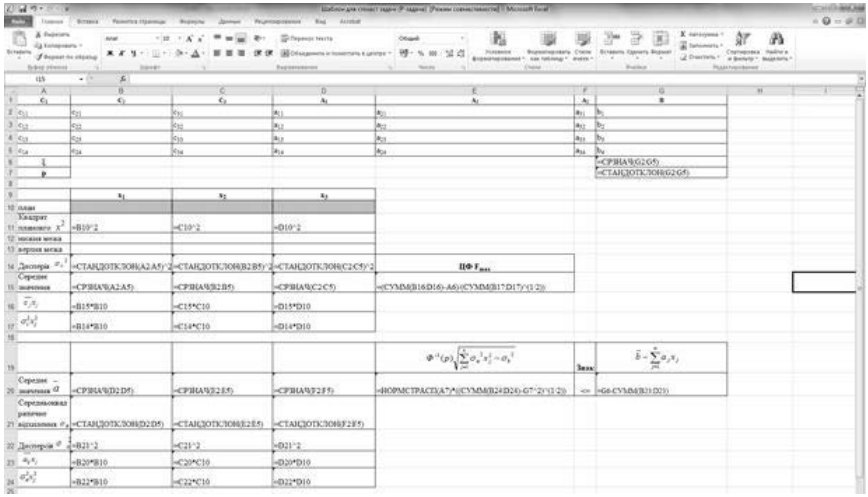
$$A_1 = \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \\ a_{14} \end{pmatrix},\ A_2 = \begin{pmatrix} a_{21} \\ a_{22} \\ a_{23} \\ a_{24} \end{pmatrix},\ A_3 = \begin{pmatrix} a_{31} \\ a_{32} \\ a_{33} \\ a_{34} \end{pmatrix},\ B = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}.$$

Ряди розподілу включають дані показника або за певний проміжок часу, або для сукупності однорідних об'єктів. Наприклад, ряд розподілу ціни на хліб може бути представлений у вигляді цін на цей вид продукту харчування або за останні кілька років, або для сукупності хлібопекарських підприємств. Функція Лапласа від числа $\Phi(p)$ визначається, виходячи із формули $\Phi(p) = F(p) - 0{,}5$, де $F(p)$ – це функція стандартного нормального інтегрального розподілу від числа $p$. У середовищі *MS Excel* значення функції стандартного нормального розподілу $F(p)$ визначається за допомогою функції НОРМСТРАСП (*p*). Форми введення умови *M*-задачі та *P*-задач стохастичного програмування наведено на рис. 10.1, 10.2.



| | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| 1 | $c_2$ | $c_3$ | $a_1$ | $a_2$ | $a_2$ | B |
| 2 | $c_{21}$ | $c_{31}$ | $a_{11}$ | $a_{21}$ | $a_{31}$ | $b_1$ |
| 3 | $c_{22}$ | $c_{32}$ | $a_{12}$ | $a_{22}$ | $a_{32}$ | $b_2$ |
| 4 | $c_{23}$ | $c_{33}$ | $a_{13}$ | $a_{23}$ | $a_{33}$ | $b_3$ |
| 5 | $c_{24}$ | $c_{34}$ | $a_{14}$ | $a_{24}$ | $a_{34}$ | $b_4$ |
| 6 | | | | | | =СРЗНАЧ(G2:G5) |
| 7 | | | | | | =СТАНДОТКЛОН(G2:G5) |
| 8 | | | | | | |
| 9 | $x_1$ | $x_2$ | $x_3$ | | | |
| 10 | | | | | | |
| 11 | =B10^2 | =C10^2 | =D10^2 | | | |
| 12 | | | | | | |
| 13 | | | | **ЦФ $F_{max}$** | | |
| 14 | =СРЗНАЧ(A2:A5) | =СРЗНАЧ(B2:B5) | =СРЗНАЧ(C2:C5) | =СУММПРОИЗВ(B14:D14;B10:D10) | | |
| 15 | | | | | | |
| 16 | **Середні значення та середньоквадратичні відхилення норм витрат $a_j$** | | | $\Phi^{-1}(p)\sqrt{\sum_{j=1}^{n}\sigma_a^2 x_j^2 - \sigma_b^2}$ | **Знак** | $\overline{b} - \sum_{j=1}^{n} a_j x_j$ |
| 17 | =СРЗНАЧ(D2:D5) | =СРЗНАЧ(E2:E5) | =СРЗНАЧ(F2:F5) | =НОРМСТРАСП(A7)*((СУММ(B21:D21)-G7^2)^(1/2)) | <= | =G6-СУММ(B20:D20) |
| 18 | =СТАНДОТКЛОН(D2:D5) | =СТАНДОТКЛОН(E2:E5) | =СТАНДОТКЛОН(F2:F5) | | | |
| 19 | =B18^2 | =C18^2 | =D18^2 | | | |
| 20 | =B17*B10 | =C17*C10 | =D17*D10 | | | |
| 21 | =B19*B11 | =C19*C11 | =D19*D11 | | | |
| 22 | | | | | | |
| 23 | | | | | | |

*Рисунок 10.1.* **Форма введення умови задачі стохастичного моделювання в MS Excel (*M*-задача)**



| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | $c_2$ | $c_2$ | $c_3$ | $a_1$ | $a_2$ | $a_2$ | B |
| 2 | $c_{21}$ | $c_{21}$ | $c_{31}$ | $a_{11}$ | $a_{21}$ | $a_{31}$ | $b_1$ |
| 3 | $c_{22}$ | $c_{22}$ | $c_{32}$ | $a_{12}$ | $a_{22}$ | $a_{32}$ | $b_2$ |
| 4 | $c_{23}$ | $c_{23}$ | $c_{33}$ | $a_{13}$ | $a_{23}$ | $a_{33}$ | $b_3$ |
| 5 | $c_{24}$ | $c_{24}$ | $c_{34}$ | $a_{14}$ | $a_{24}$ | $a_{34}$ | $b_4$ |
| 6 | $\overline{x}$ | | | | | | =СРЗНАЧ(G2:G5) |
| 7 | $p$ | | | | | | =СТАНДОТКЛОН(G2:G5) |
| 8 | | | | | | | |
| 9 | | $x_1$ | $x_2$ | $x_3$ | | | |
| 10 | план | | | | | | |
| 11 | Квадрат плановых $x^2$ | =B10^2 | =C10^2 | =D10^2 | | | |
| 12 | минимальная мощность | | | | | | |
| 13 | верхняя мощность | | | | | | |
| 14 | Дисперсия $\sigma_a^2$ | =СТАНДОТКЛОН(A2:A5)^2 | =СТАНДОТКЛОН(B2:B5)^2 | =СТАНДОТКЛОН(C2:C5)^2 | **ЦФ $F_{max}$** | | |
| 15 | Средние значения | =СРЗНАЧ(A2:A5) | =СРЗНАЧ(B2:B5) | =СРЗНАЧ(C2:C5) | =(СУММ(B16:D16)-A6)/(СУММ(B17:D17)^(1/2)) | | |
| 16 | $\overline{a_j} x_j$ | =B15*B10 | =C15*C10 | =D15*D10 | | | |
| 17 | $\sigma_a^2 x_j^2$ | =B14*B10 | =C14*C10 | =D14*D10 | | | |
| 18 | | | | | | | |
| 19 | | | | | $\Phi^{-1}(p)\sqrt{\sum_{j=1}^{n}\sigma_a^2 x_j^2 - \sigma_b^2}$ | **Знак** | $\overline{b} - \sum_{j=1}^{n} a_j x_j$ |
| 20 | Средние значения $\overline{a}$ | =СРЗНАЧ(D2:D5) | =СРЗНАЧ(E2:E5) | =СРЗНАЧ(F2:F5) | =НОРМСТРАСП(A7)*((СУММ(B24:D24)-G7^2)^(1/2)) | <= | =G6-СУММ(B21:D21) |
| 21 | Среднеквадратичное отклонение $\sigma_a$ | =СТАНДОТКЛОН(D2:D5) | =СТАНДОТКЛОН(E2:E5) | =СТАНДОТКЛОН(F2:F5) | | | |
| 22 | Дисперсия $\sigma^2$ | =B21^2 | =C21^2 | =D21^2 | | | |
| 23 | $a_j x_j$ | =B20*B10 | =C20*C10 | =D20*D10 | | | |
| 24 | $\sigma_a^2 x_j^2$ | =B22*B10 | =C22*C10 | =D22*D10 | | | |
| 25 | | | | | | | |

*Рисунок 10.2***. Форма введення умови задачі стохастичного моделювання в MS Excel (*P* -задача)**

### *Список питань до самоконтролю:*

**1.** Зміст та особливості задач оптимізації.

**2.** Види задач оптимізації. Приклади.

**3.** Сутність та приклади критерію оптимальності. Формальне представлення критерію оптимальності.

**4.** Що таке «інструментальні змінні» в задачах оптимізаційного моделювання.

**5.** Поняття обмежуючих умов в оптимізаційних задачах та їх представлення у формальному вигляді.

**6.** Зміст і компоненти математичної моделі оптимізації.

**7.** Формальна постановка задачі оптимізації плану виробництва.

**8.** Характеристика задач стохастичного моделювання.

**9.** Математичний вираз стохастичних цільової функції та обмежень.

**10.** Зміст М-задачі та Р-задачі стохастичного моделювання.

**11.** Детермінований вигляд М-задачі та Р-задачі стохастичного моделювання.

### *Список рекомендованої літератури:*

1. Бережна Л.В., Снитюк О.І. Економіко – математичні методи та моделі в фінансах. – К: Кондор, - 2009. – 301 с.
2. Бродський Ю. Б., Малютіна В. П. Економіко – математичне моделювання. Конспект лекцій // Житомир: ЖНАЕУ, 2010. – 116 с.
3. Бродський Ю. Б., Сайкевич М. І. Економіко-математичні методи та моделі: навч. посіб. // Житомир: Вид-во «Житомирський національний агроекологічний університет», 2016. – 148 с.
4. Економіко-математичне моделювання: навч. посіб. / За ред. О. Т. Іващука. – Тернопіль : ТНЕУ «Економічна думка», 2008. – 704 с.
5. Николюк О. М., Бродський Ю. Б., Молодецька К. В. Оптимізаційні методи і моделі: підручник // Житомир: вид-во «Житомирський національний агроекологічний університет», 2019. – 144 с.
6. Оптимізаційні методи та моделі: підр. / Л. В. Забуранна [та ін.]. – К. : ЦП «Компринт», 2014. – 372 с.

## Розділ 11. СТАТИСТИЧНІ МЕТОДИ МОДЕЛЮВАННЯ РИЗИКІВ

### 11.1. Основи поняття ризику

Ризик є природною складовою функціонування суспільства у всіх сферах його діяльності. В економіці розгляд господарського суб'єкта ринку(фірми, підприємства, об'єднання, холдингу) як виробничої системи, тобто як сукупності взаємодіючих між собою, а також з довкілям елементів, змушує визнати, що його діяльність носить невизначений (стохастичний) характер. Тому рішення, що приймається власником виробничої системи, завжди пов'язані з ризиком. Стохастичний характер в діяльності суб'єктів притаманний і в інших сферах людської діяльності.

Ризик поняття комплексне, категоріальне. *Ризик як історична категорія* – це небезпека пов'язана з настанням небажаних подій. Приклад, небезпека ураження блискавкою повітряного судна., Виникає питання про кількісну характеристику ризику, тобто кількісну характеристику цієї події (міру події). Теорія ймовірності якраз і займається кількісним відображенням події, по – перше, з точку зору якісного аналізу подій, неважливо яких, бажаних чи небажених, по – друге – випадкових, абстрактних подій. Стохастичний характер системи, дає можливість її імовірністної формалізації – побудови імовірністної моделі ризику ( наприклад, моделі ризику суб'єктів ринку).

В основі теорії імовірності лежить поняття випадкової події. Сама подія в теорії імовірності розглядається, в контексті з поняттям простору елементарних подій (позначають Ω.). В сучасній теорії ризику (ризикології), на відміну теорії імовірності, розглядаються цілком реальні випадкові події.

*Ризик як ймовірністна категорія* – це можливість (настання ) появи з іншої події елементарної випадкової події, яку називають *ризиковою,*

Якщо звернутись до поняття такої операції в теорії ймовірності, як слідування із однієї події іншої, то, позначивши ризикову подію через $B$, а небажану подію через $A$ , за структурою ризик, як ймовірнісна категорія, може бути визначений за імплікацією подій $A \Rightarrow B$ (із події $A$ слідує подія $B$)

Серед елементарних ризикових випадкових подій виділяють важливий їх клас – *наслідки*. В теорії ризику такими наслідками є: шкода, втрати, (втрати прибутку) втрати певного рівня, перевитрати, збитки, банкрутство, програш, відхилення від цілі в небажану сторону, виграш, вигода, прибуток, прибуток певного рівня, доход, відхилення від цілі в бажану сторону і т.д. Шкода, втрати, перевитрати, збитки, банкрутство, небажане відхилення від цілі і т.д. складають в теорії ризику поняття *небажаних наслідків* (ризикових подій), а виграш, вигода, прибуток, доход, відхилення від цілі в бажану сторону і т.д. складають поняття *бажаних наслідків* (ризикових подій). Взагалі, ризикова подія називається просто ризиком.

Як правило, наслідок має кількісне відображення (величина втрат, збитків, прибутку і т.д.). З позицій теорії імовірності кількісне відображення наслідку $B$ є значення випадкової величини $X(B)$ на події $B$.Значення функції $X(B)$ назвемо результатом наслідку $B$ , його мірою. Коли $X(B)=0$ – говорять про наслідок з нульовим результатом.

Наслідки – це характеристика певної множини випадкових подій в просторі елементарних подій $\Omega$. Саме поняття ризику як категорії – *це можливість настання(появи) з невизначеним результатом, випадкової (випадкових) ризикової (ризикових) події (подій) у відносинах, процесах і явищах суспільства.*

Приклад ризику (наслідку, назвемо його $B$) - втрата вкладених коштів клієнтом банку. Для події $B$ (небажана подія) існує протилежна подія $\overline{B}$ - збереження вкладу (бажана подія).

В ризикології розглядаються складні події, серед елементарних подій яких присутні ризикові події (наслідки), Якщо

ризикова подія *B* слідує із іншої випадкової події *A*, то в цьому випадку випадкова подія *A* називається *причиною* ризикової події *B* або просто *причиною ризику*, Говорять, що "ризик події *B* повязаний з подією *A*". В широкому розумінні, причина ризику (активне його начало) – це випадкове явище, подія, що спонукає іншу подію (ризикову), яка може потенційно відбутися. Наприклад, в якості причини ризику (втрати вкладу клієнтом банку), можна розглядати випадкову подію *A* – банкрутство банку, що має наслідок *B*– втрата вкладених коштів клієнтом банку. Причина сама по собі ще не визначає наслідок, вона залежить від факторів, завдяки яким вона виникає. В свою чергу, для виникнення наслідку необхідне органічне поєднання причин і факторів. Обєднання (сукупність) факторів називають *ситуацією*. Як відомо, в більшості випадків термін "ситуація", визначається як комбінація, сукупність різних обставин і умов(факторів), що створюють певну обстановку для дії .

Сукупність факторів випадкова подія (*C*), їх називають *факторами ризику* (умовами, обставинами ризику). Обєднання факторів ризику, що визначають причину ризику, і причини ризику для даної ризикової події, називають *ризиковою ситуацією*.

З приведених вище міркувань можна констатувати, ризик комплексне структурне поняття, характеризується наступною структурною моделлю - імплікацією випадкових подій, $C \Rightarrow A \Rightarrow B$. Фактором (обставиною) виникнення банкрутства банку (подія *C*) може бути, наприклад конкурентна боротьба банків (подія *C*-конкурентний фактор). Очевидно, $C \Rightarrow A$. Конкурентна боротьба банків створює ризикову ситуацію. Фактор конкуренції на ринку (подія *C*) для торгового підприємства спонукає таку подію (причину ризику), як скорочення обсягів збуту, що в свою чергу може привести до негативного наслідку – зниження доходу підприємства.

Ризикові ситуації бувають:

1. *Визначені* - наявність повної інформації про сукупність різних обставин і умов(факторів),що створюють певну ризикову ситуацію (повна визначеність).

2. *Невизначені* - а) така інформація відсутня повністю (повна невизначеність); б) така інформація відсутня частково (часткова невизначеність).

Повна невизначеність характерна тим, що імовірність ризикової події невідома із-за відсутності необхідної інформації– це такий вид визначеності, який характеризується близькою до нуля імовірністю прогнозованості настання подій. Наприклад, в умовах повної невизначеності суб'єкти підприємницької діяльності позбавлені прогнозувати перспективи свого розвитку і ринку в цілому.

Повна визначеність характеризується близькою до одиниці імовірністю прогнозованості настання події, тобто суб'єкт підприємницької діяльності на сто відсотків може прогнозувати свою стратегію на ринку і ринок в цілому.

Часткова невизначеність характеризується тим, що ймовірність настання події знаходиться між 0 і 1. Даний вид невизначеності носить конкретний, практичний характер і є предметом дослідження в теорії ризику.

Невизначені ситуації називають просто "невизначеністю". Таким чином, виходячи із структурної моделі ризику: $C \Rightarrow A \Rightarrow B$ можна зробити висновок, що ризик є *наслідком невизначеності*.

*Об'єктом ризику* будемо називати управлінську(керовану) систему, ефективність наслідків і умови функціонування якої наперед точно невизначена.

*Суб'єкт ризику* - одиниця управлінської системи, яка зацікавлена в результатах управління об'єктом ризику і має компетентність прийняття рішення по відношенню об'єкта ризику.

*Джерело ризику* – це предмети, які породжують невизначеність об'єкта ризику.

Джерелом ризику, наприклад в економіці може бути:

- *господарська (підприємницька) діяльність;*
- *політична діяльність;*
- *особистість;*
- *природні явища.*

Наприклад, поміщаючи вклад на депозит в банку, для вкладника ризик - втрата вкладу внаслідок банкрутства банку, Об'єкт ризику – грошові кошти вкладника, які він помістив в банк. Суб'єкт ризику – їх власник, або вкладник.Джерело ризику – банк.

*Страхові фон*ди.

В ринковій економіці виникають два питання:

- яким чином фактор невизначеності впливає на поведінку людей;

- яким чином схильність до ризику розподіляється серед окремих індивідів.

Відповідь на перше питання зводиться до того, що люди в своїй більшості не люблять ризикувати і готові для уникнення ризику заплатити за нього гроші. Таким чином ризик в ринковій економіці виступає товаром.

Відповідь на друге питання полягає в тому , що індивідуальне відношення до ризику продиктоване існуванням трьох типів суб'єктів ризику :

- не схильні до ризику, тобто противники ризику,
- нейтральні до ризику,
- любителі (схильні до) ризику.

Статистика говорить, що найбільш чисельна група – особи, не схильні до ризику. Важливість фактору ризику обумовлює його стати товаром (є особи, які готові заплатити за нього). Таким чином можна говорити про ринок ризиків. Споживачі такого товару є страхові компанії. Розраховуються страхові компанії за такий товар страховими полісами Одержані грошові кошти від продажу полісів формують страховий фонд. Кожний, будь – то фізична чи юридична

особа, одержуючи страховий поліс, звільнюється від ризику, передаючи його на плечі страховій компанії. Страховий поліс виступає об'єктом ризику, суб'єкт – власник полісу, джерело ризику – страхова компанія.

### 11.2. Класифікація ризиків

Будемо розглядати множину подій (їх називають діяльністю суб'єктів господарювання), що мають безпосереднє відношення до суб'єкта господарювання.. Сукупність ризикових подій для суб'єкта господарювання називають його *ризиковою діяльністю*. Це можуть бути ризикові події, джерелом виникнення яких є сам суб'єкт господарювання, в цьому випадку говорять про *ризикові події суб'єктивного характеру*, так і зовнішні ризикові події відносно суб'єкта господарювання – *ризикові події об'єктивного характеру*. Господарський ризик– *можливість появи суб'єктивно-об'єктивного характеру, з невизначеним результатом, ризикових подій в економічних відносинах, процесах і явищах суспільства.*

Оскільки, ризикові події можуть бути як об'єктивного характеру, так і суб'єктивного, то відповідно можна розглядати для суб'єкта господарської діяльності об'єктивну і суб'єктивну невизначеність. Слід зауважити, що фактори, які діють на господарські ризики, бувають зовнішні і внутрішні. Під зовнішніми слід розуміти ті умови довкілля, в яких відбувається господарська діяльність і стан яких, субєкт господарської діяльності не може змінити. Наприклад природні фактори, демографічні фактори. Внутрішні - пов'язані з виробничо – фінансовою діяльністю підприємства.

Фактори в структурі господарського ризику бувають зовнішні і внутрішні. Зовнішні фактори діляться в свою чергу на дві групи:

1. *Фактори прямої дії*;
2. *Фактори опосередкованої дії*.

Фактори прямої дії безпосередньо впливають на рівень ризику, до них відносяться:

- правова база – це законодавство, що регламентує підприємницьку діяльність;
- непередбачені дії органів державної влади і місцевого самоврядування;
- податкові органи;
- рівень підприємництва, поведінка партнерів;
- конкуренція підприємств;
- соціально – політична ситуація, корупція і рекет

Фактори опосередкованої дії звичайно не впливають на рівень господарського ризику настільки помітно, як фактори прямої дії. Основні із них:

1) стан економіки;
2) рівень інфляції в країні;
3) політичний стан.

Неменш багаточисельні і внутрішні фактори, що впливають на господарський ризик. виділяють три групи таких факторів:

1) філософія фірми;
2) принцип діяльності фірми;
3) ресурси і їх використання.

*Об'єктивна невизначеність*, коли є часткова або відсутня інформація про внутрішні фактори суб'єкта господарювання, що утворюють ризикову ситуацію.

*Суб'єктивна невизначеність*, коли є часткова або відсутня інформація про зовнішні фактори суб'єкта господарювання, що утворюють ризикову ситуацію.

Аналіз і ефективність управління ризиком суттєво залежить від класифікації ризиків. Складність класифікації ризиків в їх багатоаспектності.. Найбільш практичний підхід до визначення об'єктів класифікації ризиків можна здійснити за їх структурою:

$$C \Rightarrow A \Rightarrow B \qquad (11.1)$$

де: *С* – фактори ризику, *А* - причини і джерела ризикових подій, *В*-ризикові події (наслідки). Таким чином, класифікація ризиків випадкових подій повинна здійснюватися за причинами (факторами), джерелами і результатами їх наслідків.

За масштабом джерел виникнення ризиків. ризики умовно можна розглядати на чотирьох рівнях: на рівні країни - *країнні* ризики, на рівні регіону- *регіональні* ризики, на рівні галузі - *галузеві* ризики, на рівні підприємства, фірми, організації - *мікроризики***.** Країнні, регіональні, галузеві складають одне пог\няття – *макроризики.*

Класифікація ризиків починається з поділу їх за результатами наслідків на *чисті* **(**прості, статичні, об'єктивні ) і *спекулятивні* (динамічні, суб'єктивні).Чисті ризики – це ризики з негативним або нульовим результатом, спекулятивні – з негативним, нульовим або позитивним результатом. Спекулятивні ризики називають просто *фінансовими* ризиками,

Сукупність подій з можливими наслідками (результатами) для підприємства, фірми, організації називають їх *діяльністю*. Це можуть бути події, джерелом виникнення яких є само підприємство, так і зовнішні події відносно підприємства з можливими наслідками для нього. Так як в основі поняття ризику лежить можливість випадкової ризикової події, то класифікацію ризиків для підприємства доцільно здійснювати за видами їх діяльності . Пропонується загальна схема ризиків за видами діяльності (Рис.11.1)

Стосовно схеми ризиків діяльності підприємства, можна дати їх детальну класифікацію і характеристики. Наприклад, *виробничий ризик* - це ризик повязаний із виробництвом продукції, товарів і послуг, із здійсненням любих видів виробничої діяльності. За наслідками виробничий ризик може бути як спекулятивним так і чистим.

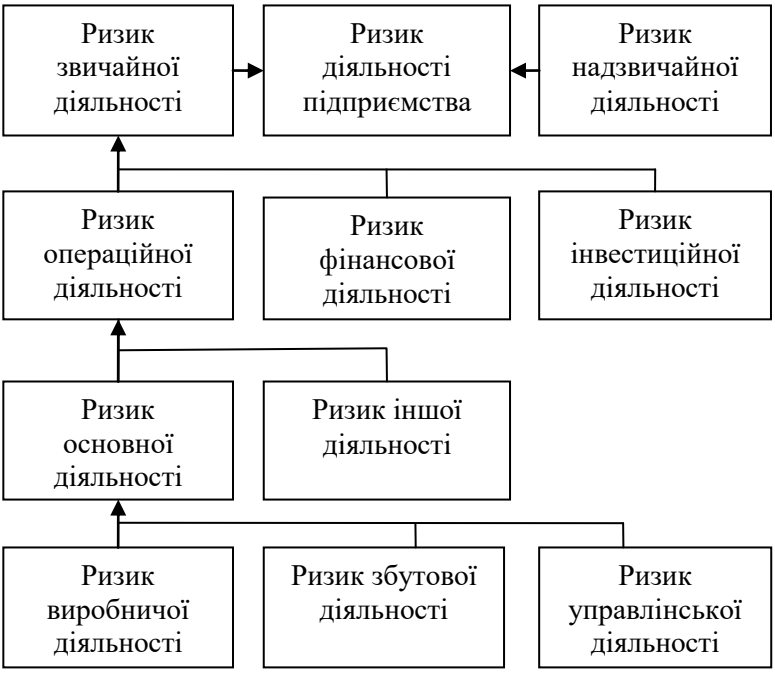



*Рисунок 11.1*. **Схема ризиків діяльності підприємства**

*Спекулятивні (фінансові) виробничі ризики* виникають, коли підприємці стикаються з проблемами використання матеріальних (наприклад, сировини), і виробничих ресурсів (наприклад, як результат зміни робочого часу може привести до його втрат (негативний результат)), зміни собівартості, використання нових методів виробництва, зміни ходу економічної діяльності щодо запланованих основних економічних показників, що визначається економічними прорахунками менеджерів підприємства. або до його скорочення (позитивний результат). *Чистий виробничий ризик* - це ризик пов'язаний зі збитками від зупинки виробництва внаслідок

загибелі, пошкодження основних і оборотних фондів (обладнання, сировини, транспорту).

*Ризики збутової діяльності* підприємства (збутові ризики) виникають під час збуту продукції (послуг). *Спекулятивні збутові ризики* виникають по причині: зниження об'ємів реалізації внаслідок зміни кон'юктури або інших обставин, підвищення закупівельних цін товарів, які не блоковано умовами договору, зростання витрат обігу. *Чисті збутові ризики* – це ризики повязані зі збитками по причині затримки платежів, відмови від платежів в період транспортування товару, непостачання товару.

Можна розглядати більш широкий клас ризиків ніж збутові ризики – це маркетингові ризики, які теж підлягають класифікації.

Детальну характеристику ризиків по групах можна знайти наприклад в [8].

Якщо стати на суб'єктивну точку зору ризику(тобто на ризикових подіях суб'єктивного характеру), то можна дати поняття *ділового ризику* (*господарського*)– як категорії що визначає здійснення діяльності в умовах невизначеності; при якій існує ймовірність відхилення від досягнення постановленої цілі. Тобто це ризик визначений за ризиковою подією (наслідком) - відхилення від досягнення постановленої цілі Якщо такою ціллю є прибуток, то такий діловий ризик назвемо *підприємницьким*. Джерелом виникнення підприємницьких ризиків є підприємницька діяльність. Тому класифікація цих ризиків здійснюється за видами підприємницької діяльності – *виробничої, комерційної, фінансової, посередницької, страхової*.

Якщо видом підприємницької діяльності є страхування, воно полягає в тому, що підприємець за певну плату гарантує споживачу (страхувальнику) компенсацію можливих втрат майна, то природно в цьому випадку ризик назвати *страховим***.** Для страхової компанії страховий ризик – це ризик появи непередбачуваної умовами страхування події, в результаті якої страхова компанія повинна

виплатити застрахованому страхову суму. Наслідком цієї події є збитки, що зумовлені неефективною страховою діяльністю, як на етапі укладання договору страхування, так і на етапі – перестрахування, формування страхових резервів. Звичайно, існують ризики страхування і для особи (фізичної чи юридичної), що страхується, це вже ризик втраченої вигоди, тобто невдалого проведення страхування.

По відношенню суб'єкта господарської діяльності, як джерела ризиків, ризики можна поділити на *зовнішні і внутрішні*. Джерелом зовнішніх ризиків є зовнішнє середовище по відношенню фірми. Підприємство (підприємець) на довкілля впливати не може, тільки прогнозувати і враховувати. Це ризики не пов'язані з діяльністю підприємства.

Внутрішні ризики пов'язані з діяльністю підприємства, їх джерелом є сама фірма. Підприємницькі ризики, що входять до внутрішніх ризиків складають групу *систематичних* ризиків. Це ризики неефективного менеджменту, помилкової маркетингової політики. Серед зовнішніх ризиків важливе місце посідає група *несистематичних* ризиків їх ще називають *ринковими* або *бетта ризиками* це – валютні, процентні та кредитні ризики разом взяті.

Несистематичні ризики виникають у випадку любого виду підприємницької діяльності. Дана група ризиків є тим максимальним набором ризиків, перевищення яких означає зниження ефективності суб'єкта господарської діяльності.

Взаємозв'язок систематичного й несистематичного ризиків можна представити графічно (рис. 11.2).

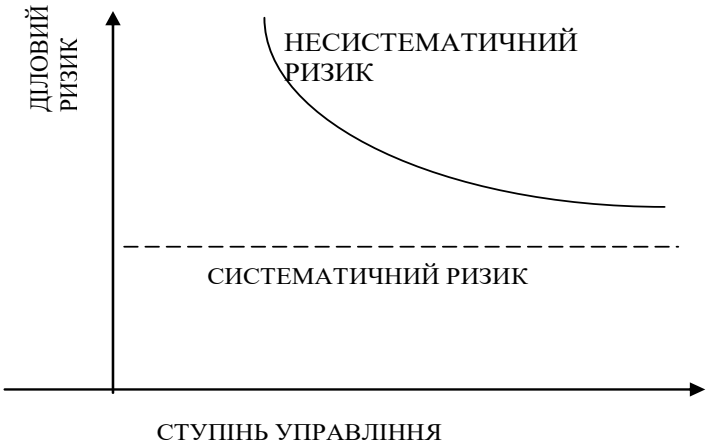



*Рисунок 11.2.* **Взаємозв'язок систематичного й несистематичного ризиків**

На графіку можна помітити, що по мірі здійснення функцій управління, система управління в змозі зменшити несистематичний ризик**.**

### 11.3. Поняття аналізу ризику

Призначення аналізу ризику – одержання необхідних даних як для себе так і для потенційних партнерів для прийняття рішень стосовно участі в певній діяльності (проекті) з ціллю передбачити способи захисту від можливих негативних наслідків. Коли говорять про необхідність врахування ризику в певному виді діяльності (проекті) мають на увазі інтереси суб'єктів ризику, які беруть участь в цій діяльності (проекті): замовник, інвестор, підрядник або продавець, покупець, а також страхова компанія.

Рівень ризику – це величина: 1) потенційних вірогідних втрат, шкоди, збитків; 2) потенційного рівня доходу, прибутку . Рівень ризику є випадковою величиною.

Зоною ризику називають діапазон загальних втрат (доходу, прибутку, їх імовірності), в границях якого втрати (доход, прибуток, їх імовірність) знаходяться в межах встановленого рівня ризику. Щодо певного виду підприємницької діяльності, то під зоною ризику розуміють діапазон втрат (доходу, прибутку, їх імовірності), в межах якого даний вид підприємницької діяльності ще можливий.

Найпростішу схему аналізу ризику, або створення його моделі, можна подати в такому вигляді (рис. 11.3).

Дамо пояснення до запропонованої схеми, аналізу (моделі )ризиків. Аналіз (модель) ризиків можна розділити на два взаємодоповнюючих види – якісний і кількісний. Завдання якісного аналізу – це ідентифікація ризиків, виявлення і аналіз факторів, що змінюють вид ризику. Ідентифікація вважає встановлення переліку основних видів господарських ризиків, , подальший розподіл ризиків на систематичні і несистематичні, в завершенні – формування загального портфеля господарських ризиків.

Приклад якісного аналізу проектного ризику в будівництві. Як відомо, період, за який реалізується поставлена мета, називають життєвим циклом проекту. Існує традиційна послідовність етапів проектного циклу: ідентифікація, розробка експертизи, фінансове забезпечення, реалізація проекту, експлуатація. Згідно цієї схеми і здійснюється аналіз, в тому числі якісний, ризику проєкту.

Ідентифікувати ризики потрібно на стадії проектування будівництва, експлуатації, ліквідності. На стадії будівництва можуть виникнути:

- ризик неправильного вибору підрядчика,
- ризик не поставки устаткування,
- ризик нестачі фінансових ресурсів.


*Рисунок 11.3.***Схема аналізу ризику**

Самострахування такого ризику здійснювалось шляхом одержання інформації про якість виконання робіт, введення об'єктів в експлуатацію в строк підрядчиком. Для вибраного БМУ "Промбуд" перераховані параметри виявились найкращими.

Ризик непоставки устаткування страхується відповідними санкціями, що передбачаються в договорах з постачальниками. Ризик нестачі фінансових ресурсів мінімізується можливістю одержання кредитів в банку.

На стадії експлуатації можна виділити такі ризики:

- ризик невиходу на експлуатаційну потужність у встановлені строки;

- ризик збуту продукції.

Перший ризик обмежений простою технологією виробництва. Другий ризик страхується високими потребами на продукцію в регіоні, збільшенням увізного мита, ставок акцизного мита, ставок акцизного збору, ПДВ на аналогічну імпортну продукцію.

Ризик ліквідності на стадії будівництва зумовлений використанням типових проектів на будівлі та споруди. Після завершення будівництва, до монтажу устаткування, ризик ліквідності може бути усунутий високою ліквідністю проекту через можливість пристосування будівель і споруд на переорієнтацію випуску інших видів продукції. Після монтажу устаткування ризик ліквідності може проявитися у специфічності устаткування, що не уможливлює переорієнтації на інше виробництво.

Завдання кількісного аналізу – оцінка ступеня імовірності окремого виду господарських ризиків, визначення можливих фінансових втрат по певному виду ризику.

В кількісному аналізі ризику використовуються наступні методи:

- ймовірнісно– статистичний,
- метод зон,
- метод експертних оцінок,
- аналітичний,
- аналогій.

### 11.4. Ймовірнісно-статистична модель аналізу ризику

На практиці використовуються наступні імовірнісно-статистичні методи аналізу ризиків:

- оцінка ймовірності.
- аналіз ймовірності за законами розподілу
- імітаційне моделювання ризиків.

Статистичний метод аналізу ризиків застосовується у випадку, коли суб'єкт ризику володіє достатньою базою аналітичної і статистичної інформації. Перевагою статистичного методу є те що він дозволяє здійснювати аналіз і оцінку різних варіантів розвитку подій і враховувати різні фактори ризику.

*Метод оцінки* ймовірності дозволяє дати спрощену статистичну оцінку ймовірності рішення шляхом розрахунків долі виконаних і невиконаних рішень в загальній сумі прийнятих рішень. Застосування відповідних інструментів теорії ймовірності до виміру ризику залежить від характеру можливості випадкових подій. Такими подіями є події з несприятливими (небажаними) наслідками, тобто збитки, шкода, витрати, перевитрати і т. д.

### 11.4.1. Оцінка ризику за математичним сподіванням (середньою величиною)

Приклад оцінки ризику за середньою (сподіваною) величиною. Вкладник помістив на депозит банку суму $x_1$ грн. $p$ - ймовірність втрати вкладу його банкрутство, внаслідок конкуренції банків. Потрібно знайти величину втрат для вкладника банку внаслідок його банкрутства.

Кількісному аналізу ризику повинен передувати якісний його аналіз. За блок-схемою якісного аналізу: об'єкт ризику – депозит, суб'єкт ризику – вкладник банку, джерело ризику – банк, причина ризику - конкуренція банків. Для вкладника банку його банкрутство з ймовірністю $p$ може мати два випадкові наслідки (дві елементарні випадкові події): втрата депозиту (неповернення вкладу), небажаний наслідок і збереження депозиту (вкладу), бажаний наслідок. Вважається, що втрата депозиту, так і його збереження відбувається внаслідок банкрутства банку, тобто з ймовірністю $p$. Якщо ризикова подія $X$ – втрата вкладу розміром $x_1$ ум. од., вона відбувається з ймовірністю $p$, як і банкрутство банку, то протилежна ризикова подія

$\overline{X}$ - збереження вкладу розміром $x_2=0$ ум.од.з ймовірністю *1-p*. Таким чином випадкова ризикова подія втрата депозиту має дискретний закон розподілу (Таблиця 11.1), представляє собою модель ризику втрат депозиту.

*Таблиця 11.1.* **Розподіл втрат депозиту**

| X | $x_1$ | $x_2$ |
|---|---|---|
| P | p | 1-p |

Кількісна модель ризику – це математичне сподівання випадкової величини *X*, яке знаходиться за формулою:

$$risk = I\ X = px_1 + (1-p)x_2 = px_1 \qquad (11.2)$$

Таким чином, *величина ризику можливої випадкової події*– це математичне сподівання її небажаного наслідку. Якщо результат наслідку має дискретний закон розподілу, то величина ризику представляє собою добуток міри небажаної (небажаного наслідку) ризикової події на ймовірність її (його) появи.

Аналогічно, якщо ризикова подія *Y* – збереження вкладу розміром $y_1$ ум. од., з ймовірністю, (очевидно) $q = 1-p$, то протилежна ризикова подія $\overline{Y}$ - збереження вкладу розміром $y_2=0$ ум. од. з ймовірністю *1-p(*з ймовірністю збереження банку*)*. Таким чином, випадкова величина збереження депозиту має дискретний закон розподілу:

*Таблиця 11.2***. Розподіл збереження депозиту**

| Y | $y_1$ | $y_2$ |
|---|---|---|
| Q | q | 1-q |

Математичне сподівання випадкової величини *Y*:

$$risk = MY = qy_1 + (1-q)\,y_2 = qy_1 \qquad (11.3)$$

Математичне сподівання *MY* називають ціною ризику випадкової величини *Y*. *Ціна ризику можливої випадкової події* – це математичне сподівання її бажаного (сприятливого) наслідку.

Ризикова ситуація може мати ряд наслідків, серед яких міри небажаних наслідків позначимо $x_j$ ($i$ = 1, . . . ,n), відповідні ймовірності їх появи $p_j$. В цьому випадку величина ризику в абсолютному вимірі *є* ***су****мою математичних сподівань небажаних наслідків (сподіваний програш).*

$$risk = \sum_{j=1}^{n} MX_j = \sum_{j=1}^{n} x_j p_j \qquad (11.4)$$

Коли випадкова величина *Y* (наприклад розміри депозитів в ряді банків) має ряд наслідків, $y_j$ – міра *j*- того бажаного наслідку (депозиту), а $q_j$ – його імовірність збереження, що співпадає з ймовірністю збереження банку), то ціна ризику (сподіваний виграш) в абсолютному виразі обчислюється за формулою:

$$risk = \sum_{j=1}^{n} MY_j = \sum_{j=1}^{n} y_j q_{j\,j} = \sum_{j=1}^{n} y_j (1 - p_j) \qquad (11.5)$$

Слід мати на увазі, формули (11.2) - (11.5) визначають середні (сподівані)втрати і виграш із ряду можливих випадків.

Як приклад, якщо помістити вклад 1000 грн. в комерційний банк, ймовірність банкрутства якого 0,05, то ризик втрати депозиту становитиме за формулою (11.1):

risk = 1000×0.05 + 0×0,95 = 1000×0.05 =50 грн,

тобто середні втрати депозиту (із двох можливих наслідків) становлять 50 грн. За формулою (6.2) ціна ризику crisk = 1000×0.95 + 0×0.05 = 1000×0.95 =950грн. В середньому збереження депозиту (середній виграш, або сподіваний виграш), внаслідок збереження банку, складатиме 950 грн.

Нехай у формулі (11.4) $X_j$ – запланований рівень підприємницького доходу (рівень вкладених коштів), $p_j$ – імовірність

втрати *j*- того підприємницького доходу (вкладених коштів). Якщо $X_j$ – виступає як міра небажаного наслідку, тоді $q_j = 1- p_j$ – імовірність збереження підприємницького доходу $X_j$, тому формула (11.3) матиме наступний вигляд:

$$risk = \sum_{j=1}^{n} x_j p_j = \sum_{j=1}^{n} x_j (1-q_j) = \sum_{j=1}^{n} x_j - \sum_{j=1}^{n} x_j q_j \qquad (11.6)$$

Формула(14.6) визначає, що величина ризику підприємницького доходу є різниця між запланованим доходом і ціною ризику. В свою чергу ціна ризику підприємницького доходу є різниця між величиною запланованого доходу і величиною ризику:

$$crisk = \sum_{j=1}^{n} x_j q_j = \sum_{j=1}^{n} x_j - risk \qquad (11.7)$$

Якщо, крім того підприємець має змогу одержати доход зверх запланованого $x_j$ з імовірністю $s_j$, то ціна ризику (виграш) складатиметься із запланованого доходу (тобто доходу без ризику) мінус величина ризику і плюс очікуваний дохід зверх запланованого, тобто:

$$risk = \sum_{j=1}^{n} x_j q_j = \sum_{j=1}^{n} x_j - risk + \sum_{j=1}^{n} x_j s_j \qquad (11.8)$$

Величину $\sum x_j s_j$ називають премією за ризик.

*Задача 11.1*. Не американська компанія продала в кредит товар на суму 3 млн. доларів. На дату представлення кредиту курс національної валюти становив 2 долари, на дату повернення кредиту – 2.1 долара. Знайти величину ризику, ціну ризику від проведеної угоди, вказати його вид, джерело, об'єкт і суб'єкт ризику.

Спочатку здійснимо ідентифікацію ризиків згідно класифікаційної схеми, (рис. 11.1) ризикової діяльності компанії. Компанія обтяжена двома видами ризиків іншої операційної діяльності – валютним і кредитним. Кредитний ризик повязаний із неповерненням боргу покупцем, який і є джерелом ризику для компанії. Обєктом кредитного ризику є кредитна сума в розмірі 3 млн.дол. Субєктом кредитного ризику є компанвя, шо видала кредит. Величина кредитного ризику в абсолютному виразі становить 3 млн.дол.

Джерелом валютного ризику є коливання обмінного курсу національної валюти, обєктом валютного ризику є виділена сума кредиту, субєкт ризику – компанія, що видала кредит. Оскільки валютний ризик повязаний з надходженням коштів при проведенні торговельної операції, то стосовно класифікації валютних ризиків маємо операційний вид валютного ризику.

Знайдемо величину і ціну операційного валютного ризику. Розглянемо декілька підходів до його знаходження.

На основі класичного визначення імовірності.

Імовірність виграшу від зміни курсу національної валюти становить:

$$q = 2 / 2{,}1 = 0{,}952381;$$

ціна ризику (виграш) за формулою 11.1 становить:

$$3 \times 0{,}952381 = 2{,}857143 \text{ дол.}$$

Імовірність втрат від зміни курсу національної валюти становить:

$$p = 1 - q = 0{,}047619;$$

величина ризику: $3 \times 0{,}047619 = 0{,}142857$ млн. дол.

На основі визначення прямих можливих фінансових втрат (збитків) від проведеної операції, що є характерним для будь-якого об’єкта.

В момент видачі кредиту його вартість в національній валюті складала $3 \times 2 = 6$ млн. одиниць. В момент повернення кредиту, за новим курсом національної валюти, його вартість зросла до:

$$3 \times 2{,}1 = 6{,}3 \text{ млн. одиниць.}$$

Можливі збитки, тобто величина валютного ризику в національній валюті становитимуть 6,3 – 6 = 0,3 млн. дол. В доларах валютний ризик складатиме 0,3 / 2,1 = 0,142857 млн. дол.

Загальна величина ризику (загальний ризик) від проведеної угоди – це сума кредитного і валютного ризиків.

$$3 + 0{,}142857 = 3{,}142857 \text{ млн. дол.}$$

Як підсумок, результати розв'язання задачі 11.1 говорять про те, що коли об'єктом ризику є актив суб'єкта господарювання, то величиною ризику небажаної для його активу події (небажаного наслідку) є добуток величини активу на ймовірність появи небажаної (небажаного) для нього події(наслідку). Так же само, величиною ризику (або ціною ризику) бажаної для активу події (бажаного наслідку) є добуток величини активу на ймовірність бажаної (бажаного) для нього події (наслідку). Алгебраїчна сума величини ризику активу і його ціни ризику збігається з величиною активу.

### 14.4.2. Оцінка ризику стосовно обумовленого результату

Середня величина (математичне сподівання) представляє собою узагальнену кількісну характеристику і не дозволяє прийняти рішення стосовно обумовленого результату, середнього (сподіваного) результату, або на користь якого-небудь варіанту, коли ризик розглядати з точки зору ситуації вибору. Для остаточного прийняття рішення (наприклад, згідно основного правила стратегії ризик-менеджменту) необхідно обчислити максимальну коливаність показників, тобто визначити міру коливаності можливого результату, що представляє собою ступінь відхилення очікуваного значення від обумовленої величини. Для цього застосовується такий статистичний інструмент як середнє квадратичне відхилення. Дію цього інструменту розглядається в наступних задачах.

*Задача 11.2.* Бажання влаштуватись на роботу фірми супроводжується мати дані про розмір заробітних плат співробітників на фірмі. Дані розмірів заробітних плат співробітників фірми приведені в таблиці 11.3.

*Таблиця 11.3.* **Розмір заробітних плат**

| Співробітник | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Зарплата | 1200 | 1100 | 1300 | 1400 | 1500 | 1600 |
| Співробітник | 7 | 8 | 9 | 10 | 11 | 12 |
| Зарплата | 1550 | 1350 | 1450 | 1420 | 1520 | 1700 |

Потрібно визначити середню коливаність (зону ризику) заробітної плати по фірмі.

Об'єкт ризику – заробітна плата. Суб'єкт ризику – співробітник. Джерело ризику – фірма. Структура даних – просторова.

Влаштування на роботу зумовлює бажання мати інформацію про найбільш очікуваний розмір, яким є середня заробітна плата ( в даному прикладі вона рівна $\overline{X}$ =1424,16, відхилення її розміру від розміру кожного працівника фірми, тобто знання відхилень $x_j - \overline{x}$, а точніше їх середнього значення $\frac{\sum_{j=1}^{n}(x_j - \overline{x})}{n}$ Така міра, як середня сума відхилень значень ознаки від середньої арифметичної, не може бути по тій причині, що сума відхилень завжди рівна нулю. Як правило, влаштування на роботу на фірму зумовлює бажання мати інформацію про найбільш очікуваний розмір, яким є середня заробітна плата ( в даному прикладі вона рівна $\overline{X}$ =1424,16), і зокрема, відхилення її розміру від розміру зарплати кожного працівника фірми, тобто знання відхилень $x_j - \overline{x}$, а точніше їх середнього значення: Така міра, як середня сума відхилень не може

бути по тій причині, що сума відхилень завжди рівна нулю. Можна визначити середнє лінійне відхилення за формулою :

$$\frac{\sum_{j=1}^{n}\left|x_j-\overline{x}\right|}{n}.$$

Цей показник не є обмеженим зверху тому не може слугувати за показник порівняльного аналізу степеня розсіювання значень двох ознак (навколо їх середніх значень, (наприклад, розсіювання заробітних плат навколо середніх двох фірм). Такий показник повинен бути безрозмірним (відносним)і не перевищувати одиниці. Вибір такого показника можна здійснити наступним чином, знайти долю зміни значень ознаки в загальній зміні всіх значень ознаки (в її структурі) за формулою: $\frac{\left|x_i-\overline{x}\right|}{\sum_{i=1}^{n}\left|x_i-\overline{x}\right|}$ (частки відхилення заробітної плати в загальній зміні всіх заробітних плат), а потім усереднити всі ці значення. Але це приводить до константи: $\frac{1}{n}\sum_{i=1}^{n}\frac{\left|x_i-\overline{x}\right|}{\sum_{i=1}^{n}\left|x_i-\overline{x}\right|}=\frac{1}{n}$, яка не може слугувати таким показником. Тому степінь відхилення всіх значень ознаки відносно середнього може бути виражена показником

$$d_s=\frac{\sum_{j=1}^{n}\left|x_j-\overline{x}\right|}{n\overline{\sigma}} \qquad (11.9)$$

Для порівняльного аналізу ступеню розсіювання годиться середнє квадратичне відхилення, (відхилення в квадратичній

метриці), $\sigma = \sqrt{\dfrac{\sum_{j=1}^{n}(x_j - \overline{x})^2}{n}}$. В даному прикладі σ=169,3 дол. Ризик відхилення від середнього розміру заробітної плати по фірмі в середньому становить 169,3 дол. Коливаність заробітної плати від середньої по фірмі знаходиться в межах від $\overline{x} - \sigma$ = 1424,16 - 169,3 = 1254,86 ум. од. до 1424,16+169,3 =1593,46 ум. од. Песимістичний рівень заробітної плати по фірмі в середньому складає 1254,86 дол., оптимістичний –1593,46 дол.

У формулі обчислення середнього квадратичного відхилення вважається не важливим, яке за знаком відхилення ознаки $x_j$ від її середньої арифметичної $\overline{x}$. На практиці у ряді задач відхилення $x_j - \overline{x}$ в більшу і меншу сторону не завжди бажані. Наприклад, якщо ознака відображає прибуток і його значення $x_j$ є меншою від (сподіваного, запланованого) середнього прибутку $x_j < \overline{x}$, то це є ознакою несприятливої ситуації (у випадку вибіркових спостережень небажаної випадкової події). В той же час додатне відхилення вказує на те, що величина прибутку більша за його середнє (сподіване, заплановане) значення і це наприклад, для інвестора є сприятливою ситуацією (у випадку вибіркових спостережень небажаної випадковою подією).

За несприятливої ситуації застосовується *від'ємне семіквадратичне відхилення* за формулою:

$$SV^{-} = \sqrt{\frac{1}{\sum_{j=1}^{k}\alpha_j p_j}\sum_{j=1}^{k}\alpha_j p_j (x_j - \overline{x})^2},$$

$$\text{де } \alpha_j = \begin{cases} 0, & \textit{коли } x_j \geq \overline{x}; \\ 1 & \textit{коли } x_j < \overline{x} \end{cases},$$

$$p_j = \frac{n_j}{n}, \sum_{j=1}^{k} n_j = n \qquad (11.10)$$

Від'ємне семіквадратичне відхилення тепер виражатиме коливність ознаки, або ризик втрати за несприятливої ситуації. В задачі 11.2:

$$p_j = \frac{1}{12},\ \alpha = (1,1,1,1,0,0,0.1,0,1,0,0),$$

$$SV^{-} = 542.81 \text{ ум. од.}$$

Таким чином, ризик зниження заробітної плати відносно середньої (очікуваної), для частини працюючих із зарплатою меншою середньої, (в даному випадку, для половини працюючих на фірмі) становить в середньому 542,81 ум. од.

За сприятливої ситуації, коли очікуване значення ознаки менше її поточних значень, використовують *додатне семіквадратичне відхилення,* яке обчислюється за формулою:

$$SV^{+} = \sqrt{\frac{1}{\sum_{j=1}^{k} \beta_j p_j} \sum_{j=1}^{k} \beta_j p_j (x_j - \bar{x})^2} \qquad (11.11)$$

де $\beta_j = \begin{cases} 0, & коли\ x_j \le \bar{x}; \\ 1 & коли\ x_j > \bar{x} \end{cases}$, $p_j = \frac{n_j}{n}, \sum_{j=1}^{k} n_j = n$.

Додатне семіквадратичне відхилення тепер виражатиме коливність ознаки, або ризик виграшу за сприятливої ситуації. В задачі 11.2, $p_j = \frac{1}{12}$, $\beta = (0,0,0,0,1,1,1.0,1,0,1,1)$., додатне семіквадратичне відхилення становить $SV^{+} = 490.77$. Таким чином, ризик підвищення заробітної плати відносно середньої (очікуваної\, запланованої) в половині працюючих становить: 490, 77 ум. од.

Обчислення абсолютної величини ризику може бути обумовлено не тільки за сподіваним значенням, але за контрактом,

планом. Пропонується в задачі 14.13 управління ризиком обсягу резерву сировини стосовно обумовленого (очікуваного, середнього) результату її поставок.

*Задача 11.3* За контрактом існує домовленість між постачальником і підприємством, що сировину постачають рівними порціями в розмірі 300 тон через кожні 17 діб (тобто 21 раз на рік). За попередні півроку зафіксовані такі відхилення від встановленої величини (17 діб, табл.11.4).

*Таблиця 11.4.* **Кількість діб між поставками сировини**

| Номер поставки | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Кількість днів між поставками | 16 | 17 | 17 | 18 | 18 | 19 | 17 | 18 | 18 | 18 | 47 |

Визначити ризик обсягу резерву сировини (комплектуючих) і управління ним, спираючись на дані відхилень, шо мали місце в минулому (задокументовані) від встановлених термінів постачання.

За класифікаційною схемою ризик обсягу резерву сировини належить до виробничих чистих ризиків. Знаходимо вибіркову дисперсію:

$((16-17)^2+(17-17)^2+(17-17)^2+(18-17)^2+(18-17)^2+(19-17)^2+(17-17)^2+(18-17)^2+(18-17)^2+(18-17)^2+(17-17)^2)/11 = 10 / 11 = 0{,}91.$

Середнє квадратичне відхилення становить: $\sqrt{0{,}91} = 0{,}954$. Максимальна коливаність результату, $17 \pm 0{,}954$, або (16,14; 17,954), тобто період між поставками становить від 16 до 18 днів. Ризикова величина - 18 днів.

Величина резерву, який служить для зменшення ризику, встановлюється шляхом перемноження добового споживання

сировини (300 / 17) на кількість діб (300 / 17)18, визначених як середнє квадратичне відхилення фактичного періоду постачання від обумовленого за контрактом мінус величина (300) сировини за контрактом. (300 / 17)18 - 300 = 17,647 тон. Для уникнення ризику 17 добовий обсяг резерву повинен становити 17,647 тон.

За достатньої бази даних використання середньоквадратичного відхилення недостатньо обґрунтовано. Наступна задача покликана виправити ряд його недоліків. Кількісний аналіз ризику в абсолютному виразі в задачі 11.3 проведено за просторовою структурою даних. Аналіз ризику за часовою структурою даних приведено в задачі 11.4.

*Задача 11.4.* Підприємець планує отримати прибутки за місячними інвестованими коштами. В таблиці 11.5, наведено динаміку прибутків за місяцями протягом року.

*Таблиця 11.5.* **Прибуток інвестора за місяцями**

| Місяць | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Прибуток (грн..) | 1200 | 1100 | 1300 | 1400 | 1500 | 1600 |
| Місяць | 7 | 8 | 9 | 10 | 11 | 12 |
| Прибуток (грн..) | 1550 | 1350 | 1450 | 1420 | 1520 | 1700 |

Потрібно визначити ризик запланованого прибутку і його коливаність (зону ризику) для інвестора.

В даній задачі: об'єктом ризику виступає прибуток, а суб'єктом ризику – інвестор, джерело ризику –об'єкт інвестування, наслідки ризику – відхилення прибутків від запланованого, структура статистичних даних - часова. Можна провести кількісний аналіз ризику за розрахунком середнього розміру прибутку, (він становить 1254,86 ум. од.) і середнього квадратичного відхилення $\sigma$ = 169,3 ум. од. Такий підхід кількісного моделювання ризику за

середнім квадратичним відхиленням $\sigma$ не є виправданим, він повинен враховувати якісний аналіз ризику за схемою: фактори →причини, джерела→наслідки ($C \Rightarrow A \Rightarrow B$). Фактором впливу на формування ризику в даній задачі є час . Час універсальний фактор, він стосується всіх факторів, в якому знаходять прояв фактори, що впливають на результуючий фактор (прибуток) . Таким середньоквадратичним відхиленням, що враховує час, є середнє квадратичне відхилення розрахункових (вирівняних) значень прибутків за лінійною трендовою моделлю, якщо таку модель можна побудувати. Побудована трендова лінійна модель в середовищі SPSS має наступні значення параметрів:

- коефіцієнт детермінації *R*- квадрат 0,54164;
- рівень значимості лінійної моделі *F* – значимість 0,006466913;
- варіація пояснювальної дисперсії: S = 170309,3;
- середнє квадратичне відхилення:

$$-\sigma = \sqrt{\frac{170309.3}{12}} = 119.32 \text{ ум. од.}$$

Дана трендова модель є лінійною, Значимість лінійної моделі визначається за критерієм Фішера, він менший 0,05.

Ризик відхилення від середнього розміру прибутку для інвестора вимірюється за моделлю середнім квадратичним відхиленням σ = 119, 3 ум. од. Коливаність прибутку знаходиться в межах від $\overline{x} - \sigma$ = 1424,16 – 119,3 =1304,86 ум. од. до 1424,16+199,3 =1543,86 ум. од. Песимістичний рівень прибутку1304,86 ум. од, оптимістичний 1543,86 ум. од. Аналогічно можна оцінити ризики і за семі квадратичними відхиленнями. Зниження ризику, якщо його порівняти з ризиком задачі 11.2, за трендовою моделлю складає 169,3-119,3=50 ум. од. Загальне пояснення такого зниження ризику за моделлю пояснюється перевищенням варіації результуючого показника над варіацією його вирівняних значень.

Якщо за критерієм *F* лінійна трендова модель виявиться лінійно не значимою. то потрібно переходити до побудови нелінійної моделі, за якою оцінка ризику за середнім квадратичним відхиленням стає неможливою. Залишається задовольнятись звичайною формулою його розрахунку, що не є точним розрахунком.

За змістом задачі в прикладі 11.3 інвестора цікавить отримати величину ризику в цілому, незважаючи на порядок надходження прибутків в часі. Не можна стверджувати, що середнє квадратичне відхилення 119,3 ум.од., тобто ризик є оптимальним. У випадку просторової структури даних, за змістом задачі, можна запровадити фактор часу і таким чином прийти до задачі 11.3.

Використання середнього квадратичного відхилення можливе при прийнятті рішень на користь якого-небудь результату.

*Задача 11.5.* Приватний детектив стоїть перед дилемою: піти на роботу державним службовцем в поліцію, або лишитись приватним детективом. Коли стане інспектором поліції, то буде одержувати 100 ум. од.на тиждень. Коли посвариться з начальством (імовірність цього 0,5), то буде отримувати лише по безробіттю 50 ум. од. Коли буде займатись приватним зиском, то при успішному розкритті справ (а це відбувається в 8 випадках із 10), то клієнт дає гонорар 90 ум. од.; коли потерпить невдачу, то – 15 ум. од. Потрібно прийняти правильне управлінське рішення по варіантам працевлаштування.

Вихідні дані доцільно занести в наступну таблицю:

*Таблиця 11.6.* **Статистичні дані варіантів працевлаштування**

| Варіант працевлаш-тування | В кращому випадку | | В гіршому випадку | |
|---|---|---|---|---|
| | імовірність | дохід | імовірність | дохід |
| Інспектор в поліції | 0,5 | 100 | 0,5 | 50 |
| Приватний детектив | 0,8 | 90 | 0,2 | 15 |

Очікуваний дохід (ризик) за обома варіантами працевлаштування один і той же: $X_1$ = 100*0.5 + 50*0.5 = 75 ум. од. $X_2$ = 90*0.5 + 15*0.2 = 75 ум. од.

Значення дисперсій:

$$D_1 = 0.5(100 - 75)^2 + 0.5(50 - 75)^2 = 625;\ D_2 = 0{,}8(90 - 75)^2 + +0{,}2(75 - 15)^2 = 900.$$

Середні квадратичні відхилення:

$$\sigma_1 = \sqrt{D_1} = \sqrt{625} = 25\text{ум.од.};$$
$$\sigma_2 = \sqrt{D_2} = \sqrt{900} = 30\text{ ум. од.}$$

Так як $\sigma_1 < \sigma_2$, то менший ризик – варіант працевлаштування інспектором в поліції. Число $\sigma_1 = 25$ означає ступінь коливаності навколо середнього доходу. Діапазон цієї коливаності доходів, $75 \pm 25$ або (50;100). Інтервал (0;25), за визначенням, представляє собою зону ризику. При $\sigma_2 = 30$ маємо, $75 \pm 30$, або (40;105), зона ризику інтервал (0;30). Другий варіант більш ризиковий, так як можливий найменший очікуваний дохід 40 ф. ст., а за першим варіантом – 50 ф.ст. Але другий варіант більш виграшний – 105 ф. ст. проти 100 ф. ст.

В даній задачі вимір ризику здійснювався за дискретним законом розподілу лише для двох значень випадкових величин. Малий об'єм вибіркових даних не позволяє його уточнення за моделлю. Розглянемо вимір ризику для значно більшого числа значень випадкових величин.

*Задача 11.6.* Потрібно прийняти управлінське рішення про вибір найбільш стабільного партнера шляхом проведення статистичного аналізу даних про строки розрахунків зі своїми трьома покупцями для складання договору на наступний період. Оцінити міру ризику продовження контракту з кожним із партнерів. Зробити висновки. Дані про строки розрахунків з постійними покупцями за попередні 10 місяців показано в таблиці 11.7.

*Таблиця 11.7.* **Дані про строки розрахунків за отриману продукцію протягом попередніх 10 місяців.**

| Замовник | Місяць | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| ВАТ «Україна» | 70 | 39 | 58 | 75 | 80 | 120 | 70 | 42 | 50 | 80 |
| ПП Сидоренко | 50 | 63 | 32 | 89 | 61 | 45 | 31 | 51 | 55 | 50 |
| ЗАТ «Полісся» | 60 | 70 | 30 | 10 | 30 | 58 | 70 | 40 | 70 | 60 |

Для розв'язання задачі перейдемо від вихідних даних до таблиці з ймовірностями кожної величини (табл. 11.8).

*Таблиця 11.8.* **Дані про ймовірності строків розрахунків за продукцію протягом 10 місяців**

| Замовник | Місяць | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| ВАТ «Україна» | 0,2 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | - | 0,1 | 0,1 | - |
| ПП Сидоренко | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 |
| ЗАТ «Полісся» | 0,2 | 0,3 | 0,2 | 0,1 | - | 0,1 | - | 0.1 | - | - |

Терміни розрахунку для кожного партнера знаходяться за дискретним розподілом:

$$\bar{x} = \sum_{i=1}^{n} x_i p_i ,$$

де: $x_i$ — фактичні значення показника терміну розрахунків;
$p_i$ — імовірність появи досліджуваного показника.

$$\bar{x_1} = 70 \times 0{,}2 + 39 \times 0{,}1 + 58 \times 0{,}1 + 75 \times 0{,}1 + 80 \times 0{,}2 + 120 \times 0{,}1 + 42 \times 0{,}1 + 50 \times 0{,}1 = 68 \text{ днів};$$

$$\bar{x_2} = 50 \times 0{,}1 + 63 \times 0{,}1 + 32 \times 0{,}1 + 89 \times 0{,}1 + 61 \times 0{,}1 + 45 \times 0{,}1 + 31 \times 0{,}1 + 51 \times 0{,}1 + 55 \times 0{,}1 = 48 \text{ днів};$$

$$\bar{x_3} = 60 \times 0{,}2 + 70 \times 0{,}3 + 30 \times 0{,}2 + 10 \times 0{,}1 + 58 \times 0{,}1 + 40 \times 0{,}1 = 50 \text{ } днів.$$

Тобто перший партнер в середньому розраховувався через 68 днів за отриману продукцію, другий – через 48 дні, а третій – через 50 днів. Визначимо дисперсії $D$ цих показників та їх середні квадратичні відхилення $\sigma$.

$$D_1 = (70-68)^2 0{,}2 + (39-68)^2 0{,}1 + (58-68)^2 0{,}1 + (75-68)^2 0{,}1 + (80-68)^2 0{,}2 + (120-68)^2 0{,}1 + (42-68)^2 0{,}1 + (50-68)^2 0{,}1 = 497{,}3$$

$$D_2 = (50-48)^2 0{,}1 + (63-48)^2 0{,}1 + (32-48)^2 0{,}1 + \\ +(89-48)^2 0{,}1 + \\ +(61-48)^2 0{,}1 + (45-48)^2 0{,}1 + (31-48)^2 0{,}1 + (51-48)^2 0{,}1 \\ + (55-48)^2 = 270{,}4$$

$$D_3 = (60-50)^2 0,2 + (70-50)^2 0,3 + (30-50)^2 0,2 + (10-50)^2 0,1 + \\ (58-50)^2 0,1 + (40-50)^2 0,1 = 400,5$$

$$\sigma_1 = \sqrt{497{,}3} = 22{,}3 \ днів,$$
$$\sigma_2 = \sqrt{270{,}4} = 16{,}4 днів,$$
$$\sigma_3 = \sqrt{400{,}5} = 20{,}0 \ днів.$$

Таким чином, продовження контрактів про постачання продукції знаходиться в зоні допустимого ризику для усіх трьох партнерів. Проте, другий партнер (ПП Сидоренко) розраховувався протягом досліджуваного періоду в середньому швидше (на 48 днів) порівняно з іншими партнерами (ВАТ «Україна» - на 68 день, ЗАТ «Полісся»- на 53 день. Діапазони коливаності для ВАТ «Україна» 68±22 (46,90); ПП Сидоренко 48±17 (31,65); ЗАТ «Полісся» 50±20(30,70). Найбільш оптимістична ситуація розрахунків належить ЗАТ «Полісся» - 90 днів, Найбільш песимістична оцінка для ВАТ «Україна» - 46 днів.

Дану задачу можна узагальнити на випадок інвестиційних проектів. Нехай маємо два інвестиційні проекти *A* і *B* в які можна вкласти кошти. Нехай проект *A* в майбутньому забезпечить середній дохід $X_1$ з відхиленням $\sigma_1$, проект *B* – середній дохід $X_2$ з відхиленням $\sigma_2$, то можливі випадки:

1. $X_1 = X_2$, $\sigma_1 < \sigma_2$ , висновок – вибирають проект А (менш ризиковий);
2. $X_1 > X_2$, $\sigma_1 < \sigma_2$, висновок – вибирають проект А (менш ризиковий);

3. $X_1 > X_2$, $\sigma_1 = \sigma_2$, висновок – вибирають проект $A$ (менш ризиковий);

Наступні два випадки створюють невизначену ситуацію:

*4.* $X_1 > X_2$, $\sigma_1 > \sigma_2$;

*5.* $X_1 = X_2$, $\sigma_1 = \sigma_2$.

Для прийняття рішень у такому випадку потрібні інші статистичні інструменти виміру ризику. За допомогою семіквадратичного відхилення можна дати відповідь стосовно п'ятого пункту. Найменш ризиковим вважається той проект, в якого за несприятливої ситуації семіквадратичне відхилення виявиться найменшим. Практична реалізація п'ятого випадку приведена в наступній задачі.

*Задача 11.7.* Результати спостережень за нормами прибутку портфелів цінних паперів $A$ і $B$ протягом минулих п'яти періодів наведено в таблиці

*Таблиця 11.9.* **Норми прибутків цінних паперів**

| Період | Норма прибутку | |
|---|---|---|
| | $R_a$ | $R_b$ |
| 1 | 5 | 3 |
| 2 | 3 | 5 |
| 3 | 2 | 6 |
| 4 | 3 | 5 |
| 5 | 7 | 1 |

Інвестор має можливість придбати лише один з портфелів. Потрібно вибрати найменш ризиковий портфель.

Найменш ризиковим буде той портфель, у якого найменше семіквадратичне відхилення в разі несприятливої ситуації. Спочатку переконаємось у виконанні умов пятого випадку. Для цього обчислюємо середні норми прибутку по кожному портфелю:

$M(R_a) = 1/5(5+3+2+3+7) = 4$; $M(R_b) = 1/5(3+5+6+5+1) = 4$;

середні квадратичні відхилення їх норм прибутку:

$$\sigma_a = \sqrt{1/5((5-4)^2+(3-4)^2+(2-4)^2+(3-4)^2+(7-4)^2)} = \sqrt{3};$$
$$\sigma_b = \sqrt{1/5((3-4)^2+(5-4)^2+(6-4)^2+(5-4)^2+(1-4)^2)} = \sqrt{3}.$$

Таким чином, не можна надати перевагу жодному портфелю.

Переходимо до визначення семіквадратичного відхилень за несприятливої ситуації. В нашому випадку, для портфеля *A:*

$\alpha_1=0$, $\alpha_2=1$, $\alpha_3=1$, $\alpha_4=1$, $\alpha_5=0$; $p_1 = p_2 = \ldots = p_5 = 1/5$;

$\sum \alpha_j p_j = 3/5$;

$SV^-(A) = \sqrt{(5/3)(1/5)(0(5-4)^2 +1(3-4)^2 +1(2-4)^2 +1(3-4)^2 +0(7-4)^2 )} = \sqrt{2}$;

для портфеля *B*:

$\alpha_1=1$, $\alpha_2=0$, $\alpha_3=0$, $\alpha_4=0$, $\alpha_5=1$; $p_1 = p_2 = \ldots = p_5 = 1/5$;

$\sum \alpha_j p_j = 2/5$;

$S\sigma^-(B) = \sqrt{(5/2)(1/5)(1(3-4)^2 +0(5-4)^2 +0(6-4)^2 +0(5-4)^2 +1(1-4)^2 )} = \sqrt{5}$.

Оскільки, $SV^-(A) < SV+^-(B)$, то для інвестора більш привабливий портфель *A*, оскільки він менш ризиковий.

### 11.4.3. Вимір ризику у відносному виразі

Невизначена ризикова ситуація 4, ($X_1 > X_2$, $\sigma_1 > \sigma_2$)), коли середні доходи і середні квадратичні відхилення в обох інвестиційних проектів мають однаковий сенс, вимагає порівняльного аналізу ефективності ризиків за схемою результат / затрати. В загальному випадку ризик у відносному виразі, за геометричним способом визначення імовірності, обчислюється за формулою:

$$k = \frac{X}{K} \qquad (11.12)$$

де: *X*- міра небажаного наслідку (результат) ризикової ситуації,

*K* – база (затрати), відносно якої розглядається небажаний наслідок. Величини *X* і *K* випадкові величини.

Якщо у формулі (11.12) в якості бази береться об'єм фінансових ресурсів, то маємо вимір фінансового ризику у відносному виразі.

В якості бази для підприємця береться:

1) його майно;

2) загальні затрати на ресурси за даним видом підприємницької діяльності;

3) очікуваний дохід (прибуток) від даного підприємницької діяльності.

Стосовно підприємства за базу береться:

1) вартість основних фондів і оборотних коштів;

2) заплановані сумарні затрати за даним видом діяльності, мається на увазі як поточні затрати так і капітальні вкладення;

3) розрахунковий прибуток.

Якщо у формулі (11.12) в якості міри небажаного наслідку взяти максимально можливий обсяг збитків підприємства, а за базу – всі його активи, то формула (11.12) визначає ризик банкрутства підприємства.

Якщо у формулі (11.12) в якості міри небажаного наслідку взяти залучені кошти, а за базу – власні кошти, то отримаємо ризик фінансової стійкості підприємства.

Якщо у формулі (11.12) в якості міри небажаного наслідку взяти величину оборотних коштів, а за базу короткострокові зобов'язання підприємства, то отримаємо величину ризику неплатоспроможності підприємства.

Якщо у формулі (11.12) віднести величину ризику (ціну ризику) наслідку в абсолютному вимірі суб'єкта господарської

діяльності до його активу, то отримаємо ступінь, або коефіцієнт ризику програшу (виграшу).

Якщо у формулі (11.12) небажаного наслідку взяти середнє квадратичне відхиленням *σ(X)* випадкової величини (ризик), а в якості бази її математичне сподівання (сподіване, середнє значення), то отримаємо один із статистичних інструментів виміру ризику у відносному виразі - коефіцієнт варіації *V(X)* випадкової величини *X*, який за заданим середнім квадратичним відхиленням *σ(X)* і математичним сподіванням *M(X)* визначається за наступною формулою:

$$V(X) = \frac{\sigma(X)}{M(X)} \qquad (11.13)$$

Зміст коефіцієнта варіації: він визначає середнє квадратичне відхилення випадкової величини (неважливо в бажану чи небажану сторону з одинаковою імовірністю) на одиницю її середнього (математичного сподівання) значення, тобто визначає ризик одиниці очікуваного результату. Наприклад, коли випадкова величина *X* – дохід від інвестування, то *σ(X)* визначатиме втрати (або премію за ризик) від однієї гривні очікуваного(середнього ) доходу інвестицій.

В залачі 11.6 визначимо коефіцієнт ризику для кожного партнера за формулою (14.13):

$$V_1 = \frac{22.3}{68} * 100\% = 32.8\%;$$

$$V_2 = \frac{16.4}{48} * 100\% = 34.1\%;$$

$$V_3 = \frac{20,0}{50} * 100\% = 40,0\%.$$

Ризикова ситуація 4, за схемою ($X_1 > X_2$, $\sigma_1 > \sigma_2$) присутня для всіх трьох пар партнерів, тому їх порівняльний аналіз можна здійснити за коефіцієнтом ризику. Величина ризику для ЗАТ

«Полісся» є найбільшою і становить 40%, тобто має найбільше значення в зоні допустимого ризику. Менш ризикованими є розрахунки з першим партнером (ВАТ «Україна»), де коефіцієнт ризику становить 32,8%. Найбільш стабільними та найменш ризикованими є взаємовідносини з другим партнером (ПП Сидоренко), для якого коефіцієнт ризику становить 34,1%.

Невизначений випадок 5 ($X_1 = X_2$, $\sigma_1 = \sigma_2$), коли однакові середні результати і ризики в абсолютному виразі, може бути реалізований з допомогою коефіцієнта семіваріації: той проект вважається менш ризиковим, в якого коефіцієнт семіваріації менший.

$$V^-(X) = \frac{SV^-(X)}{M(X)} \qquad V^+(X) = \frac{SV^+(X)}{M(X)} \tag{14.14}$$

Цей випадок також може бути реалізований і на суб'єктивному відношенні особи, що приймає рішення, до ризику. Вибір при цьому визначається тим, якою величиною додаткового середнього прибутку компенсується для особи збільшення ризику. Суб'єктивне відношення до ризику враховується в теорії Неймана-Моргенштерна.

### 14.4.4. Комплексний аналіз ризику

Комплексний аналіз ризику передбачає поєднання аналізу ризику в абсолютному і відносному виразі. Пропонується розгляд відповідних прикладів.

*Задача 11.8.* Для підприємства, що займається виробництвом меблів, знайти ризик збільшення собівартості стола для підприємства в зв'язку з можливим збільшенням цін на матеріали і деталі, потрібних для його виготовлення. Вихідні дані наведено в таблиці 11.10.

*Таблиця 11.10.* **Прямі матеріальні витрати на виготовлення одного стола**

| Артикул | Питомі витрати матеріалів | Ціна матеріалів(грн..) | Прогноз збільшення ціни |
|---|---|---|---|
| Дерево | 0,1 м$^3$ | 200 м$^3$ | 1% |
| Каркас | 6 м | 2.9 м | 3% |
| Ніжки | 4 м | 2,0 м | 2% |
| Клей | 500 г. | 2,0 кг | 23% |
| Фарба | 500 г. | 2б, кг | 32% |

Непрямі витрати підприємства наведено в таблиці 11.11.

Таблиця 11.11. Непрямі витрати підприємства

| Статті витрат | Сума, грн | Прогнозовані витрати |
|---|---|---|
| Утримання виробничого приміщення | 2152 | 1% |
| Поточний ремонт виробничого приміщення | 3560 | 3% |
| Витрати на охорону праці та техніку безпеки | 800 | 2% |
| Малоцінні та швидкозношувані предмети | 1100 | 4% |

Собівартість реалізованої продукції (робіт, послуг) складається з:

1) виробничої собівартості (робіт, послуг) продукції, яка була реалізована протягом аналізованого періоду;

2) нерозподілених постійних загальновиробничих витрат;

3) наднормативних виробничих витрат. До виробничої собівартості продукції (робіт, послуг) включаються:

а) прямі матеріальні витрати; до них відносяться: вартість сировини, основних матеріалів, придбаних напівфабрикатів і комплектуючих виробів, допоміжних і інших матеріалів, палива й енергії, тари і тарних матеріалів.

б) прямі витрати на оплату праці; зарплата і прямі виплати робітникам, зайнятим у виробництві продукції (робіт, послуг) за окладами і тарифами, компенсаційні виплати, оплата відпусток.

в) інші прямі витрати; до їх складу відносять: відрахування на соціальні заходи, амортизацію, плата за оренду земельних і майнових паїв і інші виробничі витрати – витрати від браку, що складають вартість остаточно забракованої продукції (виробів, напівфабрикатів), витрати на виправлення браку за мінусом справедливої вартості остаточно забракованої продукції (виробів, напівфабрикатів), суми відшкодування робітниками, що допустили брак, і суми відшкодування від постачальників за неякісні матеріали і комплектуючі вироби.

г) загальновиробничі витрати.

До складу загальновиробничих витрат включаються:

1. Витрати на управлінням виробництвом (оплата праці апарату управління цехами, дільницями, тощо, відрахування на соціальні заходи й медичне страхування апарату управління цехами, дільницями; витрати на оплату службових відряджень персоналу цехів, дільниць тощо).

2. Амортизація основних засобів загальновиробничого (цехового, дільничного, лінійного) призначення.

3. Амортизація нематеріальних активів загальновиробничого (цехового, дільничного, лінійного) призначення.

4. Витрати на утримання, експлуатацію та ремонт, страхування, операційну оренду засобів, інших необоротних активів загальновиробничого призначення.

5. Витрати на удосконалення технології й організації виробництва (оплата праці та відрахування на соціальні заходи

працівників, зайнятих удосконаленням технологій й організації виробництва, поліпшенням якості продукції, підвищенням її надійності, довговічності, інших експлуатаційних характеристик у виробничому процесі; витрати матеріалів, купівельних комплектуючих виробів і напівфабрикатів, оплата послуг сторонніх організацій).

6. Витрати на опалення, освітлення, водопостачання, водовідведення та інше утримання виробничих приміщень.

7. Витрати на обслуговування виробничого процесу (оплата праці загальновиробничого персоналу, відрахування на соціальні заходи, медичне страхування робітників та апарату управління виробництвом; витрати на здійснення технологічного контролю за виробничими процесами та якістю продукції, робіт, послуг).

8. Витрати на охорону праці, техніку безпеки і охорону довкілля.

9. Інші витрати (втрати від браку, оплата простоїв тощо).

Загальновиробничі витрати поділяються на постійні і змінні, які установлюються підприємством, а також перечень і склад статей калькуляції виробничої собівартості.

Як було зазначено спекулятивний виробничий ризик виникає, коли підприємство стикається з проблемою зміни собівартості продукції. Причина ризику збільшення собівартості стола для підприємства – підвищення цін на матеріали. Наслідки (ризикові події) – це перевитрати на дерево (наслідок 1.); перевитрати на каркас (наслідок 2.); перевитрати на ніжки (наслідок 3.); перевитрати на клей (наслідок 4.); перевитрати на фарбу (наслідок 5.); перевитрати зарплати (наслідок 6.) і збільшення непрямих витрат (наслідок 7.). Як бачимо, ризик збільшення собівартості продукції включає в себе ціновий ризик на вході підприємства плюс непрямі витрати. Таким чином, величина ризику становитиме:

$$risk = 0.1 \times 200 \times 0.01 + 6 \times 2{,}9 \times 0{,}03 + 4 \times 2 \times 0.02 +$$

$$+0.5 \times 2 \times 0.23 + 0.5 \times 2 \times 0.32 + 28 - 27.73 + \\ + \frac{2152 \times 0.01 + 3560 \times 0.03 + 800 \times 0.02 + 1100 \times 0.04 + 300 - 250}{1000} = 1.94 \text{ грн.}$$

Для визначення ступеня ризику знайдемо величину активу, тобто собівартість стола без ризику,

$$aktiv = 0.1 \times 200 + 6 \times 2.9 + 4 \times 2 + 0.5 \times 2 + 0.5 \times 2 + 27.73 + (2152 + 3560 + 800 + 1100) / 1000 + 250 / 1000 = 82.992 \text{ грн.}$$

Ступінь ризику знаходиться за формулою:,

$$P = \frac{risk}{aktiv} = \frac{1{,}94}{82{,}99} = 0.0232 \ (2{,}3\%) \qquad (11.15)$$

Розглянемо застосування деяких теорем теорії ймовірності. В якості прикладу використаємо формули повної імовірності у вимірі ризику в абсолютному виразі.

### 14.4.5. Вимір ризику з використанням оцінки ймовірності

Приклад на застосування формули повної ймовірності у абсолютному вимірі ризику.

Розглянемо приклад використання формули Байєса у відносному вимірі ризику, тобто визначення його ступеня.

*Задача 11.9*. На думку експертів фірми "Зоря", конкурент може піти на випуск нової, сильно конкурентоспроможної продукції, імовірність чого вони оцінили на рівні 70%. Ця імовірність викликає у керівництва фірми "Зоря" тривогу, але не достатню щоб піти на крайні міри відповіді. Прийнято рішення зібрати додаткову інформацію про наміри конкурента.

Експерти фірми "Зоря" вважають, що для випуску нової продукції конкурент з 90% -ю імовірністю піде на розширення своїх виробничих потужностей. Звичайно, він може почати розширяти виробничі площі і за іншими причинами, але імовірність останнього експерти оцінили на рівні всього 20%.

Керівництву фірми "Зоря" стало відомо про початок нового будівництва конкурентом. Потрібно визначити ризик загрози для

керівництва фірми “Зоря” з боку конкурента у зв’язку з випуском нової продукції?

Потрібно перш за все, знайти ступінь ризику розширення виробничих площ конкурента. Домовимось в позначеннях наступних подій:

$A$ – початок нового будівництва (розширення виробничих площ) конкурентом;

гіпотеза $H_1$ – перехід на випуск нової продукції;

гіпотеза $H_2$ – переходу на випуск нової продукції не буде;

$A \mid H_1$ – розширення виробничих площ внаслідок переходу на випуск нової продукції;

$A \mid H_2$ - розширення виробничих площ без переходу на випуск нової продукції;

$H_1|A$ – перехід конкурента на випуск нової продукції в результаті одержання інформації про начало в нього розширення виробничих площ. Визначення імовірності останньої події визначається за формулою Байєса (формулою переоцінки імовірності гіпотези):

$$P(H_1 / A) = \frac{P(\mathrm{A}/ \mathrm{H}_1)P(H_1)}{P(A)} \qquad (11.16)$$

Імовірності $P(A \mid H_1)$ і $P(A \mid H_2)$ задані в умові прикладу – відповідно, 0,9 і 0,2; а також імовірність гіпотез:

$$P(H_1) = 0{,}7,\ P(H_2) = 0{,}3;$$

ймовірність $P(A)$ - початку розширення виробничих площ у конкурента незалежно від того, буде випуск нової продукції, чи ні знаходиться за формулою повної імовірності:

$$P(A) = P(A \mid H_1)\, P(H_1) + P(A|H_2)\, P(H_2) = 0{,}9\times 0{,}7 + 0{,}2\times 0{,}3 = 0{,}69.$$

Після підстановки у формулу (11.16) одержимо:

$$P(H_1|A) = 0,9 \times 0,7 / 0,69 = 0,913.$$

Це саме той рівень імовірності (ступінь ризику для керівництва фірми "Зоря"), коли потрібно приймати рішення про відповідні міри на загрозу конкурента.

Особливий варіант розрахунку ризику пов'язаний з розоренням (банкрутством). В загальному випадку цей ризик породжується мінусовою величиною відхилення економічної випадкової величини $X$ від її сподіваного значення $MX$, або планового значення α ($X - MX < 0$, $X - \alpha < 0$), яке не лишає інвестору (менеджеру) можливостей компенсації відхилення. Ступінь ризику розорення визначається як імовірність відхилення :

$$P(X - MX < 0), P(X - \alpha < 0),$$

ступінь ціни ризику:

$$– 1 - P(X - MX < 0), 1- P(X - \alpha < 0).$$

Для оцінки таких ризиків можуть застосовуватися основні теореми і нерівності теорії імовірності.

1. Застосування леми Маркова. Лема Маркова говорить, що коли випадкова величина X не приймає від'ємних значень, то для любого числа α справедлива нерівність $P(X > \alpha) \leq M(X) / \alpha$. Наприклад, фірма за перший день продала 80 штук деякого товару, за другий день – 120 штук, за третій – 100 штук. Знайдемо ступінь ризику (ймовірність, відсоток продажу), коли фірма має намір продати не менше 110 штук товару за четвертий день.

В середньому щодня фірма продавала товару:

$$M(X) = (80+120+100)/3 = 100 \text{ штук}.$$

За лемою Маркова, $P(X > 110) \leq 105 / 110 = 0.9$.

Таким чином, ймовірність продажу товару не менше 110 штук складає не більше 0,9 (90%), а ймовірність ризику не менше 0,1 (10%). Коли вартість товару складала наприклад 5 дол., то рівень можливих втрат складе не менше 550×0,1 = 55 дол.

2.*Застосування нерівності Чебишева*.

Нерівність Чебишева дозволяє знаходити верхню границю ймовірності того, що випадкова величина $X$ відхиляється в обидві сторони від свого середнього значення $X$ (математичного сподівання $M(X)$) на величину δ: ймовірність того, що випадкова величина відхиляється за модулем від свого математичного сподівання більше, ніж на задану величину $\delta$, не перевищує її дисперсії ($\sigma^2$), поділеної на $\delta^2$,

$$P\left(\left|X-\bar{X}\right|>\delta\right)\le\frac{\sigma^2}{\delta^2} \qquad (11.17)$$

Зауважимо, що дисперсія $\sigma^2$ у формулі (6.14) має бути меншою середнього квадратичного відхилення, оскільки імовірність не перевищує одиниці. В якості $X$ можна брати очікуваний дохід або прибуток**,** ринковий процент, тоді $\bar{X}$ – поточний дохід, прибуток або ефективність (норма прибутку, дохідність). Як правило, нерівність Чебишева слугує для виявлення граничних можливостей (шансів) інвестора, підприємця.

Звернемося до попереднього прикладу. Імовірність реалізувати скажімо 120, 145, і т. п. штук товару за лемою Маркова однакова, не перевищує 0,9. Насправді це не так, реалізувати, як свідчить практика, все більшу кількість товари складніше, тобто ймовірність реалізації товару з ростом обсягів продажу повинна зменшуватись. Для того, щоб в цьому переконатись у формулі (11.17) мають иісце наступні перетворення:

$$P(\left|X-\bar{X}\right|>\delta)=P((X>\bar{X}+\delta)\cup(\mathrm{X}<\bar{X}-\delta))=$$
$$2P(X>\bar{X}+\delta)\le\frac{\sigma^2}{\delta^2}=P(X>\bar{X}+\delta)\le\frac{\sigma^2}{2\delta^2} \qquad (11.18)$$

Очевидно,

$$\sigma^2=((80\text{-}100)^2+(120\text{-}100)^2+(100\text{-}100)^2)/3=266.$$

Тепер визначимося у виборі відхилення. Його вибір диктується змістом оцінки (за формулою 11.18), вона має сенс, коли $\sigma^2 / 2\delta^2$, звідки: $\delta > \sigma/\sqrt{2}$.

В нашому прикладі, $\delta > \sqrt{266}/2 = 11{,}5$, в "штуках", тому: $\delta \ge 12$. У формулі (11.18) послідовно покладемо, $\delta$ = 12; 13; 15, тоді одержимо при $X$ = 100:

$$P(X>112) \le 266 / 2\times144 = 0{,}92;\ P(X>113) \le 266 / 2\times169 = 0{,}78;$$
$$P(X>115) \le 266 / 2\times225 = 0{,}59.$$

Ймовірність ризику становить:

100-0,92=0,08 (8%); 100-0,78=0,22 (22%); 100-0,59=0,41(41%).

Відповідно величина ризику складає:

550×0,08=44 грн.; 550×0,22=121 грн.; 550×0,41=225 грн.

Таким чином, оцінка рівня імовірності ризику за лемою Маркова значно нижча, ніж за нерівністю Чебишева.

Великою заслугою леми Маркова і нерівності Чебишева є те, що вони придатні для довільної кількості спостережень і будь-якому

закону розподілу імовірностей. Платою за відсутністю жорстких обмежень є деяка невизначеність оцінок рівня ймовірності. Наприклад в попередньому прикладі здійснити оцінку ймовірності продажу товару для відхилень $\delta$ = 1, 2, 3, . . . ,11 неможливо. Невизначеність оцінок істотно знижується, коли можна допустити наявність закону нормального розподілу. Як відомо, при чисельності спостережень не менше 30, коли випадкова величина розподілена за нормальним законом, можна використати формулу:

$$P(X - M(X) > \delta) = 1 - F(t), \qquad (11.19)$$

де: $F(t)$ – інтегральна функція Лапласа.

За числа спостережень менше 30, розрахунок можна здійснити за формулою:

$$P(X - M(X)) > \delta) = 1 - S(t), \qquad (11.20)$$

де: $S(t)$ – інтегральна функція Стьюдента.

Значення цих функцій знаходяться в багатьох комп'ютерних программах при $t = \delta / \sigma$, де $\sigma$ – стандартна помилка. При числі спостережень, більшому за 30:

$$\sigma = \sqrt{\sigma_b^2},$$

де: $\sigma_b^2$ – вибіркова дисперсія.

За числа спостережень, меншого 30:

$$\sigma_{\hat{a}} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - M(X))^2}{n-1}}. \qquad (11.21)$$

Зауважимо, що для позитивних відхилень:

$$P(X > M(X) + \delta) = (1 - F(t)) / 2 (= ((1 - S(t)) / 2).$$

Перевірка наявності нормального закону розподілу, особливо за малими вибірками, в економіці часто приводить до негативного результату. Тому краще перевіряти за вибірковими даними наявність закону розподілу близького до нормального. Інструментом такої перевірки можуть слугувати коефіцієнти асиметрії і ексцесу визначені за Блісом, або Айвазяном.

### *Список питань для самоконтролю:*

**1.** Охарактеризуйте структуру і схему якісного аналізу ризику.

**2.** Поясніть сутність класифікації ризиків, зокрема чистих і спекулятивних. Привести приклади.

**3.** Поясніть метод середніх величин в кількісному аналізі ризику.

**4.** Охарактеризуйте особливості застосування імовірнісних методів в статистичному моделюванні ризику.

### *Список рекомендованої літератури по темі:*

1. Балджи М.Д. Економічний ризик та методи його вимірювання. Навчальний посібник. Харьків: Промарт, 2015. 300 с.

2. Вітлінський В. В. Аналіз, моделювання та управління економічним ризиком: навч.-метод. посібник для самостійного вивчення дисципліни / В. В. Вітлінський, П. І. Верченко. К. : КНЕУ, 2000. 292 с.
3. Вітлінський В. В. Аналіз, оцінка і моделювання економічного ризику / В. В.Вітлінський. К. : ДЕМІУР, 1996. 212 с..
4. Вітлінський В. В. Ризикологія в економіці та підприємництві : Монографія / В. В. Вітлінський, Г. І. Великоіваненко. К. : КНЕУ, 2002. 490 с.
5. Донець Л. І. Економічний ризик і методи його вимірювання / Л. І. Донець. – К. : Центр навчальної літератури, 2006. 312 с.
6. Економічний ризик: методи оцінки та управління [Текст] : навч. посібник / [Т. А. Васильєва, С. В. Лєонов, Я. М. Кривич та ін.] ; під заг. ред. д-ра екон. наук, проф. Т. А. Васильєвої, канд. екон. наук Я. М. Кривич. Суми : ДВНЗ "УАБС НБУ", 2015. 208 с
7. Клебанова Т. С. Теория экономического риска : учебн. пособ. / Т. С. Клебанова, Е. В. Раевнева. Х. : Изд. ХГЭУ, 2001. 132 с
8. Устенко О.Л. Теория экономического риска: Монография.О.Л.Устенко. - К.: МАУП.1997. 164 с.
9. Ястремський О. І. Моделювання економічного ризику / О. І. Ястремський. К. : Либідь, 1992. 80с.

# Розділ 12. МЕТОД ГОЛОВНИХ КОМПОНЕНТ

## 12.1. Сутність аналізу головних компонент

Метод головних компонент (МГК) є універсальним методом і допомагає побачити які змінні схожі між собою, а які відрізняються, дозволяє виявити найсуттєвіші та приховані закономірності у великих і складних масивах інформації (навіть у випадках сильно корельованих змінних).

Метод було винайдено К. Пірсоном у 1901 р. для перевірки валідності психометричних шкал та доповнено у 1933 р. Г. Готелінгом. Суть аналізу головних компонент полягає в тому, щоб спроєктувати багатовимірний простір даних на двовимірну площину таким чином, щоб будь-які приховані особливості в наборі даних стали видимими.

Англійський статистик К. Пірсон був початківцем у дослідженні методів зниження розмірності ознакового простору. Наразі метод набув широкого використання в різних сферах людського життя. Тривалий час метод головних компонент розглядався як підвид факторного аналізу, але в останній час все більше дослідників їх відокремлюють. МГК є скоріше геометричним, аніж статистичним, так як реалізується на основі лінійних перетворень простору для знаходження нових змінних. На відміну від МГК, факторний аналіз припускає наявність в сукупності прихованих змінних, які неможливо виміряти, але можна дослідити і описати на основі аналізу взаємозв'язків вихідних змінних.

Метод головних компонент (англ. principal component analysis — PCA ) – статистичний метод, який використовується для аналізу взаємозв'язків між великою кількістю змінних і пояснення цих взаємозв'язків з точки зору меншої кількості змінних (головних компонент), з мінімальною втратою інформації (рис. 12.1). МГК є дуже гнучким інструментом і дозволяє аналізувати набори даних, які можуть містити, наприклад, мультиколінарність, відсутні значення,

категоричні дані та неточні вимірювання. Мета полягає в тому, щоб виявити важливу інформацію з даних і виразити цю інформацію як набір підсумкових індексів - *головних компонент* (principal components (PC))**.**

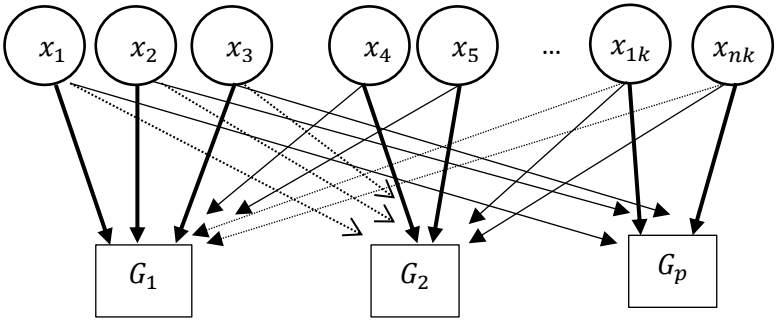



*Рисунок 12.1.* **Модель методу головних компонент**[1]

Основні причини, які спонукають використовувати МГК і знижувати розмірність ознакового простору:

- Ознакова множина настільки велика, що обчислення вимагають більше часу, аніж потрібно для прийняття рішення. Прикладом може слугувати реклама у сфері роздрібної інтернет мережі. Інтернет-магазини в лічені секунди пропонують своїм потенційним клієнтам персоналізовані рекомендовані товари. Якщо ви повернетесь на сайт, то отримаєте ще більше і більш точних пропозицій. В даному випадку, ознак на основі яких можна надати рекомендації неймовірно багато для прийняття правильного рішення. Виникає необхідність скоротити кількість випадкових

[1] Єріна А. М. Статистичне моделювання та прогнозування [Текст]: підручник / А. М. Єріна, Д. Л. Єрін; Держ. ВНЗ "Київ. нац. екон. ун-т ім. Вадима Гетьмана". Київ: КНЕУ, 2014. 348 с.

змінних, виокремити множину головних змінних і при цьому витратити якомога менше часу.

- Розрахунки з використанням великих масивів інформації є занадто енергозатратними та економічно невигідними.

- Інформаційна база практично завжди має зайві, випадкові дані (шум), які значно погіршують якість моделі і результати дослідження. Виключення з ознакової множини шуму методом головних компонент ґрунтується на припущенні, що дисперсія шуму менша відносно дисперсії самих даних. Після лінійного перетворення компоненти, які мають достатньо малі дисперсії будуть вважатись шумом і їх можна виключати з моделі, припускаючи, що це не вплине на результати дослідження.

- Досить часто дані спостережень мають високу розмірність, а в дійсності є функцією всього декількох латентних (прихованих) змінних

Отже, можна виокремити основні задачі використання методу головних компонент:

1. Зменшення кількості змінних. Зниження розмірності вихідного ознакового простору в новий без втрати інформативності.
2. Вимір безмірного. Побудова нових узагальнюючих показників.
3. Візуалізація багатовимірних спостережень (проєкція даних).
4. Виявлення прихованих закономірностей в даних, які не можливо виявити в процесі аналізу окремих змінних.
5. Спрощення моделей, розрахунків та інтерпретації.
6. Збільшення точності аналізу (обернено пропорційної кількості головних компонент).
7. Виявлення структури взаємозв’язків між змінними, зокрема знаходження груп взаємозалежних змінних.

8. Подолання мультиколінеарності змінних у регресійному аналізі (шляхом об'єднання мультиколінеарних змінних в один фактор впливу).

9. Інше.

МГК – це тип лінійного перетворення певного набору даних. Лінійна трансформація вписує початкову сукупність даних в нову систему координат таким чином, що найбільша дисперсія знаходиться на першій координаті, а кожна наступна координата є ортогональною до попередньої і має меншу дисперсію. Цим способом вихідний масив $x$ взаємопов'язаних змінних трансформується у набір $p$ некорельованих основних компонент з ознаками початкової сукупності даних.

Якщо у вихідній сукупності даних багато змінних корелюють між собою, всі вони будуть сприяти формуванню однієї і тієї ж головної компоненти. Кожна головна компонента включає певний відсоток від загального обсягу варіації вихідних даних.

Якщо у досліджуваній сукупності змінні мають тісний взаємозв'язок, то вся множина взаємопов'язаних ознак може бути описана всього декількома основними компонентами. При збільшенні кількості основних компонент буде описуватись все більше і більше вихідних даних. Додавання кожної наступної компоненти буде робити оцінку вихідного набору даних більш точною, але і більш громіздкою.

Перш ніж перейти до формул, розглянемо найпростіший приклад лінійної трансформацій для двох змінних x та y, значення яких нанесено на рис. 12.2. Припустимо, що ми можемо обирати як будуть виглядати координати і розташовувати осі як нам буде зручно. Для того щоб обрати нові осі знайдемо спочатку напрямок максимального обсягу даних.

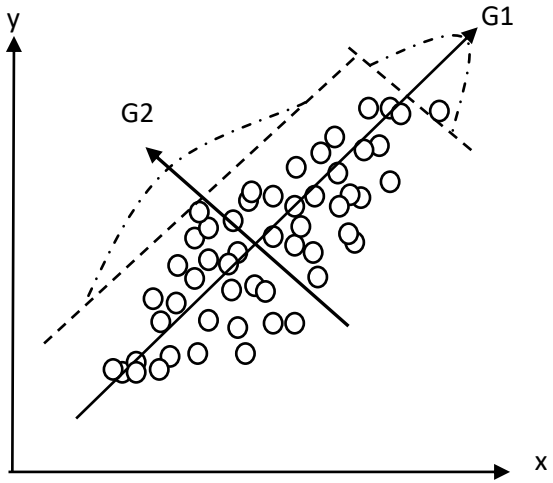



*Рисунок 12.2.* **Діаграма розсіювання для умовних змінних x та y**

За діаграмою розсіювання розташуємо першу вісь у бік цього максимального розкиду, а іншу - перпендикулярно до неї і обертаємо її навколо першої осі, поки варіація уздовж нової осі не буде максимальною. Чому саме перпендикулярно? Тому що, головні компоненти не повинні корелювати між собою і графічно це можна відобразити як перпендикуляр.

Якщо простір багатовимірний, то цю процедуру проводять повторно в бік наступного найбільшого розкиду даних, але кожна наступна вісь повинна бути ортогональна вихідній. Пошук головних компонент буде повторюватися до тих пір, поки всі виміри не будуть «використані». В більшості випадків така процедура призводить до значного зменшення розмірності даних (наприклад, 100000 ознак скорочується до 100), що дозволяє швидко і якісно обробити інформацію на звичайному комп'ютері.

Отже, лінійна трансформація дозволяє нам створити нові змінні. Алгоритм працює так, що спочатку обираються максимально мінливі змінні (збирається максимальна мінливість), тому перша головна компонента передає максимальну варіацію між точками даних і містить мінімальну похибку, а потім переходить до інших. Ваги підбираються так, щоб розкид точок був максимально можливим за умови, що сума квадратів ваг кожного фактора включеного в першу компоненту, дорівнює одиниці. Оскільки компоненти описують конкретні напрямки в просторі даних, то кожна компонента залежить від певних величин - від кожної з вихідних змінних: кожна компонента є лінійною комбінацією всіх вихідних змінних.

В результаті підбору потрібного розміщення осей буде отримано отримали дві нові змінні – компоненти *G1* та *G2* і їх кількість дорівнює кількості початкових ознак (х та у). Різниця між початковими змінними в тому, що перша компонента містить більше інформації, аніж друга. За рис. 12.2 можна побачити, що розкид точок у першої головної компоненти більший, а отже і дисперсія більше, ніж у другої. Компоненти ортогональні, тобто не корелюють між собою.

У багатовимірному просторі, при виокремленні головних компонент, частка загальної дисперсії буде зменшуватись з кожною наступною компонентою. Процес розрахунку компонент закінчують у той момент, коли розмір залишкової дисперсії залишається достатньо малим або якщо приріст факторної дисперсії за допомогою головних компонент різко падає.

Зазвичай, для візуалізації та подальшого аналізу достатньо виділити 2-3 головні компоненти, які в першому випадку (2 ГК) подаються у двовимірному вигляді, у другому (3ГК) – трьох вимірному.

Вимоги до головних компонент:

1. Максимальна інформативність.

2. Взаємна ортогональність (не повинні корелювати між собою).
3. Мінімальне спотворення геометричної структури вихідних даних.

МГК дає гарні результати якщо: нормальний розподіл вихідних даних; дублюється інформація у вихідних даних (виключення); наявні неінформативні дані (виключення); наявні однотипні змінні (агрегування).

### 12.2. Теоретичні основи методу головних компонент

Аналіз головних компонент призначений для перетворення вихідної ознакової множини *k* в множину *k*-нових ознак (так званих, головних компонент) і включає такі кроки:

I. Стандартизація діапазону неперервних початкових змінних.
II. Перевірка розподілу ознак на нормальність.
III. Розрахунок кореляційної матриці для виявлення взаємозв'язків, видалення лінійних звязків, перевірка якості моделі.
IV. Розрахунок власних векторів та власних значень кореляційної матриці для ідентифікації основних компонент.
V. Визначення кількості головних компонент, щоб вирішити, які основні компоненти слід зберегти, а які ігнорувати.
VI. Обертання факторної структури, у разі неможливості пояснення попередньо отриманих компонент.
VII. Пояснення результатів аналізу, візуалізація.

Вихідна інформація для дослідження описується множиною спостережень $\boldsymbol{i}$ та показників *j* в $\boldsymbol{n}$ рядках (кількість спостережень) та k стовпчиках (кількість змінних) (табл. 12.1).

*Таблиця 12.1.* **Ознакова множина для ідентифікації головних компонент**

| | | | | | |
|---|---|---|---|---|---|
| $x_{11}$ | $x_{12}$ | … | $x_{1j}$ | … | $x_{1k}$ |
| $x_{21}$ | $x_{22}$ | … | $x_{2j}$ | … | $x_{2k}$ |
| … | … | … | … | … | |
| $x_{i1}$ | $x_{i2}$ | … | $x_{ij}$ | … | $x_{ik}$ |
| … | … | … | … | … | |
| $x_{n1}$ | $x_{n2}$ | … | $x_{nj}$ | … | $x_{nk}$ |

Ознакова множина, здебільшого, формується но основі теоретичних знань про об'єкт дослідження, тобто висувається гіпотеза щодо природи латентних властивостей явища. Але досить часто метод використовується для недосліджуваних явищ і в такому випадку береться максимально можливий обсяг інформації з метою видобутку корисної інформації і пошуку прихованих закономірностей.[2]

Вимоги до ознакової множини (вибірки) для проведення МГК:

- дані за всіма ознаками повинні підпорядковуватись нормальному закону розподілу;
- ознаки повинні мати взаємозв'язки;
- зв'язок між ознаками повинен бути лінійний;
- коефіцієнти кореляції між ознаками не повинні дорівнювати 0 та 1 (тобто, змінні у яких відсутній або лінійний зв'язок виключаються з аналізу);
- вибірки повинні мати рівні (гомогенні) дисперсії;
- мінімальний обсяг спостережень – 50, оптимальний –100 і більше;

[2] Єріна А. М. Статистичне моделювання та прогнозування [Текст]: підручник / А. М. Єріна, Д. Л. Єрін; Держ. ВНЗ "Київ. нац. екон. ун-т ім. Вадима Гетьмана". Київ: КНЕУ, 2014. 348 с.

- кількість спостережень повинна перевищувати кількість показників.

Досить легко помітити, що вихідна інформація подібна до матриці, тому запишемо її у вигляді матриці розміром $k \times n$, де $k < n$.

$$X = \begin{pmatrix} x_{11} \ \dots \ x_{1j} \ \dots \ x_{1k} \\ \dots \dots \dots \dots \dots \dots \dots \\ x_{i1} \ \dots \ x_{ij} \ \dots \ x_{ik} \\ \dots \dots \dots \dots \dots \dots \dots \\ x_{n1} \ \dots \ x_{nj} \ \dots \ x_{nk} \end{pmatrix}, \qquad (12.1)$$

де: $x_{ij}$ – значення $j$-го показника у $i$-го спостереження ($i$ = 1, 2, …, $n$; $j$ = 1, 2, …, $k$).

Метод головних компонент є досить чутливим до дисперсії вихідних змінних і дає кращі результати, якщо дані мають однакову природу та однакові одиниці виміру. На результати дослідження можуть вплинути, на перший погляд, такі незначні деталі, як зміна розмірності одиниць вимірювання та різні одиниці виміру змінних. Отже, якщо існують великі відмінності між діапазонами початкових показників, змінні з більшими варіаційним розмахом будуть домінувати над тими, що мають менший варіаційний розмах (наприклад, змінна, яка коливається в межах від 0 до 1000, буде домінувати над змінною, яка коливається від 0 до 100; грн – тис.грн, см – мм, тощо), що призведе до упереджених результатів. Вирішення цієї проблеми полягає у перетворенні вихідних даних в порівняльні масштаби, щоб кожна з них однаково сприяла аналізу. Це гарантує те, що аналіз буде зосереджений на походженні наших основних компонент і на результати не вплинуть однакова природа даних та різні одиниці виміру.

Тому *першим кроком МГК* є стандартизація діапазону неперервних початкових змінних $x_{11}$, $x_{12}$, …, $x_{nk}$ у стандартизовані змінні $z_{11}$, $z_{12}$, …, $z_{nk}$ за формулою:

$$z_{ij} = \frac{x_{ij} - \overline{x_j}}{\sigma_j}, \qquad (12.2)$$

де: $\sigma_j$- середньоквадратичне відхилення *j*-го показника,

$x_{ij}$ – значення *j*-го показника у *i*-го спостереження,

$\overline{x_j}$ – середня арифметична проста *j*-го показника.

Математичні властивості стандартизованих змінних:

1) середня арифметична стандартизованої змінної дорівнює нулю ($\bar{z}_j = 0$);

2) середньоквадратичне відхилення та дисперсія стандартизованої змінної дорівнює одинці $\sigma^2 = \sigma = 1$.

Саме ці властивості дозволяють зробити результати дослідження на основі МГК адекватними реальному процесу та без втрати інформативності. Перша властивість доводить що, припущення, яке ми застосували на початку розділу, щодо переносу системи координат у зручне місце для аналізу, є правильним: з точки зору геометрії початок координат у стандартизованих показників зсувається у центр діаграми розсіювання. Друга властивість підтверджує, що інформативність сукупності не втрачається – всі стандартизовані показники мають однакову дисперсію, яка вважається мірою інформативності, і загальний обсяг інформації дорівнює $k\left(\sum_{j=1}^{k} \sigma_j^2 = k\right)$.[3]

Після стандартизації показників початкова матриця набуває вигляду:

[3] Єріна А. М. Статистичне моделювання та прогнозування [Текст]: підручник / А. М. Єріна, Д. Л. Єрін; Держ. ВНЗ "Київ. нац. екон. ун-т ім. Вадима Гетьмана". Київ: КНЕУ, 2014. 348 с.

$$Z = \begin{pmatrix} z_{11} \dots z_{1j} \dots z_{1k} \\ \dots \dots \dots \dots \dots \dots \dots \\ z_{i1} \dots z_{ij} \dots z_{ik} \\ \dots \dots \dots \dots \dots \dots \dots \\ z_{n1} \dots z_{nj} \dots z_{nk} \end{pmatrix} \quad (12.3)$$

Візуально процес переносу початку координат у центр сукупності для двовимірного простору після процедури стандартизації вихідної інформації представлено на рис. 12.3.

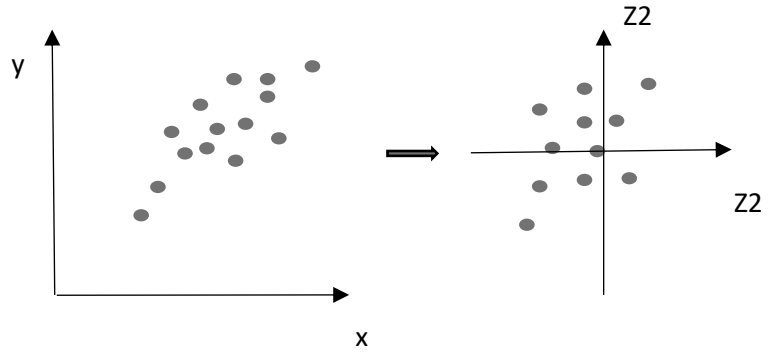



*Рисунок 12.3.* **Процес перенесення початку координат після стандартизації змінних**

Другим кроком МГК є розрахунок кореляційної матриці.

Кореляційна матриця на основі стандартизованих змінних набуває вигляду:

$$R = \frac{1}{n} Z^T Z \quad (12.4)$$

і складається з парних коефіцієнтів кореляції:

$$r_{jl} = \frac{1}{n}\sum_{i=1}^{n} z_{ij}z_{il} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{il}-\bar{\bar{x}}_l)}{\sigma_j\sigma_l} \quad (12.5)$$

де: $j$, $l$ = 1, 2, …, $k$.

За умови, що $j = l$ на головній діагоналі матриці R розташовані:

$$r_{jj} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}{{\sigma_j}^2} = 1. \quad (12.6)$$

МГК набуває вигляду:

$$z_{ij} = \sum_{p=1}^{k} a_{jp}G_{ip}, \quad (12.7)$$

де: $a_{ip}$ – вага, факторне навантаження $p$-ї головної компоненти на $j$-у ознаку,

$G_{ip}$ – значення $p$-ї головної компоненти для $i$-го спостереження,

де $p$ = 1, 2, …, $k$.

Запишемо модель МГК (12.7) у вигляді матриці:

$$Z = GA^T \quad (12.8)$$

Матриця значень головних компонент розміром $k \times n$ набуває вигляду:

$$G = \begin{pmatrix} Gz_{11} \dots G_{1j} \dots G_{1k} \\ \dots \dots \dots \dots \dots \dots \dots \\ G_{i1} \dots G_{ij} \dots G_{ik} \\ \dots \dots \dots \dots \dots \dots \dots \\ G_{n1} \dots G_{nj} \dots G_{nk} \end{pmatrix} \quad (12.9)$$

А матриця факторних навантажень розміром $k \times k$:

$$A = \begin{pmatrix} a_{11} \dots a_{1p} \dots a_{1k} \\ \dots \dots \dots \dots \dots \dots \dots \\ a_{i1} \dots a_{ip} \dots a_{ik} \\ \dots \dots \dots \dots \dots \dots \dots \\ a_{k1} \dots a_{kp} \dots a_{kk} \end{pmatrix} \quad (12.10)$$

Матриця факторних навантажень має важливе значення для формування висновків щодо головних компонент, які є лінійними комбінаціями вихідних ознак. За такої умови використовують лише ті вихідні ознаки, у яких $|a_{jp}| > 0{,}5$. Лінійна комбінація для першої головної компоненти $(G_1)$ і вихідних ознак $x_{11}$, $x_{12}$, …, $x_{nk}$: $G_1 = a_{11}\,x_1 + a_{11}\,x_1 + \dots + a_{1k}\,x_k$ , а для стандартизованих ознак $z_{11}$, $z_{12}$, …, $z_{nk}$: $G_1 = a_{11}\,z_1 + a_{11}\,z_1 + \dots + a_{1k}\,z_k$. За аналогією записують лінійну комбінацію і для всіх інших головних компонент.

Беручи до уваги те, що елементи матриці G стандартизовані можна записати властивості головних компонент: $\overline{G_p} = \frac{1}{n}\sum_{i=1}^{n} G_{ip} = 0$ , $\sigma^2_{Gip} = \sum_{i=1}^{n} G^2_{ip} = 1$. Отже, головні компоненти незалежні і в геометричному плані ортогональні. Зазначені властивості дозволяють зробити висновок, що:

$$\frac{1}{n}G^T G = E, \tag{12.11}$$

де $E$ – одинична матриця розміром $k \times k$ :

$$E = \begin{pmatrix} 1\ 0 & \dots & 0 \\ 0\ 1 & \dots & 0 \\ \dots & \dots & \dots . \\ 0\ 0 & \dots & 1 \end{pmatrix} \tag{12.12}$$

Враховуючи властивості головних компонент та вираз (12.11) парний коефіцієнт кореляції для всіх $j$=1, 2, …, $k$ та $p$ = 1, 2, …, $k$ можна записати наступним чином:

$$r_{z_j G_p} = a_{jp}\ , \tag{12.13}$$

Вираз (12.13) допомагає пояснити складові елементи матриці факторних навантажень А: $a_{jp}$ – характеризує тісноту зв’язку між ознакою $z_j$ та головною компонентою $G_p$.

Коефіцієнтам $a_{jp}$ присутні ті ж самі властивості, що і будь якій мірі щільності зв’язку:

- $a_{jp}$ коливаються в межах від 0 до $\pm$ 1;
- якщо $a_{jp} = \pm 1$ - лінійний функціональний зв'язок;
- якщо $a_{jp} = 0$ - зв’язок відсутній.

З урахуванням формул (12.7) та (12.13) отримаємо трансформовану формулу дисперсії стандартизованих показників:

$$\sigma_j^2 = \sum_{p=1}^{k} a_{jp}^2 = 1 \tag{12.14}$$

Відповідно до властивостей стандартизованих показників їх дисперсія дорівнює одиниці ($\sigma^2 = 1$). Отже, дисперсія стандартизованого показника $z_j$ згідно з (12.14) представлена складовими, які характеризують питому вагу внеску в неї всіх k головних компонент.

Повний внесок p-ї головної компоненти в сумарну дисперсію всіх k ознак становить:

$$\lambda_k = \sum_{j=1}^{k} a_{jp}^2 \qquad (12.15)$$

Підставивши матричне вираження моделі головних компонент (12.8) у кореляційну матрицю (12.4), виконаємо одну з головних умов МГК – виразимо кореляційну матрицю через матрицю факторних навантажень[4]:

$$R = AA^T \qquad (12.16)$$

Саме на цьому етапі потрібно провести оцінку якості отриманої кореляційної матриці задля розуміння можливості подальшого аналізу і виключення з сукупності лінійно пов'язаних змінних. Адже модель головних компонент добре спрацьовує лише в тому випадку, якщо між ознаками є зв'язки. Обмеженням для проведення обох тестів є виключення з моделі факторів з лінійним зв'язком (коефіцієнт кореляції дорівнює одиниці). Перевірка

[4] Єріна А. М. Статистичне моделювання та прогнозування [Текст]: підручник / А. М. Єріна, Д. Л. Єрін; Держ. ВНЗ "Київ. нац. екон. ун-т ім. Вадима Гетьмана". Київ: КНЕУ, 2014. 348 с.

виконується за тими ж самими критеріями, що використовують і в факторному аналізі:

1. *Тест сферичності Бартлетта (Bartlett's Test of Sphericity),* який перевіряє кореляційну матрицю на те чи є вона одиничною, що виключає можливість проведення аналізу. Головною умовою проведення тестування є нормальність багатовимірного розподілу, тому важливо переглянути графіки нормального розподілу по кожному показнику до проведення аналізу. Тест перевіряє нульову гіпотезу (H0)*:* кореляційна матриця дорівнює одиничній і ознаки не взаємопов'язані між собою; (H1): матриці не співпадають, між ознаками існує зв'язок. МГК можна застосовувати якщо буде відхилена нульову гіпотеза, висновок буде статистично значимий при достатньо високому значенні критерію і ймовірність випадкового виникнення такого ж та більш екстремального результату, помилки 1-роду, менше 0,05 ($p \leq 0,05$).

2. *Критерій адекватності вибірки Кайзера-Мейера-Олкіна (Kaiser-Meyer-Olkin measure, КМО*), який показує можливість застосовування МГК та факторного аналізу до обраного набору даних. Його рівень коливається в межах від 0 до 1 (чим ближче до одиниці, тим кращі будуть результати аналізу).

Марія Норіус запропонувала наступні правила пояснення рівня показника КМО[5]:

≥0,9 – безумовна адекватність;
≥0,8 – висока адекватність;
≥ 0,7 – середня адекватність;
≥ 0,6 – задовільна адекватність;
≥ 0,5 – низька адекватність;
≤ 0,5 аналіз не застосовується.

[5] Norusis M.J. SPSS Professional Statistics, Version 6.1 (SPSS for Windows 6.1). Chicago:SPSS, 1994. 397p.

Але навіть низький рівень критеріїв для МГК не є перешкодою для застосування, що не можна сказати про факторний аналіз. Тобто вказана перевірка є бажаною, але не обов'язковою, що пояснюється геометричним базисом методу головних компонент

І тепер можемо перейти до наступного етапу аналізу - визначення власних значень і власних векторів кореляційної матриці.

*Третім кроком МГК є* визначення власних значень і власних векторів кореляційної матриці.

Враховуючи той факт, що для кожної симетричної кореляційної матриці R завжди є така ортогональна матриця U, яка виконує умову[6]:

$$U^T R U = \Lambda, \tag{12.17}$$

де: $\Lambda$ – ддіагональна матриця власних значень, $U$ ортогональна матриця власних векторів.

Діагональна матриця власних значень $\Lambda$ розміром $k \times k$ вигляду:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_k \end{pmatrix} \tag{12.18}$$

Ортогональна матриця власних векторів $U$ розміром $k \times k$ вигляду:

[6] Бізнес-аналітика багатовимірних процесів: навчальний посібник [Електронний ресурс] / Т. С. Клебанова, Л. С. Гур'янова, Л. О. Чаговець та ін. Харків : ХНЕУ ім. С. Кузнеця, 2018. 272 с.

$$U = \begin{pmatrix} u_{11} \dots u_{1p} \dots u_{1k} \\ \dots \dots \dots \dots \dots \dots \dots \\ u_{i1} \dots u_{ip} \dots u_{ik} \\ \dots \dots \dots \dots \dots \dots \dots \\ u_{k1} \dots u_{kp} \dots u_{kk} \end{pmatrix} \tag{12.19}$$

Всі елементи кореляційної матриці R позитивні величини, тому всі власні значення $\lambda_{\text{p}} > 0$ для будь-яких $p = 1,2, \dots k$.

Елементи матриці власних значень $\Lambda$ упорядковані у порядку зменшення: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{\text{p}} \geq \lambda_k \geq 0$.

Власні значення $\lambda_1 \geq \cdots \geq \lambda_{\text{p}} \geq \lambda_k$ розраховують як корінь з рівняння:

$$\lceil \Lambda E - R \rceil = 0 \tag{12.20}$$

Власний вектор $\text{P}_{\text{p}}$, який відповідає власному значенню $\lambda_p$ кореляційної матриці R знаходимо з рівняння:

$$\left(\lambda_p E - R\right) P_p = 0 \tag{12.21}$$

Після розрахунків власних значень та власних векторів отримаємо новий ортогональний простір ознак, розмірність якого повністю співпадає з вхідною ознаковою множиною (рис. 12.4).

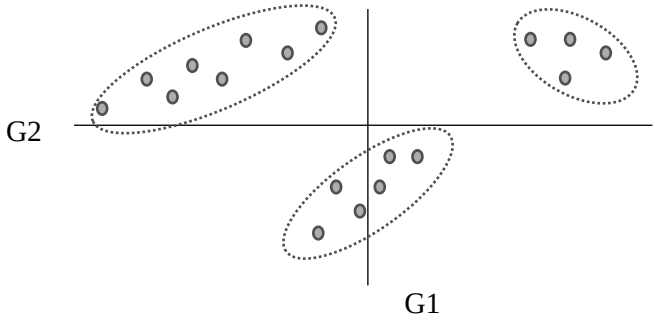



*Рисунок 12.4.* **Графічне відображення нових змінних (головних компонент) у ортогональному просторі з чітко вираженими групами, які не були видимі в початковій системі координат**

До того ж, нові змінні розташовані в системі координат ранжовано – у порядку зменшення їх інформативності (зменшення дисперсії).

Враховуючи те, що матриці R та $\Lambda$ подібні, а сума ддіагональних елементів матриці $R = k$ отримаємо:

$$\sum_{p=1}^{k} \lambda_p = k \qquad (12.22)$$

Таким чином, повний внесок р-ї компоненти в загальну дисперсію вихідних ознак (за аналогією до формули 12.15) розраховується за формулою:

$$\lambda_p = \sum_{j=1}^{k} a_{jp}^2 \qquad (12.23)$$

Питому вагу внеску *p*-ї головної компоненти розраховується за формулою:

$$\frac{\lambda_p}{k} \cdot 100\% \qquad (12.24)$$

Сумарний внесок $k$ перших компонент (повноту факторизації) розраховують за формулою:

$$\frac{\sum_{p=1}^{m} \lambda_p}{k} \cdot 100\% \qquad (12.25)$$

Відповідно до вищезазначеного, перша компонента $G_1$ враховує найбільш можливу дисперсію в сукупності даних, а остання - найменшу (ранжуються за $\lambda_1 \geq \cdots \geq \lambda_{\mathrm{p}} \geq \lambda_k$). Виділення першої головної компоненти за максимальним вкладом у дисперсію ознак означає, що знаходиться такий напрямок у просторі ознак, якому відповідає максимальна дисперсія, тобто максимальний розкид точок у системі координат. Звичайно, можна зробити дисперсію $\lambda_p$ якомога більшою, обравши найбільші значення для ваг $\mathrm{a}_{11}, \mathrm{a}_{12,} \cdots, \mathrm{a}_{1\mathrm{p}}$, але для запобігання цьому, ваги розраховуються таким чином щоб сума їх квадратів дорівнювала 1.

$$a_{11}^2 + a_{12}^2 + \cdots + a_{1p}^2 = 1 \qquad (12.26)$$

Далі розраховується друга головна компонента, яка не корелює з першою і враховує наступну найбільшу дисперсію, тобто найбільшу диференціацію об'єктів, не пояснену першою компонентою. І цей процес продовжується до тих пір, поки не буде

розраховано загальну кількість k головних компонент, що дорівнює початковій кількості змінних. На цьому етапі сума дисперсії усіх головних компонент буде дорівнювати сумі дисперсії усіх вихідних змінних ($\sigma^2$), тобто вся вихідна інформація буде пояснена або врахована:

$$\sigma^2 = \sum_{p=1}^{k} \lambda_p = \sum_{p=1}^{k} a_{jp}^2 \qquad (12.27)$$

Зазвичай основна інформація міститься в перших компонентах, внесок інших у сумарну дисперсію досить незначний. Для дослідження використовують m перших головних компонент, внесок яких в сумарну дисперсію становить 60-70%. Більш детально вибір кількості компонент розглянемо в наступному кроці аналізу.

Результати аналізу методу головних компонент записуються у вигляді таблиці (табл.12.2), яка представляє собою матрицю факторних навантажень $a_{jp}^2$ - коефіцієнтів переходу від стандартизованих показників до нових – компонент ($G_p$). Нумерація факторних навантажень відрізняється від більшості матриць: перший нижній індекс відповідає порядковому номеру стовпчика таблиці, а другий – порядковому номеру рядка.

Коефіцієнти $a_{jp}^2$ в рядках матриці показують внесок кожної компоненти у формування варіації $j$-ї ознаки. Для кожної ознаки характерна своя факторна структура. Чим менше внесок головної компоненти у варіацію $j$-ї ознаки, тим простішою є її факторна структура.[7]

[7] Єріна А. М. Статистичне моделювання та прогнозування [Текст]: підручник / А. М. Єріна, Д. Л. Єрін; Держ. ВНЗ "Київ. нац. екон. ун-т ім. Вадима Гетьмана". Київ: КНЕУ, 2014. 348 с.

*Таблиця 12.2.* **Матриця переходу до нових змінних та показники якості моделі**

| Стандартизовані показники $\boldsymbol{z_j}$ | Головні компоненти $\boldsymbol{G_p}$ | | | | | | Дисперсія стандартизованих показників $z_k$ $\boldsymbol{\sigma^2}$ |
|---|---|---|---|---|---|---|---|
| | $\boldsymbol{G_1}$ | $\boldsymbol{G_2}$ | … | $\boldsymbol{G_l}$ | … | $\boldsymbol{G_p}$ | |
| $\boldsymbol{z_1}$ | $a_{11}^2$ | $a_{21}^2$ | … | $a_{l1}^2$ | … | $a_{p1}^2$ | 1 |
| $\boldsymbol{z_2}$ | $a_{12}^2$ | $a_{22}^2$ | … | $a_{l2}^2$ | … | $a_{p2}^2$ | 1 |
| … | … | … | … | … | … | … | 1 |
| $\boldsymbol{z_k}$ | $a_{1k}^2$ | $a_{2k}^2$ | … | $a_{lk}^2$ | … | $a_{pk}^2$ | 1 |
| Дисперсія $\boldsymbol{G_p}$ | $\lambda_1$ | $\lambda_2$ | … | $\lambda_l$ | … | $\lambda_p$ | $\sum_{p=1}^{k} \lambda_p$ |

*Четвертим кроком МГК* є визначення кількості головних компонент для зменшення ознакового простору без втрати інформативності.

Зазвичай загальна дисперсія ознак розкладається таким чином, що більша частина вхідної інформації стискається або входить до перших головних компонент, тобто перші кілька компонент пояснюють більшу частку цієї дисперсії, а решта майже нічого. Організація інформації в основних компонентах таким чином дозволить нам зменшити розмірність без вагомої втрати інформації. На цьому етапі потрібно обрати найбільш інформативні компоненти і проігнорувати менш інформативні. Візуально цей процес відображено на рисунку 12.5, на якому пунктирною лінією «відсічено» найбільш інформативні головні компоненти.

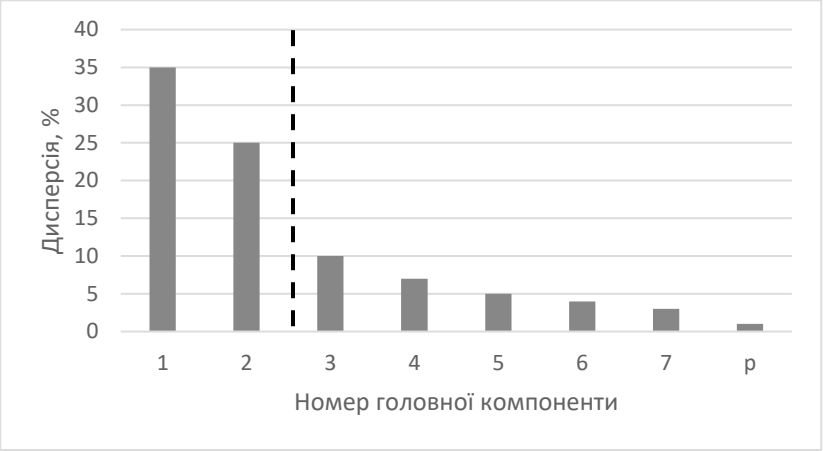



*Рисунок 12.5.* **Відсоток дисперсії для кожної головної компоненти**

Важливо розуміти, що компоненти не мають інтерпретації і не мають ніякого практичного значення, оскільки вони побудовані як лінійні комбінації початкових змінних. З точки зору геометрії, головні компоненти - це напрямки даних, які пояснюють максимальну дисперсію, тобто лінії, які захоплюють більшу частину інформації даних. Зв'язок між дисперсією та інформацією тут полягає в тому, що чим більше дисперсія, що переноситься лінією, тим більша дисперсія точок даних уздовж неї, і чим більше дисперсія по лінії, тим більше інформації вона має.

У таблиці 12.2 сірим кольором виділено найменш інформативні компоненти з максимальними номерами, які в подальшому дослідженні не будуть використовуватись. За рахунок відсікання цих компонент буде зменшуватись розмірність. Незафарбовані $l$ стовпчиків змінних $G_1, G_2, \dots, G_l$ $(l < k)$ називаються головними компонентами. У такий спосіб модель аналізу головних компонент (12.7) на цьому етапі набуває вигляду:

$$z_{ij} = \sum_{p=1}^{k} a_{jl} G_{il} \qquad (12.28)$$

Сумарна дисперсія головних компонент, які залишились після процедури зменшення розмірності, менша дисперсії ознакової множини ($k$). Отже, зменшуючи кількість компонент ми втрачаємо певну частину інформації. Частку втраченої інформації можна розрахувати користуючись формулою (12.25):

$$1 - \frac{\sum_{p=1}^{m} \lambda_p}{k} \cdot 100\% \qquad (12.29)$$

Оскільки метою аналізу головних компонент є зменшення розмірності, тобто зосередження уваги на кількох головних компонентах, вченими було запропоновано кілька критеріїв для визначення того, скільки головних компонент доцільно досліджувати, а скільки ігнорувати.

*Критерій Кайзера-Гутмана (критерій Кайзера, Kaiser, 1960)* застосовується для кореляційної матриці $R$ і базується на власних значеннях кореляційної матриці $\lambda_p$. За критерієм Кайзера обирають головними компонентами $G_p$ ті, у яких $\lambda_p \geq 1$, тому що їх інформаційне наповнення більше ($\sigma_{Gp}^2 \geq 1$), аніж у проігнорованих/"відсічених" компонент ($\sigma_{Gp}^2 < 1$). У випадку використання коваріаційної матриці – "відсікаються" компоненти, власні числа яких менше ніж середня дисперсія.

*Другий критерій* базується на внеску компонент у загальну дисперсію. Кумулятивний внесок перших головних компонент повинен бути не меншим попередньо визначеної частки вихідної дисперсії $k$ (наприклад, 90%). У цьому випадку використовують внесок кумулятивних часток перших головних компонент в сумарну дисперсію:

$$\frac{\lambda_1}{k}; \frac{\lambda_1+\lambda_2}{k}; \ldots; \frac{\lambda_1+\lambda_2+\cdots+\lambda_l}{k} \qquad (12.30)$$

*Критерий Кеттеля (Cattell,1966), графічний критерій «кам'янистого осипу» (в окремих працях «кам'яного обвалуу»).* Геологічний термін "кам'яний осип", "скельний осип" означає нагромадження щебню, уламків гірських порід біля підніжжя скельного схилу. Критерій передбачає побудову простого графіка на основі значень $\lambda_p$ таблиці 12.2: на осі $x$ відображається порядковий номер компоненти $1, 2, 3, \ldots, p$, а на осі $y$ – значення $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_p$. Для зменшення розмірності ознакової множини Кеттель, грунтуючись на методі Монте-Карло, запропонував знайти на графіку безперервного падіння власних значень точку, починаючи з якої значення $\lambda_p$ зліва направо максимально сповільнюються і справа від неї залишається лише "факторіальний осип". Ця точка і є номером останньої компоненти, яку будемо залишати для подальшої інтерпретації, а решту компонент вважати випадковим шумом. За рисунком 12.6 можна обрати дві або три компоненти.

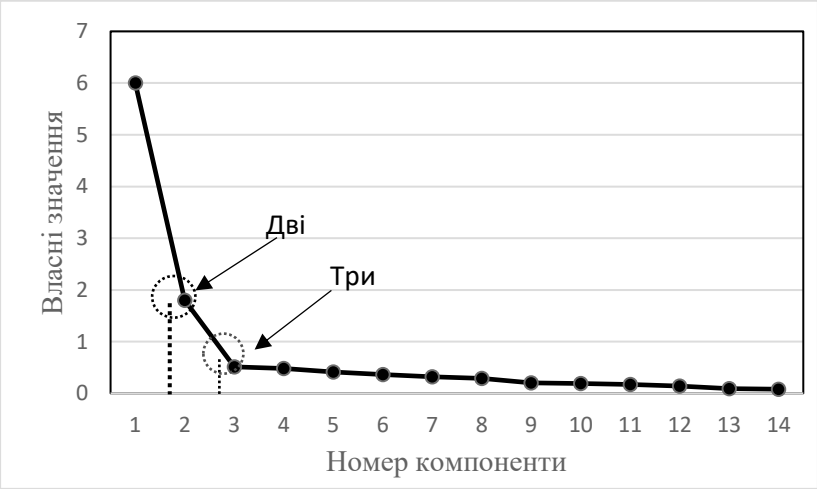



*Рисунок 12.6.* **Графік «кам'янистого осипу»**

*Четвертий критерій* — ігнорування останніх компонент, дисперсії яких приблизно однакові.[8]

Який критерій краще використовувати? Так як різні критерії лишають різну кількість компонент (одні залишають їх досить мало, а інші занадто багато) перед дослідником постають питання: який же критерій обирати і яку кількість компонент залишати. На практиці, зазвичай, не зупиняються на одному критерії, а обирають декілька з більшою та меншою кількістю головних компонент і поміж них обирають один, який залишає максимально пояснюваний набір компонент.

*П'ятим кроком МГК* – оцінка якості моделі, пояснення головних компонент.

Оцінка якості отриманої моделі головних компонент відбувається за результуючою таблицею (табл. 12.3).

Якість проведеного аналізу методом головних компонент оцінюють за показниками внизу таблиці: повнота факторизації та внеску головних компонент в сумарну дисперсію, на основі якого можна розрахувати частку втрати інформації при видаленні частини компонент (видалені компоненти помічені сірим кольором в стовпчику).

Для пояснення головних компонент використовують матрицю факторних навантажень. Факторні навантаження по суті є коефіцієнтами кореляції між вихідними ознаками та головними компонентами, і для інтерпретації використовують їх максимальні значення по модулю.

Досить легко інтерпретувати компоненти, в яких показники корелюють лише з однією компонентою, тобто вихідні ознаки мали взаємозв'язки. І набагато складніше піддається поясненню ситуація, якщо головні компоненти мають багато невисоких факторних

[8] Steven M. Holand. (2019). Principal components analysis (PCA), Department of Geology, University of Georgia, Athens, GA 30602-2501. URL: https://strata.uga.edu/software/pdf/pcaTutorial.pdf

навантажень і ознаки практично однаково корелюють з різними головними компонентами. В останньому випадку, для покращення інтерпретації головних компонент, доречно провести процедуру обертання.

*Таблиця 12.3.* **Показники якості моделі**

| Стандартизовані показники $\boldsymbol{z_j}$ | Головні компоненти $\boldsymbol{G_p}$ | | | | | | Дисперсія стандартизованих показників $z_k$ $\boldsymbol{\sigma^2}$ |
|---|---|---|---|---|---|---|---|
| | $\boldsymbol{G_1}$ | $\boldsymbol{G_2}$ | … | $\boldsymbol{G_l}$ | … | $\boldsymbol{G_p}$ | |
| $\boldsymbol{z_1}$ | $a_{11}^2$ | $a_{21}^2$ | … | $a_{l1}^2$ | … | $a_{p1}^2$ | … |
| $\boldsymbol{z_2}$ | $a_{12}^2$ | $a_{22}^2$ | … | $a_{l2}^2$ | … | $a_{p2}^2$ | … |
| … | … | … | … | … | … | … | … |
| $\boldsymbol{z_k}$ | $a_{1k}^2$ | $a_{2k}^2$ | … | $a_{lk}^2$ | … | $a_{pk}^2$ | … |
| Дисперсія $\boldsymbol{G_p}$ | $\lambda_1$ | $\lambda_2$ | … | $\lambda_l$ | … | $\lambda_p$ | $\sum_{p=1}^{k} \lambda_p$ |
| Внесок p-ї головної компоненти в сумарну дисперсію, % (повнота факторизації) | $\frac{\lambda_1}{k}$ | $\frac{\lambda_2}{k}$ | … | $\frac{\lambda_l}{k}$ | … | $\frac{\lambda_p}{k}$ | - |
| Внесок головних компонент в сумарну дисперсію, % | $\frac{\lambda_1}{k}$ | $\frac{\lambda_1 + \lambda_2}{k}$ | … | … | … | 100 | - |

При поясненні головних компонент виникає два випадки:

1) найбільші абсолютні значення факторних навантажень мають однакові знаки (+ або -) – це головна компонента розміру; в такому випадку ми маємо справу зі спільною латентною характеристикою у всіх об'єктів дослідження, яку і потрібно пояснювати;

2) найбільші абсолютні значення факторних навантажень мають різні знаки – це головна компонента форми; в такому випадку головна компонента диференціює об'єкти дослідження за наявністю у них двох властивостей, в якомусь сенсі протилежних. У більшості випадків, після застосування процедури обертання головні компоненти форми перетворюються на компоненти розміру.

Пояснення головної компоненти відбувається на основі матриці факторних навантажень.

Факторні навантаження при проведенні аналізу ранжують в порядку його зменшення і для економічної інтерпретації обираються лише ті вихідні ознаки, які задовольняють умову: $a_{jp} =\geq 0{,}5$ (мають найбільший вплив на головну компоненту).

Досліджуємо зміст компоненти, виявляємо спільну (латенту) характеристику для ознак групи, шукаємо те, що може їх об'єднувати. Ця спільна риса або назва групи властивостей отримує назву, яка і буде назвою компоненти.

Наприклад, інформаційною базою для дослідження факторів впливу на валютні курси слугувало 16 факторів. Методом головних компонент було виділено дві головні компоненти, які після аналізу змісту факторів в кожній компоненті отримали назву структурні показники та грошово-кредитні показники.

*Шостим кроком МГК* є вимірювання головних компонент. Розрахунок головних компонент є завершальним етапом переходу від вихідних показників до нових змінних – головних компонент. Ця процедура має назву факторне шкалювання.

Кількісні характеристики головних компонент для $i$-го спостереження ($i = 1, 2, \dots, n$) задаються матрицею значень головних компонент $G$, яку знаходять з формули (12.8):

$$G = Z(A^T)^{-1} = ZP\Lambda^{-1/2}, \quad (12.31)$$

де: $Z$ – матриця стандартизованих показників,
$(A^T)^{-1}$ – обернена матриця факторних навантажень,
$P$ – власний вектор,
$\Lambda$ – матриця власних значень.

При зменшенні розмірності сукупності, на четвертому кроці аналізу, кількість компонент $p$ було зменшено до $l$ головних компонент ($l < p$) тому формулу (13.31) можна записати таким чином:

$$G = \lambda^{-1} Z A^T \quad (12.32)$$

Або алгебраїчно це можна записати:

$$G_i = \sum_{i=1}^{l} z_{ji} \frac{a_{jp}}{\lambda_p} \quad (12.33)$$

Відбувається підсумовування стандартизованих значень ознаки з вагами, пропорційними факторним навантаженням (до обертання). Ділення факторних навантажень на $\lambda_p$забезпечує нульове математичне сподівання та одиничну дисперсію для значень головних компонент $G$. Відємні або додатні знаки перед оцінками головної компоненти свідчать про те, що рівень компоненти у *i*-тої одиниці сукупності вищий або нижчий за середній.[9]

[9] Єріна А. М. Статистичне моделювання та прогнозування [Текст]: підручник / А. М. Єріна, Д. Л. Єрін; Держ. ВНЗ "Київ. нац. екон. ун-т ім. Вадима Гетьмана". Київ: КНЕУ, 2014. 348 с.

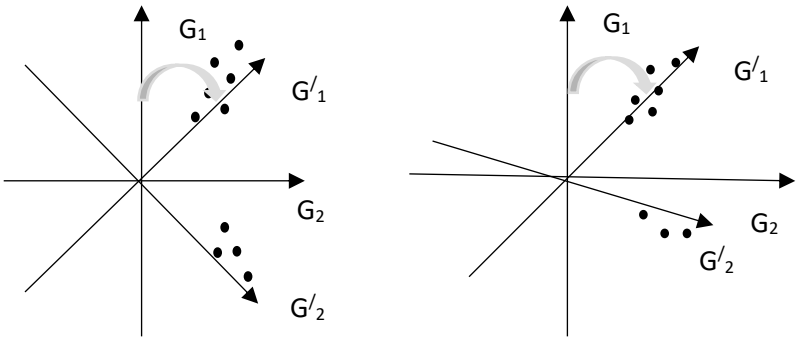
Іноді виникає ситуація коли після проведення аналізу МГК і виділення головних компонент дослідник не досягає очікуваного результату – виділені компоненти не піддаються поясненню. В такому випадку потрібно провести додаткові дії: повернути осі для кращої кореляції між ознаками, провести процедуру обертання.

*Сьомий крок МГК* - обертання факторної структури з метою максимізації дисперсії нової компоненти та покращення інтерпретації результатів аналізу. Процедура обертання не змінює розташування точок, повертаються лише осі навколо цих точок до моменту найбільшої їх концентрації біля осей. В результаті отримуємо нові координати (нові факторні навантаження), де точки показують однозначну належність певній компоненті, що значно полегшує їх пояснення. При використанні обертання, загальний відсоток дисперсії головних компонент не змінюється, але може перерозподілитись між ними по іншому. Наприклад, при дослідженні певного ознакового простору було обрано 2 головні компоненти, які пояснювали 75% загальної варіації (50% та 25%) та пояснити її зміст не було можливості; після використання процедури обертання було отримано нові зрозумілі компоненти з часткою загальної дисперсії 38% та 37% (обидві 75%).

Існує два підходи до обертання:

1. *Графічний*. Використовується для простої конфігурації даних: коли головна компонента має одне чітко виражене факторне навантаження, а інші - незначні, тобто обумовлені впливом одного фактору. В такому випадку дослідник може самостійно визначити розташування осей і зробити обертання до потрібної точки; коли на графіку виокремлюються певні групи і система координат повертає в скупчення цих груп; коли є гіпотеза про спільну властивість для об’єднання факторів і осі повертаються до потрібних точок.

2. *Аналітичний*. Застосовується у випадках, якщо графічний підхід не дав очікуваних результатів або кількість факторів в компоненті достатньо велика і візуально не можливо оцінити куди повертати осі. В залежності від кута нахилу осей (прямого або іншого) розрізняють дві групи методів обертання (рис. 12.7): ортогональне та косокутне.




Ортогональне Косокутне

*Рисунок 12.7.* **Аналітичні методи обертання факторних структур**

*Джерело*: власні розробки автора

*Ортогональне обертання*, до якого входять методи: варімакс (Varimax), квартімакс (Quartimax), еквімакс (Equimax), біквартімакс (Biquartimax) та інші. Обертання системи координат відбувається під прямим кутом, головні компоненти залишаються ортогональні/незалежні одна до одної. Найчастіше в дослідженнях використовується метод варімакс, який суттєво спрощує пояснення головних компонент.

*Косокутне обертання*: прямий облімін (Direct oblimin), квартімін (Quartimin), промакс (Promax) та інші. В групі цих методів система координат змінює кут нахилу від прямого до потрібного з урахуванням обмежень конкретного методу. Головні компоненти перестають бути незалежними (ортогональними). Косокутне

обертання, зазвичай, показує набагато кращі результати, так як на відміну від ортогонального надає можливість обирати кут нахилу осей для мінімізації відхилень точок факторних навантажень від осі. Але даний вид обертання використовують досить рідко, що пояснюється складністю інтерпретації отриманих результатів дослідження.

Якщо в процесі косокутного обертання отримують нові ортогональні структури, можна стверджувати, що вони насправді незалежні. Методи цієї групи, зазвичай, використовують для великих даних (big data).

Обертання факторної структури виконується шляхом множення початкової матриці факторного навантаження на матрицю перетворення. Матриця, отримана як добуток транспонованої матриці перетворення, на саму матрицю перетворення відповідає матриці факторних навантажень. Обмеження для матриці перетворення: 1) косокутне обертання: по діагоналі матриці повинні бути 1, тобто добуток однакових навантажень дорівнює одиниці, а поза діагональні елементи не повинні перевищувати одиницю; 2) ортогональне обертання: добуток різних факторних навантажень повинен дорівнювати нулю. Матриця перетворення визначається у відповідності до певного критерію від елементів кінцевої факторної структури. Мінімізація або максимізація цього критерію і дозволяє знайти оптимальну матрицю перетворення.

Розглянемо детальніше окремі методи.

*Варімакс (Varimax, $V_m$)* – метод ортогонального обертання, який зменшує до мінімуму кількість ознак з найбільшим факторним навантаженням. Максимізуючи дисперсії квадратів навантажень отримуємо нові компоненти, факторні навантаження яких наближаються до 1 або до 0, що значно спрощує факторну структуру та полегшує пояснення.

$$V_k = \frac{k\sum_{p=1}^{k} a_{jp}^4 - \left(\sum_{p=1}^{k} a_{jp}^2\right)^2}{k^2} \quad (12.34)$$

*Квартімакс (Quartimax)* – метод протилежний до попереднього, мінімізує кількість факторів для пояснення головних компонент, чим і досягається краща інтерпретація компоненти.

$$q = \sum_{j=1}^{l} \frac{\sum_{p=1}^{k} a_{jp}^4 - \left(\sum_{p=1}^{k} a_{jp}^2\right)^2}{k} \quad (12.35)$$

*Еквімакс (*Equimax*)* та *біквартімакс (Biquartimax*) є варіацією методів варімакс та квартімакс.

В практиці ортогонального обертання найчастіше застосовується метод варімакс.

Числові значення факторних навантажень до та після обертання різняться, але сума внеску головних компонент в формування загальної дисперсії залишається однаковою. Після проведеного обертання інформація від навантажень про важливість внеску фактору в головну компоненту втрачає актуальність, ця властивість не зберігається.

Метод головних компонент можна використовувати і як самостійний метод дослідження, так і як базу для інших методів. Як самостійний метод аналізу МГК використовують для: зменшення розмірності ознакової множини, дослідження рядів динаміки, знаходження узагальнюючих показників, оцінювання взаємозв'язку між інтегральними показниками і множиною первинних ознак, ранжирування і/або класифікація одиниць досліджуваної сукупності, структурний аналіз інформації.[10]

[10] Єріна А. М. Статистичне моделювання та прогнозування [Текст]: підручник / А. М. Єріна, Д. Л. Єрін; Держ. ВНЗ "Київ. нац. екон. ун-т ім. Вадима Гетьмана". Київ: КНЕУ, 2014. 348 с.

Як перший крок для інших методів МГК використовують для: регресійного аналізу (позбавлення від мультиколінеарності) та у системах одночасних рівнянь, коли коефіцієнти регресії визначаються двокроковим МНК (спрощення розрахунків без втрати інформації) [11].

### 12.3. Реалізація методу головних компонент

Вихідною інформацією для розрахунку головних компонент слугували дані 20 країн та показники: $x_1$ – реальний ефективний валютний курс, $x_2$ – сальдо зовнішньої торгівлі, $x_3$ – валовий внутрішній продукт. На першому кроці аналізу для кожного показника було розраховано середню арифметичну просту, середньоквадратичне відхилення та проведено процедуру стандартизації вихідних показників (Z1, Z2, Z3). Ця процедура є обов'язковою, адже показники мають різні одиниці виміру, що може призвести до хибних результатів.

Другим кроком аналізу є розрахунок кореляційної матриці. Розрахунки можна проводити в пакетах Excel, Statistica, SPSS тощо. Результати розрахунків представлено в табл. 12.4.

*Таблиця 12.4.* **Кореляційна матриця**

| Фактор | $Z_1$ | $Z_2$ | $Z_3$ |
|---|---|---|---|
| Z1 | 1,000 | 0,580 | 0,155 |
| Z2 | 0,580 | 1,000 | 0,440 |
| Z3 | 0,155 | 0,440 | 1,000 |

Отримані результати свідчать про досить сильний зв'язок між показниками Z1 та Z2.

Знайдемо власні значення кореляційної матриці:

[11]Єріна А. М. Статистичне моделювання та прогнозування: навч. посібник. К.: КНЕУ, 2001. 170 с.

$$\begin{vmatrix} 1{,}000-\lambda & 0{,}580 & 0{,}155 \\ 0{,}580 & 1{,}000-\lambda & 0{,}440 \\ 0{,}155 & 0{,}440 & 1{,}000-\lambda \end{vmatrix} = 0$$

З матриці записуємо характеристичне рівняння:

$$(1{,}000-\lambda)^3 + 0{,}580 * 0{,}155 * 0{,}440 + 0{,}580 * 0{,}155 * 0{,}440 - 0{,}155 * 0{,}155 * (1{,}000-\lambda) - 0{,}580 * 0{,}580 * (1{,}000-\lambda) - 0{,}440 * 0{,}440 * (1{,}000-\lambda) = 0$$

$$\lambda^3 - 3\lambda^2 + 2{,}446\lambda - 0{,}525 = 0$$

Знаходимо корені рівняння:

$$\lambda_1 = 1{,}797, \lambda_2 = 0{,}875, \lambda_3 = 0{,}326.$$

Можемо зробити висновок, що перша головна компонента має дисперсію 1,797 ($\frac{1{,}797}{3} = 0{,}599$, пояснює 59,9% загальної мінливості ознак/дисперсії), друга – 0,875 та пояснює 29,2% загальної дисперсії, третя - 0,326 та пояснює 10,9% загальної дисперсії. Загальна дисперсія всіх трьох компонент дорівнює:

100% = 59,9+29,2+10,9.

Внесок першої компоненти в загальну дисперсію розраховується наступним чином:

$$\frac{1{,}797}{1{,}797 + 0{,}875 + 0{,}326} * 100 = 59{,}9\%$$

Отже, дві перші компоненти описують 89,1% загальної дисперсії ознак, що є достатньо вагомим внеском.

Наступним кроком аналізу є розрахунок власних векторів матриці парної кореляції, які знаходимо підставляючи по черзі в систему лінійних рівнянь знайдені значення $\lambda_1 = 1{,}797, \lambda_2 = 0{,}875, \lambda_3 = 0{,}326$. Перша система рівнянь має вигляд:

$$\begin{cases} (1{,}000 - 1{,}797) * u_{11} + 0.580 * u_{21} + 0.155 * u_{31} \\ 0.580 * u_{11} + (1{,}000 - 1{,}797) * u_{21} + 0.440 * u_{31} \\ 0.155 * u_{11} + 0.440 * u_{21} + (1{,}000 - 1{,}797) * u_{31} \end{cases}$$

Друга система рівнянь:

$$\begin{cases} (1{,}000 - 0{,}875) * u_{12} + 0.580 * u_{22} + 0.155 * u_{32} \\ 0.580 * u_{12} + (1{,}000 - 0{,}875) * u_{22} + 0.440 * u_{32} \\ 0.155 * u_{12} + 0.440 * u_{22} + (1{,}000 - 0{,}875) * u_{32} \end{cases}$$

Третя система рівнянь:

$$\begin{cases} (1{,}000 - 0{,}326) * u_{13} + 0.580 * u_{23} + 0.155 * u_{33} \\ 0.580 * u_{13} + (1{,}000 - 0{,}326) * u_{23} + 0.440 * u_{33} \\ 0.155 * u_{13} + 0.440 * u_{23} + (1{,}000 - 0{,}326) * u_{33} \end{cases}$$

Отримаємо три власних вектори:

$$u_1 = \begin{matrix} 1{,}263 \\ 1{,}468, \\ 1{,}000 \end{matrix} u_2 = \begin{matrix} -0{,}145 \\ -0{,}235, \\ 1{,}000 \end{matrix} u_3 = \begin{matrix} 1{,}308 \\ -1{,}778. \\ 1{,}000 \end{matrix}$$

Пронормуємо вектори $U_p$ за формулою $P_p = \frac{U}{U_p}$ для отримання матриці нормованих значень власних векторів.

Факторні навантаження $a_{jp}^2$ для матриці А знаходимо за формулами:

$$a_{j1}^2 = \sqrt{\lambda_1} * u_{j1},\ a_{j2}^2 = \sqrt{\lambda_2} * u_{j2},\ a_{j3}^2 = \sqrt{\lambda_3} * u_{j3},$$

де: $j = 1,2,3$, $u_{j1}, u_{j2}, u_{j3}$ – координати власних векторів$u_1, u_2, u_3$.

Запишемо у вигляді матриці $A = P\Lambda^{1/2}$:

$$A = \begin{pmatrix} 0{,}578 & -0.137 & 0.538 \\ 0.675 & -0.225 & 0.965 \\ 0.458 & 0.965 & 0.414 \end{pmatrix} \cdot \begin{pmatrix} \sqrt{1.797} & & \\ & \sqrt{0.875} & \\ & & \sqrt{0.326} \end{pmatrix}$$

В результаті розрахунку факторних навантажень отримуємо матрицю переходу до нових показників (табл. 12.5).

*Таблиця 12.5*. **Матриця переходу до нових змінних**

| Фактор | $G_1$ | $G_2$ | $G_3$ |
|---|---|---|---|
| Z1 | 0,775 | -0,121 | 0,307 |
| Z2 | 0,903 | -0,209 | -0,418 |
| Z3 | 0,618 | 0,907 | 0,239 |

В табл.12.5 відображено часткові коефіцієнти кореляції, зв'язки між стандартизованими показниками Z1, Z2, Z3 та головними компонентами, які змінюються у межах від -1 до +1.

Зробимо арифметичну перевірку отриманих показників першої головної компоненти, знаючи що $\sum a_{jp}^2 = \lambda_{\text{р}}$: $0{,}775^2 \cdot 0{,}903^2 \cdot 0{,}618^2 = 1{,}797$.

Наступним кроком є визначення кількості головних компонент. Якщо зупинитись на критерію Кайзера, то нам доведеться залишити лише одну головну компоненту: $\lambda_1 > 1$. Але вона враховує лише 59,9% загальної дисперсії і втрати інформації будуть значні. Якщо залишити першу та другу компоненти, то ми отримаємо вже 89,1% загальної дисперсії. Отже, для даного

дослідження потрібно залишити дві головні компоненти. Оцінку якості моделі наведено в табл.12.6.

*Таблиця 12.6.* **Показники оцінки якості моделі**

| Показники | $G_1$ | $G_2$ |
|---|---|---|
| $z_1$ | 0,937 | -0,042 |
| $z_2$ | 0,782 | -0,321 |
| $z_3$ | 0,281 | 0,947 |
| Дисперсія $G_p$ | 1,759 | 0,863 |
| Внесок p-ї головної компоненти в сумарну дисперсію, % (повнота факторизації) | 57,2 | 28,2 |
| Внесок головних компонент в сумарну дисперсію, % | 57,2 | 85,4 |

Далі розраховують головні компоненти:

$$G_1 = 0.785 * z_1 - 0.186 * z_2 + 0.401 * z_3$$
$$G_2 = 0.896 * z_1 - 0.315 * z_2 - 0.478 * z_3$$
$$G_3 = 0.689 * z_1 - 0.937 * z_2 + 0.347 * z_3$$

Іноді отримані головні компоненти досить важко пояснити, тому можна вдатися до процедури обертання або продовжити дослідження за допомогою факторного аналізу.

Розрахунок методу головних компонент досить громіздкий і реалізується через статистичні або математичні програми, наприклад через пакет SPSS для соціальних наук.

### 12.4. Реалізація МГК в програмі SPSS

Розглянемо на прикладі, як реалізується процедура знаходження головних компонент в статистичному пакеті SPSS для соціальних наук.

Виконаємо обов'язкову умову методу головних компонент – стандартизуємо вихідні показники. Процедуру стандартизації (знаходження z-scores) можна провести у вкладці Аналіз (Analyze) обравши модуль Дескриптивна статистика (Descriptive Statistics) → Описові (Descriptives) (рис. 12.8):

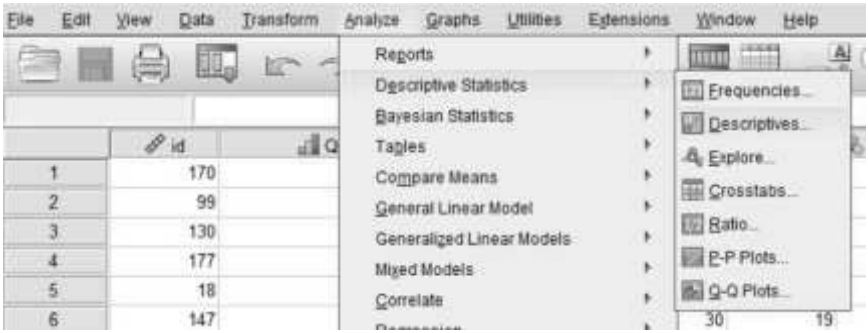



*Рисунок 12.8.* **Вибір модулю для стандартизації показників**

*Джерело*: власні розробки автора

На екрані з'явиться нове вікно, в якому потрібно обрати показникb для стандартизації та перенаправити їх стрілочкою у поле Variable(s). Стандартизувати можна як один показник так і всі відразу. Потрібно пам'ятати, що стандартизувати можна лише числові показники. Відмітьте галочкою поле Save standardized values as variables, та підтвердіть свої дії кнопкою ОК (рис. 12.9).

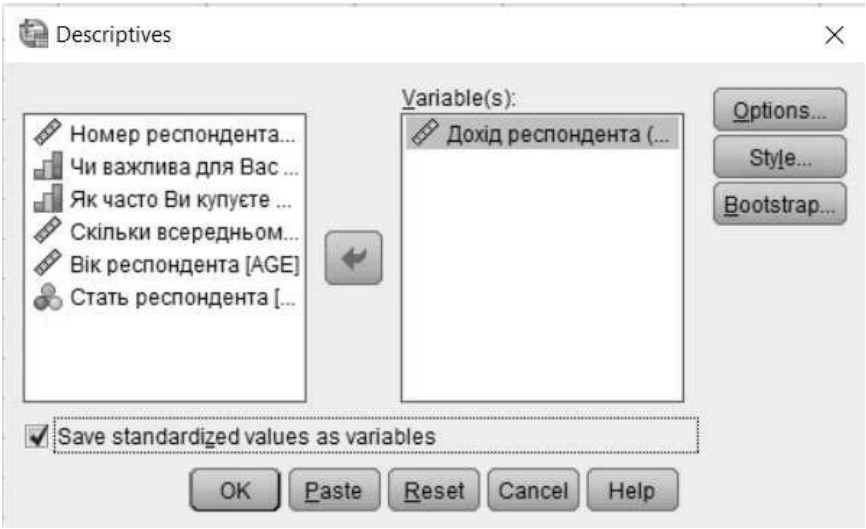



*Рисунок 12.9.* **Вибір показників для стандартизації**
*Джерело*: власні розробки автора

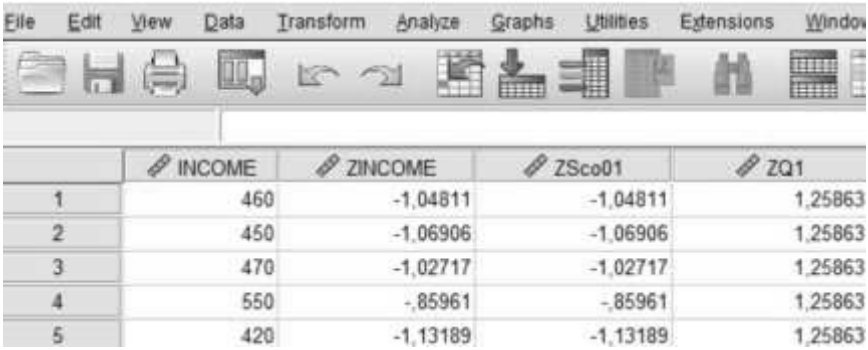
Після підтвердження вибору в новому вікні ви отримаєте показники описової статистики (рис. 12.10).

**Descriptive Statistics**

| | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| Дохід респондента (в у. о.) | 200 | 380 | 2950 | 960,42 | 477,453 |
| Valid N (listwise) | 200 | | | | |

*Рисунок 12.10.* **Показники описової статистики стандартизованого показника**
*Джерело*: власні розробки автора

Поряд з вихідними показниками у таблиці з'явиться стовпчик зі стандартизованими значеннями, в назві нового показника перед назвою додається Z (рис. 12.11).




| | INCOME | ZINCOME | ZSco01 | ZQ1 |
|---|---|---|---|---|
| 1 | 460 | -1,04811 | -1,04811 | 1,25863 |
| 2 | 450 | -1,06906 | -1,06906 | 1,25863 |
| 3 | 470 | -1,02717 | -1,02717 | 1,25863 |
| 4 | 550 | -,85961 | -,85961 | 1,25863 |
| 5 | 420 | -1,13189 | -1,13189 | 1,25863 |

*Рисунок 12.11.* **Значення стандартизованого показника**
*Джерело*: власні розробки автора

Стандартизація вихідних даних дозволяє перейти від різновимірних показників до уніфікованих, уникнути впливу розмірності на результати аналізу. Якщо ознакова множина має однакові одиниці виміру та розмірність, процедуру стандартизації можна пропустити.

Після отримання нормованих змінних можна переходити безпосередньо до аналізу головних компонент.

У вікні програми обираємо вкладку Аналіз (Analyze), модуль Зменшення розмірності даних (Dimension reduction) → Факторний аналіз (Factor). У новому вікні оберіть стандартизовані показники (Z…) та за допомогою стрілочки перемістіть їх у пусте вікно «Variables» (рис. 12.12).

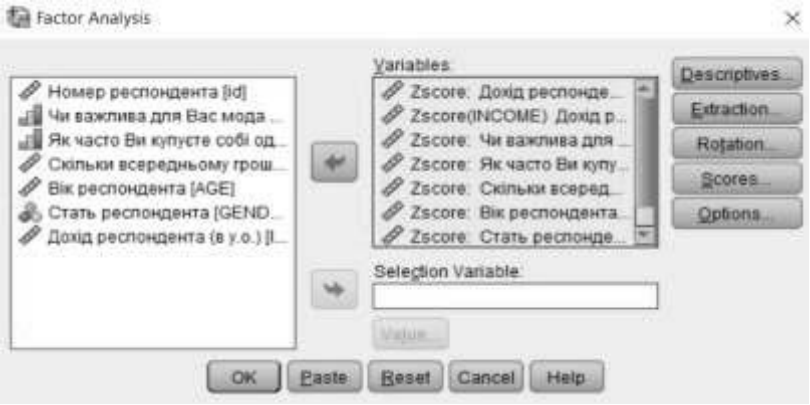



*Рисунок 12.12.* **Відбір ознак для аналізу**

*Джерело*: власні розробки автора

В розділі «Описові» (Descriptives) оберіть в кореляційних матрицях (Correlation Matrix) «Коефіцієнти» і «КМО та критерій сферичності Бартлетта», натисніть продовжити (рис. 12.13).

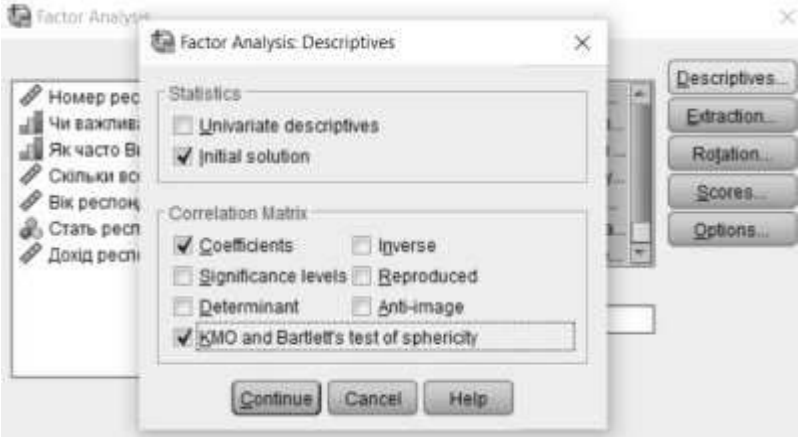



*Рисунок 12.13.* **Вибір кореляційної матриці та показників якості моделі**

*Джерело*: власні розробки автора

Переходимо до наступної владки «Вилучення» де обираємо Метод → Метод головних компонент (Principal components), Графік власних значень «кам'яного осипу» (Scree plot), та показник відбору кількості головних компонент – «Опираючись на власні значення» (Based on Eigenvalue). Натискаємо: продовжити (рис. 12.14).

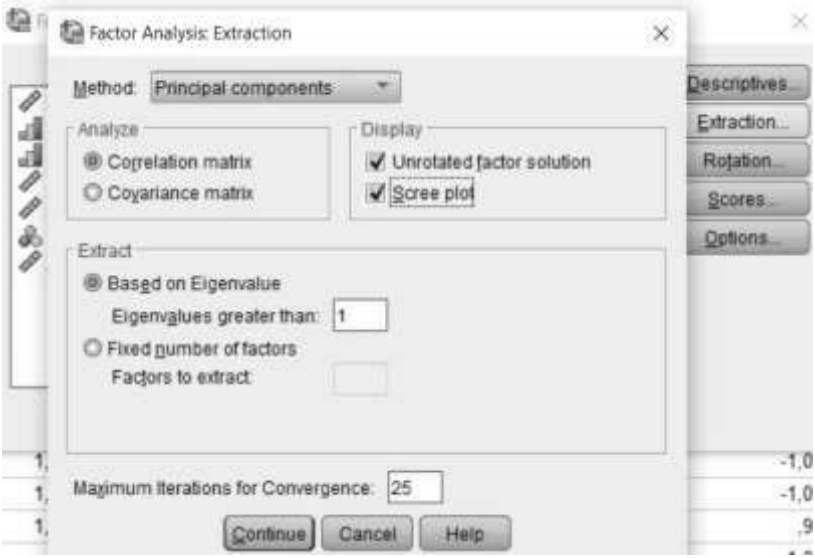



*Рисунок 12.14.* **Відбір кількості головних компонент**
*Джерело*: власні розробки автора

За потреби, в наступних вкладках можна обрати метод обертання («Обертання»), зберегти нові змінні для подальшого аналізу («Значення факторів») та вибрати метод представлення коефіцієнтів («Параметри»).

Після відбору всіх необхідних критеріїв, підтверджуємо кнопкою ОК і отримуємо в новому вікні результативну таблицю (рис. 12.15).

**Correlation Matrix**[a]

| | Zscore: Дохід респондента (в у.о.) | Zscore (INCOME) Дохід респондента (в у.о.) | Zscore: Чи важлива для Вас мода в одязі, взутті, аксесуарах? | Zscore: Як часто Ви купуєте собі одяг? | Zscore: Скільки всередньому грошей (в у.о.) Ви витрачаєте на одяг за один похід в магазин? | Zscore: Ви респондента | Zscore: Стать респондента |
|---|---|---|---|---|---|---|---|
| Zscore: Дохід респондента (в у.о.) | 1,000 | 1,000 | ,168 | ,242 | ,753 | ,200 | ,183 |
| Zscore(INCOME): Дохід респондента (в у.о.) | 1,000 | 1,000 | ,168 | ,242 | ,753 | ,200 | ,183 |
| Zscore: Чи важлива для Вас мода в одязі, взутті, аксесуарах? | ,168 | ,168 | 1,000 | ,514 | ,259 | -,553 | ,429 |
| Zscore: Як часто Ви купуєте собі одяг? | ,242 | ,242 | ,514 | 1,000 | ,177 | -,342 | ,320 |
| Zscore: Скільки всередньому грошей (в у.о.) Ви витрачаєте на одяг за один похід в магазин? | ,753 | ,753 | ,259 | ,177 | 1,000 | ,081 | ,234 |
| Zscore: Ви респондента | ,200 | ,200 | -,553 | -,342 | ,081 | 1,000 | -,315 |
| Zscore: Стать респондента | ,183 | ,183 | ,429 | ,320 | ,234 | -,315 | 1,000 |

*Рисунок 12.15*. **Кореляційна матриця**
*Джерело*: власні розробки автора

Критерії для оцінки якості моделі (рис. 12.16).

**KMO and Bartlett's Test**

| | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | ,639 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 402,843 |
| | df | 15 |
| | Sig. | ,000 |

*Рисунок 12.16*. **Оцінка якості моделі**
*Джерело*: власні розробки автора

Показники для відбору кількості головних компонент представлено рис. 12.17 та рис.12.18.

**Total Variance Explained**

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 3,008 | 42,973 | 42,973 | 3,008 | 42,973 | 42,973 |
| 2 | 2,072 | 29,594 | 72,568 | 2,072 | 29,594 | 72,568 |
| 3 | ,699 | 9,979 | 82,547 | | | |
| 4 | ,579 | 8,273 | 90,820 | | | |
| 5 | ,360 | 5,138 | 95,958 | | | |
| 6 | ,283 | 4,042 | 100,000 | | | |
| 7 | -5,557E-18 | -7,938E-17 | 100,000 | | | |

Extraction Method: Principal Component Analysis.

*Рисунок 12.17.* **Факторизація моделі**
*Джерело*: власні розробки автора

З рис. 12.18 можна побачити, що для подальшого аналізу та пояснення потрібно обрати дві головні компоненти.

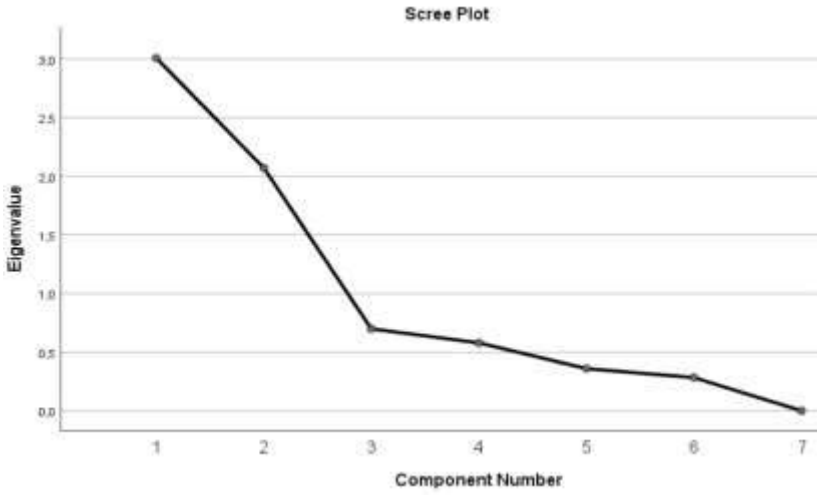



*Рисунок 12.18.* **Графік «кам'яного осипу»**
*Джерело*: власні розробки автора

На рис. 12.19 можна побачити матрицю факторних навантажень і оцінити вплив факторів в кожній компоненті.

| | Component | |
|---|---|---|
| | 1 | 2 |
| Чи важлива для Вас мода в одязі, взутті, аксесуарах? | ,804 | -,306 |
| Як часто Ви купуєте собі одяг? | ,707 | -,175 |
| Скільки всередньому грошей (в у.о.) Ви витрачаєте на одяг за один похід в магазин? | ,579 | ,702 |
| Вік респондента | -,511 | ,686 |
| Стать респондента | ,662 | -,137 |
| Дохід респондента (в у. о.) | ,523 | ,775 |

Extraction Method: Principal Component Analysis.

*Рисунок 12.19.* **Матриця факторних навантажень**

*Джерело*: власні розробки автора

Перша компонента має тісний зв'язок з першою, другою та п'ятою ознакою і її можна ідентифікувати як «схильність до моди». Друга пов'язана з третьою, четвертою та шостою ознакою і можна ідентифікувати як «матеріальний добробут та мода». Цікаво, що в першій компоненті відіграє важливу роль стать, то в другій – вік.

Статистичний пакет для соціальних наук SPSS дозволяє проаналізувати головні компоненти швидко та змістовно, використати додаткові процедури та показники якості моделі. На основі результатів дослідження є можливість продовження дослідження в інших модулях програми, або ж перенести результати МГК в потрібні офісні програми.

### *Список питань до самопідготовки:*

**1.** Яка головна мета методу головних компонент?

**2.** Чим відрізняється аналіз головних компонент від кластерного аналізу?

**3.** Як розраховуються стандартизовані змінні? В чому полягає сутність стандартизації і з якою метою її проводять?

**4.** В чому полягає сутність факторних навантажень?

**5.** З якою метою використовують критерії КМО та сферичності Бартлетта?

**6.** Які потрібно виконати кроки для знаходження головних компонент?

**7.** Які компоненти називають головними?

**8.** Які критерії використовують для визначення кількості головних компонент?

**9.** Що таке «графік кам'яного осипу» і з якою метою його використовують?

**10.** Для чого використовують методи обертання факторної структури? Назвіть декілька найбільш розповсюджених методів.

**11.** Чи є відмінності в поясненні компонент до та після обертання?

**12.** Процедура обертання в методі головних компонент використовується завжди?

**13.** За результатами факторного рішення власні значення п'яти компонент становили: $\lambda_1$ = 2,52; $\lambda_2$ =1,12; $\lambda_3$=0,85; $\lambda_4$ =0,42; $\lambda_5$ 0,09. Розрахуйте внесок кожної компоненти у сумарну дисперсію ознак, виокремте головні компоненти.

**14.** За результатами аналізу головних компонент на 5 показниках економічного стану фірми виділена одна компонента з дисперсією $\lambda$ = 3,64 і власним вектором $U$ = (0,9; 0,8; 1,1; 1,0; 0,9). Визначте факторні навантаження кожного показника та оцініть адекватність моделі головних компонент.

**15.** Власні значення трьох виокремлених компонент становили: : $\lambda_1$ = 1,319; $\lambda_2$ =0,949; $\lambda_3$=0,732. Розрахуйте показники повноти факторизації та виділіть за двома критеріями кількість головних компонент.

**16.** За результатами обстеження екологічної ситуації в містах, було проведено дослідження техногенного навантаження на оточуюче середовище і отримано такі показники:

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|---|
| 1 | 0,23 | 0,006 | 0,8 | 0,03 |
| 2 | 0,18 | 0,004 | 1,3 | 0,05 |
| 3 | 0,21 | 0,005 | 2,2 | 0,04 |
| 4 | 0,27 | 0,011 | 1,9 | 0,07 |
| 5 | 0,24 | 0,010 | 2,8 | 0,09 |
| 6 | 0,17 | 0,003 | 1,7 | 0,07 |
| 7 | 0,19 | 0,004 | 1,1 | 0,05 |

$x_1$ – середній рівень забруднення атмосферного повітря пилом (мг/м$^3$);

$x_2$- середній рівень забруднення атмосферного повітря ангідридом сірчаним (мг/м$^3$);
$x_3$ - середній рівень забруднення атмосферного повітря азотом діоксиду (мг/м$^3$);
$x_4$ - середній рівень забруднення атмосферного повітря окисом вуглецю (мг/м$^3$);

За наведеними даними проведіть аналіз головних компонент в програмі Statistica або SPSS. Визначити: 1) факторні навантаження кожного показника; 2) власні значення компонент та їх внесок в сумарну дисперсію; 3) надайте пояснення отриманим компонентам.

**17.** Аналіз показників фінансового стресу було проведено за даними динамічних рядів. Виокремлено три головні компоненти, факторні навантаження яких становлять:

| | Факторні навантаження | | |
|---|---|---|---|
| | $G_1$ | $G_2$ | $G_3$ |
| Індекс Доу-Джонса | 0,664 | 0,198 | 0,721 |
| Валютний курс | 0,635 | 0,341 | -0,687 |
| Ціна навфти | 0,376 | -0,921 | -0,092 |

1) Поясніть зміст факторних навантажень.
2) Розрахуйте внесок кожної компоненти в сумарну дисперсію.
3) Виокремити за одним з критеріїв головні компоненти.
4) Надайте економічне пояснення отриманим компонентам.

**18.** Використовуючи модель головних компонент, проаналізуйте рівень життя в регіонах. Для розрахунків використайте пакет Statistica або SPSS. Надайте економічну інтерпретацію отриманим головним компонентам.

| Показник | Регіон | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Кількість дітей, народжених жінкою за все життя | 1,7 | 1,5 | 1,4 | 1,1 | 1,3 | 1,8 | 1,2 |
| На 1000 осіб працездатного віку припадає непрацездатних | 752 | 825 | 874 | 900 | 849 | 924 | 851 |
| Коефіцієнт дитячої смертності | 12,1 | 13,9 | 12,7 | 12,9 | 12,1 | 11,2 | 12,8 |
| Очікувана тривалість життя, років | 73,4 | 68,5 | 69,7 | 70,5 | 65,3 | 73,8 | 68,3 |

**19.** За наведеними даними про зовнішньоторговельну діяльність з окремими країнами провести процедуру виділення головних компонент з метою покращення результатів кластерного аналізу.

| Країни | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| Німеччина | 1754,2 | 5445,0 | 538,5 | 357,7 | 1688,6 |
| Кіпр | 79,7 | 20,5 | 275,4 | 231,9 | 984,9 |
| Великобританія | 480,0 | 798,9 | 585,1 | 487,1 | 2088,8 |
| Польща | 2724,6 | 3453,8 | 296,3 | 150,2 | 789,9 |
| Нідерланди | 1676,1 | 643,7 | 170,5 | 119,4 | 6120,6 |
| Італія | 2469,5 | 1625,0 | 128,4 | 43,9 | 326,4 |
| Угорщина | 1326,4 | 1152,3 | 147,2 | 35,5 | 783,1 |
| Франція | 419,1 | 1563,8 | 117,8 | 83,5 | 1320,6 |

| | | | | | |
|---|---|---|---|---|---|
| Австрія | 535,1 | 484,5 | 161,6 | 109,1 | 1267,1 |
| Чехія | 715,2 | 869,5 | 71,2 | 65,2 | 111,9 |

$x_1$ – експорт товарів, млн ум. гр. од.;

$x_2$ - імпорт товарів, млн ум. гр. од.;

$x_3$ - експорт послуг, млн ум. гр. од.;

$x_4$ - імпорт послуг, млн ум. гр. од.;

$x_5$ - прямі інвестиції (акціонерний капітал) в Україну із країн ЄС, млн ум. гр. од.

Розрахувати:

1) провести процедуру стандартизації для вихідних показників;
2) факторні навантаження та пояснити їх зміст;
3) показники якості моделі та виділити найбільш значущі показники;
4) провести процедуру кластеризації з урахуванням отриманих нових змінних.

Надати змістовну економічну інтерпретацію результатам аналізу.

**20.** Знайти внесок головних компонент в сумарну дисперсію (за двома факторами разом та за кожним окремо), факторні навантаження яких становлять:

| Факторне навантаження | Показник | | | | | | |
|---|---|---|---|---|---|---|---|
| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ |
| $a_{1k}^2$ | 0,7 | 0,4 | 0,8 | 0,9 | 0,5 | 0,8 | 0,5 |
| $a_{1k}^2$ | 0,2 | 0,6 | 0,4 | 0,1 | 0,2 | 0,5 | 0,6 |

З метою покращення інтерпретації головних компонент проведіть процедуру факторного обертання методом варімакс. Оцініть адекватність моделі.

**21.** Діяльність компаній у сфері інтернет торгівлі характеризується показниками (балів);

| № | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| 1 | 85 | 76 | 80 | 80 | |
| 2 | 80 | 74 | 75 | 85 | |
| 3 | 75 | 79 | 41 | 90 | |
| 4 | 90 | 87 | 86 | 97 | |
| 5 | 70 | 71 | 57 | 80 | |
| 6 | 95 | 93 | 89 | 98 | |
| 7 | 83 | 75 | 74 | 80 | |
| 8 | 73 | 74 | 68 | 81 | |
| 9 | 86 | 75 | 54 | 82 | |
| 10 | 94 | 92 | 73 | 97 | |

$x_1$ – критерій користувальницької доступності веб-сайту[12];

$x_2$ – критерій релевантності інтернет магазину в пошукових системах;

$x_3$ – критерій змістовності веб-сайту;

$x_4$ – критерій маркетингової активності веб сайту;

$x_5$ – критерій лояльності покупців.

За даними таблиці: розрахувати кореляційну матрицю; визначити факторні навантаження та кількість головних компонент; показники якості моделі; зробити висновки.

[12] Меркулова, Т. В., Лубенець, С. В., & Янголенко, А. А. (2019). Комплексна оцінка ефективності інтернет-магазинів в електронній комерції. Вісник Харківського національного університету імені В. Н. Каразіна серія «Економічна», (96), 43-54. URL: https://periodicals.karazin.ua/economy/article/view/13373/12653

### *Список рекомендованої літератури по темі:*


1. Бізнес-аналітика багатовимірних процесів: навчальний посібник [Електронний ресурс] / Т. С. Клебанова, Л. С. Гур'янова, Л. О. Чаговець та ін. Харків : ХНЕУ ім. С. Кузнеця, 2018. 272 с.
2. Єріна А. М. Статистичне моделювання та прогнозування [Текст]: підручник / А. М. Єріна, Д. Л. Єрін; Держ. ВНЗ "Київ. нац. екон. ун-т ім. Вадима Гетьмана". Київ: КНЕУ, 2014. 348с.
3. Ковтун Н.В. Теорія статистики: підручник / Н. В. Ковтун. К.: Знання, 2012. 399 с.
4. Cronk B. C. (2019). How to Use SPSS: A Step-By-Step Guide to Analysis and Interpretation (11th Edition). Routledge. 228.
5. Field A. (2018). Discovering Statistics Using IBM SPSS Statistics. Fifth Edition. SAGE Publications Ltd. 1104.
6. George D., Mallery P. (2019). IBM SPSS Statistics 26 Step by Step: A Simple Guide and Reference (16th edition) . Routledge. 402.
7. Pallant J. (2016). SPSS Survival Manual (6th edition). NY:Open University Press. 368.
8. Zaino J. (2018). Adventures in Social Research: Data Analysis Using IBM SPSS Statistics (10th Edition). SAGE Publications, Inc, 512.
9. Terrell S.R. (2021). Statistics Translated. A Step-by-Step Guide to Analysing and Interpreting Data. Guilford Press. 433.

# Розділ 13. МОДЕЛІ ВІДНОВЛЕННЯ СТАТИСТИЧНИХ ДАНИХ

## 13.1. Причини, механізм породження і наслідки пропусків даних. Типи пропусків

Проблема пропусків даних нерозривно пов'язана з якістю результатів статистичного спостереження. Один із способів вирішення цієї проблеми полягає в тому, щоб повністю виключити записи, які мають пропуски, що призводить до зменшення вибірки, а отже, впливає на достовірність результатів. Водночас неправильна імпутація пропусків також може вплинути на довірчі інтервали. Наявність пропусків у даних, як і аналіз тільки повних спостережень (після виключення спостережень з пропусками), може призвести до отримання зміщених результатів і, як наслідок, – до викривлення висновків за результатами дослідження і прийняття неправильних рішень. До основних причин пропусків належать:

- виключення суб'єкта (об'єкта) із дослідження внаслідок невідповідності певним вимогам спостереження;
- наявність несприятливої події;
- відсутність результату у суб'єкта;
- відсутність реєстрації;
- помилки дослідників.

Можна визначити такі діапазони пропусків даних:

1. Менше 5% пропусків вважається несуттєвим і не впливає на результати дослідження.
2. Втрата 20% і більше даних ставлять під сумнів результати дослідження.

Отже, чим більша частка відсутніх даних, тим менша надійність висновків і тим складніше довести достовірність результатів дослідження.

Знання (або незнання) механізму, що приводить до відсутності значень, є ключовим при виборі методу аналізу та інтерпретації результатів. Іноді цей механізм управляється статистиком. Наприклад, ми можемо вважати, що вибірковому обстеженню притаманні пропуски, так як значення частини змінних в обстеженні (змінних плану) присутні у всіх об'єктів популяції, а досліджувані змінні «пропущені» у об'єктів, які не включені до вибірки. Тут механізм породження пропусків – процес вилучення вибірки. Якщо об'єкти витягуються з популяції випадково, то механізм управляється дослідником (при успішній реалізації плану) і його можна назвати «ігнорованим». Якщо правило вилучення вибірки не дотримується або для деяких об'єктів вибірки значення відсутні, то механізм породження пропусків не настільки ясний. У цьому випадку аналіз залежить від припущень про механізм утворення пропусків, які слід явно обумовлювати.

Метод подвійного вибору (double sampling) в теорії вибіркових обстежень – ще один приклад, коли структура пропусків підконтрольна досліднику. Витягується велика вибірка, і певні базові характеристики регулюються. Потім з цієї вибірки випадково витягується підвибірка, для якої вимірюються додаткові змінні. Одержані дані утворюють монотонну структуру. Методи регресії, які використовуються для аналізу таких даних, можна розглядати як методи обробки даних з пропусками, хоча зазвичай ці методи розглядають під іншим кутом зору.

Цензурування – приклад ситуації, коли механізм породження пропусків може бути некерованим, але відомим статистику. Прикладом є час настання події. Для деяких об'єктів вибірки час події цензуруваний, оскільки подія не встигла настати до закінчення експерименту. Якщо відома точка (час) цензурування, то ми маємо часткову інформацію про те, що час настання неспостережуваної події більше часу цензурування. Таку інформацію треба враховувати при аналізі, щоб уникнути зсувів.

Багато методів обробки явно не враховують механізм породження пропусків. Мається на увазі, що цей механізм ігнорується. Однак механізм пропусків можна вводити в статистичну модель, включаючи в неї розподіл індикаторів присутності, рівних 1 для присутнього значення і 0 – для пропуску. У загальному випадку механізмом пропусків можна знехтувати.

Існують різні способи вирішення проблем пропусків даних і найпростіший із них – це виключення суб'єкта з розрахунків. Однак наслідками такого підходу є скорочення обсягу вибірки, зниження суттєвості статистичних висновків і зміна довірчого інтервалу (наприклад, звуження внаслідок недооцінки варіативності). Саме тому при роботі з пропущеними даними важливою є ідентифікація характеру пропуску. Це, своєю чергою, вимагає застосування відповідних методів оброки даних з пропущеними значеннями: виключення, заповнення, зважування і моделювання. Всі ці методи дають різні результати за різних обсягів і характеру пропусків.

Досить низький рівень культури обробки даних з пропусками знаходить своє відображення у сучасному стані статистичного програмного забезпечення. Більшість статистичних програмних засобів, в яких передбачена можливість урахування наявності пропусків у даних, містить лише прості методи та їхні модифікації. Загалом ці методи малоефективні, призводять до зміщення оцінок результатів дослідження, а також до порушення рівня значущості критеріїв та інших викривлень статистичних висновків.

Перш ніж застосовувати будь-який метод обробки даних з пропусками необхідно оцінити з яким типом пропусків ми маємо справу.

Природа пропусків, їхня структура та тип включених в аналіз змінних є визначальними факторами при виборі методу та моделі аналізу даних із наявними пропусками.

Першим виміром постає тип пропусків, який характеризує структуру наявних даних по відношенню до відсутніх значень та

визначає конкретний розподіл відсутніх спостереження (що представлено матрицею M) за рядками ($i = 1, \ldots . n$) та змінними ($j = 1, \ldots . p$) що формують аналітичний набір даних.

Як правило, вирізняють 2 основні шаблони пропущених даних:

- Немонотонний (довільний);
- Монотонний шаблон пропущених даних.

На практиці найбільш поширеним є довільний (немонотонний) шаблон пропущених даних. Він характеризується відсутністю конкретної структури пропусків у наявних даних. У такому випадку відсутні значення розподіляються між рядками та змінними несистематично. Зазвичай дані із подібним шаблоном пропусків можна проаналізувати із використанням ланцюга Маркова Монте-Карло (MCMC) або за допомогою методу повної умовної специфікації (*англ.* – fully conditional specification – FCS).

Прикладом подібної структури пропусків може слугувати ситуація, коли вимірювання є відсутнім у конкретний момент часу, натомість у всі минулі чи наступні моменти часу дані присутні. У сфері клінічних досліджень дана ситуація можлива, коли певне значення параметру пацієнта, що оцінюється (тиск, пульс, кількість еритроцитів тощо) відсутнє у певний момент під час проходження лікування, натомість у період після його завершення під час контрольного візиту (*англ.* – follow-up visit) значення параметру, що є об'єктом дослідження, наявне.

Графічне зображення немонотонного шаблону пропусків наведено у табл. 13.1, де X – це наявні непропущені дані, а O – відсутні значення.

*Таблиця 13.1.* **Графічне зображення немонотонної структури пропусків**

| Номер спостереження | Змінна | | | |
|---|---|---|---|---|
| | V1 | V2 | V3 | V4 |
| 1 | X | O | O | X |
| 2 | X | X | O | X |
| 3 | O | X | X | O |
| 4 | X | O | O | X |
| 5 | O | X | X | X |

Наступним типом структури пропущених значень є монотонні пропуски.

Деякі процеси збору даних створюють більш структуровану або систематизовану схему пропусків. Наприклад, при медичному дослідженні, що складається з декількох фаз, відсутні дані можуть мати місце, коли для цілої фази збору даних (наприклад, взяття крові або отримання медичних записів), що є необхідними для подальшого дослідження, потрібна спеціальна згода пацієнта. Без його згоди щодо участі у цьому типі збору даних, усі змінні будуть мати пропущенні значення на цій фазі дослідження для конкретного пацієнта. Результатом подібної ситуації з точки зору структури даних є монотонна модель пропусків, подібна до тієї, що проілюстрована у табл. 13.2.

*Таблиця 13.2.* **Графічне зображення монотонної структури пропусків**

| Номер спостереження | Змінна | | | |
|---|---|---|---|---|
| | V1 | V2 | V3 | V4 |
| 1 | X | X | X | X |
| 2 | X | X | X | O |
| 3 | X | X | O | O |
| 4 | X | O | O | O |
| 5 | O | O | O | O |

Монотонні зразки відсутніх даних, як правило, аналізують за допомогою методів обчислення, що вимагають більш простих припущень, ніж підходи, які необхідні для ефективного оброблення даних, що мають немонотонну структуру пропусків. Графічно монотонну структуру пропусків можна представити у наступному вигляді.

Отже, структура пропущених значень є важливою характеристикою набору даних із пропусками, яка в подальшому суттєво впливає на вибір методу імпутації даних та складність цього процесу.

Виділяють три типи пропусків: MCAR (missing completely at random, повністю випадкові), MAR (missing at random, випадкові), MNAR (missing not at random, невипадкові).

Дані, відсутні повністю випадковим чином (MCAR), математично можна виразити у такий спосіб:

$$(R|Y_0, Y_m) = P(R),$$

де: $Y_0$ – наявні дані, $Y_m$ – відсутня частина матриці даних.

Це означає, що ймовірність відсутності даних однакова для всіх суб'єктів незалежно від спостережуваного аспекту в даних *Y*. Припущення про MCAR передбачає, що пропущеними даними можна знехтувати без упередженості, оскільки результати будуть подібними до результатів дослідження без пропусків.

Дані відсутні у випадковому порядку (MAR) характеризуються тим, що ймовірність такої відсутності залежить від наявних змінних, а не від відсутніх. Математично це можна записати так:

$$(R|Y_0, Y_m) = P(R|Y_0).$$

Концепція MAR передбачає, що якщо основні характеристики та проміжні вимірювання будуть схожі при усунені та комплектації,

то очікувані результати будуть аналогічні, тому відсутні результати можуть бути змодельовані на основі відомих результатів.

Дані вважаються пропущеними невипадково (MNAR), коли імовірність пропуску даних залежать від самих пропущених даних. Математично це можна записати у такий спосіб:

$$(R|Y_0, Y_m) = P(R|Y_m).$$

Цей механізм пропусків також називають неігнорованим, оскільки результати будуть зміщені, якщо ігнорується процес, який призводить до пропусків даних. Припущення щодо MNAR означає, що рішення про відмову може бути засноване на події, що не спостерігається, тому результати для відмов відрізняються від даних для учасників, які мають подібні характеристики.

### 13.2. Методи обробки даних з пропусками

Методи обробки даних з пропусками можна розділити на чотири групи: методи виключення, заповнення, зважування та методи, засновані на моделюванні. Метод виключення (обробка наявних даних) може бути задовільним за незначної кількості пропусків. Однак іноді він призводить до зсуву і зазвичай не є дуже ефективним.

Метод зважування передбачає, що заповнення пропусків у кожній підгрупі середніми у підгрупі та зважування наявних значень за їх часткою в кожній підгрупі ведуть до однакових оцінок середнього за сукупністю, проте оцінки вибіркової дисперсії при цьому різні.

Метод заповнення включає безумовну та умовну імпутацію середнього значення, заповнення пропусків з послідовним підбором, заповнення пропусків з упередженим підбором. При безумовній імпутації відсутні значення замінюються середнім з наявних значень для інших суб'єктів. Умовна імпутація середнього значення (метод Бака) передбачає заміну на умовні середні значення на основі

використання лінійних регресійних моделей оцінки пропущених значень. Заповнення пропусків з послідовним підбором (метод LOCF) застосовується у випадку, коли змінна, значення якої імпутуються, упорядкована. У цьому випадку пропущені значення замінюється значенням найближчого попереднього в цій послідовності суб'єкта. Заповнення пропусків з упередженим підбором (метод hot-deck) передбачає заміну всіх пропущених значень за допомогою відомих значень, взятих у іншого суб'єкта з аналогічними характеристиками.

В основі методу множинної імпутації лежить байєсівський підхід з використанням алгоритму Монте-Карло та ланцюжків Маркова. Сутність методу полягає в тому, щоб одночасно генерувати задану кількість значень пропущеної величини замість того, щоб замінювати пропущену інформацію одним значенням.

До більш простих методів належать: аналіз повних значень, метод поодинокої імпутації тощо.

*Аналіз повних значень.* Дана стратегія передбачає видалення спостереження, якщо будь-яка змінна містить пропуски, і проведення аналізу лише на основі наявних даних. Дана стратегія, як правило, використовується за замовчуванням у більшості статистичних пакетах.

*Переваги стратегії*: Вона може використовуватися при будь-якому виді статистичного аналізу, і для цього не потрібні спеціальні обчислювальні методи.

*Недоліки:* У випадку значної частки пропусків видалення великої кількості спостережень може привести до зміщених оцінок і неправильних висновків. Дана стратегія добре працює, коли дані мають повністю випадкових характер пропусків (MCAR), що рідко трапляється на практиці.

*Група методів поодинокої імпутації.* Суть даної стратегії в тому, що пропуски замінюються певним значенням наявних даних або значенням, розрахованим на основі наявних даних, в залежності

від конкретного підходу. Одним із найбільш популярних методів є *одинична імпутація на основі середнього значення*. У такому випадку всі пропуски змінної заповнюються розрахунковим значенням середньої.

У випадку здійснення повторюваних у часі вимірювань популярним підходом є *одинична імпутація на основі останнього значення* в межах кожної одиниці сукупності. У такому випадку пропущені дані замінюються останнім наявним непропущеним значенням для кожної окремої одиниці сукупності.

Ще однією варіацією даної стратегії виступає *заповнення пропусків на основі найгіршого значення* параметру отриманого в попередні чи наступні (у випадку немонотонній структурі пропусків) моменти часу для кожної одиниці сукупності. При застосуванні даного методу важливим є розуміння природи показника, що аналізується, і визначення напряму його зміни: зростання показника свідчить про кращий чи гірший результат? Даний підхід передбачає найбільш песимістичний сценарій, використання якого не завжди є виправданим.

*Метод умовної імпутації середнього* застосовується у випадку використання регресійної моделі, коли одна незалежна змінна, містить пропуски. У цьому випадку значення цієї змінної регресується по всім іншим незалежним змінним, і потім використовується рівняння для передбачення пропущених значень змінної.

Загальним недоліком усіх перерахованих простих методів є те, що вони призводять до заниження стандартних похибок, і, отже, завищення значень статистичних тестів. Основна причина цього полягає в тому, що розрахункові значення повністю визначаються моделлю, що застосовується до наявних даних, і не враховує мінливість відсутніх значень.

До більш складних методів належать: метод максимальної правдоподібності, метод імпутації hot-deck, множинна імпутація та інші.

*Метод максимальної правдоподібності* (*англ.* – Maximum Likelihood (ML)). Даний метод передбачає отримання матриці варіації-коваріації для змінних у моделі на основі всіх доступних даних, а потім використання цієї матриці для оцінки наявної регресійної моделі.

Існує два основних методи ML:

a.Алгоритм прямої максимальна правдоподібності передбачає пряму максимізацію багатовимірної функції нормальної правдоподібності для передбачуваної лінійної моделі. Перевагою підходу є отримання ефективних оцінок з коректними стандартними похибками. Недоліком є використання спеціалізованого програмного забезпечення, яке може бути складним та трудомістким.

b. Алгоритм очікуваної максимізації (*англ.* – Expectation-Maximization (EM)) надає оцінки середніх значень та матрицю коваріації, які можна використовувати для отримання послідовних оцінок параметрів, що цікавлять. Він базується на етапі очікування та етапі максимізації, які повторюються кілька разів до отримання оцінок максимальної ймовірності. Це вимагає великого обсягу вибірки і щоб дані випадково відсутні (MAR). Перевагою алгоритму є його досить просте використання у статистичних пакетах, яке не вимагає вказування додаткових опцій для аналізу. Недоліком методу є те, що він може використовуватися лише для лінійних та лог-лінійних моделей, адже за їх межами не розроблено ні теорії, ні програмного забезпечення для застосування алгоритму.

*Метод імпутації Hot-deck.* Даний підхід використовується у Бюро перепису населення США. Даний метод передбачає заповнення пропущених значень змінної для кожної конкретної одиниці сукупності на основі вибраної випадковим чином іншої

одиниці сукупності, що має однакові значення інших змінних. Іншими словами, пропуск заповнюється на основі наявного значення іншого суб'єкта, який є близьким до даного за набором інших характеристик.

*Множинна імпутація* даних є надзвичайно важливим інструментом при вирішенні проблеми наявності пропущених даних. Основними перевагами даного методу є:

1. Множинна імпутація базується на конкретній моделі, яка враховує особливості структури пропусків та тип змінних. Зазвичай модель заповнення пропусків має за основу явний розподіл, але використання і неявних методів (таких як метод найближчого сусіда) також може застосовуватися на практиці. Дана особливість може гарантувати статистичну прозорість та цілісність процесу імпутації.

2. Множинна імпутація є стохастичною. Даний інструмент заповнює пропуски на основі вибору параметрів моделі та похибок із передбачуваного розподілу відсутніх даних.

3. Множинна імпутація є багатовимірною. Це означає, що беруться до уваги не лише властивості розподілу непропущених значень кожної окремої змінної, але і взаємозв'язок між змінними, що включені в модель імпутації.

4. Даний метод використовує кілька незалежних повторень у процесі заповнення пропусків, що дає змогу оцінити мінливість оцінок параметрів, які пов’язані з імпутацією відсутніх значень. Це варіація, яка є доповненням до варіації вхідних даних та варіації вибіркової сукупності.

5. Множинна імутація є надійною проти несуттєвих відхилень від теоретичних припущень.

6. Даний механізм є досить поширеним у різних статистичних програмних пакетах (таких як SAS, STAT, R studio тощо), що робить застосування цього методу досить простим та зручним для користувача.

Множинна імпутація має такі ж оптимальні властивості, що і метод максимальної правдоподібності, але не вимагає деяких властивих йому припущень.

Недоліком даного методу є складність його використання та отримання різних результатів при використанні різних випадкових чисел.

Множинна імпутація являє собою не просто техніку для введення даних, вона є дієвим інструментом для отримання оцінок і правильних статистичних висновків – від результатів простої описової статистики до параметрів складних багатовимірних моделей.

Загальний підхід до аналізу даних із застосуванням множинної імпутації передбачає 3 основні етапи:

1. Створення $M$ наборів даних на основі конкретної моделі;
2. Аналіз $M$ наборів даних, згенерованих на попередньому кроці, за допомогою різних статистичних процедур та методів (ANCOVA, логістична регресія тощо) для кожного із $M$ наборів даних окремо;
3. Поєднання результатів із кроку 2 для отримання оцінок відповідного параметру.

*Крок 1* передбачає визначення змінних та формулювання припущень щодо розподілу пропусків у моделі заповнення відсутніх значень, та застосовує конкретні алгоритми для генерації імпутованих даних, а також виведення повних наборів даних для кожної із $m = 1 \dots M$ ітерацій процесу. Отже, цей крок можна розділити на такі 3 етапи: визначення моделі імпутації на основі розподілу даних, потім імпутація відсутніх значень та отримання $M$ повних наборів даних.

*Крок 2* полягає в оцінці параметрів, що є об'єктом аналізу, окремо по кожному набору даних. Модель, що використовується для аналізу на основі імпутованих значень, є аналогічною до моделі, у

якій використовуються лише наявні дані. Результати, отримані при кожному виді аналізу, будуть відрізнятися, оскільки дані, на основі яких проводиться оцінювання параметрів, є різними

*Крок 3* передбачає використання процедур чи алгоритмів об'єднання результатів, отриманих на попередньому кроці (у SAS даною процедурою є PROC MIANALYZE), що передбачає виведення агрегованих оцінок параметрів та їх стандартних похибок, довірчих інтервалів та інших результатів аналізу для перевірки гіпотез дослідження та для формулювання висновків на основі отриманих результатів.

Отже, першим кроком до проведення імпутації є визначення типу моделі, вибір якої відбувається у першу чергу в залежності від наявного шаблону відсутніх значень, а також з урахуванням типу змінної, що імпутується, і типу змінних, що включаються у модель. У таблиці 13.3 наведений перелік методів множинної імпутації в залежності від згаданих вище факторів.

*Таблиця 13.3.* **Методи імпутації даних з урахуванням структури пропусків та типів змінних**

| Структура пропущених даних | Тип змінної, значення якої імпутуються | Метод множинної імпутації |
|---|---|---|
| Монотонний | Неперервний | • Метод лінійної регресії,<br>• Метод співставлення із прогнозованим середнім значенням (*англ.* predictive mean matching) |
| | Бінарний/ порядковий | • Логістична регресія |
| | Номінальний | • Дискримінантна функція |

Продовження таблиці 13.3

| Структура пропущених даних | Тип змінної, значення якої імпутуються | Метод множинної імпутації |
|---|---|---|
| Немонотонний | Неперервний | Якщо незалежні змінні є неперервними:<br>• Метод Монте-Карло марковських ланцюгів із формуванням монотонного типу пропусків<br>• Метод Монте-Карло марковських ланцюгів із імпутацією повних даних |
| | Неперервний | Якщо незалежні змінні моделі мають змішаний тип<br>• Регресія на основі методу повної умовної специфікації<br>• Метод співставлення із прогнозованим середнім на основі методу повної умовної специфікації |
| | Бінарний/ порядковий | • Логістична регресія на основі методу повної умовної специфікації |
| | Номінальний | • Дискримінантна функція основі методу повної умовної специфікації |

*Метод лінійної регресії* є базовим методом, який використовується при імпутуванні пропущених значень для монотонних та немонотонних пропусків. Останній варіант передбачає незначні перетворення для урахування незначних ітераційних циклів алгоритму. Даний метод складається із двох кроків: *P*-кроку (від англ. predict – передбачати) та *I*-кроку (від англ.

iterate – повторювати). Перший крок базується на отриманні оцінок із застосуванням стандартної регресійної моделі:

$$Y_j = \beta_0 + \beta_1 X_1 + \beta_n X_n + \varepsilon. \quad (13.1)$$

На основі цього рівняння будуть отримані поточні оцінки параметрів регресії $\beta = \{\beta_0, \beta_1, \beta_n\}$ та залишкова дисперсія $\widehat{\sigma_j^2}$, а також $V_j$ – обернена матриця суми квадратів та перехресних добутків з регресії $Y_j$ на основі $Y_1$, $Y_n$. Ці оцінки регресії визначають апостеріорний розподіл для параметрів моделі регресії.

Базуючись на оцінках апостеріорних параметрів регресійної моделі, імпутація на $I$-кроці заповнює $Y_j$. Перший крок у цій імпутації полягає в тому, щоб витягти значення випадкових параметрів з їх спільного апостеріорного розподілу. При цьому спершу значення для залишкової дисперсії береться з його апостеріорного (за умови, що апріорне неінформативний) розподілу:

$$\sigma_{*j}^2 = \frac{\widehat{\sigma_j^2}(n_j - k - 1)}{g}, \quad (13.2)$$

де $n_j$ – кількість непропущених значень для змінної, яку імпутують;

$k$ – кількість параметрів (за винятком вільного члена рівняння) у моделі;

$g$ – випадкове значення із χ-квадрат розподілу.

Після цього коефіцієнти регресії отримують як:

$$\beta_* = \hat{\beta} + \sigma_{*j}^2 \sqrt{V} Z, \quad (13.3)$$

де $\sqrt{V}$ – квадратний корінь із верхнього трикутника при розкладанні Холеського;

$Z - k + 1$-вимірний вектор незалежних випадкових нормально розподілених змінних.

Після цього відсутні значення замінюються на:

$$Y_{j*} = \beta_{0*} + \beta_{1*} X_1 + \beta_{n*} X_n + z\sigma_{*j}, \quad (13.4)$$

де $X_1, X_2$ – значення незалежних змінних;

$Z$ –змодельоване нормальне відхилення.

*Метод співставлення із прогнозованим середнім* (англ. Predictive mean matching) є іншим альтернативним методом множинної імпутації даних, який використовується для заповнення пропущених значень неперервної змінної. Він подібний до методу регресії, за винятком того, що для кожного відсутнього значення визначається непропущене спостережуване значення, яке є найближчим до прогнозованого зі згенерованої моделі регресії. Метод співставлення із прогнозованим середнім гарантує, що імпутовані значення є правдоподібними, і даний метод може бути більш доречними, ніж лінійна регресія, якщо припущення про нормальність порушено.

*Метод логістичної регресії* використовується для імпутації бінарних (0/1) або порядкових класифікаційних змінних. Фактичний процес заповнення пропущених значень відбувається за аналогією до простої лінійної регресії, описаної вище, із використанням 2-крокового алгоритму ($P$-крок та $I$-крок)

На $P$-етапі логістична регресія застосовується до непропущених значень залежної та незалежних змінних, які включаються до моделі. У результаті отримується відповідна модель логістичної регресії для оцінювання ймовірності того, що відсутнє значення належить до кожної категорії бінарної або порядкової змінної. Описана модель має вигляд:

$$logit\left(p\left(Y_j = 1\right)\right) = \log\left(\frac{p}{1-p}\right) = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_n \qquad (13.5)$$

Описана модель логістичної регресії дає оцінки параметрів логістичної регресії $\beta = \{\beta_0, \beta_1, \beta_n\}$ та матрицю коваріації для параметрів $V$. Як і при лінійній регресії, апостеріорний розподіл параметрів логістичної регресії вважається багатовимірним нормальним. Для інтерпретації результатів проводиться обернене перетворення вигляду:

$$p\left(Y_j = 1\right) = \frac{\exp(Y_{1:n}\beta_*)}{1-\exp(Y_{1:n}\beta_*)} \ . \qquad (13.6)$$

*Метод Монте-Карло марковських ланцюгів* (*англ.* – Markov chain Monte Carlo (MCMC)). У ситуації, коли проблема відсутніх даних є багатовимірною, наявна немонотонна структура відсутніх значень, і модель може включати змінні різного типу (безперервні, номінальні, бінарні, порядкові), аналітично важко або неможливо оцінити справжнє вираження сумісного апостеріорного розподілу, $p(\theta|Y_{obs})$. Саме для таких випадків статистиками були розроблені ітеративні методи моделювання, які б дозволили б апроксимувати отримані результати.

Ланцюг Маркова являє собою послідовність випадкових величин, в якій розподіл кожного елемента залежить від значення попереднього. У MCMC створюється ланцюг Маркова, достатньо довгий, щоб розподіл елементів стабілізувався до загального розподілу. Шляхом багаторазового моделювання етапів ланцюга, відбувається моделювання оцінки на основі стаціонарного розподілу. У байєсівському висновку інформація про невідомі параметри виражається у формі апостеріорного розподілу ймовірностей. MCMC застосовується як метод дослідження апостеріорних розподілів за байєсівським висновком. Тобто за допомогою MCMC можна змоделювати весь сумісний апостеріорний розподіл невідомих величин і отримати оцінки апостеріорних параметрів, які представляють інтерес.

Припускаючи, що дані відносяться до багатовимірного нормального розподілу, процес імпутації даних відбувається на основі повторення наступних кроків:

• $I$-крок: на основі вектору оціненого середнього значення та матриці коваріацій відбувається моделювання відсутніх значення для кожного окремого спостереження незалежно. Тобто, на даному кроці відбувається моделювання $Y_{mis}$ з умовного розподілу $Y_{mis}$ з урахуванням $Y_{obs}$.

• $P$-крок: на цьому етапі відбувається моделювання вектору апостеріорного середнього значення генеральної сукупності і

матриці коваріацій на основі оцінок з повної вибірки. Потім ці нові оцінки знову використовуються на $I$-кроці.

Отже, кожен із зазначених методів застосовується виходячи зі структури наявних пропусків та типу змінних.

Існують також певні методи, які використовуються при аналізі чутливості для того, щоб пересвідчитись, що отримані в ході первинного аналізу результати задовільнять цілі дослідження, а обраний метод чи модель дійсно описує реальну ситуацію дослідження. До таких методів належать, наприклад, модель змішаної структури (англ. – Pattern mixture model) та так звана модель переломної точки (англ. – Tipping point analysis).

Наступним кроком після визначення вигляду моделі постає вибір кількості необхідних імпутацій. Найбільш поширеним підходом до даного процесу є використання поняття відносної ефективності множинної імпутації.

Рубіном у 1987 році було встановлено, що відносна ефективність $(RE)$ використання скінченої кількості імпутацій $(m)$ замість нескінченого числа для отримання повністю ефективної імпутації в одинцях дисперсії приблизно залежить від $m$ та величини $\lambda$ у наступному співвідношенні:

$$RE = (1 + \frac{\lambda}{\mathrm{m}})^{-1}, \qquad (13.7)$$

де $m$ −кількість імпутацій;

$\lambda$ − частка пропущених даних.

На основі проведених ним розрахунків, було отримані показники відносної ефективності імпутації, наведені у таблиці 13.4.

*Таблиця 13.4.* **Значення відносної ефективності імпутації в залежності від частки відсутніх даних та кількості імпутацій**

| m | λ | | | | |
|---|---|---|---|---|---|
| | 10% | 20% | 30% | 50% | 70% |
| 3 | 0,9677 | 0,9375 | 0,9031 | 0,8571 | 0,8108 |
| 5 | 0,9804 | 0,9615 | 0,9434 | 0,9091 | 0,8772 |
| 10 | 0,9901 | 0,9804 | 0,9709 | 0,9524 | 0,9346 |
| 20 | 0,9950 | 0,9901 | 0,9852 | 0,9756 | 0,9662 |

Тобто, дослідник, вибираючи кількість необхідних імпутацій даних у дослідженні виходить із припущення про приблизний відсоток пропусків у даних, що будуть аналізуватися, і прийнятного для дослідження рівня відносної ефективності.

З таблиці видно, що якщо частка відсутньої інформації незначна, потрібна лише невелика кількість ітерацій заповнення пропущених значень. Наприклад, якщо частка пропусків становить усього 30%, то для забезпечення ефективності у 91% потрібно лише три імпутації, а для ефективності у 94% – п’ять імпутацій.

## 13.3. Практика використання методів обробки даних з пропусками

Розглянемо результати використання різних методів імпутації, які були протестовані на прикладі вибірки суб’єктів, хворих на анорексію (дод. 13.1). Вибірка включає 72 пацієнта. Стратифікаційна змінна – вид лікування. Залежна зміна – вага після лікування. Незалежна змінна – вага до лікування.

Для простоти аналізу когнітивне поведінкове лікування та сімейне лікування були об'єднані в одну групу, позначену методом лікування 1, частка яких становить 36,11% і контрольна група була позначена 0, частка яких становить 63,89%.

Описова статистика бази даних для дослідження наведена у табл. 13.5, що є стандартним способом узагальнення даних у клінічних випробуваннях.

Отже, середня вага до лікування становила 37,38 кг зі стандартним відхиленням ± 2,351 кг. Середня вага після лікування збільшилась до 38,63 кг зі стандартним відхиленням ± 3,645 кг. Для того, щоб дослідити, як група лікування (treatm) та вага до дослідження (prewgt) вплинули на вагу після дослідження, була побудована модель лінійної регресії для оригінальних даних (табл. 13.5).

*Таблиця 13.5.* **Описова статистика для даних**

| | Вага до лікування, кг | Вага після лікування, кг |
|---|---|---|
| Кількість осіб, n | 72 | 72 |
| Середнє вага, $\bar{x}$ | 37,38 | 38,63 |
| Стандартне відхилення, σ | 2,351 | 3,645 |
| Мінімум | 31,8 | 32,3 |
| Медіана | 37,33 | 38,12 |
| Максимум | 43,0 | 47,0 |

Для цієї вибірки програмними засобами були симульовані різні варіанти пропусків даних (дод. 13.3 – 13.7). Мета дослідження – оцінювання ефективності лікування анорексії. Для того, щоб проаналізувати, як група лікування та вага до дослідження визначають вагу після дослідження, була побудована модель лінійної регресії для оригінальних даних (табл. 13.6, 13.7):

*Таблиця 13.6.* **Параметри кореляційно-регресійної моделі впливу виду лікування та ваги до лікування на вагу після лікування**

| Компоненти моделі | Коефіцієнт | Стандартна помилка | *t*-критерій Стьюдента | *p*-value |
|---|---|---|---|---|
| Вільний член | 20,2 | 6,143 | 3,29 | 0,0016 |
| Група лікування | 2,61 | 0,803 | 3,26 | 0,0017 |
| Вага до лікування | 0,45 | 0,165 | 2,72 | 0,0084 |

Отже, рівняння регресії має вигляд:

$$Y = 20{,}2 + 2{,}61x_1 + 0{,}45x_2,$$

де: $y$ – вага після лікування; $x_1$ – група лікування; $x_2$ – вага до лікування.

*Таблиця 13.7.* **Показники адекватності кореляційно-регресійної моделі**

| Показник | Ступені свободи | Сума квадратів відхилень | Дисперсія | *F*-критерій Фішера | *p*-value | Середньо-квадратична помилка, кг | Коефіцієнт варіації, % | $R^2$ |
|---|---|---|---|---|---|---|---|---|
| Регресія | 2 | 216,0 | 108,02 | 10,25 | 0,0001 | x | x | x |
| Залишок | 69 | 727,1 | 10,54 | x | x | x | x | x |
| Усього | 71 | 943,1 | x | x | x | 3,246 | 8,402 | 0,229 |

Усі змінні були статистично значущими, оскільки *p*-value нижче за 0,05. Середньоквадратична помилка моделі становить 3,25,

тобто з імовірністю 68% реальна величина пост-ваги з прогнозу становить ±3,25 кг. Включені в модель змінні пояснюють 22,91% від загальної варіації результуючого показника.

Регресійна модель, побудована на цьому масиві, слугувала базою для порівняння моделей, отриманих різними методами на основі імпутованих даних. Більш ефективним буде вважатися той метод імпутації, який забезпечить максимальну подібність результатів моделювання.

Результати відновлення даних MCAR при 10% пропусків представлені в табл. 13.8, 13.9 (тут і надалі у дужках – стандартна помилка).

*Таблиця 13.8.* **Результати застосування різних методів при 10% пропусків даних**

| Метод | Вільний член рівняння | Вага до лікування | | Група лікування | |
|---|---|---|---|---|---|
| | | Оцінка параметра | *p*-value | Оцінка параметра | *p*-value |
| Повні дані | 20,20 (6,14) | 0,45 (0,17) | 0,0084 | 2,61 (0,80) | 0,0017 |
| Аналіз повних випадків | 21,63 (6,51) | 0,41 (0,17) | 0,0221 | 2,74 (0,88) | 0,0028 |
| Аналіз наявних випадків | 21,57 (6,66) | 0,41 (0,18) | 0,024 | 2,69 (0,87) | 0,003 |
| Заповнення безумовним середнім | 22,90 (6,08) | 0,38 (0,16) | 0,0215 | 2,38 (0,79) | 0,0038 |
| Заповнення умовним середнім | 18,76 (6,46) | 0,49 (0,17) | 0,006 | 2,47 (0,84) | 0,0046 |

Продовження таблиці 13.8

| Метод | Вільний член рівняння | Вага до лікування | | Група лікування | |
|---|---|---|---|---|---|
| | | Оцінка параметра | *p*-value | Оцінка параметра | *p*-value |
| Заповнення пропусків з послідовним підбором | 20,71 (6,06) | 0,44 (0,16) | 0,0093 | 2,57 (0,79) | 0,0018 |
| Заповнення пропусків з упередженим підбором | 24,89 (6,54) | 0,33 (0,18) | 0,0647 | 2,30 (0,85) | 0,0088 |

Середньоквадратичні помилки, розраховані за методом безумовного середнього та за методом заповнення пропусків з послідовним підбором, були найближчими до оригінальної моделі, всі інші методи завищували цю оцінку.

Найбільш схожий коефіцієнт детермінації з вихідними даними, як це видно з табл. 13.9, мав місце при застосуванні методу заповнення пропусків з послідовним підбором. Зміна стандартних помилок була мінімальною, *p*-value – більшими за всіма методами, за винятком методу заповнення умовним середнім для змінної “вага до лікування”. Також слід зазначити, що при застосуванні методу заповнення пропусків з упередженим підбором цей фактор став неістотним. Оцінки параметрів, *p*-value та коефіцієнт детермінації, визначені в результаті застосування методу заповнення пропусків з послідовним підбором, були найближчими до значень, отриманих при побудові регресії на повних даних.

*Таблиця 13.9.* **Середньоквадратична помилка та коефіцієнт детермінації для різних методів при 10% пропусків даних**

| Метод | Середньоквадратична помилка, кг | Коефіцієнт детермінації, % |
|---|---|---|
| Повні дані | 3,25 | 22,9 |
| Аналіз повних випадків | 3,34 | 21,6 |
| Аналіз наявних випадків | 3,34 | 21,4 |
| Заповнення безумовним середнім | 3,21 | 19,3 |
| Заповнення умовним середнім | 3,41 | 21,6 |
| Заповнення пропусків з послідовним підбором | 3,20 | 22,6 |
| Заповнення пропусків з пристрасним підбором | 3,45 | 15,1 |

Результати імпутації даних при 25% пропусків представлені в табл. 13.10, 13.11.

*Таблиця 13.10.* **Результати застосування різних методів при 25% пропусків даних**

| Метод | Вільний член рівняння | Вага до лікування | | Група лікування | |
|---|---|---|---|---|---|
| | | Оцінка параметра | *p*-value | Оцінка параметра | *p*-value |
| Повні дані | 20,20 (6,14) | 0,45 (0,17) | 0,0084 | 2,61 (0,80) | 0,0017 |
| Аналіз повних випадків | 15,97 (7,60) | 0,57 (0,20) | 0,0071 | 2,06 (0,96) | 0,0366 |
| Аналіз наявних випадків | 17,79 (7,14) | 0,53 (0,19) | 0,0082 | 1,94 (0,93) | 0,0429 |
| Заповнення безумовним середнім | 24,11 (5,53) | 0,37 (0,15) | 0,0156 | 1,41 (0,72) | 0,055 |
| Заповнення умовним середнім | 14,44 (6,05) | 0,63 (0,16) | 0,0002 | 0,96 (0,79) | 0,2299 |
| Заповнення пропусків з послідовним підбором | 10,60 (5,50) | 0,71 (0,15) | <,0001 | 1,67 (0,72) | 0,0227 |
| Заповнення пропусків з пристрасним підбором | 20,42 (6,77) | 0,47 (0,18) | 0,0116 | 1,87 (0,88) | 0,0036 |

*Таблиця 13.11.* **Середньоквадратична помилка та коефіцієнт детермінації для різних методів при 25% пропусків даних**

| Метод | Середньоквадратична помилка, кг | Коефіцієнт детермінації,% |
|---|---|---|
| Повні дані | 3,25 | 22,91 |
| Аналіз повних випадків | 3,25 | 21,37 |
| Аналіз наявних випадків | 3,26 | 20,95 |
| Заповнення безумовним середнім | 2,92 | 14,11 |
| Заповнення умовним середнім | 3,20 | 20,80 |
| Заповнення пропусків з послідовним підбором | 2,90 | 31,77 |
| Заповнення пропусків з пристрасним підбором | 3,57 | 15,61 |

Найбільша недооцінка середньоквадратичної помилки мала місце при застосуванні методів заповнення пропусків з послідовним підбором та заповнення безумовним середнім. Для коефіцієнта детермінації найменша оцінка була при застосуванні методу заповнення безумовним середнім значенням, а переоцінка – найбільшою при методі заповнення пропусків з послідовним підбором. Фактор "вага до лікування" залишався статистично значущим, тоді як фактор "група лікування" став неістотним при застосуванні методів заповнення умовним середнім значенням і безумовним середнім значенням. Для змінної "група лікування" оцінка параметра значно змінилася, особливо при методі заповнення умовним середнім, коли вона була майже втричі меншою. Найбільша

різниця (зі зміною 0,26) з оцінкою ваги до лікування відбулася при застосуванні методу заповнення пропусків із послідовним підбором.

Отримані оцінки параметрів та *p*-value в результаті застосування методу "аналіз наявних випадків" були більш наближені до значень, отриманих при побудові регресії на повних даних. Вплив пропущених даних на результати аналізу найкраще проявляється при 50% відсутніх даних (табл. 13.12, 13.13).

*Таблиця 13.12.* **Результати застосування різних методів при 50% пропусків даних**

| Метод | Вільний член рівняння | Вага до лікування | | Група лікування | |
|---|---|---|---|---|---|
| | | Оцінка параметра | *p*-value | Оцінка параметра | *p*-value |
| A | B | C | D | E | F |
| Повні дані | 20,20 (6,14) | 0,45 (0,17) | 0,0084 | 2,61 (0,80) | 0,0017 |
| Аналіз повних випадків | 23,67 (7,37) | 0,37 (0,20) | 0,0754 | 3,40 (1,12) | 0,0046 |
| Аналіз наявних випадків | 14,48 (7,79) | 0,60 (0,21) | 0,0069 | 3,94 (1,02) | 0,0005 |
| Заповнення безумовним середнім | 24,85 (4,36) | 0,36 (0,12) | 0,0031 | 1,97 (0,57) | 0,0009 |
| Заповнення умовним середнім | 20,61 (6,73) | 0,44 (0,18) | 0,0175 | 3,38 (0,88) | 0,0003 |
| Заповнення пропусків з послідовним підбором | 10,65 (4.97) | 0,72 (0,13) | <0,0001 | 1,04 (0.65) | 0,1143 |

Продовження таблиці 13.12

| A | B | C | D | E | F |
|---|---|---|---|---|---|
| Заповнення пропусків з пристрасним підбором | 23,01 (6,13) | 0,38 (0,16) | 0,0214 | 2,49 (0,80) | 0,0027 |

При застосуванні методу аналізу повних випадків вага до лікування стала статистично неістотною, причому *p*-значення майже в 9 разів перевищує вихідний набір даних. Фактор "група лікування" став неістотним при застосуванні методу заповнення пропусків з послідовним підбором, *p*-значення зменшувалося при застосуванні інших методів. Найбільш точна оцінка ваги до лікування була отримана в результаті методу умовного середнього (0,44), найменш схожа з оригіналом оцінка спостерігалася за підходу "заповнення пропусків з послідовним підбором". Для змінної "група лікування" оцінки значно відрізнялися.

Якщо заповнення безумовним середнім та оцінки за методом заповнення пропусків з послідовним підбором були меншими (на 0,64 та 1,57 відповідно), то використання інших методів призвело до надмірної переоцінки ефекту лікування. Оцінка вільного члена рівняння при методі заповнення пропусків з послідовним підбором зменшилась удвічі, а при методі аналізу повних випадків – в 1,4 рази.

Єдиним методом, який переоцінював середньоквадратичну помилку, було заповнення умовним середнім; за інших методів середньоквадратична помилка недооцінена. При методі аналізу наявних випадків коефіцієнт детермінації майже вдвічі перевищував результат початкового аналізу. Виділимо метод заповнення безумовним середнім, результати застосування якого найбільш наближені до моделі, побудованої на повних даних (див. табл. 13.8).

*Таблиця 13.13.* **Середньоквадратична помилка та коефіцієнт детермінації для різних методів при 50% пропусків даних**

| Метод | Середньоквадратична помилка, кг | Коефіцієнт детермінації, % |
|---|---|---|
| Повні дані | 3,25 | 22,91 |
| Аналіз повних випадків | 3,08 | 36,96 |
| Аналіз наявних випадків | 2,89 | 44,48 |
| Заповнення безумовним середнім | 2,31 | 26,14 |
| Заповнення умовним середнім | 3,56 | 25,32 |
| Заповнення пропусків з послідовним підбором | 2,63 | 33,43 |
| Заповнення пропусків з пристрасним підбором | 3,24 | 20,06 |

Результати застосування методів імпутації для пропущених даних MAR для 10% пропусків представлено в табл. 13.14. Зміна в оцінці параметрів для вільного члена та двох інших факторів була мінімальною. Для факторів у регресійній моделі *p*-value знизилися лише при застосуванні методу максимальної правдоподібності. Для двох інших методів значення *p*-value зросло, проте не настільки, щоб зробити ці фактори не значущими.

*Таблиця 13.14.* **Результати застосування різних методів при 10% пропусків даних**

| Метод | Вільний член рівняння | Вага до лікування | | Група лікування | |
|---|---|---|---|---|---|
| | | Оцінка параметра | *p*-value | Оцінка параметра | *p*-value |
| Повні дані | 20,20 (6,14) | 0,45 (0,17) | 0,0084 | 2,61 (0,80) | 0,0017 |

Продовження таблиці 13.14

| Метод | Вільний член рівняння | Вага до лікування | | Група лікування | |
|---|---|---|---|---|---|
| | | Оцінка параметра | *p*-value | Оцінка параметра | *p*-value |
| Метод зважування | 21,68 (6,62) | 0,42 (0,18) | 0,0237 | 2,53 (0,86) | 0,0046 |
| Множинна імпутація | 20,62 (6,48) | 0,45 (0,18) | 0,0111 | 2,45 (0,85) | 0,0042 |
| Метод максимальної правдоподібності | 20,98 (5,75) | 0,44 (0,15) | 0,0059 | 2,50 (0,75) | 0,0014 |

Середньоквадратична помилка та коефіцієнт детермінації не були реалізовані в методі множинної імпутації, тому вони були обраховані як середнє для кожної імпутації. Саме при застосуванні методу множинної імпутації середньоквадратична помилка та коефіцієнт детермінації були максимально близькі до оригінальних значень, обрахованих за повними даними (табл. 13.15).

*Таблиця 13.15.* **Середньоквадратична помилка та коефіцієнт детермінації для різних методів при 10% пропусків даних**

| Метод | Середньоквадратична помилка, кг | Коефіцієнт детермінації, % |
|---|---|---|
| Повні дані | 3,25 | 22,91 |
| Метод зважування | 3,35 | 24,15 |
| Множинна імпутація | 3,20 | 22,38 |
| Метод максимальної правдоподібності | 3,04 | 24,05 |

Найбільш точна оцінка вільного члена рівняння за 50% пропусків була визначена при застосуванні методу максимальної

правдоподібності, при цьому була знижена стандартна помилка (табл. 13.16). При застосуванні методу множинної імпутації фактор "вага до лікування" став статистично не значущим, а фактор "група лікування" є статистично не значущим при застосуванні методу зважування.

*Таблиця 13.16.* **Результати застосування різних методів при 50% пропусків даних**

| Метод | Вільний член рівняння | Вага до лікування | | Група лікування | |
|---|---|---|---|---|---|
| | | Оцінка параметра | *p*-value | Оцінка параметра | *p*-value |
| Повні дані | 20,20 (6,14) | 0,45 (0,17) | 0,0084 | 2,61 (0,80) | 0,0017 |
| Метод зважування | 18,86 (8,75) | 0,51 (0,24) | 0,0448 | 1,74 (1,31) | 0,1923 |
| Множинна імпутація | 21,07 (8,91) | 0,45 (0,25) | 0,0769 | 1,94 (1,02) | 0,0588 |
| Метод максимальної правдоподібності | 19,92 (4,65) | 0,48 (0,13) | 0,0003 | 1,78 (0,61) | 0,0046 |

Табл. 13.17 показує, що при застосування методу зважування значно (майже у два рази) переоцінюється середньоквадратична помилка, водночас коефіцієнт детермінації майже близький до оригінального значення. Незважаючи на те, що середньоквадратична помилка була знижена при застосуванні методу максимальної правдоподібності, коефіцієнт детермінації був переоцінений.

Отже, при 10% пропусків MCAR оцінки параметрів і *p*-value для двох факторів, отримані із застосуванням першої групи методів, наближені до результатів, отриманих на повних даних. Середньоквадратичні помилки, розраховані за методом безумовного середнього і методом заповнення пропусків з послідовним підбором,

близькі до результатів, отриманих на оригінальних даних, всі інші методи завищували цю оцінку. Найбільш схожим з вихідними даними був коефіцієнт детермінації при застосуванні методу заповнення пропусків з послідовним підбором.

Таблиця 13.17. **Середньоквадратична помилка та коефіцієнт детермінації для різних методів при 50%-му пропуску даних**

| Метод | Середньоквадратична помилка, кг | Коефіцієнт детермінації, % |
|---|---|---|
| Повні дані | 3,25 | 22,91 |
| Метод зважування | 5,13 | 19,78 |
| Множинна імпутація | 3,51 | 16,50 |
| Метод максимальної правдоподібності | 2,46 | 27,60 |

При 25% пропусків MCAR для коефіцієнта детермінації найменша оцінка мала місце при застосуванні методу заповнення безумовним середнім значенням, а переоцінка була найнижчою при методі заповнення пропусків з послідовним підбором. З іншими підходами зміна була мінімальною. Отже, отримані оцінки параметрів і *p*-value в результаті застосування методу аналізу наявних випадків більш наближені до значень, отриманих при побудові регресії на повних даних.

При 50% пропусків MCAR фактор ваги став незначним при застосуванні методу аналізу повних спостережень. Найбільш точна оцінка змінної отримана в результаті методу умовного середнього, найменш точна – послідовним підбором. Зазначимо також метод заповнення безумовним середнім, результати застосування якого найбільш наближені до первинних даних.

За результатами імпутації 10% і 50% пропущених даних MAR різними способами зміна в оцінці параметрів для вільного члена і двох інших факторів мінімальна. При застосуванні методу

множинної імпутації середньоквадратична помилка і коефіцієнт детермінації максимально близькі до результатів, отриманих на основі повних даних.

До процесу імпутації слід підходити дуже обережно і проблема імпутації має вирішуватися в кожному конкретному випадку на основі ретельного аналізу існуючої бази даних з урахуванням не тільки особливостей самих даних і обсягу пропусків, а й специфіки конкретного дослідження.

## 13.4. Програмна реалізація відновлення пропущених даних: порівняльний аналіз

Розглянемо процедури по відновленню даних, які пропонуються в програмно-аналітичних середовищах, зазначених вище.

Так, в пакеті IBM® SPSS Statistics допускаються два види пропущених значень:

- Пропущені значення, які визначаються системою (System-defined missing values): якщо в матриці даних є незаповнені чисельні комірки, система SPSS самостійно ідентифікує їх як пропущені значення. Цей факт відображається в матриці даних за допомогою коми ",".
- Пропущені значення, що задаються користувачем (User-defined missing values): якщо в певних випадках у змінних відсутні значення, користувач може за допомогою кнопки Missing оголосити ці значення як пропущені.

Процедура Аналіз пропущених значень в SPSS виконує три основні функції:

1) Описує структуру пропущених даних: Де розташовані пропущені значення? Наскільки широку область вони охоплюють? Чи є тенденція до пропуску значень в декількох спостереженнях у

пар змінних? Чи приймають дані крайні значення? Чи носять пропуски випадковий характер?

2) Оцінює середні, середньоквадратичні відхилення, коваріації і кореляції для різних методів обробки пропущених значень: построкове та попарне видалення, регресія, оцінка максимальної правдоподібності. Попарний метод виводить також частоти повних пар спостережень.

3) Імпутує на місце пропущених значень оціночні значення, використовуючи метод регресії, оцінку максимальної правдоподібності або більш точний метод множинної імпутації.

В R середовищі пропущені дані позначаються символом NA (not available – немає в наявності). Неприпустимі значення (наприклад, ділення на 0) позначаються як NaN (not a number – не є числом). На відміну від SAS, в R використовується однакові позначення для пропущених значень в текстових і числових даних. Крім того, в R є кілька функцій, призначених для виявлення пропущених значень. Функція is.na надає можливість перевірити дані на наявність пропущених значень. Можна виділити наступну класифікацію методів обробки пропущених значень:

1) Видалення пропущених значень (по рядкам та попарно);

2) Оцінка максимальної правдоподібності;

3) Заміщення пропущених значень (одиночна та множина імпутація).

В програмі Statistica порожнім коміркам проставляється деякий спеціальний код – Missing Data Code – Код пропущених даних, значення якого за замовчуванням дорівнює 99999.

Спосіб використання пропущених даних можна підібрати індивідуально для кожної процедури аналізу. Там, де це можливо, користувачеві надано вибір способу обробки пропущених даних: видалення їх з обчислень через по рядках або попарно, заміна на

середні значення, а також їх перетворення або інтерполяція (наприклад, в модулі Часові ряди).

Щоб дізнатися про конкретні способи використання пропущених даних в певних процедурах, потрібно натиснути кнопку довідки або клавішу f1 у відповідному діалоговому завдання аналізу.

В програмі SAS пропущені числові дані позначаються як «.», а текстові залишаються порожніми. Як правило, процедури SAS, обробляють та аналізують дані, оминаючи відсутні значення. Тобто за замовчуванням використовується метод видалення по рядках або попарного видалення залежно від процедури.

В програмно-аналітичному середовищі SAS можна імпутувати пропуски за допомогою всіх існуючих методів: підстановка середнього по вибірці, метод хот-дек, регресійний аналіз, оцінка максимальної правдоподібності, підстановка за допомогою факторного аналізу та метод множинної імпутації. Процедури proc mi, proc mianalyze – потужний інструмент для обробки та відновлення даних в SAS.

Задля ілюстрації роботи розглянутих вище процедур було використано дані про пацієнтів хворих на анорексію, що були розглянуті в п.13.3. Описова статистика даних для дослідження наведена в табл. 13.5, що є стандартним способом узагальнення даних у клінічних випробуваннях.

Для проведення порівняльного аналізу програмної реалізації відновлення пропусків за допомогою програмного коду випадковим чином було симульовано набори даних, де для змінної вага після лікування пропущено 5%, 10% та 25%, 50% значень відсутніх спостережень.

Розглядаючи повні дані як генеральну сукупність, а симульовані з пропусками як вибірку, було пораховано середнє значення на пропущених даних та довірчі межі (табл. 13.18).

Розраховані довірчі межі включають в себе середнє значення, пораховане на повних даних – 39,68 кг та 36,79 кг для пацієнтів, які

отримали та не отримали лікування відповідно. Проте, чим більша частка пропущених даних, тим більше відхиляється середня обчислена на повних даних від середньої, що отримана на неповних даних (особливо при 50% пропусків). Зважаючи на те, що 5% пропусків майже не призвело до суттєвих відхилень середньої ваги в обох групах, то процедуру відновлення даних було застосовано для 10% пропусків і вище.

*Таблиця 13.18.* **Середня вага після лікування, кг**

| % пропусків | Отримали лікування | | Контрольна група | |
|---|---|---|---|---|
| | Середня вага, $\bar{x}$ | Довірчі межі | Середня вага, $\bar{x}$ | Довірчі межі |
| 0 | 39,68 | x | 36,79 | x |
| 5 | 39,64 | (38,45; 40,83) | 36,65 | (35,75; 37,55) |
| 10 | 39,79 | (38,57; 41,01) | 36,80 | (35,87; 37,73) |
| 25 | 39,51 | (38,18; 40,84) | 37,19 | (36,25; 38,12) |
| 50 | 41,27 | (39,57; 42,98) | 37,11 | (35,91; 38,31) |

При застосуванні методу покрокового видалення було отримано такі ж самі результати, оскільки даний метод включає тільки ті спостереження, що не містять пропущених даних. Оскільки пропуски містяться лише в одній змінній, то в даному випадку застосовувати метод попарного видалення не має сенсу. Якщо необхідно провести аналіз враховуючи інші наявні зміні, то тоді доречно застосувати метод попарного видалення.

Метод безумовної імпутації (підстановка середнього по вибірці) можна використати в програмі Statistica, SAS або R. Проте з'ясувалось, що в програмі Statistica даний метод присутній не в усіх модулях. Так, наприклад, при використанні модулю множинної регресії, зазначений метод можна застосувати. В модулі описової статистики наявні лише два найпростіші методи. При незначному

обсягу даних, працюючи в Statistica, дослідник може лише вручну підставити середнє замість пропусків.

Метод хот-дек (заповнення пропусків з упередженим підбором) можна реалізувати лише в програмі SAS. Результати двох вищезазначених методів наведено в таблиці 13.19.

*Таблиця 13.19.* **Середня вага після лікування після проведення процедури відновлення пропусків в Statistica та SAS, кг**

| Метод безумовної імпутації | | | | |
|---|---|---|---|---|
| % пропусків | Отримали лікування | | Контрольна група | |
| | Середня вага, $\bar{x}$ | Довірчі межі | Середня вага, $\bar{x}$ | Довірчі межі |
| 0 | 39,68 | x | 36,79 | x |
| 10 | 39,72 | (38,58; 40,86) | 37,11 | (36,28; 37,94) |
| 25 | 39,37 | (38,31; 40,43) | 37,74 | (37,07; 38,41) |
| 50 | 40,33 | (39,54; 41,12) | 38,14 | (37,32; 38,96) |
| Метод хот-дек | | | | |
| % пропусків | Отримали лікування | | Контрольна група | |
| | Середнє вага, $\bar{x}$ | Довірчі межі | Середнє вага, $\bar{x}$ | Довірчі межі |
| 10 | 39,90 | (38,74; 41,06) | 36,46 | (35,57; 37,35) |
| 25 | 39,40 | (38,19; 40,62) | 37,67 | (36,69; 38,65) |
| 50 | 40,00 | (38,9; 41,09) | 38,53 | (37,34; 39,72) |

Для групи пацієнтів, які отримали лікування, довірчі межі включають середнє значення, обраховане на повних даних. В другій контрольній групі бачимо, що метод безумовної імпутації погано спрацював для 25% та 50% пропусків, метод хот-дек погано відпрацював для 50% пропущених даних. Можемо зробити висновок, що вище зазначені методи краще застосовувати при відносно незначних обсягах даних – до 25%.

Більш потужні методи, такі як метод регресійного аналізу, множинної імпутації, оцінка максимальної правдоподібності було реалізовано в SAS та SPSS. Результати обчислень наведені в таблиці 13.20, 13.21.

*Таблиця 13.20.* **Середня вага після лікування після проведення процедури відновлення пропусків в SAS, кг**

| Метод регресійного аналізу | | | | |
|---|---|---|---|---|
| % пропусків | Отримали лікування | | Контрольна група | |
| | Середня вага, $\bar{x}$ | Довірчі межі | Середня вага, $\bar{x}$ | Довірчі межі |
| 0 | 39,68 | x | 36,79 | x |
| 10 | 39,67 | (38,51; 40,83) | 36,78 | (35,92; 37,65) |
| 25 | 39,66 | (38,50; 40,82) | 36,79 | (35,93; 37,66) |
| 50 | 39,66 | (38,49; 40,82) | 36,78 | (35,91; 37,65) |
| Оцінка максимальної правдоподібності | | | | |
| % пропусків | Отримали лікування | | Контрольна група | |
| | Середня вага, $\bar{x}$ | Довірчі межі | Середня вага, $\bar{x}$ | Довірчі межі |
| 10 | 39,67 | (38,51; 40,83) | 36,78 | (35,92; 37,65) |
| 25 | 39,67 | (38,50; 40,83) | 36,79 | (35,92; 37,66) |
| 50 | 39,66 | (38,50; 40,83) | 36,78 | (35,91; 37,65) |
| Метод множинної імпутації | | | | |
| % пропусків | Отримали лікування | | Контрольна група | |
| | Середня вага, $\bar{x}$ | Довірчі межі | Середня вага, $\bar{x}$ | Довірчі межі |
| 10 | 39,67 | (39,32; 40,03) | 36,78 | (36,53; 37,04) |
| 25 | 39,67 | (39,31; 40,02) | 36,78 | (36,53; 37,04) |
| 50 | 39,66 | (39,31; 40,02) | 36,78 | (36,53; 37,04) |

*Джерело: власні розрахунки в середовищі SAS*

*Таблиця 13.21.* **Середня вага після лікування після проведення процедури відновлення пропусків в SPSS, кг**

| Метод регресійного аналізу | | | | |
|---|---|---|---|---|
| % пропусків | Отримали лікування | | Контрольна група | |
| | Середня вага, $\bar{x}$ | Довірчі межі | Середня вага, $\bar{x}$ | Довірчі межі |
| 0 | 39,68 | x | 36,79 | x |
| 10 | 39,85 | (38,70; 41,00) | 37,15 | (36,15; 38,15) |
| 25 | 39,69 | (38,53; 40,85) | 37,06 | (36,19; 37,93) |
| 50 | 39,89 | (38,79; 40,99) | 37,88 | (36,50; 39,27) |
| Оцінка максимальної правдоподібності | | | | |
| % пропусків | Отримали лікування | | Контрольна група | |
| | Середня вага, $\bar{x}$ | Довірчі межі | Середня вага, $\bar{x}$ | Довірчі межі |
| 10 | 39,67 | (38,51; 40,83) | 36,78 | (35,92; 37,65) |
| 25 | 39,67 | (38,51; 40,83) | 36,78 | (35,92; 37,65) |
| 50 | 39,67 | (38,51; 40,83) | 36,78 | (35,92; 37,65) |
| Метод множинної імпутації | | | | |
| % пропусків | Отримали лікування | | Контрольна група | |
| | Середня вага, $\bar{x}$ | Довірчі межі | Середня вага, $\bar{x}$ | Довірчі межі |
| 10 | 39,67 | (39,17; 40,17) | 36,78 | (36,42; 37,15) |
| 25 | 39,67 | (39,17; 40,17) | 36,78 | (36,42; 37,15) |
| 50 | 39,48 | (38,98; 39,98) | 37,12 | (36,68; 37,57) |

В результати реалізації даних методів дають можливість говорити про наступне:

1) метод регресійного аналізу краще реалізований в *SAS, оскільки* середнє на відновлених даних не тільки практично не відрізняється від середнього, отриманого на повних даних, а й від результатів, які отримані іншими, більш потужними методами;

*2)* на 50% відсотках пропусків очевидно, що *SAS дає більш точні оцінки, ніж* SPSS; більш того, результати, що отримані в *SAS різними методами, практично не відрізняються один від одного, що свідчить про їхню однакову потужність;*

*3) як бачимо, що в SPSS при 50% пропусків доцільніше користуватися оцінками максимальної правдоподібності ніж методом множинної імпутації або методом регресійного аналізу; більше того, регресійний аналіз в SPSS не дає стійких результатів, а тому їм слід користуватися дуже обережно незалежно від кількості пропусків.*

Таким чином, найзручнішим та простим засобом відновлення пропусків є ППП Statistica, але способи відновлення даних обмежені наявними програмними функціями. Тобто, Statistica допоможе впоратися з пропущеними даними при незначному обсязі пропусків (до 10%). Наступним за простотою програмним середовищем оброблення пропущених даних йде SPSS, яка пропонує більш широкий спектр методів відновлення даних, порівняно зі Statistica, і водночас пропонує більш зрозуміліший інтерфейс для користувача, порівняно з мовою програмування R чи SAS. Проте найпотужнішими програмами по відновленню даних залишаються R та SAS, які не тільки надають можливість обробляти великі масиви даних зі значною часткою пропусків, а й застосовувати різні методи відновлення даних від найпростіших до найскладніших процедур. Проте робота в середовищі R та SAS має свою специфіку і потребує знання мови програмування і вимагає спеціальної фахової підготовки.

Водночас слід зазначити, що жодне з розглянутих програмно-аналітичних середовищ не має вбудованих процедур обробки категоріальних даних. Є певні підходи, які можна реалізувати за аналогією для упорядкованих категорій в програмних середовищах R та SAS проте це не покриває всі потреби аналізу соціально-економічних явищ та процесів, які реалізовані у вигляді опитувань і

результати яких здебільшого представлені у вигляді відповідей на питання. Навіть якщо були використані цифри для кодування відповідей, оскільки код має тільки цифрову форму представлення і умовно відображає кількісну цифрову послідовність проте код не є числом по суті і йому не притаманні властивості числа, а значить методи, які ми можемо застосувати для кількісних даних не можуть бути поширені на категоріальні.

### *Список завдань до самоконтролю:*

***Завдання 1***.

Знайти відповідність між наявними даними та типом пропущених даних (MAR, MCAR, MNAR).

А.

| № об’єкта | Стать | Місто | Заробітна плата |
|---|---|---|---|
| 1 | Ч | Київ | |
| 2 | Ж | Дніпро | $25,796 |
| 3 | Ч | Одеса | |
| 4 | Ж | Одеса | $3,520 |
| 5 | Ж | Київ | $15,312 |
| 6 | Ж | Харків | $17,027 |
| 7 | Ч | Харків | |
| 8 | Ж | Одеса | $36,309 |
| 9 | Ч | Дніпро | |
| 10 | Ч | Дніпро | |

Б.

| № об’єкта | Стать | Місто | Заробітна плата |
|---|---|---|---|
| A | B | C | D |
| 1 | Ч | Київ | $11,975 |
| 2 | Ж | Дніпро | $25,796 |
| 3 | Ч | Одеса | $31,503 |
| 4 | Ж | Одеса | $3,520 |
| 5 | Ж | Київ | $15,312 |

| | | | |
|---|---|---|---|
| 6 | Ж | Харків | $17,027 |
| A | B | C | D |
| 7 | Ч | Харків | $63,697 |
| 8 | Ж | Одеса | $36,309 |
| 9 | Ч | Дніпро | |
| 10 | Ч | Дніпро | |

C.

| № об’єкта | Стать | Місто | Заробітна плата |
|---|---|---|---|
| 1 | Ч | Київ | $11,975 |
| 2 | Ж | Дніпро | $25,796 |
| 3 | Ч | Одеса | |
| 4 | Ж | Одеса | $3,520 |
| 5 | Ж | Київ | |
| 6 | Ж | Харків | $17,027 |
| 7 | Ч | Харків | |
| 8 | Ж | Одеса | $36,309 |
| 9 | Ч | Дніпро | $53,703 |
| 10 | Ч | Дніпро | |

***Завдання 2***.

За допомогою якої процедури (процедур) можна відновити пропущені дані методом множинної імпутаціїї в SAS?

A) proc mixed, proc reg;

Б) proc mi, proc mianalyze;

C) proc corr;

Д) proc mean.

***Завдання 3***.

Використовуючи базу даних sashelp.class згенеруйте 15% пропусків для зміної *Height* та *Weight*. За допомогою методу

одиничної імпутації на основі середнього значення відновіть пропущені дані та порівняйте результати з повним набором даних.

***Завдання 4.***

Використовуючи змінні *Salary* (1987 Salary in $ Thousands), *CrRuns* (Career Runs), *Yrmajor* (Years in the Major Leagues), *CrHits* (Career Hits) з бази даних *sashelp.baseball* згенеруйте 20% пропусків для змінної *CrRuns*. Відновіть пропущені дані за допомогою методу доступних спостережень, методу заповнення пропусків з упередженим підбором та методу умовної імпутації середнього. Порівняйте результати між собою та з повним набором даним.

***Список рекомендованої літератури по темі:***

1. Fichman, M., Cummings, J. M. Multiple Imputation for Missing Data: Making the Most of What you Know. *Organizational Research Methods*. 2003. Vol. 6. 282–308.
2. Guidelines on Missing Data in Confirmatory Clinical Trials. European Medicines Agency, 2010. EMA/CPMP/EWP/1776/99 Rev. 1. URL: *http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/09/WC500096793.pdf*
3. Missing data in SAS. URL:*https://stats.idre.ucla.edu/sas/modules/missing-data-in-sas/*
4. Molenberghs, G., Kenward, M. G. Missing Data in Clinical Studies. Chichester: John Wiley & Sons Ltd, 2007. 526 p. URL: *https://doi.org/10.1002/9780470510445.ch9*
5. The Prevention and Treatment of Missing Data in Clinical Trials / R. J. Little et al. *The New England Journal of Medicine.* 2012. Vol. 367, № 14. URL: *http://www.nejm.org/doi/pdf/10.1056/nejmsr1203730*

6. SAS 9.4 Product Documentation. URL: *https://support.sas.com/documentation/94/*
7. Schafer J. L. Multiple imputation: A primer. *Statistical Methods in Medical Research*. 1999. Vol. 8, № 1. P. 3–15.
8. StatSoft, Inc. (2012). Электронный учебник по статистике. Москва, StatSoft. WEB: *http://www.statsoft.ru/home/textbook/default.htm.*
9. Strategies for dealing with missing data in clinical trials: from design to analysis / J. D. Dziura et al. *Yale Journal of Biology and Medicine.* 2013. 86 (3). 343–358. URL: *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3767219/*
10. Аналіз пропущених значень. URL:*https://www.ibm.com/support/knowledgecenter/ru/SSLVMB_sub/statistics_mainhelp_ddita/spss/mva/idh_miss.html*
11. Злоба Е., Яцкив И. Статистические методы восстановления пропущенных данных. *Computer Modelling & New Technologies*. 2002. Vol. 6, № 1. P. 51–61.
12. Ілюстрований самовчитель по SPSS. URL: *http://www.datuapstrade.lv/rus/spss/*
13. Кутлалиев А. Метод множественного восстановления данных. URL: https://www.hse.ru/mirror/pubs/lib/data/access/ram/ticket/4.pdf
14. Шипунов А. Б., Балдин Е. М., Волкова П. А.. Наглядная статистика. 2014. URL: *https://cran.r-project.org/doc/contrib/Shipunov-rbook.pdf*

# ДОДАТКИ

# ТА СТАТИСТИЧНІ ТАБЛИЦІ

## Додатки до розділу 3

## Таблиця ймовірностей для стандартного нормального розподілу

| *z-значення* | Ймовірність | *z-значення* | Ймовірність | *z-значення* | Ймовірність | *z-значення* | Ймовірність | *z-значення* | Ймовірність | *z-значення* | Ймовірність |
|---|---|---|---|---|---|---|---|---|---|---|---|
| –2.00 | 0.0228 | –1.00 | 0.1587 | 0.00 | 0.5000 | 0.00 | 0.5000 | 1.00 | 0.8413 | 2.00 | 0.9772 |
| –2.01 | 0.0222 | –1.01 | 0.1562 | –0.01 | 0.4960 | 0.01 | 0.5040 | 1.01 | 0.8438 | 2.01 | 0.9778 |
| –2.02 | 0.0217 | –1.02 | 0.1539 | –0.02 | 0.4920 | 0.02 | 0.5080 | 1.02 | 0.8461 | 2.02 | 0.9783 |
| –2.03 | 0.0212 | –1.03 | 0.1515 | –0.03 | 0.4880 | 0.03 | 0.5120 | 1.03 | 0.8485 | 2.03 | 0.9788 |
| –2.04 | 0.0207 | –1.04 | 0.1492 | –0.04 | 0.4840 | 0.04 | 0.5160 | 1.04 | 0.8508 | 2.04 | 0.9793 |
| –2.05 | 0.0202 | –1.05 | 0.1469 | –0.05 | 0.4801 | 0.05 | 0.5199 | 1.05 | 0.8531 | 2.05 | 0.9798 |
| –2.06 | 0.0197 | –1.06 | 0.1446 | –0.06 | 0.4761 | 0.06 | 0.5239 | 1.06 | 0.8554 | 2.06 | 0.9803 |
| –2.07 | 0.0192 | –1.07 | 0.1423 | –0.07 | 0.4721 | 0.07 | 0.5279 | 1.07 | 0.8577 | 2.07 | 0.9808 |
| –2.08 | 0.0188 | –1.08 | 0.1401 | –0.08 | 0.4681 | 0.08 | 0.5319 | 1.08 | 0.8599 | 2.08 | 0.9812 |
| –2.09 | 0.0183 | –1.09 | 0.1379 | –0.09 | 0.4641 | 0.09 | 0.5359 | 1.09 | 0.8621 | 2.09 | 0.9817 |
| –2.10 | 0.0179 | –1.10 | 0.1357 | –0.10 | 0.4602 | 0.10 | 0.5398 | 1.10 | 0.8643 | 2.10 | 0.9821 |
| –2.11 | 0.0174 | –1.11 | 0.1335 | –0.11 | 0.4562 | 0.11 | 0.5438 | 1.11 | 0.8665 | 2.11 | 0.9826 |
| –2.12 | 0.0170 | –1.12 | 0.1314 | –0.12 | 0.4522 | 0.12 | 0.5478 | 1.12 | 0.8686 | 2.12 | 0.9830 |
| –2.13 | 0.0166 | –1.13 | 0.1292 | –0.13 | 0.4483 | 0.13 | 0.5517 | 1.13 | 0.8708 | 2.13 | 0.9834 |
| –2.14 | 0.0162 | –1.14 | 0.1271 | –0.14 | 0.4443 | 0.14 | 0.5557 | 1.14 | 0.8729 | 2.14 | 0.9838 |
| –2.15 | 0.0158 | –1.15 | 0.1251 | –0.15 | 0.4404 | 0.15 | 0.5596 | 1.15 | 0.8749 | 2.15 | 0.9842 |
| –2.16 | 0.0154 | –1.16 | 0.1230 | –0.16 | 0.4364 | 0.16 | 0.5636 | 1.16 | 0.8770 | 2.16 | 0.9846 |
| –2.17 | 0.0150 | –1.17 | 0.1210 | –0.17 | 0.4325 | 0.17 | 0.5675 | 1.17 | 0.8790 | 2.17 | 0.9850 |
| –2.18 | 0.0146 | –1.18 | 0.1190 | –0.18 | 0.4286 | 0.18 | 0.5714 | 1.18 | 0.8810 | 2.18 | 0.9854 |
| –2.19 | 0.0143 | –1.19 | 0.1170 | –0.19 | 0.4247 | 0.19 | 0.5753 | 1.19 | 0.8830 | 2.19 | 0.9857 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| –2.20 | 0.0139 | –1.20 | 0.1151 | –0.20 | 0.4207 | 0.20 | 0.5793 | 1.20 | 0.8849 | 2.20 | 0.9861 |
| –2.21 | 0.0136 | –1.21 | 0.1131 | –0.21 | 0.4168 | 0.21 | 0.5832 | 1.21 | 0.8869 | 2.21 | 0.9864 |
| –2.22 | 0.0132 | –1.22 | 0.1112 | –0.22 | 0.4129 | 0.22 | 0.5871 | 1.22 | 0.8888 | 2.22 | 0.9868 |
| –2.23 | 0.0129 | –1.23 | 0.1093 | –0.23 | 0.4090 | 0.23 | 0.5910 | 1.23 | 0.8907 | 2.23 | 0.9871 |
| –2.24 | 0.0125 | –1.24 | 0.1075 | –0.24 | 0.4052 | 0.24 | 0.5948 | 1.24 | 0.8925 | 2.24 | 0.9875 |
| –2.25 | 0.0122 | –1.25 | 0.1056 | –0.25 | 0.4013 | 0.25 | 0.5987 | 1.25 | 0.8944 | 2.25 | 0.9878 |
| –2.26 | 0.0119 | –1.26 | 0.1038 | –0.26 | 0.3974 | 0.26 | 0.6026 | 1.26 | 0.8962 | 2.26 | 0.9881 |
| –2.27 | 0.0116 | –1.27 | 0.1020 | –0.27 | 0.3936 | 0.27 | 0.6064 | 1.27 | 0.8980 | 2.27 | 0.9884 |
| –2.28 | 0.0113 | –1.28 | 0.1003 | –0.28 | 0.3897 | 0.28 | 0.6103 | 1.28 | 0.8997 | 2.28 | 0.9887 |
| –2.29 | 0.0110 | –1.29 | 0.0985 | –0.29 | 0.3859 | 0.29 | 0.6141 | 1.29 | 0.9015 | 2.29 | 0.9890 |
| –2.30 | 0.0107 | –1.30 | 0.0968 | –0.30 | 0.3821 | 0.30 | 0.6179 | 1.30 | 0.9032 | 2.30 | 0.9893 |
| –2.31 | 0.0104 | –1.31 | 0.0951 | –0.31 | 0.3783 | 0.31 | 0.6217 | 1.31 | 0.9049 | 2.31 | 0.9896 |
| –2.32 | 0.0102 | –1.32 | 0.0934 | –0.32 | 0.3745 | 0.32 | 0.6255 | 1.32 | 0.9066 | 2.32 | 0.9898 |
| –2.33 | 0.0099 | –1.33 | 0.0918 | –0.33 | 0.3707 | 0.33 | 0.6293 | 1.33 | 0.9082 | 2.33 | 0.9901 |
| –2.34 | 0.0096 | –1.34 | 0.0901 | –0.34 | 0.3669 | 0.34 | 0.6331 | 1.34 | 0.9099 | 2.34 | 0.9904 |
| –2.35 | 0.0094 | –1.35 | 0.0885 | –0.35 | 0.3632 | 0.35 | 0.6368 | 1.35 | 0.9115 | 2.35 | 0.9906 |
| –2.36 | 0.0091 | –1.36 | 0.0869 | –0.36 | 0.3594 | 0.36 | 0.6406 | 1.36 | 0.9131 | 2.36 | 0.9909 |
| –2.37 | 0.0089 | –1.37 | 0.0853 | –0.37 | 0.3557 | 0.37 | 0.6443 | 1.37 | 0.9147 | 2.37 | 0.9911 |
| –2.38 | 0.0087 | –1.38 | 0.0838 | –0.38 | 0.3520 | 0.38 | 0.6480 | 1.38 | 0.9162 | 2.38 | 0.9913 |
| –2.39 | 0.0084 | –1.39 | 0.0823 | –0.39 | 0.3483 | 0.39 | 0.6517 | 1.39 | 0.9177 | 2.39 | 0.9916 |
| –2.40 | 0.0082 | –1.40 | 0.0808 | –0.40 | 0.3446 | 0.40 | 0.6554 | 1.40 | 0.9192 | 2.40 | 0.9918 |
| –2.41 | 0.0080 | –1.41 | 0.0793 | –0.41 | 0.3409 | 0.41 | 0.6591 | 1.41 | 0.9207 | 2.41 | 0.9920 |
| –2.42 | 0.0078 | –1.42 | 0.0778 | –0.42 | 0.3372 | 0.42 | 0.6628 | 1.42 | 0.9222 | 2.42 | 0.9922 |
| –2.43 | 0.0075 | –1.43 | 0.0764 | –0.43 | 0.3336 | 0.43 | 0.6664 | 1.43 | 0.9236 | 2.43 | 0.9925 |
| –2.44 | 0.0073 | –1.44 | 0.0749 | –0.44 | 0.3300 | 0.44 | 0.6700 | 1.44 | 0.9251 | 2.44 | 0.9927 |
| –2.45 | 0.0071 | –1.45 | 0.0735 | –0.45 | 0.3264 | 0.45 | 0.6736 | 1.45 | 0.9265 | 2.45 | 0.9929 |
| –2.46 | 0.0069 | –1.46 | 0.0721 | –0.46 | 0.3228 | 0.46 | 0.6772 | 1.46 | 0.9279 | 2.46 | 0.9931 |
| –2.47 | 0.0068 | –1.47 | 0.0708 | –0.47 | 0.3192 | 0.47 | 0.6808 | 1.47 | 0.9292 | 2.47 | 0.9932 |

| –2.48 | 0.0066 | –1.48 | 0.0694 | –0.48 | 0.3156 | 0.48 | 0.6844 | 1.48 | 0.9306 | 2.48 | 0.9934 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| –2.49 | 0.0064 | –1.49 | 0.0681 | –0.49 | 0.3121 | 0.49 | 0.6879 | 1.49 | 0.9319 | 2.49 | 0.9936 |
| –2.50 | 0.0062 | –1.50 | 0.0668 | –0.50 | 0.3085 | 0.50 | 0.6915 | 1.50 | 0.9332 | 2.50 | 0.9938 |
| –2.51 | 0.0060 | –1.51 | 0.0655 | –0.51 | 0.3050 | 0.51 | 0.6950 | 1.51 | 0.9345 | 2.51 | 0.9940 |
| –2.52 | 0.0059 | –1.52 | 0.0643 | –0.52 | 0.3015 | 0.52 | 0.6985 | 1.52 | 0.9357 | 2.52 | 0.9941 |
| –2.53 | 0.0057 | –1.53 | 0.0630 | –0.53 | 0.2981 | 0.53 | 0.7019 | 1.53 | 0.9370 | 2.53 | 0.9943 |
| –2.54 | 0.0055 | –1.54 | 0.0618 | –0.54 | 0.2946 | 0.54 | 0.7054 | 1.54 | 0.9382 | 2.54 | 0.9945 |
| –2.55 | 0.0054 | –1.55 | 0.0606 | –0.55 | 0.2912 | 0.55 | 0.7088 | 1.55 | 0.9394 | 2.55 | 0.9946 |
| –2.56 | 0.0052 | –1.56 | 0.0594 | –0.56 | 0.2877 | 0.56 | 0.7123 | 1.56 | 0.9406 | 2.56 | 0.9948 |
| –2.57 | 0.0051 | –1.57 | 0.0582 | –0.57 | 0.2843 | 0.57 | 0.7157 | 1.57 | 0.9418 | 2.57 | 0.9949 |
| –2.58 | 0.0049 | –1.58 | 0.0571 | –0.58 | 0.2810 | 0.58 | 0.7190 | 1.58 | 0.9429 | 2.58 | 0.9951 |
| –2.59 | 0.0048 | –1.59 | 0.0559 | –0.59 | 0.2776 | 0.59 | 0.7224 | 1.59 | 0.9441 | 2.59 | 0.9952 |
| –2.60 | 0.0047 | –1.60 | 0.0548 | –0.60 | 0.2743 | 0.60 | 0.7257 | 1.60 | 0.9452 | 2.60 | 0.9953 |
| –2.61 | 0.0045 | –1.61 | 0.0537 | –0.61 | 0.2709 | 0.61 | 0.7291 | 1.61 | 0.9463 | 2.61 | 0.9955 |
| –2.62 | 0.0044 | –1.62 | 0.0526 | –0.62 | 0.2676 | 0.62 | 0.7324 | 1.62 | 0.9474 | 2.62 | 0.9956 |
| –2.63 | 0.0043 | –1.63 | 0.0516 | –0.63 | 0.2643 | 0.63 | 0.7357 | 1.63 | 0.9484 | 2.63 | 0.9957 |
| –2.64 | 0.0041 | –1.64 | 0.0505 | –0.64 | 0.2611 | 0.64 | 0.7389 | 1.64 | 0.9495 | 2.64 | 0.9959 |
| –2.65 | 0.0040 | –1.65 | 0.0495 | –0.65 | 0.2578 | 0.65 | 0.7422 | 1.65 | 0.9505 | 2.65 | 0.9960 |
| –2.66 | 0.0039 | –1.66 | 0.0485 | –0.66 | 0.2546 | 0.66 | 0.7454 | 1.66 | 0.9515 | 2.66 | 0.9961 |
| –2.67 | 0.0038 | –1.67 | 0.0475 | –0.67 | 0.2514 | 0.67 | 0.7486 | 1.67 | 0.9525 | 2.67 | 0.9962 |
| –2.68 | 0.0037 | –1.68 | 0.0465 | –0.68 | 0.2483 | 0.68 | 0.7517 | 1.68 | 0.9535 | 2.68 | 0.9963 |
| –2.69 | 0.0036 | –1.69 | 0.0455 | –0.69 | 0.2451 | 0.69 | 0.7549 | 1.69 | 0.9545 | 2.69 | 0.9964 |
| –2.70 | 0.0035 | –1.70 | 0.0446 | –0.70 | 0.2420 | 0.70 | 0.7580 | 1.70 | 0.9554 | 2.70 | 0.9965 |
| –2.71 | 0.0034 | –1.71 | 0.0436 | –0.71 | 0.2389 | 0.71 | 0.7611 | 1.71 | 0.9564 | 2.71 | 0.9966 |
| –2.72 | 0.0033 | –1.72 | 0.0427 | –0.72 | 0.2358 | 0.72 | 0.7642 | 1.72 | 0.9573 | 2.72 | 0.9967 |
| –2.73 | 0.0032 | –1.73 | 0.0418 | –0.73 | 0.2327 | 0.73 | 0.7673 | 1.73 | 0.9582 | 2.73 | 0.9968 |
| –2.74 | 0.0031 | –1.74 | 0.0409 | –0.74 | 0.2296 | 0.74 | 0.7704 | 1.74 | 0.9591 | 2.74 | 0.9969 |
| –2.75 | 0.0030 | –1.75 | 0.0401 | –0.75 | 0.2266 | 0.75 | 0.7734 | 1.75 | 0.9599 | 2.75 | 0.9970 |
| –2.76 | 0.0029 | –1.76 | 0.0392 | –0.76 | 0.2236 | 0.76 | 0.7764 | 1.76 | 0.9608 | 2.76 | 0.9971 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| –2.77 | 0.0028 | –1.77 | 0.0384 | –0.77 | 0.2206 | 0.77 | 0.7794 | 1.77 | 0.9616 | 2.77 | 0.9972 |
| –2.78 | 0.0027 | –1.78 | 0.0375 | –0.78 | 0.2177 | 0.78 | 0.7823 | 1.78 | 0.9625 | 2.78 | 0.9973 |
| –2.79 | 0.0026 | –1.79 | 0.0367 | –0.79 | 0.2148 | 0.79 | 0.7852 | 1.79 | 0.9633 | 2.79 | 0.9974 |
| –2.80 | 0.0026 | –1.80 | 0.0359 | –0.80 | 0.2119 | 0.80 | 0.7881 | 1.80 | 0.9641 | 2.80 | 0.9974 |
| –2.81 | 0.0025 | –1.81 | 0.0351 | –0.81 | 0.2090 | 0.81 | 0.7910 | 1.81 | 0.9649 | 2.81 | 0.9975 |
| –2.82 | 0.0024 | –1.82 | 0.0344 | –0.82 | 0.2061 | 0.82 | 0.7939 | 1.82 | 0.9656 | 2.82 | 0.9976 |
| –2.83 | 0.0023 | –1.83 | 0.0336 | –0.83 | 0.2033 | 0.83 | 0.7967 | 1.83 | 0.9664 | 2.83 | 0.9977 |
| –2.84 | 0.0023 | –1.84 | 0.0329 | –0.84 | 0.2005 | 0.84 | 0.7995 | 1.84 | 0.9671 | 2.84 | 0.9977 |
| –2.85 | 0.0022 | –1.85 | 0.0322 | –0.85 | 0.1977 | 0.85 | 0.8023 | 1.85 | 0.9678 | 2.85 | 0.9978 |
| –2.86 | 0.0021 | –1.86 | 0.0314 | –0.86 | 0.1949 | 0.86 | 0.8051 | 1.86 | 0.9686 | 2.86 | 0.9979 |
| –2.87 | 0.0021 | –1.87 | 0.0307 | –0.87 | 0.1922 | 0.87 | 0.8078 | 1.87 | 0.9693 | 2.87 | 0.9979 |
| –2.88 | 0.0020 | –1.88 | 0.0301 | –0.88 | 0.1894 | 0.88 | 0.8106 | 1.88 | 0.9699 | 2.88 | 0.9980 |
| –2.89 | 0.0019 | –1.89 | 0.0294 | –0.89 | 0.1867 | 0.89 | 0.8133 | 1.89 | 0.9706 | 2.89 | 0.9981 |
| –2.90 | 0.0019 | –1.90 | 0.0287 | –0.90 | 0.1841 | 0.90 | 0.8159 | 1.90 | 0.9713 | 2.90 | 0.9981 |
| –2.91 | 0.0018 | –1.91 | 0.0281 | –0.91 | 0.1814 | 0.91 | 0.8186 | 1.91 | 0.9719 | 2.91 | 0.9982 |
| –2.92 | 0.0018 | –1.92 | 0.0274 | –0.92 | 0.1788 | 0.92 | 0.8212 | 1.92 | 0.9726 | 2.92 | 0.9982 |
| –2.93 | 0.0017 | –1.93 | 0.0268 | –0.93 | 0.1762 | 0.93 | 0.8238 | 1.93 | 0.9732 | 2.93 | 0.9983 |
| –2.94 | 0.0016 | –1.94 | 0.0262 | –0.94 | 0.1736 | 0.94 | 0.8264 | 1.94 | 0.9738 | 2.94 | 0.9984 |
| –2.95 | 0.0016 | –1.95 | 0.0256 | –0.95 | 0.1711 | 0.95 | 0.8289 | 1.95 | 0.9744 | 2.95 | 0.9984 |
| –2.96 | 0.0015 | –1.96 | 0.0250 | –0.96 | 0.1685 | 0.96 | 0.8315 | 1.96 | 0.9750 | 2.96 | 0.9985 |
| –2.97 | 0.0015 | –1.97 | 0.0244 | –0.97 | 0.1660 | 0.97 | 0.8340 | 1.97 | 0.9756 | 2.97 | 0.9985 |
| –2.98 | 0.0014 | –1.98 | 0.0239 | –0.98 | 0.1635 | 0.98 | 0.8365 | 1.98 | 0.9761 | 2.98 | 0.9986 |
| –2.99 | 0.0014 | –1.99 | 0.0233 | –0.99 | 0.1611 | 0.99 | 0.8389 | 1.99 | 0.9767 | 2.99 | 0.9986 |
| –3.00 | 0.0013 | –2.00 | 0.0228 | –1.00 | 0.1587 | 1.00 | 0.8413 | 2.00 | 0.9772 | 3.00 | 0.9987 |

**Додатки до розділу 8**

## Положення про ПФТС-індекс

ПФТС-індекс є офіційним показником Першої фондової торговельної системи. Індекс розраховується на основі простих акцій підприємств, що пройшли лістинг в ПФТС.

***Назва індексу:*** ПФТС-індекс.

***Період розрахунку індексу (поточний період)***

Індекс може розраховуватися в режимі реального часу, на погодинній, щоденній та щотижневій основі.

*Погодинний індекс*. Погодинний ПФТС-індекс розраховується в кінці кожної години протягом торгової сесії.

*Щоденний індекс*. Щоденний ПФТС-індекс розраховується кожного робочого дня в кінці торгової сесії.

*Щотижневий індекс.* Щотижневий індекс розраховується в кінці кожного робочого тижня. Якщо тиждень є неповним/або має більше робочих днів, ніж звичайний робочий тиждень, індекс розраховується з/або без урахування цих днів.

***Перелік акцій підприємств, що входять до індексу***

До Переліку входять акції, що мають найбільші показники ліквідності.

Перелік акцій, що входять до індексу, переглядається кожного місяця.

Індекс в поточному періоді розраховується на основі переліку підприємств, що визначаються за попередній місяць.

Критеріями для обрання підприємств у список індексу є:

1) підприємства-емітенти повинні пройти лістинг ПФТС та належати до першого або другого рівнів Списку ПФТС;

2) до Переліку акцій підприємств, що входять до індексу, відбираються акції, за якими в ПФТС було зареєстровано найбільшу кількість угод, та за якими середня різниця між

найкращою ціною купівлі та найкращою ціною продажу (спред) за останній місяць не перевищувала 40 %.

Якщо підприємство, що входить до індексу, за будь-яких причин виключається зі Списку ПФТС, воно автоматично виключається з Переліку підприємств, що входять у лістинг. За таких умов індекс за наступний розрахунковий період розраховується без виключеного підприємства, а за базовий період береться список попереднього місяця без виключеного підприємства.

***Базові значення***

"Базовий період – це період, з якого починається розрахунок індексу. Базовий період корегується зі зміною переліку акцій підприємств, що входять до індексу. ПФТС-індекс починає розраховуватися з 01.10.97.

Базове значення індексу – це значення індексу в базовому періоді. Базове значення індексу корегується зі зміною переліку підприємств, що входять до індексу. Для ПФТС-індексу базове значення становить 100".

***Формула розрахунку індексу***

Індекс розраховується за принципом ринкового зважування, що використовує метод арифметичної середньої. Під час розрахунку індексу враховуються всі угоди, що Були зареєстровані в ПФТС та задовольняють наведеним нижче умовам.

Формула індексу

$$I_{\text{pfts}} = I_{\text{pfts}_\text{O}} \frac{\sum MC_{i,t}}{\sum MC_{i,0}},$$

де $I_{\text{pfts}_\text{O}}$ – базове значення індексу;

$\sum_i MC_{i,t}$ - сума ринкових капіталізацій усіх акцій з Переліку акцій

індексу в поточному періоді. Капіталізація розраховується за формулою:

$$MC_{i,t} = Q_i, \omega \times Pl_{i,t},$$

де $Q_i$ – кількість звичайних акцій, випущених цим емітентом. Така методика не розподіляє державну та недержавну частку акцій; $Pl_{i,t}$ – ціна останньої угоди *i*-тої акції в поточному періоді, якщо вона задовольняє таку умову:

$$B_{i,t} \leq PI_{i,t} \leq A_{i,t},$$

де $B_{i,t}$ – значення найкращої (найвищої) ціни купівлі; $A_{i,t}$ – значення найкращої (найнижчої) ціни продажу.

Якщо ціна останньої угоди за поточний період не відповідає наведеній вище умові, то для щотижневого ПФТС-індексу за основу береться ціна попередньої угоди за поточний період, що відповідає наведеній вище умові. Для погодинного, щоденного та іншого ПФТС-індексу в разі відсутності зареєстрованої угоди за цією акцією за поточний період береться ціна, що розраховується за формулою

$$Pl_{i,t} = \frac{B_{i,t} + A_{i,t}}{2},$$

де $B_{i,t}$ – значення найкращої (найвищої) ціни купівлі на закриття торгової сесії ПФТС; $A_{i,t}$ – значення найкращої (найнижчої) ціни продажу на закриття торгової сесії ПФТС;

3) $\sum MC_{i,b}$ – сума ринкових капіталізацій усіх акцій із Переліку підприємств індексу в базовому періоді.

***Статистична база***

Індекс розраховується на основі офіційних результатів або поточних торгів в ПФТС.

***Поновлення Списку індексу***

Для того щоб запобігти різкому скачку індексу в разі зміни Переліку акцій, що входять до індексу, у поточному періоді індекс розраховується за новим Переліком згідно з формулою

$$I_{\text{pfts}_{t-1}} \cdot \frac{\sum_i MC_{i,t}}{\sum_i MC_{i,t-1}},$$

де: 1) $I_{\text{pfts}_{t-1}}$ – базове значення індексу, розраховане на ($t$–1) період з новим Переліком акцій індексу;

2) $\sum MC_{i,t}$ – сума ринкових капіталізацій всіх акцій з нового Переліку акцій індексу в поточному періоді;

3) $\sum MC_{i,t-1}$ – сума ринкових капіталізацій усіх акцій з нового Переліку акцій індексу в базовому ($t$–1) періоді.

Об'єктивність та правильність розрахунку ПФТС-індексу забезпечується чіткою методикою та тим, що первинна інформація для розрахунку індексу є відкритою та рівнодоступною для всіх зацікавлених осіб.

## Додатки до розділу 13

Додаток 13.1

**Перелік пацієнтів за групою лікування та вагою**

| № пацієнта | Група лікування | Вага до лікування, кг | Вага після лікування, кг |
|---|---|---|---|
| 1 | 0 | 36,6 | 36,4 |
| 2 | 0 | 40,6 | 36,3 |
| 3 | 0 | 41,6 | 39,2 |
| 4 | 0 | 33,6 | 39,1 |
| 5 | 0 | 35,4 | 34,5 |
| 6 | 0 | 40,1 | 35,4 |
| 7 | 0 | 39,6 | 34,1 |
| 8 | 0 | 34,1 | 39,3 |
| 9 | 0 | 36,6 | 33,3 |
| 10 | 0 | 35,6 | 38,4 |
| 11 | 0 | 35,2 | 35,1 |
| 12 | 0 | 40,2 | 36,1 |
| 13 | 0 | 36,9 | 40,6 |
| 14 | 0 | 35,4 | 36,9 |
| 15 | 0 | 32,0 | 37,1 |
| 16 | 0 | 35,1 | 35,1 |
| 17 | 0 | 38,6 | 38,2 |
| 18 | 0 | 39,0 | 34,2 |
| 19 | 0 | 38,1 | 36,1 |
| 20 | 0 | 36,2 | 33,1 |
| 21 | 0 | 38,8 | 40,1 |
| 22 | 0 | 38,3 | 38,4 |
| 23 | 0 | 36,1 | 36,9 |
| 24 | 0 | 35,2 | 36,8 |

| № пацієнта | Група лікування | Вага до лікування, кг | Вага після лікування, кг |
|---|---|---|---|
| 25 | 0 | 32,8 | 40,0 |
| 26 | 0 | 40,4 | 35,7 |
| 27 | 1 | 36,5 | 37,3 |
| 28 | 1 | 38,5 | 38,8 |
| 29 | 1 | 37,0 | 36,9 |
| 30 | 1 | 37,5 | 37,1 |
| 31 | 1 | 36,2 | 34,7 |
| 32 | 1 | 40,2 | 47,0 |
| 33 | 1 | 43,0 | 44,6 |
| 34 | 1 | 34,6 | 42,4 |
| 35 | 1 | 36,7 | 33,3 |
| 36 | 1 | 36,5 | 37,2 |
| 37 | 1 | 38,6 | 43,9 |
| 38 | 1 | 40,5 | 43,2 |
| 39 | 1 | 36,9 | 37,4 |
| 40 | 1 | 34,7 | 32,9 |
| 41 | 1 | 31,8 | 41,2 |
| 42 | 1 | 36,5 | 32,3 |
| 43 | 1 | 37,8 | 38,7 |
| 44 | 1 | 37,6 | 37,0 |
| 45 | 1 | 39,8 | 40,4 |
| 46 | 1 | 38,2 | 38,1 |
| 47 | 1 | 39,2 | 37,5 |
| 48 | 1 | 34,7 | 34,3 |
| 49 | 1 | 36,4 | 37,5 |
| 50 | 1 | 39,8 | 45,5 |
| 51 | 1 | 37,8 | 38,6 |
| 52 | 1 | 36,2 | 37,9 |
| 53 | 1 | 38,3 | 38,4 |

| № пацієнта | Група лікування | Вага до лікування, кг | Вага після лікування, кг |
|---|---|---|---|
| 54 | 1 | 36,7 | 43,6 |
| 55 | 1 | 39,6 | 39,3 |
| 56 | 1 | 38,0 | 43,2 |
| 57 | 1 | 37,8 | 42,8 |
| 58 | 1 | 39,0 | 41,5 |
| 59 | 1 | 37,4 | 41,7 |
| 60 | 1 | 39,3 | 45,5 |
| 61 | 1 | 36,1 | 34,8 |
| 62 | 1 | 34,9 | 34,8 |
| 63 | 1 | 42,7 | 46,1 |
| 64 | 1 | 33,3 | 43,0 |
| 65 | 1 | 36,5 | 34,1 |
| 66 | 1 | 37,0 | 35,3 |
| 67 | 1 | 37,2 | 43,3 |
| 68 | 1 | 35,2 | 41,1 |
| 69 | 1 | 37,9 | 42,0 |
| 70 | 1 | 40,8 | 42,5 |
| 71 | 1 | 39,0 | 41,6 |
| 72 | 1 | 39,6 | 44,5 |

Джерело: Rdatasets.URL: https://vincentarelbundock.github.io/Rdatasets/datasets.html

Додаток 13.2

**Програмний код для опрацювання первинних даних**

```
data anorexia_trial;
set anorexia;
ID=input(var1,BEST12.);

/*for simplicity only two treatments*/
```

```
if treat in ('FT' 'CBT') then treatm=1;
else if treat='Cont' then treatm=0;
else treatm='';
format prewgt 6.1 postwgt 6.1;
prewgt=prewt*0.45359237;
/*converting lbs into kg*/
postwgt=postwt*0.45359237;
label ID='Identifier'
treatm='Treatment'
/*0 is control, 1 is medication*/
postwgt='Post weight'
prewgt='Preweight'
run;
drop var1 treat prewt postwt;

/*general statistics*/
proc means data=anorexia_trial N mean min max
std median; var prewgt postwgt;
run;
proc freq data=anorexia_trial; tables treatm;
run;

/*regression model*/
proc reg data=anorexia_trial; model postwgt=treatm
prewgt;
run;
```

Додаток 13.3

**Програмний код для створення пропущених даних MCAR**

```
data help;
call streaminit(114); do ID = 1 to 72;
```

```
    miss1=rand('uniform',0,1);
    miss2=rand('uniform',0,1);
    miss3=rand('uniform',0,1); output;
 end;
 run;

proc sort data=help;
by miss1;
run;

data help1;
set help;
if _N_ LT 8 then miss1=1; else miss1=0;
run;

proc sort data=help1;
by miss2;
run;

data help2;
set help1;
if _N_ LT 19 then miss2=1; else miss2=0;
run;

proc sort data=help2;
by miss3;
run;

data help3;
set help2;
    if _N_ LT 37 then miss3=1;
       else miss3=0;
run;

proc sort;
by ID;
```

```
run;

data anorexia_mcar;
merge anorexia_trial help3; by ID;
       format postwgt1 postwgt2 postwgt3 6.1;
       if miss1=1 then postwgt1=.;
             else postwgt1=postwgt;
             if miss2=1 then postwgt2=.;
             else postwgt2=postwgt;
             if miss3=1 then postwgt3=.;
             else postwgt3=postwgt
run;
label postwgt1='Postweight 10%'
postwgt2='Postweight 25%'
postwgt3='Postweight 50%';
drop miss1 miss2 miss3 postwgt;
```

Додаток 13.4

**Пропущені дані за механізмом MCAR**

| № пацієнта | Група лікування | Вага до лікування, кг | Вага після лікування, кг | | |
|---|---|---|---|---|---|
| | | | 10% пропусків | 25% пропусків | 50% пропусків |
| 1 | 0 | 36,6 | | 36,4 | 36,4 |
| 2 | 0 | 40,6 | 36,3 | 36,3 | |
| 3 | 0 | 41,6 | 39,2 | 39,2 | 39,2 |
| 4 | 0 | 33,6 | 39,1 | 39,1 | 39,1 |
| 5 | 0 | 35,4 | 34,5 | | |
| 6 | 0 | 40,1 | 35,4 | | |
| 7 | 0 | 39,6 | 34,1 | | |
| 8 | 0 | 34,1 | 39,3 | | 39,3 |
| 9 | 0 | 36,6 | | | 33,3 |
| 10 | 0 | 35,6 | 38,4 | 38,4 | |
| 11 | 0 | 35,2 | 35,1 | 35,1 | 35,1 |
| 12 | 0 | 40,2 | 36,1 | 36,1 | |

| № пацієнта | Група лікування | Вага до лікування, кг | Вага після лікування, кг | | |
|---|---|---|---|---|---|
| | | | 10% пропусків | 25% пропусків | 50% пропусків |
| 13 | 0 | 36,9 | 40,6 | 40,6 | |
| 14 | 0 | 35,4 | 36,9 | 36,9 | |
| 15 | 0 | 32,0 | 37,1 | 37,1 | 37,1 |
| 16 | 0 | 35,1 | 35,1 | 35,1 | 35,1 |
| 17 | 0 | 38,6 | 38,2 | 38,2 | |
| 18 | 0 | 39,0 | 34,2 | 34,2 | 34,2 |
| 19 | 0 | 38,1 | 36,1 | | |
| 20 | 0 | 36,2 | 33,1 | | |
| 21 | 0 | 38,8 | | 40,1 | 40,1 |
| 22 | 0 | 38,3 | 38,4 | | 38,4 |
| 23 | 0 | 36,1 | | 36,9 | 36,9 |
| 24 | 0 | 35,2 | 36,8 | 36,8 | 36,8 |
| 25 | 0 | 32,8 | 40,0 | | 40,0 |
| 26 | 0 | 40,4 | 35,7 | 35,7 | 35,7 |
| 27 | 1 | 36,5 | 37,3 | 37,3 | 37,3 |
| 28 | 1 | 38,5 | 38,8 | 38,8 | |
| 29 | 1 | 37,0 | 36,9 | 36,9 | 36,9 |
| 30 | 1 | 37,5 | 37,1 | 37,1 | |
| 31 | 1 | 36,2 | 34,7 | 34,7 | |
| 32 | 1 | 40,2 | 47,0 | 47,0 | |
| 33 | 1 | 43,0 | 44,6 | | 44,6 |
| 34 | 1 | 34,6 | 42,4 | 42,4 | |
| 35 | 1 | 36,7 | 33,3 | 33,3 | |
| 36 | 1 | 36,5 | 37,2 | 37,2 | |
| 37 | 1 | 38,6 | 43,9 | 43,9 | 43,9 |
| 38 | 1 | 40,5 | 43,2 | 43,2 | 43,2 |
| 39 | 1 | 36,9 | 37,4 | 37,4 | |

| № пацієнта | Група лікування | Вага до лікування, кг | Вага після лікування, кг | | |
|---|---|---|---|---|---|
| | | | 10% пропусків | 25% пропусків | 50% пропусків |
| 40 | 1 | 34,7 | 32,9 | | |
| 41 | 1 | 31,8 | 41,2 | | 41,2 |
| 42 | 1 | 36,5 | 32,3 | 32,3 | 32,3 |
| 43 | 1 | 37,8 | 38,7 | 38,7 | 38,7 |
| 44 | 1 | 37,6 | 37,0 | 37,0 | |
| 45 | 1 | 39,8 | | 40,4 | 40,4 |
| 46 | 1 | 38,2 | 38,1 | | |
| 47 | 1 | 39,2 | 37,5 | 37,5 | |
| 48 | 1 | 34,7 | | 34,3 | |
| 49 | 1 | 36,4 | 37,5 | | |
| 50 | 1 | 39,8 | 45,5 | 45,5 | 45,5 |
| 51 | 1 | 37,8 | 38,6 | 38,6 | 38,6 |
| 52 | 1 | 36,2 | 37,9 | 37,9 | |
| 53 | 1 | 38,3 | 38,4 | 38,4 | |
| 54 | 1 | 36,7 | 43,6 | 43,6 | 43,6 |
| 55 | 1 | 39,6 | | | 39,3 |
| 56 | 1 | 38,0 | 43,2 | 43,2 | |
| 57 | 1 | 37,8 | 42,8 | | 42,8 |
| 58 | 1 | 39,0 | 41,5 | 41,5 | |
| 59 | 1 | 37,4 | 41,7 | 41,7 | 41,7 |
| 60 | 1 | 39,3 | 45,5 | 45,5 | 45,5 |
| 61 | 1 | 36,1 | 34,8 | 34,8 | 34,8 |
| 62 | 1 | 34,9 | 34,8 | 34,8 | |
| 63 | 1 | 42,7 | 46,1 | 46,1 | 46,1 |
| 64 | 1 | 33,3 | 43,0 | 43,0 | |
| 65 | 1 | 36,5 | 34,1 | 34,1 | |
| 66 | 1 | 37,0 | 35,3 | 35,3 | |

| № пацієнта | Група лікування | Вага до лікування, кг | Вага після лікування, кг | | |
|---|---|---|---|---|---|
| | | | 10% пропусків | 25% пропусків | 50% пропусків |
| 67 | 1 | 37,2 | 43,3 | 43,3 | 43,3 |
| 68 | 1 | 35,2 | 41,1 | 41,1 | |
| 69 | 1 | 37,9 | 42,0 | | |
| 70 | 1 | 40,8 | 42,5 | 42,5 | 42,5 |
| 71 | 1 | 39,0 | 41,6 | 41,6 | |
| 72 | 1 | 39,6 | 44,5 | | 44,5 |

Додаток 13.5

**Метод повних спостережень**

```
data cc_1;
set anorexia_mcar;
if nmiss(postwgt1)=0;
run;
proc reg data=cc_1;
model postwgt1= treatm prewgt;
run;
```

**Метод доступних спостережень**

```
proc corr data=anorexia_mcar cov outp=ac_1;
var postwgt1 treatm prewgt;
run;

proc reg data=ac_1;
model postwgt1=treatm prewgt;
run;
```

**Метод безумовної імпутації**

```
proc means data=anorexia_mcar mean;
 var postwgt1 postwgt2 postwgt3; output
out=mean1;
run;

data means;
set mean1;
where _STAT_='MEAN';
drop _TYPE_ _FREQ_ _STAT_;
do ID=1 to 72;
m_postwgt1=postwgt1;
m_postwgt2=postwgt2;
m_postwgt3=postwgt3;
output;
end;
drop postwgt1 postwgt2 postwgt3;
run;
data unconditional;
merge anorexia_mcar means;
by ID;
format unpostwgt1 unpostwgt2 unpostwgt3 6.1;
if postwgt1=. then unpostwgt1=m_postwgt1;
else unpostwgt1=postwgt1;
if postwgt2=. then unpostwgt2=m_postwgt2;
else unpostwgt2=postwgt2;
if postwgt3=. then unpostwgt3=m_postwgt3;
else unpostwgt3=postwgt3;
drop m_postwgt1 m_postwgt2 m_postwgt3;
run;

proc reg data=unconditional;
```

```
model unpostwgt1=treatm prewgt;
run;
```

**Метод умовної імпутації**

```
proc mi data=anorexia_mcar nimpute=1
seed=37887 out=cond_1; fcs nbiter=1;
var postwgt1 treatm prewgt;
run;

proc reg data=cond_1;
model postwgt1= treatm prewgt;
run;
```

**Заповнення пропусків з послідовним підбором**

```
data locf;
set anorexia_mcar;
format postw1 postw2 postw3 6.1;
if postwgt1=. then postw1=prewgt;
else postw1=postwgt1;
if postwgt2=. then postw2=prewgt;
else postw2=postwgt2;
if postwgt3=. then postw3=prewgt;
else postw3=postwgt3;
run;
proc reg data=locf;
model postwgt1=treatm prewgt;
run;
```

**Заповнення пропусків з упередженим підбором**

```
proc surveyimpute data=anorexia_mcar
method=HOTDECK;
   var postwgt1;
```

```
   output out=MarginalHotDeck1;
run;

proc reg data=MarginalHotDeck1;
model postwgt1= treatm prewgt;
run;
```

Додаток 13.6

**Програмний код для створення пропущених даних MAR**

```
data anorexia_mar1;
set anorexia_trial;
if prewgt le 35 or prewgt ge 40 then miss1=1;
/*for 10% and 25%*/ else miss1=0;
if prewgt le 37 or prewgt ge 40 then miss2=1;
/*for 50%*/ else miss2=0;
run;

proc sort data=anorexia_mar1;
by descending miss1 postwgt;
run;

/*10% and 25% missing*/
data anorexia_mar2;
set anorexia_mar1;
if _N_ LT 8 then postwgt1=.;
else postwgt1=postwgt;
if _N_ LT 19 then postwgt2=.;
else postwgt2=postwgt;
run;

/*50% missing*/
proc sort data=anorexia_mar2;
```

```
by miss2;
run;
data anorexia_mar;
set anorexia_mar2;
if _N_ LT 37 then postwgt3=.;
else postwgt3=postwgt;
label postwgt1='Postweight 10%'
postwgt2='Postweight 25%'
postwgt3='Postweight 50%';
drop miss1 miss2 postwgt;
run;
proc sort;
by ID;
run;
```

Додаток 13.7

**Пропущені дані за механізмом MAR**

| № пацієнта | Група лікування | Вага до лікування, кг | Вага після лікування, кг | | |
|---|---|---|---|---|---|
| | | | 10% пропусків | 25% пропусків | 50% пропусків |
| 1 | 0 | 36,6 | 36.4 | 36.4 | 36.4 |
| 2 | 0 | 40,6 | | | |
| 3 | 0 | 41,6 | 39.2 | | 39.2 |
| 4 | 0 | 33,6 | 39.1 | | 39.1 |
| 5 | 0 | 35,4 | 34.5 | 34.5 | 34.5 |
| 6 | 0 | 40,1 | | | |
| 7 | 0 | 39,6 | 34.1 | 34.1 | |
| 8 | 0 | 34,1 | 39.3 | | 39.3 |
| 9 | 0 | 36,6 | 33.3 | 33.3 | 33.3 |
| 10 | 0 | 35,6 | 38.4 | 38.4 | 38.4 |
| 11 | 0 | 35,2 | 35.1 | 35.1 | 35.1 |
| 12 | 0 | 40,2 | | | |
| 13 | 0 | 36,9 | 40.6 | 40.6 | 40.6 |

| № пацієнта | Група лікування | Вага до лікування, кг | Вага після лікування, кг | | |
|---|---|---|---|---|---|
| | | | 10% пропусків | 25% пропусків | 50% пропусків |
| 14 | 0 | 35,4 | 36.9 | 36.9 | 36.9 |
| 15 | 0 | 32,0 | 37.1 | | |
| 16 | 0 | 35,1 | 35.1 | 35.1 | 35.1 |
| 17 | 0 | 38,6 | 38.2 | 38.2 | |
| 18 | 0 | 39,0 | 34.2 | 34.2 | |
| 19 | 0 | 38,1 | 36.1 | 36.1 | |
| 20 | 0 | 36,2 | 33.1 | 33.1 | 33.1 |
| 21 | 0 | 38,8 | 40.1 | 40.1 | |
| 22 | 0 | 38,3 | 38.4 | 38.4 | |
| 23 | 0 | 36,1 | 36.9 | 36.9 | 36.9 |
| 24 | 0 | 35,2 | 36.8 | 36.8 | 36.8 |
| 25 | 0 | 32,8 | 40.0 | | 40.0 |
| 26 | 0 | 40,4 | | | |
| 27 | 1 | 36,5 | 37.3 | 37.3 | 37.3 |
| 28 | 1 | 38,5 | 38.8 | 38.8 | |
| 29 | 1 | 37,0 | 36.9 | 36.9 | 36.9 |
| 30 | 1 | 37,5 | 37.1 | 37.1 | |
| 31 | 1 | 36,2 | 34.7 | 34.7 | 34.7 |
| 32 | 1 | 40,2 | 47.0 | 47.0 | 47.0 |
| 33 | 1 | 43,0 | 44.6 | | 44.6 |
| 34 | 1 | 34,6 | 42.4 | | 42.4 |
| 35 | 1 | 36,7 | 33.3 | 33.3 | 33.3 |
| 36 | 1 | 36,5 | 37.2 | 37.2 | 37.2 |
| 37 | 1 | 38,6 | 43.9 | 43.9 | |
| 38 | 1 | 40,5 | 43.2 | | 43.2 |
| 39 | 1 | 36,9 | 37.4 | 37.4 | 37.4 |
| 40 | 1 | 34,7 | | | |
| 41 | 1 | 31,8 | 41.2 | | 41.2 |
| 42 | 1 | 36,5 | 32.3 | 32.3 | 32.3 |
| 43 | 1 | 37,8 | 38.7 | 38.7 | |
| 44 | 1 | 37,6 | 37.0 | 37.0 | |

| № пацієнта | Група лікування | Вага до лікування, кг | Вага після лікування, кг | | |
|---|---|---|---|---|---|
| | | | 10% пропусків | 25% пропусків | 50% пропусків |
| 45 | 1 | 39,8 | 40.4 | 40.4 | |
| 46 | 1 | 38,2 | 38.1 | 38.1 | |
| 47 | 1 | 39,2 | 37.5 | 37.5 | |
| 48 | 1 | 34,7 | | | |
| 49 | 1 | 36,4 | 37.5 | 37.5 | 37.5 |
| 50 | 1 | 39,8 | 45.5 | 45.5 | |
| 51 | 1 | 37,8 | 38.6 | 38.6 | |
| 52 | 1 | 36,2 | 37.9 | 37.9 | 37.9 |
| 53 | 1 | 38,3 | 38.4 | 38.4 | |
| 54 | 1 | 36,7 | 43.6 | 43.6 | 43.6 |
| 55 | 1 | 39,6 | 39.3 | 39.3 | |
| 56 | 1 | 38,0 | 43.2 | 43.2 | |
| 57 | 1 | 37,8 | 42.8 | 42.8 | |
| 58 | 1 | 39,0 | 41.5 | 41.5 | |
| 59 | 1 | 37,4 | 41.7 | 41.7 | |
| 60 | 1 | 39,3 | 45.5 | 45.5 | |
| 61 | 1 | 36,1 | 34.8 | 34.8 | 34.8 |
| 62 | 1 | 34,9 | | | |
| 63 | 1 | 42,7 | 46.1 | 46.1 | 46.1 |
| 64 | 1 | 33,3 | 43.0 | | 43.0 |
| 65 | 1 | 36,5 | 34.1 | 34.1 | 34.1 |
| 66 | 1 | 37,0 | 35.3 | 35.3 | |
| 67 | 1 | 37,2 | 43.3 | 43.3 | |
| 68 | 1 | 35,2 | 41.1 | 41.1 | 41.1 |
| 69 | 1 | 37,9 | 42.0 | 42.0 | |
| 70 | 1 | 40,8 | 42.5 | | 42.5 |
| 71 | 1 | 39,0 | 41.6 | 41.6 | |
| 72 | 1 | 39,6 | 44.5 | 44.5 | |

## Таблиця критичних значень для кореляційного відношення $\eta^2$ і коефіцієнта детермінації $R^2$ ($\alpha$=0,05)

| $k_2$ | $k_1$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 10 | 20 |
| 3 | 0,771 | 865 | 903 | 924 | 938 | 947 | 959 | 967 | 983 |
| 4 | 658 | 776 | 832 | 865 | 887 | 902 | 924 | 937 | 967 |
| 5 | 569 | 699 | 764 | 806 | 835 | 854 | 885 | 904 | 948 |
| 6 | 500 | 632 | 704 | 751 | 785 | 811 | 847 | 871 | 928 |
| 7 | 444 | 575 | 651 | 702 | 739 | 768 | 810 | 839 | 908 |
| 8 | 399 | 527 | 604 | 657 | 697 | 729 | 775 | 807 | 887 |
| 9 | 362 | 488 | 563 | 628 | 659 | 692 | 742 | 777 | 867 |
| 10 | 332 | 451 | 527 | 582 | 624 | 659 | 711 | 749 | 847 |
| 11 | 306 | 420 | 495 | 550 | 593 | 628 | 682 | 722 | 828 |
| 12 | 283 | 394 | 466 | 521 | 564 | 600 | 655 | 696 | 809 |
| 14 | 247 | 345 | 417 | 471 | 514 | 550 | 607 | 650 | 773 |
| 16 | 219 | 312 | 378 | 429 | 477 | 507 | 564 | 609 | 740 |
| 18 | 197 | 283 | 348 | 394 | 435 | 470 | 527 | 573 | 709 |
| 20 | 179 | 259 | 318 | 364 | 404 | 432 | 495 | 540 | 680 |
| 22 | 164 | 238 | 294 | 339 | 377 | 410 | 466 | 511 | 653 |
| 24 | 151 | 221 | 273 | 316 | 353 | 385 | 440 | 484 | 628 |
| 26 | 140 | 206 | 256 | 297 | 332 | 363 | 417 | 461 | 605 |
| 28 | 130 | 193 | 240 | 279 | 314 | 344 | 396 | 439 | 583 |
| 30 | 122 | 182 | 227 | 264 | 297 | 326 | 373 | 419 | 563 |
| 32 | 115 | 171 | 214 | 250 | 282 | 310 | 360 | 401 | 544 |
| 34 | 108 | 162 | 203 | 238 | 268 | 296 | 344 | 384 | 526 |
| 36 | 102 | 153 | 102 | 226 | 256 | 282 | 329 | 368 | 509 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 38 | 097 | 146 | 184 | 218 | 245 | 271 | 316 | 355 | 493 |
| 40 | 093 | 139 | 176 | 207 | 234 | 259 | 304 | 342 | 479 |
| 50 | 075 | 113 | 143 | 170 | 194 | 216 | 254 | 288 | 416 |
| 60 | 063 | 095 | 121 | 144 | 165 | 184 | 218 | 249 | 368 |
| 80 | 047 | 072 | 093 | 110 | 127 | 142 | 170 | 196 | 298 |
| 100 | 038 | 058 | 075 | 090 | 103 | 116 | 140 | 161 | 251 |
| 120 | 032 | 049 | 063 | 075 | 087 | 098 | 119 | 137 | 217 |
| 200 | 019 | 030 | 038 | 046 | 053 | 060 | 073 | 086 | 139 |
| 400 | 010 | 015 | 019 | 023 | 027 | 031 | 038 | 044 | 074 |

## Таблиця критичних значень для критерію Фішера ($\alpha$=0,05)

| $k_2$ | $k_1$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 10 | 20 |
| 1 | 161,4 | 199,5 | 215,7 | 224,6 | 230,2 | 234,0 | 238,9 | 242,0 | 248,0 |
| 2 | 18,51 | 19,00 | 19,16 | 19,25 | 19,30 | 19,33 | 19,37 | 19,39 | 19,44 |
| 3 | 10,13 | 9,45 | 9,28 | 9,12 | 9,01 | 8,94 | 8,84 | 8,78 | 8,66 |
| 4 | 7,71 | 6,94 | 6,59 | 6,39 | 6,26 | 6,16 | 6,04 | 5,96 | 5,80 |
| 5 | 6,61 | 5,79 | 5,41 | 5,19 | 5,05 | 4,95 | 4,82 | 4,74 | 4,56 |
| 6 | 5,99 | 5,14 | 4,76 | 4,53 | 4,39 | 4,28 | 4,15 | 4,06 | 3,87 |
| 7 | 5,59 | 4,74 | 4,35 | 4,12 | 3,97 | 3,87 | 3,73 | 3,63 | 3,44 |
| 8 | 5,32 | 4,46 | 4,07 | 3,84 | 3,69 | 3,58 | 3,44 | 3,34 | 3,15 |
| 9 | 5,12 | 4,26 | 3,86 | 3,63 | 3,48 | 3,37 | 3,23 | 3,13 | 2,93 |
| 10 | 4,96 | 4,10 | 3,71 | 3,48 | 3,33 | 3,22 | 3,07 | 2,97 | 2,77 |
| 11 | 4,82 | 3,98 | 3,59 | 3,63 | 3,20 | 3,09 | 2,95 | 2,86 | 2,65 |
| 12 | 4,75 | 3,88 | 3,49 | 3,26 | 3,11 | 3,00 | 2,85 | 2,76 | 2,54 |
| 14 | 4,60 | 3,74 | 3,34 | 3,11 | 2,96 | 2,85 | 2,70 | 2,60 | 2,39 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 16 | 4,49 | 3,63 | 3,24 | 3,01 | 2,85 | 2,74 | 2,59 | 2,49 | 2,28 |
| 18 | 4,41 | 3,55 | 3,16 | 2,93 | 2,77 | 2,66 | 2,51 | 2,41 | 2,19 |
| 20 | 4,35 | 3,49 | 3,10 | 2,87 | 2,71 | 2,60 | 2,45 | 2,35 | 2,12 |
| 30 | 4,17 | 3,32 | 2,92 | 2,69 | 2,53 | 2,42 | 2,27 | 2,16 | 1,93 |
| 40 | 4,08 | 3,23 | 2,84 | 2,61 | 2,45 | 2,34 | 2,18 | 2,12 | 1,84 |
| 60 | 4,00 | 3,15 | 2,76 | 2,52 | 2,37 | 2,25 | 2,10 | 2,04 | 1,7Ь |
| 120 | 3,92 | 3,07 | 2,68 | 2,45 | 2,29 | 2,17 | 2,02 | 1,90 | 1,65 |

## Квантилі $\chi^2$ розподілу

| *f* | ***1*** **- α** | | | | | |
|---|---|---|---|---|---|---|
| | 0,025 | 0,050 | 0,10 | 0,90 | 0,95 | 0,975 |
| A | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 0,01 | 0,04 | 0,02 | 2,71 | 3,84 | 5,02 |
| 2 | 0,05 | 0,10 | 0,21 | 4,61 | 5,99 | 7,38 |
| 3 | 0,22 | 0,35 | 0,58 | 6,25 | 7,82 | 9,35 |
| 4 | 0,48 | 0,71 | 1,06 | 7,78 | 9,49 | 11,14 |
| 5 | 0,83 | 1,15 | 1,61 | 9,24 | 11,07 | 12,03 |
| 6 | 1,24 | 1,64 | 2,20 | 10,65 | 12,59 | 14,45 |
| 7 | 1,69 | 2,17 | 2,83 | 12,02 | 14,07 | 16,01 |
| 8 | 2,18 | 2,73 | 3,49 | 13,36 | 15,51 | 17,54 |
| 9 | 2,70 | 3,33 | 4,17 | 14,68 | 16,92 | 19,02 |
| 10 | 3,25 | 3,94 | 4,87 | 15,99 | 18,31 | 20,48 |
| 11 | 3,82 | 4,58 | 5,58 | 17,28 | 19,68 | 21,92 |
| 12 | 4,40 | 5,23 | 6,30 | 18,55 | 21,03 | 23,34 |
| 13 | 5,01 | 5,89 | 7,04 | 19,81 | 22,36 | 24,74 |
| 14 | 5,63 | 6,57 | 7,79 | 21,06 | 23,69 | 26,12 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 15 | 6,26 | 7,26 | 8,55 | 22,31 | 25,00 | 27,49 |
| 16 | 6,91 | 7,96 | 9,31 | 23,54 | 26,30 | 28,85 |
| 17 | 7,56 | 8,67 | 10,09 | 24,77 | 27,59 | 30,19 |
| 18 | 8,23 | 9,39 | 10,87 | 25,09 | 28,87 | 31,53 |
| 19 | 8,91 | 10,12 | 11,65 | 27,20 | 30,14 | 32,85 |
| 20 | 9,59 | 10,85 | 12,44 | 28,41 | 31,41 | 34,17 |
| 22 | 10,98 | 12,34 | 14,04 | 30,81 | 33,92 | 36,78 |
| 24 | 12,40 | 13,85 | 15,66 | 33,20 | 36,42 | 39,36 |
| 26 | 13,84 | 15,38 | 17,29 | 35,56 | 38,89 | 41,92 |
| 28 | 15,31 | 16,93 | 18,94 | 37,92 | 41,34 | 44,46 |
| 30 | 16,79 | 18,49 | 20,60 | 40,26 | 43,77 | 46,90 |
| 35 | 20,57 | 22,47 | 24,80 | 46,06 | 49,00 | 53,20 |
| 40 | 24,43 | 26,51 | 29,05 | 51,81 | 55,76 | 59,34 |
| 45 | 28,37 | 30,61 | 33,35 | 57,51 | 61,66 | 65,41 |
| 50 | 32,36 | 31,76 | 37,69 | 63,17 | 67,51 | 71,42 |

## Квантилі розподілу Стьюдента t

| *k* | ***1*** **-** $\alpha$ | | | *k* | ***1*** **-** $\alpha$ | | |
|---|---|---|---|---|---|---|---|
| | 0,90 | 0,95 | 0,975 | | 0,90 | 0,95 | 0,975 |
| 3 | 1,64 | 2,35 | 3,18 | 12 | 1,36 | 1,78 | 2,18 |
| 4 | 1,53 | 2,13 | 2,78 | 14 | 1,35 | 1,76 | 2,14 |
| 5 | 1,48 | 2,02 | 2,57 | 16 | 1,34 | 1,75 | 2,12 |
| 6 | 1,44 | 1,94 | 2,45 | 18 | 1,33 | 1,73 | 2,10 |
| 7 | 1,41 | 1,89 | 2,36 | 20 | 1,33 | 1,72 | 2,09 |
| 8 | 1,40 | 1,86 | 2,31 | 22 | 1,32 | 1,72 | 2,07 |
| 9 | 1,38 | 1,83 | 2,26 | 24 | 1,32 | 1,71 | 2,06 |
| 10 | 1,37 | 1,81 | 2,23 | 28 | 1,31 | 1,70 | 2,05 |
| 11 | 1,36 | 1,80 | 2,20 | $\infty$ | 1,28 | 1,64 | 1,96 |

## Квантилі нормального розподілу

| **1 -** ***a*** | 0,800 | 0,900 | 0,950 | 0,975 |
|---|---|---|---|---|
| *t* | 0,84 | 1,28 | 1,64 | 1,96 |
| $\lvert t \rvert$ | 1,28 | 1,64 | 1,96 | 2,24 |

## Критичні значення лінійного коефіцієнта кореляції ($\alpha$=0,05)

| Обсяг вибірки, | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 14 | 16 |
|---|---|---|---|---|---|---|---|---|---|
| $r_{0,95}$ | 0,88 | 0,81 | 0,75 | 0,71 | 0,67 | 0,63 | 0,58 | 0,53 | 0,50 |

## Критичні значення коефіцієнта рангової кореляції Спірмена ($\alpha$=0,05)

| Обсяг вибірки, *n* | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|
| $P_{0,95}$ | 0,90 | 0,83 | 0,71 | 0,64 | 0,60 | 0,56 | 0,53 | 0,50 |

## Критичні значення функції $K_{(\lambda)}$ А.Н. Колмогорова ($\alpha$=0,05)

| $\lambda$ | 1,23 | 1,36 | 1,63 | 1,80 | 2,03 |
|---|---|---|---|---|---|
| $K_{(\lambda)}$ | 0,9030 | 0,9505 | 0,9902 | 0,9970 | 0,9993 |

## Критичні значення кумулятивного критерію ($\alpha$=0,05)

| *n* | Для перевірки істотності тренду | | Для перевірки гіпотези про форму тренду | |
|---|---|---|---|---|
| | *T* | *t* | Лінійна функція | Парабола другого порядку |
| 6 | 2,62 | 2,08 | 0,85 | 0,51 |
| 7 | 3,11 | 2,10 | 1,01 | 0,61 |
| 8 | 3,59 | 2,09 | 1,17 | 0,70 |
| 9 | 4,07 | 2,09 | 1,32 | 0,79 |
| 10 | 4,55 | 2,09 | 1,48 | 0,89 |
| 11 | 5,02 | 2,08 | 1,63 | 0,98 |
| 12 | 5,49 | 2,08 | 1,78 | 1,06 |
| 13 | 5,96 | 2,07 | 1,93 | 1,15 |
| 14 | 6,42 | 2,07 | 2,08 | 1,23 |
| 15 | 6,89 | 2,06 | 2,23 | 1,33 |
| 16 | 7,36 | 2,06 | 2,38 | 1,42 |
| 17 | 7,82 | 2,06 | 2,58 | 1,51 |
| 18 | 8,29 | 2,05 | 2,68 | 1,59 |
| 19 | 8,76 | 2,05 | 2,83 | 1,68 |
| 20 | 9,22 | 2,04 | 2,98 | 1,77 |
| 22 | 10,20 | 2,04 | 3,28 | 1,94 |
| 24 | 11,00 | 2,04 | 3,58 | 2,11 |
| 26 | 12,00 | 2,03 | 3,88 | 2,29 |
| 28 | 12,90 | 2,03 | 4,18 | 2,46 |
| 30 | 13,90 | 2,03 | 4,47 | 2,63 |

## Значення стандартного нормального розподілу

| z | 0,00 | 0,01 | 0,02 | 0,03 | 0,04 | 0,05 | 0,06 | 0,07 | 0,08 | 0,09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0,0 | 0,0000 | 0400 | 0080 | 0120 | 0160 | 0199 | 0239 | 0279 | 0319 | 0359 |
| 0,1 | 0398 | 0438 | 0478 | 0517 | 0557 | 0596 | 0636 | 0675 | 0711 | 0753 |
| 0,2 | 0793 | 0832 | 0871 | 0910 | 0948 | 0987 | 1026 | 1064 | 1103 | 1141 |
| 0,3 | 1179 | 1217 | 1255 | 1293 | 1331 | 1368 | 1106 | 1113 | 1480 | 1517 |
| 0,4 | 1554 | 1591 | 1628 | 1664 | 1700 | 1736 | 1772 | 1808 | 1811 | 1879 |
| 0,5 | 1915 | 1950 | 1985 | 2019 | 2054 | 2088 | 2123 | 2157 | 2190 | 2224 |
| 0,6 | 2257 | 2291 | 2324 | 2357 | 2389 | 2422 | 2454 | 2486 | 2517 | 2549 |
| 0,7 | 2580 | 2611 | 2642 | 2673 | 2704 | 2734 | 2764 | 2794 | 2823 | 2852 |
| 0,8 | 2881 | 2910 | 2939 | 2967 | 2995 | 3023 | 3051 | 3078 | 3106 | 3133 |
| 0,9 | 3159 | 3186 | 3212 | 3238 | 3264 | 3289 | 3315 | 3310 | 3365 | 3389 |
| 1,0 | 3413 | 3438 | 3461 | 3485 | 3508 | 3531 | 3554 | 3577 | 3S99 | 3621 |
| 1,1 | 3643 | 3665 | 3686 | 3708 | 3729 | 3749 | 3770 | 3790 | 3810 | 3830 |
| 1,2 | 3849 | 3869 | 3888 | 3907 | 3925 | 3944 | 3962 | 3980 | 3997 | 4015 |
| 1,3 | 4032 | 4049 | 4066 | 4082 | 4099 | 4115 | 4131 | 4117 | 4162 | 1177 |
| 1,4 | 4192 | 4207 | 4222 | 4236 | 4251 | 1265 | 4279 | 4292 | 4306 | 4319 |
| 1,5 | 4332 | 4345 | 4357 | 4370 | 4382 | 4394 | 4406 | 4118 | 4429 | 1141 |
| 1,6 | 4452 | 4463 | 4474 | 4484 | 4495 | 4505 | 4515 | 4525 | 4535 | 4545 |
| 1,7 | 4554 | 4564 | 4573 | 4582 | 1591 | 4599 | 4608 | 4616 | 4625 | 4633 |
| 1,8 | 4641 | 4649 | 4656 | 4664 | 4671 | 1678 | 4686 | 4693 | 4699 | 4706 |
| 1,9 | 4713 | 4719 | 4726 | 4732 | 4738 | 4744 | 4750 | 1756 | 4761 | 4767 |
| 2,0 | 4772 | 4778 | 4783 | 47R8 | 1793 | 4798 | 1803 | 4808 | 4812 | 4817 |
| 2,1 | 4821 | 4826 | 4830 | 4834 | 4838 | 4842 | 4846 | 4850 | 1854 | 4857 |
| 2,2 | 4861 | 4864 | 4868 | 4871 | 4875 | 4878 | 4881 | 4884 | 4887 | 4890 |
| 2,3 | 4893 | 4896 | 4898 | 4901 | 4904 | 4906 | 4909 | 4911 | 4913 | 4916 |
| 2,4 | 4918 | 4920 | 4922 | 4925 | 4927 | 4929 | 4931 | 4932 | 4934 | 4936 |
| 2,5 | 4938 | 4940 | 4941 | 4943 | 4915 | 4946 | 4948 | 4919 | 4951 | 4952 |
| 2,6 | 4953 | 4955 | 4956 | 4957 | 4959 | 4960 | 4961 | 4962 | 4963 | 4964 |
| 2,7 | 4965 | 4966 | 4967 | 4968 | 4969 | 4970 | 4971 | 4972 | 4973 | 4971 |
| 2,8 | 4974 | 4975 | 4976 | 4977 | 4977 | 4978 | 4979 | 4979 | 4980 | 4981 |
| 2,9 | 4981 | 4982 | 4982 | 4983 | 4984 | 4984 | 4985 | 4985 | 4986 | 4981 |
| 3,0 | 4987 | 4987 | 1987 | 4988 | 4988 | 4989 | 4989 | 4989 | 4990 | 4990 |

# ЗМІСТ

# *Для нотаток*

Навчальний посібник

# ОСНОВИ СТАТИСТИЧНОГО МОДЕЛЮВАННЯ

За загальною редакцією С. В. Чугаєвської, Н. В. Ковтун

Підписано до друку 25.04.2022 р.
Формат 60x84/16.
Гарнітура Times New Roman
Ум. друк. арк. 17,25
Наклад 100 прим. Зам. № 3459.

**Віддруковано в ПП "Рута"**
10014, Україна,
м. Житомир, вул. Мала Бердичівська 17а
Свідоцтво суб'єкта видавничої справи
ДК №3671 від 14.01.2010
E-mail: ruta-bond@ukr.net