Т.М. ПАЯНОК, Т.М. ЗАДОРОЖНЯ

СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ

НАВЧАЛЬНИЙ ПОСІБНИК

СЕРІЯ «НА ДОПОМОГУ СТУДЕНТУ УДФСУ»

Серію «На допомогу студенту УДФСУ» засновано 2016 року.

У СЕРІЇ «НА ДОПОМОГУ СТУДЕНТУ УДФСУ» ВИЙШЛИ ДРУКОМ:

2016

«Методичні основи спеціальної фізичної та технічної підготовки студентів за розділом «Легка атлетика»

«Самостійна робота студента як одна з форм впливу на функціональну, фізичну та психологічну підготовленість»

«Організація роботи командира механізованого взводу»

«5.45-мм автомати Калашникова (АК-74, АКС-74, АК-74Н, АКС-74Н) та 5.45-мм ручні кулемети Калашникова (РПК-74, РПКС-74, РПК-74Н, РПКС-74Н)»

«Гранатомет підствольний ГП-25»

«Ручні гранати»

«Кулемети Калашникова – 7.62, ПК, УЖМ, ПКТ»

«Ручний протитанковий гранатомет РПГ-7»

«9-мм пістолет Макарова (ПМ)»

2017

«Вища та прикладна математика»

«Цивільний захист»

«Програмування мовою JAVA : практикум»

«Інформаційні системи і технології в юридичній практиці»

«Дослідження операцій : практикум»

«Чисельні методи»

«English for Students of Finance»

«Основи військової розвідки»

2018

«CASE-технології. Міждисциплінарне інформаційне моделювання»

«Економічна інформатика: практикум»

«Економічна теорія (політекономія, мікроекономіка, макроекономіка). Політекономія»

«Економічна теорія (політекономія, мікроекономіка, макроекономіка). Мікроекономіка»

«Економічна теорія (політекономія, мікроекономіка, макроекономіка). Макроекономіка»

«Охорона праці»

«Економіка і організація діяльності об'єднань підприємств»

«Основи християнської культури»

«Економіка підприємства»

«Фізика»

«Трудове право України»

2019

«Основи тактичної медицини»

«Аудит»

«Збірник задач. Вища та прикладна математика»

«Міжнародні розрахунки та валютні операції»

«Підготовка озброєння механізованого взводу до бойового застосування»

«Контролінг в управлінні підприємством»

«Актуальні питання судової експертизи (у питаннях і відповідях)»

«Інноваційний менеджмент»

«Організація навчальних занять з фізичного виховання за методом колового тренування»

«Теорія судових доказів (у таблицях і схемах)»

«Статистика»

«Фінансовий менеджмент проектів і програм»

«Ручний протитанковий гранатомет РПГ-7 (РПГ-7Д)»

«Організація роботи командира механізованого відділення»

«Психологія управління»

«Deutsch und Wirtschaft»

«Основи тактичної підготовки працівників правоохоронних органів»

«Основи кінології»

«Моделювання систем»

«UML. Уніфікована мова моделювання інформаційних систем»

«Історія держави і права України»

«Економічна теорія»

2020

«Статистичний аналіз даних»

УНІВЕРСИТЕТ ДЕРЖАВНОЇ ФІСКАЛЬНОЇ СЛУЖБИ УКРАЇНИ

Серія «На допомогу студенту УДФСУ»
Заснована 2016 року

Т. М. Паянок,
Т .М. Задорожня

СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ

Навчальний посібник

Ірпінь
2020

УДК 519.23:004(075.8)
ББК 22.172я73
П22

Рекомендовано до друку Вченою радою
Університету державної фіскальної служби України
(протокол № 12 від 28 листопада 2019 р.)

Рецензенти:

Скрипник А. В., доктор економічних наук, професор, завідувач кафедри економічної кібернетики Національного університету біоресурсів і природокористування України;

Сторожук Є. А., доктор фізико-математичних наук, професор, провідний науковий співробітник відділу динаміки та стійкості суцільних середовищ Інституту механіки ім. С. П. Тимошенка НАН України.

Паянок Т. М.

П 22 Статистичний аналіз даних : навчальний посібник / Т. М. Паянок, Т. М. Задорожня. – Ірпінь : Університет державної фіскальної служби України, 2020. – 312 с. – (Серія «На допомогу студенту УДФСУ»; т. 60).

ISBN 978-966-337-554-0

У навчальному посібнику детально поданий практичний матеріал, який дозволяє оволодіти основними поняттями та принципами сучасних методів статистичного аналізу, розкриті методи та засоби дослідження соціально-економічних показників у різних сферах. Перевагою навчального посібника є використання функцій статистичного аналізу сучасного програмного продукту з обробки великих масивів даних, а саме SPSS. Ця програма дає можливість обробляти як динамічні, так і статичні масиви даних. При цьому вона має великі переваги в обробці результатів анкетування.

УДК 519.23:004(075.8)
ББК 22.172я73

ISBN 978-966-337-554-0

Зміст

ПЕРЕДМОВА

Основним інструментом дослідження і прогнозування економічних явищ і об'єктів є економіко-математичні методи. Володіння цим інструментарієм стає обов'язковою умовою для успішного молодого дослідника. Проведення якісного статистичного аналізу на мікро- і макрорівнях дає змогу приймати адекватні й ефективні рішення в управлінні як державою, так і фінансово-господарською діяльністю суб'єктів господарювання. Це потребує глибоких знань і вмінь практичного використання принципів та інструментарію економіки, математичної статистики та сучасних прикладних програм. Обрана тема для навчального посібника є актуальною, тому що обробка значних масивів потребує великих затрат часу і трудових ресурсів.

Вивчення статистичних даних дозволяє встановити статистичні закономірності, які властиві масовим випадковим явищам. При цьому необхідно зауважити, що не завжди можна проаналізувати всю статистичну сукупність через брак ресурсів. Якщо суцільне обстеження проводити не можна або недоречно, роблять відбірку з усієї сукупності, яку досліджують на репрезентативність, з метою перенесення результатів дослідження на всю генеральну сукупність. Застосування сучасних програм в обробці великих масивів даних зменшують затрати часу і збільшують точність результатів. Програмний продукт IBM SPSS STATISTICS надає широкі можливості досліднику в обробці як статичних, так і динамічних показників. Зазвичай її використовують для обробки соціологічних досліджень за будь-якою тематикою.

Навчальний посібник містить необхідні відомості для оволодіння основними поняттями та принципами сучасних методів статистичного аналізу, повністю розкриває методи та засоби дослідження соціально-економічних показників у різних сферах за допомогою прикладної програми.

Написаний простою мовою, на доступному рівні детально ознайомлює читачів з проведенням статистичного аналізу даних. Має прикладний аспект, оскільки усі приклади розглянути за конкретними соціально-економічними показниками, які використовуються на макро- та мікрорівнях, буде цікавим як для студентів закладів вищої освіти, так і практиків, які займаються аналітичною роботою.

РОЗДІЛ 1
ФОРМУВАННЯ СТАТИСТИЧНОЇ ВИБІРКИ

1.1. Сутність статистичної вибірки

Статистичні методи працюють у контексті. Потрібно розуміти сутність і логіку досліджуваного явища.

Приклад. Маючи кількість обслуговування автомобілів однією заправкою та обсяг території, яку обслуговує одна заправка, необхідно розрахувати, скільки заправок потрібно побудувати в державі.

1. Від кількості обслуговування автомобілів однією заправкою.
2. Залежно від території, яку обслуговує одна заправка (рис. 1.1).

Рис. 1.1. Кількість заправок для забезпечення потреб держави

Висновок: залежно від мети дослідника можна обрати будь-який варіант, при цьому необхідно враховувати доречність побудови заправок у глухих місцевостях, де населення має одиниці автомобілів.

Результати статистичних досліджень необхідно сприймати з таких позицій:

1. Хто проводив дослідження?
2. Звідки йому це відомо?

3. Чого не вистачає у його дослідженні?
4. Чи не замінений об’єкт дослідження?
5. Чи є сенс у цьому?

Термінологія:

Дані (data) – факти і числа, за допомогою яких можуть прийматися рішення.

Змінні (variable) – будь-яка характеристика об’єкта.

Набір даних (data set) – дані, що відібрані для конкретного аналізу.

Статистичне спостереження може бути суцільним і вибірковим.

Переваги вибіркового обстеження:

– економія матеріальних, фінансових, трудових ресурсів;
– оперативність;
– деталізація програми дослідження у разі обмежених ресурсів дає можливість проведення поглибленого дослідження за рахунок програми дослідження;
– є єдиноможливим у випадку нескінченої генеральної сукупності або у випадку, коли дослідження пов’язано зі знищенням об’єктів, що спостерігаються;
– дозволяє знизити помилки реєстрації, тобто розбіжність між істинним і зареєстрованим значенням властивості.

Основний недолік вибіркового методу – це помилки дослідження, які називають помилками репрезентативності (представництва). Але ці помилки можна оцінити завчасно і звести до незначних величин.

Під час формування статистичного відбору вирішують такі завдання:

1. Визначають генеральну сукупність.
2. Визначають розмір вибірки.
3. Обирають метод формування вибірки.

Генеральна сукупність (N) **–** уся сукупність об’єктів, яка цікавить дослідника (рис. 1.2).

Вибірка (n) **–** частина генеральної сукупності, випадково відібрана для дослідження з метою отримання висновків про властивості генеральної сукупності.

Параметри – це числові характеристики генеральної сукупності.

Статистика – числові характеристики вибірки. Фактичне значення статистики – це оцінка параметра генеральної сукупності.

Репрезентативна вибірка являє собою генеральну сукупність.

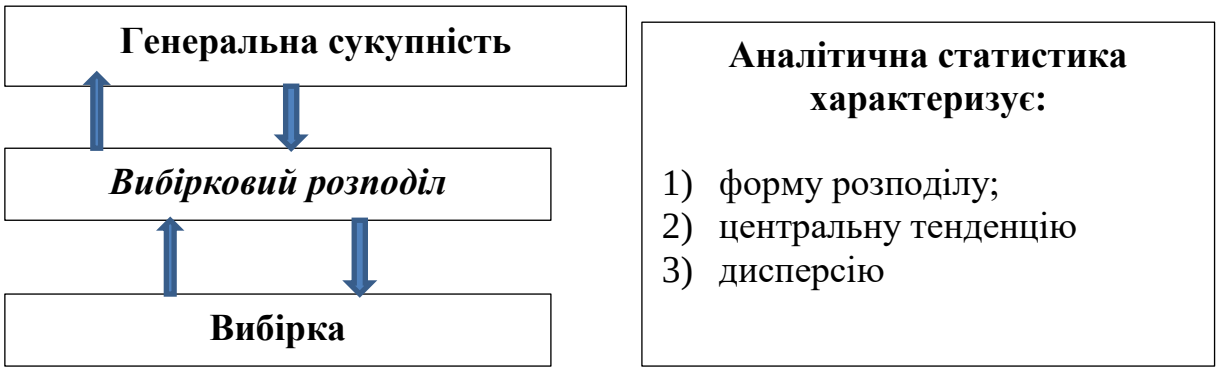


Рис. 1.2. Елементи статистичних даних

Для випадкової вибірки об'єму n з генеральної сукупності справедливі твердження (центральна гранична теорема):

1. Зі зростанням об'єму вибірки n розподіл вибіркового середнього наближається до нормального розподілу.
2. Середнє значення усіх вибіркових середніх є середнє значення генеральної сукупності ($\bar{x}$).
3. Стандартне відхилення середніх дорівнює для:

– вибірки: $S = \frac{\sigma}{\sqrt{N}}$;

– генеральної сукупності: $\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$.

Як змінюються вибіркові характеристики у разі зміни показників генеральної сукупності дає можливість прослідкувати тренажер Central Limit Theorem for Means (центральний ліміт для обґрунтованих середніх): ttps://gallery.shinyapps.io/CLT_mean/

Приклад. Нормальний розподіл генеральної сукупності (рис. 1.3–1.5)

Central Limit Theorem for Means

(центральний ліміт для обґрунтованих середніх)

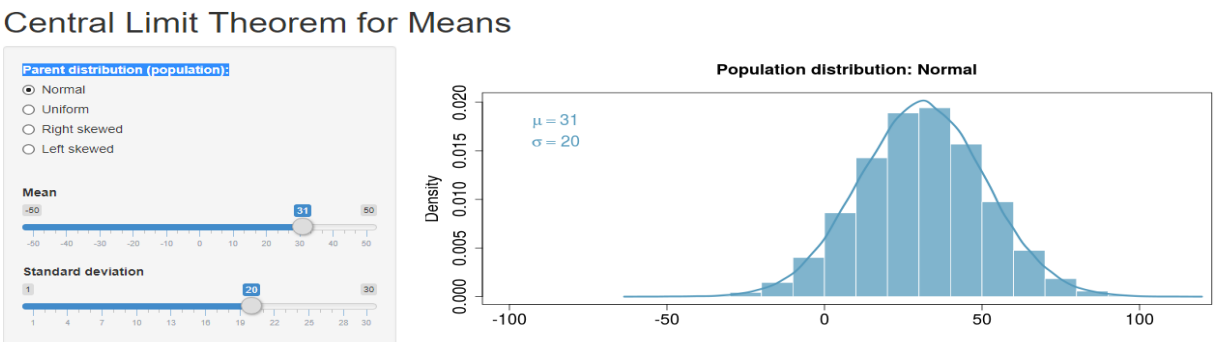


Рис. 1.3. Нормальний розподіл генеральної сукупності, середнє (mean) значення якої 31 (μ), а стандартне відхилення (standard deviation) 20 (σ)

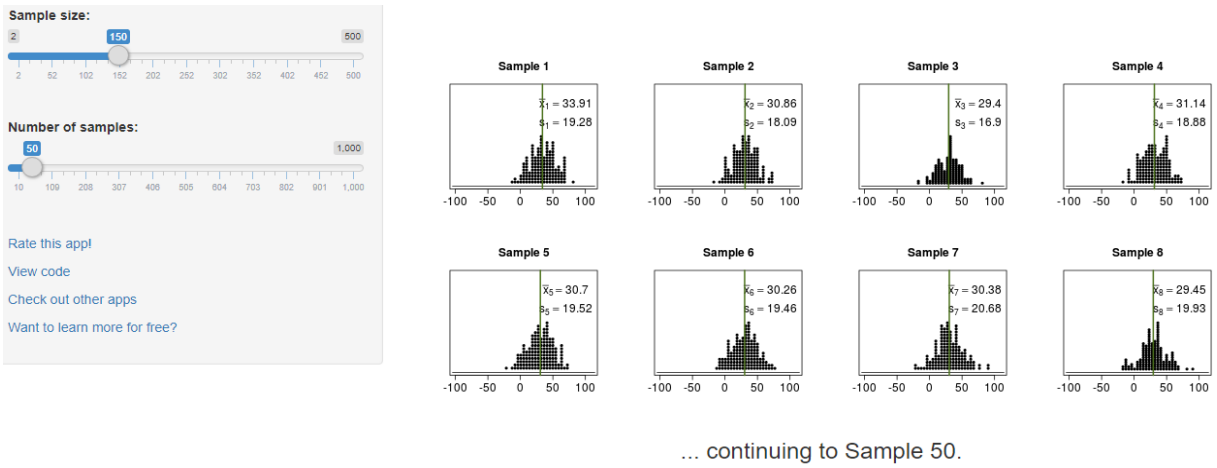


Рис. 1.4. Кількість вибіркових спостережень (number of samples) 50, обсяг вибірки 150 значень (sample size – розмір спостереження)

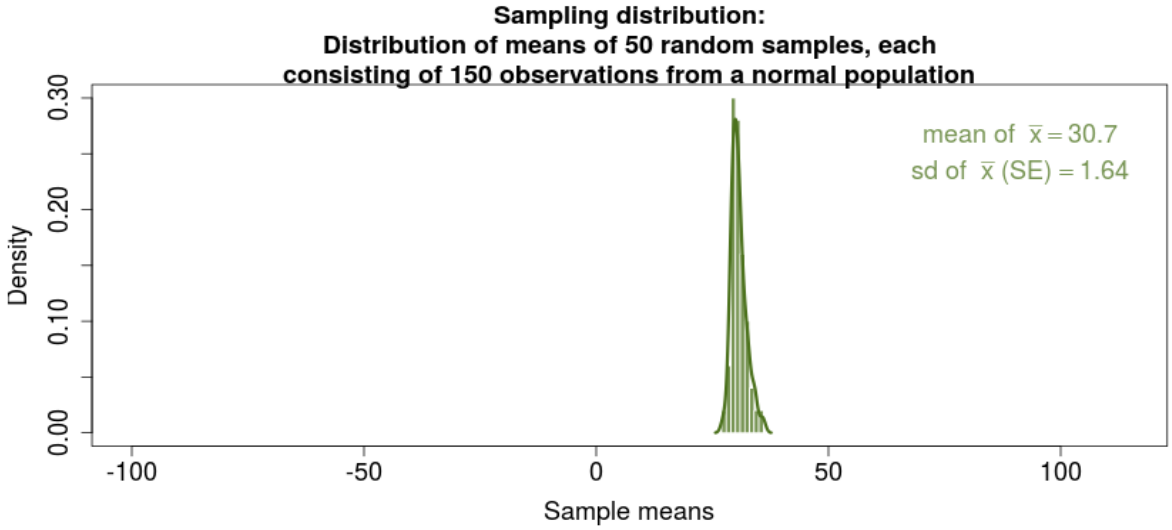


Рис. 1.5. Вибірковий розподіл: розподіл для середніх кожної з 50 випадкових вибірок, що складається зі 150 спостережень від нормального розподілу

Середнє значення вибіркового розподілу – 30,7, стандартна похибка (середньоквадратичне відхилення вибіркових середніх):

$$S = \frac{\sigma}{\sqrt{N}} = \frac{20}{\sqrt{150}} = 1{,}64 .$$

Стандартна похибка показує, наскільки вибіркове середнє відрізняється від середнього генеральної сукупності, це показник репрезентативності.

Зауваження:

1. Розподіл вибіркових середніх прямує до нормального розподілу незалежно від виду розподілів генеральної сукупності.

2. Чим сильніше розподіл генеральної сукупності відрізняється від нормального, тим більший вплив має збільшення об'єму вибірки на точність результату. Вважається, що центральна гранична теорема дає для статистичних висновків прийнятні результати, якщо об'єм вибірки більше 30.

3. Якщо генеральна сукупність має нормальний розподіл, тоді вибіркове середовище буде розподілено нормально для вибірок будь-якого об'єму.

Інші приклади низької, середньої, високої правосторонньої і лівосторонньої асиметрій наведені в додатку А.

1.2. Види формування вибірки

1. Проста випадкова вибірка передбачає, що всі елементи генеральної сукупності мають рівні шанси потрапити до статистичної вибірки.

Приклад. В університеті навчається 2 000 студентів. Необхідно скласти вибірку з 50 осіб.

Використовують можливості *Excel* СЛЧИС().

Переваги: не потребує знання структури генеральної сукупності, простий для розуміння, результати можна розповсюджувати на генеральну сукупність.

Недоліки: складно створити основу вибірки, великі затрати на проведення.

2. Систематична вибірка отримується шляхом нумерації кожного члена генеральної сукупності і потім вибором *k*-ого номеру.

Приклад. Генеральна сукупність включає 2 000 одиниць, потрібно відібрати 50. Оскільки 2 000 / 50 = 40, то обиратися буде кожний 40-й елемент. Для початку випадково обирається перший елемент вибірки серед перших 40 елементів генеральної сукупності. Якщо перший буде номер 15, тоді вибірка включатиме об'єкти з номерами 15, 55, 95 тощо, всього 50 об'єктів.

Недоліком є відсутність випадковості під час вибірки у разі об'єктів, крім першого. Може знизити репрезентативність.

3. Стратифікована вибірка передбачає, що статистична вибірка проводиться випадковим чином окремо в кожній групі генеральної сукупності зі збереженням пропорції співвідношення розмірів цих груп (рис. 1.6).

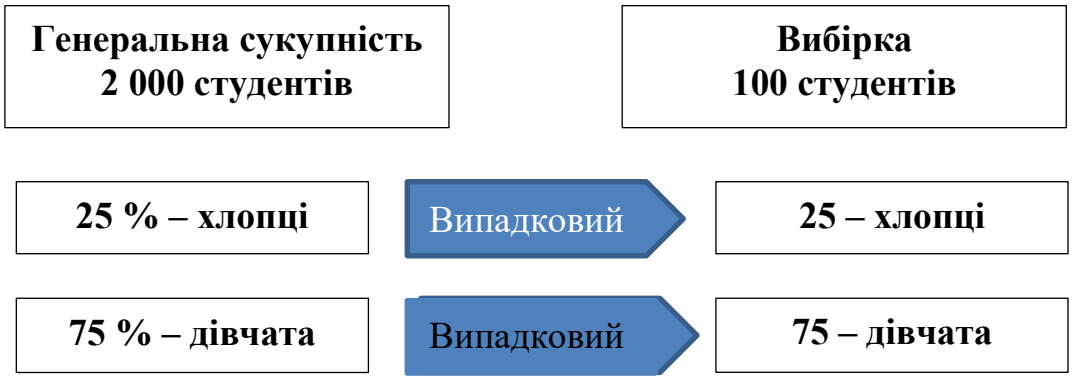


Рис. 1.6. Візуалізація стратифікованої вибірки

Переваги: включає всі важливі підгрупи сукупності, висока точність.

Недолік: передбачає більш складну організацію збору інформації на практиці. Складно вибрати змінні для стратифікації, неможливе стратифікування з урахуванням багатьох змінних, великі затрати на проведення.

Приклад. На двох факультетах навчаються 3 000 студентів, серед яких 30 % фінансистів і 70 % обліковців (рис. 1.1).

Таблиця 1.1

Проміжні розрахунки стратифікованої вибірки

Генеральна сукупність			**Випадковий**	**Вибіркова сукупність**			
Показник	30 % фінансисти	70 % обліковці		Показник	30 % фінансисти	70 % обліковці	
Чоловіки	500	900		Чоловіки	33	60	93
Жінки	400	1200		Жінки	27	80	107
Усього	900	2100		Усього	60	140	
Усього 3 000 студентів				Усього 200 студентів			

Частка студентів, яку відбирають:

3 000 / 200 = 15 разів менше потрібно відібрати студентів.

Усього:

чоловіків: 500 + 900 = 1400, 1400 / 15 = 93;

жінок: 400 + 1200 = 1600, 1600 / 15 = 107.

З 200 студентів повинно бути:

фінансистів: 200 * 30 / 100 = 60;

обліковців: 200 * 70 / 100 = 140.

Співвідношення чоловіків і жінок:

– на фінансовому факультеті 55,6 / 44,4:

500 / 900 * 100 = 55,6 400 / 900 * 100 = 44,4.

60 * 55,6 / 100 = 33,36 = 33;

60 * 44,4 / 100 = 26,64 = 27;

– на обліковому факультеті: 42,9 / 57,1:

900 / 2100 * 100 = 42,9 1200 / 2100 * 100 = 57,1.

140 * 42,9 / 100 = 60;

140 * 57,1 / 100 = 80.

4. Кластерна вибірка передбачає вибірку, в якій сукупність на першому етапі розподіляється на підгрупи, які не перетинаються між собою (кластери), а потім з цих підгруп формується випадкова вибірка (рис. 1.7).

Переваги: більш проста організація збору інформації. Легкий у застосуванні, ефективний з точки зору затрат.

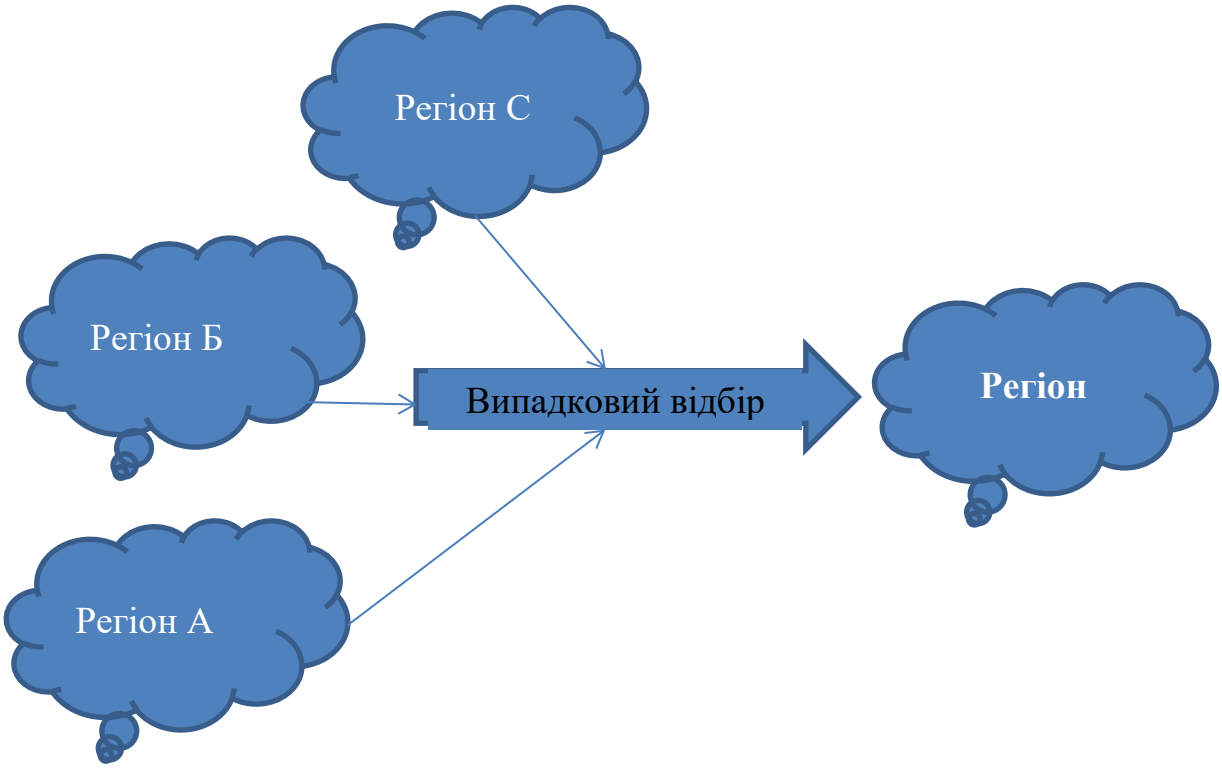


Рис. 1.7. Візуалізація кластерної вибірки

Недоліки: кластери можуть істотно відрізнятися один від одного за структурою. Низька точність, складно розрахувати й оцінити результати.

Наприклад: досліднику необхідно опитати мешканців, що живуть у квартирах провінційного міста. Якщо у межі 200 житлових будинків, дослідник може вибрати будь-які 20 і опитати їх мешканців. При цьому не враховується рівень доходу мешканців, їх вік, соціальний статус.

5. **Емпірична вибірка** (проста, квотована) – передбачає, що в коло респондентів для відбору інформації включається «перший зустрічний».

Недоліки: характеризується низьким ступенем репрезентативності.

6. **Експертний відбір**

Переваги: низька вартість, зручно, невелика тривалість.

Недоліки: не дозволяє розповсюджувати результати на генеральну сукупність, суб’єктивна думка.

7. Відбір методом «сніжного кома» – опитування проводиться шляхом передавання анкети в межах групи (наркомани, хворі на туберкульоз тощо).

Переваги: дозволяє оцінити окремі випадки для генеральної сукупності.

Недоліки: велика тривалість анкетування.

1.3. Визначення розміру вибірки

У процесі статистичного аналізу потрібно задати дві величини: **точність оцінювання** (ступінь допустимого відхилення величини, що оцінюється від істинної) і **ймовірність**, з якою гарантується результат.

Під репрезентативною вибіркою мають на увазі таку вибірку, за якою можна побудувати оцінки параметрів таким чином, щоб відхилення оцінки від істинного значення генеральної сукупності не перевищувало задану **похибку *E*** і результат гарантувався із заданою ймовірністю.

Скільки потрібно опитати осіб, щоб оцінка була отримана із заданою точністю?

Похибка оцінки = Параметр – Оцінка.

Гранична похибка – це максимально можлива похибка для взятої ймовірності *F(x)*. Довірче число *t* показує, як співвідносяться гранична та стандартна похибки. Визначається:

$$\Delta = t\mu = tS\,,$$

де μ – середня помилка, *t* – коефіцієнт довіри, коефіцієнт кратності похибки, квантиль розподілу ймовірностей. Його значення дорівнює *t*-статистиці Стьюдента (додаток А) (табл. 1.2).

1) у разі повторного випадкового відбору:

– для середньої: $\Delta_{\bar{x}} = t\sqrt{\frac{\sigma^2}{n}}$;

– для частки: $\Delta_p = t\sqrt{\frac{p \cdot q}{n}} = t\sqrt{\frac{p(1-p)}{n}}$**;**

Таблиця 1.2

Значення *t*-статистики Стьюдента для обсягу вибірки більше 500

Рівень надійності	Константа *t*
90 %	1,64
95 %	1,96
99 %	2,58

2) у разі безповторного (випадкового та механічного) відбору:

– для середньої: $\Delta_{\bar{x}} = t\sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}$;

– для частки: $\Delta_p = t\sqrt{\frac{p\cdot q}{n}\left(1-\frac{n}{N}\right)}$.

Довірчий інтервал – це розрахований на основі вибірки інтервал значень ознак, який з відомою ймовірністю містить оціночний параметр генеральної сукупності. Збільшення довірчого інтервалу зменшує точність результатів.

На практиці довіряють оцінкам, значення похибки яких не перевищує 5 % (1 %), а рівень надійності становить 95 % або 99 %.

– межі генеральної середньої $\bar{x} - \Delta_x \le \bar{x} \le \bar{x} + \Delta_x$;

– межі генеральної частки $\bar{p} - \Delta_p \le \bar{p} \le \bar{p} + \Delta_p$.

Приклад. Який у генеральній сукупності середній вік студентів першого курсу УДФСУ? Із попередніх досліджень відомо, що стандартне відхилення генеральної сукупності (σ) дорівнює 2 роки. Сформована вибірка у кількості із 60 студентів, розраховане вибіркове середнє ($\bar{x}$) 17,5 роки. Необхідно побудувати 95 % довірчий інтервал для генеральної середньої ($\bar{x}$). На перший курс вступило 550 студентів.

t – статистика Стьюдента для рівня надійності 95 % ($\alpha = 0{,}05, v = 550$) = 1,96.

$$\Delta = t\sqrt{\frac{\sigma^2}{n}} = 1{,}96\sqrt{\frac{2^2}{60}} = 0{,}51$$

$$\bar{x} - \Delta < \bar{x} < \bar{x} + \Delta$$

$$17{,}5 - 0{,}51 < \bar{x} < 17{,}5 + 0{,}51$$

$$16{,}99 < \bar{x} < 18{,}01 .$$

Відповідь: середній вік студентів першого курсу УДФСУ коливається від 16,99 до 18,01 років.

Приклад. У 25 магазинах торговельної мережі в середньому щоденно продавалося 196 одиниць продукції. Стандартне відхилення вибірки (S) дорівнює 7. Знайти 95 % довірчий інтервал для генеральної середньої.

t – статистика Стьюдента для ймовірності 95 % та 25 – 1 = 24 ступенів свободи дорівнює 1,711.

$$S = \sqrt{\frac{\sigma^2}{n}} \quad \Delta = t\sqrt{\frac{\sigma^2}{n}} = t \cdot S = 1{,}711 \cdot 7 = 11{,}977$$

$$\bar{x} - \Delta < \bar{x} < \bar{x} + \Delta$$

$$196 - 11{,}977 < \bar{x} < 196 + 11{,}977$$

$$184{,}023 < \bar{x} < 207{,}977 .$$

Відповідь: середнє значення генеральної сукупності коливатиметься від 184,023 до 207,977 одиниць продукції, тобто в середньому в усіх магазинах торговельної мережі буде продано 184–208 одиниць продукції (в кожному).

Приклад. Результати опитування 900 мешканців міста Ірпінь (50 тис. населення) показали, що 460 опитаних (51,11 %) підтримають на виборах кандидатуру чинного мера міста. Можна стверджувати, що мер міста буде обраний повторно на наступний період?

t-статистика Стьюдента для рівня надійності 95 % ($\alpha = 0{,}05, v = 550$) = 1,96.

$$\Delta = t\sqrt{\frac{p \cdot q}{n}} = 1{,}96\sqrt{\frac{0{,}5111 \cdot (1 - 0{,}5111)}{900}} = 0{,}0327$$

межі генеральної частки $p - \Delta < p < p + \Delta$

$$0{,}5111 - 0{,}0327 < p < 0{,}5111 + 0{,}0327$$

$$47{,}84\,\% < p < 54{,}38\,\% .$$

Відповідь: не можна за результатами опитування стверджувати, що більше половини виборців голосуватимуть за чинного мера на наступний період.

Приклад. З 400 студентів ННІОАА опитано 144 студента (36 % вибірка) щодо середньої успішності зі статистики. Середній бал опитаних становить 4,0, дисперсія вибірки 24. Знайти довірчі інтервали успішності всіх студентів ННІОАА, ймовірність 0,95.

t – статистика Стьюдента для рівня надійності 95 % ($\alpha = 0{,}05, v = 144$) = 1,98.

$$\Delta_{\bar{x}} = t\sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)} = 1{,}98\sqrt{\frac{24}{144}\left(1-\frac{144}{400}\right)} = 0{,}65$$

$$\bar{x} - \Delta < \bar{x} < \bar{x} + \Delta$$

$$4 - 0{,}65 < \bar{x} < 4 + 0{,}65$$

$$3{,}35 < \bar{x} < 4{,}65.$$

Серед опитаних 29 студентів склали екзамен не з першого разу. Знайти довірчі інтервали успішної здачі дисципліни «Статистика» серед усіх студентів ННІОАА, ймовірність 0,954:

$$p = \frac{m}{n} = \frac{29}{144} = 0{,}2;$$

$$\Delta_p = t\sqrt{\frac{p\cdot q}{n}\left(1-\frac{n}{N}\right)} = 1{,}98\sqrt{\frac{0{,}2\cdot(1-0{,}2)}{144}\left(1-\frac{144}{400}\right)} = 1{,}98\cdot\sqrt{0{,}001\cdot 0{,}64} = 0{,}05$$

$$\bar{p} - \Delta < \bar{p} < \bar{p} + \Delta$$

$$0{,}2 - 0{,}05 < \bar{p} < 0{,}2 + 0{,}05$$

$$0{,}15 < \bar{p} < 0{,}25$$

$$0{,}15\cdot 144 < \bar{p} < 0{,}25\cdot 144$$

$$21{,}6 < \bar{p} < 36.$$

Відповідь: з першого разу статистику можуть не скласти від 22 до 36 студентів. Середня успішність за цим предметом становить 3,35–4,65 бала.

1.4. Мінімальний обсяг вибірки

Мінімально достатній обсяг вибірки **–** це обсяг вибірки, за яким вибіркові оцінки репрезентували б основні властивості генеральної сукупності.

Обсяг вибірки визначається:

– для повторної вибірки: $n = \frac{t^2 \cdot \sigma^2}{\Delta^2}$;

– для частки: $n = \frac{t^2 \cdot pq}{\Delta^2}$;

– для безповторної вибірки: $n = \frac{t^2 \cdot \sigma^2 \cdot N}{\Delta^2 \cdot N + t^2 \cdot \sigma^2}$.

Приклад. Який розмір вибірки необхідний для статистичного дослідження? Якщо оцінка повинна бути зроблена з точністю до одного року і ймовірністю 99 %. Результати попереднього дослідження показали стандартне відхилення віку – 2 роки.

$$n = \frac{t^2 \cdot \sigma^2}{\Delta^2} = \frac{2{,}58^2 \cdot 2^2}{1^2} = 26{,}63 \approx 27.$$

Відповідь: опитати потрібно 27 респондентів.

Якщо невідомо стандартне відхилення, використовують формулу для оцінки частки.

Приклад. Знайти мінімальну вибірку, яка забезпечить 3 % похибки у разі довірчого інтервалу 95 %.

$$n = \frac{t^2 \cdot pq}{\Delta^2} = \frac{1{,}96^2 0{,}5 \cdot 0{,}5}{0{,}03^2} = 1067.$$

Відповідь: опитати потрібно 1 067 респондентів.

Приклад. Знайти статистичну похибку вибірки з 500 респондентів при рівні значущості до результатів опитування 95 %.

Максимальна статистична похибка вибірки розраховується за формулою:

$$\Delta = t \cdot \sqrt{\frac{pq}{n}}.$$

Звідси для вибірки в 500 одиниць статистична похибка вибірки дорівнюватиме:

$$\Delta = 1{,}96 \cdot \sqrt{\frac{50 \cdot 50}{500}} = 1{,}96 \cdot \sqrt{5} \approx 4{,}38\,\%.$$

Відповідь: якщо в подальшому з'ясуємо, що 40 % респондентів купують товари в інтернет-мережі, то це означатиме, що з ймовірністю 95 % у генеральній сукупності продаж в Інтернеті коливається в межах від 35,62 % (40 % – 4,38 %) до 44,38 % (40 % + 4,38 %).

Об'єм вибірки приблизно може бути розрахований за формулою:

$$n \approx \frac{1}{\Delta^2} = \frac{1}{0{,}05^2} = 400, \text{ а } \Delta_{\bar{x}} = t\sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)}.$$

Звідси:

Δ = 5 %, n = 400 *анкет*

Δ = 2,5 %, n = 1 600 *анкет*

Δ = 1 %, n = 100 *анкет*.

Якщо невідоме стандартне відхилення ні вибірки, ні генеральної сукупності, доцільно застосовувати її максимальне значення, що дасть додаткову надійність результату, а саме 0,5.

Приклад. Знайти обсяг вибірки, яка забезпечить похибку 5 % у разі довірчого інтервалу 95 %, якщо обсяг генеральної сукупності складає 2 000 осіб. Генеральна сукупність більше 500 осіб, тому для рівня надійності 95 % використовуємо значення 1,96.

$$n = \frac{t^2\sigma^2 N}{\Delta_{\bar{x}}^2 N + t^2\sigma^2} = \frac{1{,}96^2 \cdot 0{,}5^2 \cdot 2000}{0{,}05^2 \cdot 2000 + 1{,}96^2 \cdot 0{,}5^2} = 322 \text{ особи}.$$

Відповідь: потрібно опитати 322 респондента.

Зауваження:

1. Об'єм вибірки не пов'язаний з її репрезентативністю.
2. Загрожує репрезентативності вибірки не її об'єм, а зміщення, тобто відхилення від принципу випадковості.

Існують он-лайн калькулятори для визначення обсягів вибірки:

URL: http://socio-lab.vntu.edu.ua/download/Calculator.html (рис. 1.8, 1.9).

Приклад. Яку кількість мешканців достатньо було опитати щодо результатів вступної кампанії, якщо у місці проживає 50 000 осіб. Рівень надійності 95 %, а похибка 2 або 3 %.

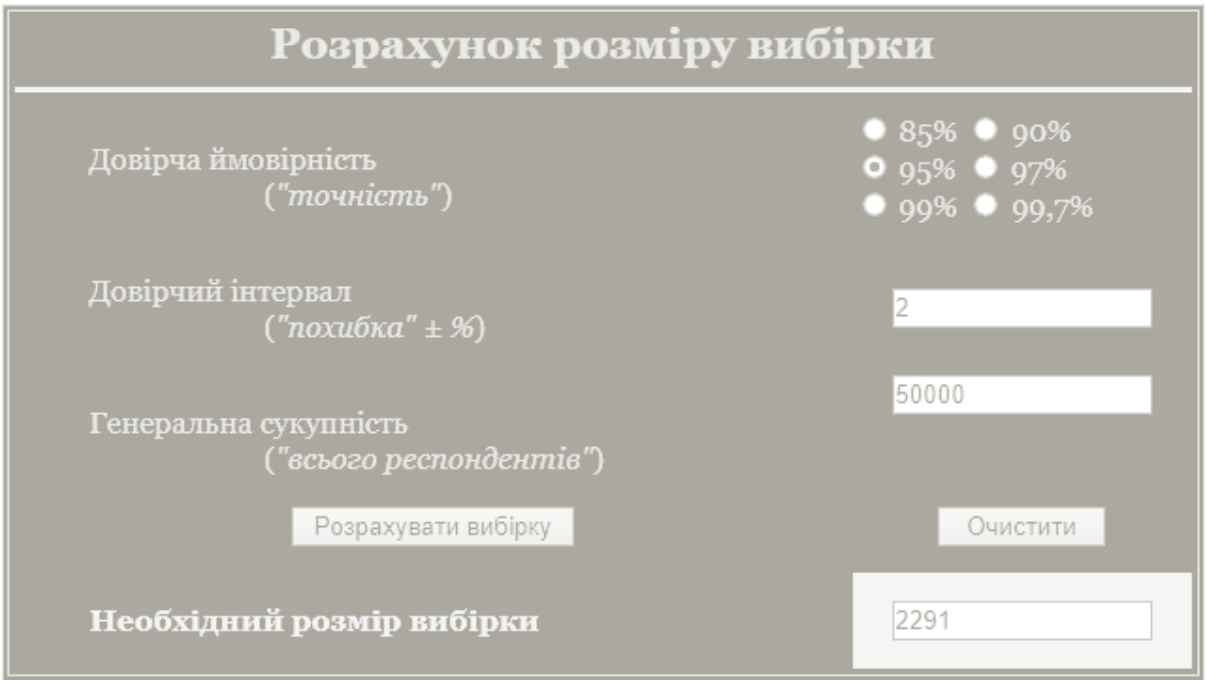


Рис. 1.8. Вибірка за допомогою он-лайн калькулятора у разі похибки 2 %

Відповідь: для похибки в 2 % достатньо опитати 2 291 особу.

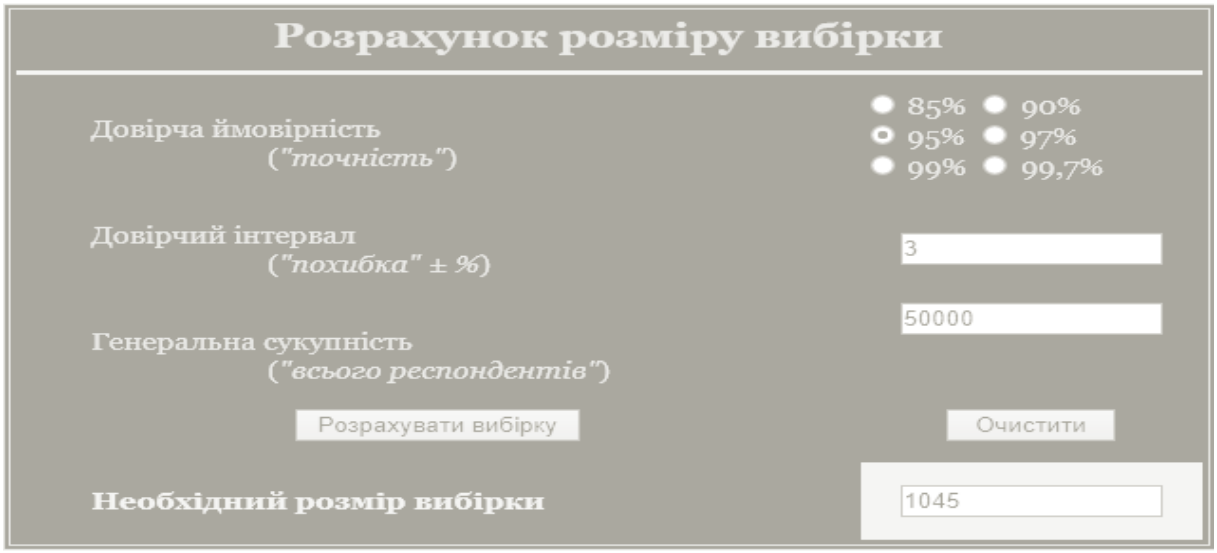


Рис. 1.9. Вибірка за допомогою он-лайн калькулятора у разі похибки 3 %

Відповідь: для похибки в 3 % достатньо опитати 1 045 осіб. Отже, зі збільшенням розміру похибки обсяг вибірки зменшується.

Зазвичай під час проведення будь-яких економіко-математичних досліджень використовують рівень надійності 95 %. Розрахуємо обсяг вибірки для цього рівня із урахуванням зміни обсягів генеральної сукупності і похибки (табл. 1.3).

Таблиця 1.3

Обсяг випадкової вибірки для рівня надійності 95 %

Обсяг генеральної сукупності		**Допустима межа похибка, %**		
	± 1 %	**± 3 %**	**± 5 %**	**± 10 %**
30	30	29	28	23
100	99	92	80	49
500	475	341	217	81
1 000	906	516	278	88
3 000	2 286	787	341	93
5 000	3 288	880	357	94
10 000	4 899	964	370	95
50 000	8057	1 045	381	96
100 000 та більше	8 763	1 056	383	96

У практичних дослідженнях (з метою мінімізації витрат) орієнтуються на те, що достатньо дослідити не більше 1 % цільової аудиторії.

Перелік питань для самоконтролю

1. Поясніть переваги вибіркового методу.
2. Назвіть завдання, які вирішує дослідник під час застосування статистичного відбору.
3. Поясніть відмінність між генеральною сукупністю і вибіркою.
4. Назвіть основні види формування вибірки.
5. У чому різниця між простою і систематичною вибірками.
6. Поясніть сутність, переваги і недоліки застосування стратифікованої вибірки. Наведіть приклади застосування.
7. Поясніть сутність, переваги і недоліки застосування кластерного відбору. Наведіть приклади застосування.
8. У чому відмінність між емпіричною вибіркою і вибіркою за методом «сніжного кому».

9. Назвіть переваги і недоліки застосування експертного відбору. Наведіть приклади застосування.

10. Для чого застосовують довірчий інтервал? Якщо дослідник його збільшує, що відбувається з результатом?

11. Назвіть алгоритм розрахунку похибки і приклади її застосування на практиці.

12. Назвіть алгоритм розрахунку мінімального обсягу вибірки.

13. Назвіть рівні надійності і поясніть їхню практичну значущість.

14. У разі відсутності стандартного відхилення вибірки і генеральної сукупності, яку величину використовувати для розрахунку мінімального обсягу вибірки.

Тести

1. Назвіть завдання, які вирішуються під час формування статистичного відбору:

а) визначення розміру вибірки:

б) усі відповіді правильні:

в) визначення генеральної сукупності:

г) вибір методу формування вибірки.

2. Частина генеральної сукупності, відібрана для дослідження з метою отримання висновків про властивості генеральної сукупності, – це:

а) змінна;

б) спостереження;

в) вибірка;

г) параметр.

3. Числові характеристики генеральної сукупності – це:

а) ознаки;

б) параметри;

в) частота;

г) вибірка.

4. Назвіть вибірку, в якій нумерується кожний член генеральної сукупності і відбирається *i*-ий номер:

а) систематична;

б) проста;

в) кластерна;

г) стратифікована.

5. Назвіть вибірку, в якій сукупність на першому етапі розподіляється на підгрупи, що не перетинаються між собою, а потім з цих груп формується випадкова вибірка:

а) емпірична;

б) кластерна;

в) стратифікована;

г) систематична.

6. Назвіть дві величини, що задаються в процесі статистичного аналізу:

а) точність оцінювання, цілісність інформації;

б) ймовірність, з якою характеризується результат, точність оцінювання;

в) ймовірність, повнота інформації;

г) точність та повнота інформації.

7. Назвіть показник, який показує, на скільки вибіркове середнє відрізняється від середнього генеральної сукупності:

а) дисперсія;

б) кореляція;

в) стандартна похибка;

г) регресія.

8. Назвіть метод, який передбачає опитування шляхом передавання анкет у межах групи:

а) метод «сніжного кому»;

б) метод «мозкового штурму»;

в) анкетування;

г) метод експертних оцінок.

9. Назвіть вид вибірки, який передбачає включення в коло респондента першого зустрічного:

а) стратифікована;

б) проста;

в) емпірична;

г) кластерна.

10. Показником репрезентативності є:

а) довірчий інтервал;

б) стандартна похибка;

в) кореляція;

г) детермінація.

11. Які оцінки значень похибки є допустимими:

а) 30 %;

б) більше 50 %;

в) менше 5 %;

г) 75 %.

12. Твердженням якої теореми виступає таке: «Середнє значення усіх вибіркових середніх є середнім значенням генеральної сукупності»:

а) центральна гранична теорема;

б) основна теорема сукупності;

в) центральна теорема генеральної сукупності;

г) основна генеральна теорема.

13. Назвіть помилки, які виникають під час вибіркового спостереження через несуцільність реєстрації даних і порушення принципів випадковості відбору:

а) репрезентативності;

б) випадкові;

в) систематичні;

г) реєстрації.

РОЗДІЛ 2
ОСНОВИ СТАТИСТИЧНОГО АНАЛІЗУ ДАНИХ

2.1. Візуальне представлення розподілів

Візуальне представлення результатів дає можливість відповісти на такі запитання:

1. Які значення типові для цього набору даних?
2. Як вирізняються між собою?
3. Чи сконцентровані дані біля якогось типового значення?
4. Який характер має ця концентрація даних?
5. Чи однаковий характер затухання для малих і великих значень?
6. Чи є в цьому наборі таке значення, яке значно відрізняється від інших, що потребує спеціальної обробки?
7. Чи можна зазначити, що це в цілому однорідний набір чи виразно виділяється наявність груп?

Різні значення властивості (випадкової величини X) називають **варіантами** (позначають їх через x).

Перший крок до осмислення статистичного матеріалу – це його впорядкування, розміщення варіанта у порядку зростання (спадання), тобто **ранжирування.**

Але для вивчення значної кількості даних цього замало. Тому розбивають варіанти на окремі інтервали, тобто проводять їхнє групування.

Кількість інтервалів m варто брати не досить великим, щоб після групування ряд не був громіздким і не зовсім малим, щоб не втратити особливостей розподілу цієї властивості.

Відповідно до **формули Стерджеса** кількість інтервалів $m = 1 + 3{,}322 \lg n$, а **величина інтервалу (інтегральна різниця,** ширина інтервалу):

$$k = \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \lg n},$$

де $x_{max} - x_{min}$ – різниця між найбільшим і найменшим значенням властивості.

Числа, які показують, скільки разів зустрічаються варіанти з цього інтервалу, називаються **частотами** (позначаються n_i), а відношення їх до загальної кількості спостережень – **частостями** або **відносними частотами,** тобто $W_i = \frac{n_i}{n}$.

Частоти і частості називають **вагами.**

Сума відносних частот дорівнює одиниці.

$$W_1 + W_2 + ... + W_k = \frac{n_1}{n} + \frac{n_2}{n} + ... + \frac{n_k}{n} = \frac{n_1 + n_2 + ... + n_k}{n} = 1.$$

Сукупність значень властивості, записаних у порядку зростання називають **варіаційним рядом.**

Статистичним розподілом вибірки називається перелік варіантів, що спостерігаються, розміщених у порядку зростання та відповідних їм частот; відносних частот.

Під час вивчення варіаційних рядів поряд з поняттям частоти використовується поняття накопиченої частоти ($n_i^{нак}$), яка показує, скільки спостерігалось варіантів значенням властивості, менших *x*.

Відношення накопиченої частоти $n_i^{нак}$ до загальної кількості спостережень *n* називається **накопиченою частістю** $W_i^{нак}$.

Накопичені частоти (частості) перебувають для кожного інтервалу послідовно додаванням частот (частостей) усіх попередніх інтервалів, включаючи цей.

Приклад. Результати 20 спостережень над кількісною властивістю генеральної сукупності записані в таблиці 2.1.

Таблиця 2.1

Вихідні дані

x_i	2	6	10	12	14
n_i	1	5	7	3	4

n = 1+5+7+3+4=20 – об'єм вибірки, відносні частоти:

$W_1 = \frac{1}{20}$; $W_2 = \frac{5}{20}$; $W_3 = \frac{7}{20}$; $W_4 = \frac{3}{20}$; $W_5 = \frac{4}{20}$

(табл. 2.2).

Таблиця 2.2

Проміжні розрахунки

x_i	2	6	10	12	14
n_i	$\frac{1}{20}$	$\frac{5}{20}$	$\frac{7}{20}$	$\frac{3}{20}$	$\frac{4}{20}$

Для задання варіаційного ряду достатньо вказати варіанти і відповідні їм частоти (частості) або накопичені частоти (частості).

Варіаційний ряд називається дискретним, якщо будь-які його варіанти відрізняються на сталу величину, і неперервним (інтервальним), якщо варіанти можуть відрізнятись одна від одної на як завгодно малу величину.

Нехай невідомо статистичний розподіл кількісної властивості *X*. Введемо позначення: n_i – кількість спостережень, під час яких спостерігалися значення властивості *X* менше *x*; *n* – загальна кількість спостережень.

Відносна частота n_i/n є функція від *x,* тобто у разі зміни *x* змінюється і функція. Оскільки ця функція отримана дослідним шляхом, то її називають емпіричною.

Емпіричною функцією розподілу (функцією розподілу вибірки) називають функцію *F**(x), яка визначає для кожного значення x відносну частоту того, що властивість (випадкова величина X) прийме значення, менше заданого *x*. $F^*(x)=\frac{n_i^{нак}}{n}$; $F^*(x)=W(X<x)=W_i^{нак}$.

На відміну від емпіричної функції розподілу *F*(x),* інтегральну функцію розподілу *F(x)* називають теоретичною функцією розподілу.

Відмінність між емпіричною та теоретичною функціями розподілу закладається в тому, що теоретична функція *F(x)* визначає для значення *x* ймовірність події $X < x$, а емпірична функція визначає відносну частоту цієї ж події, яка дорівнює *F*(x).*

Властивості емпіричної функції:

1) значення емпіричної функції належать відрізку [0; 1];

2) *F*(x)* – неспадна функція;

3) якщо x_1 – найменше значення властивості, яка спостерігається, а x_k – найбільше, то $F^*(x_1) = 0$ при $X \le x_1$ і $F^*(x_k)=1$ при $x > x_k$.

Для графічного зображення варіаційних рядів найчастіше використовують полігон, гістограму, кумулятивну криву.

Значення варіант та частот або відносних частот можна розглядати як координати точок.

$$M_1(x_1,\ n_1),\quad M_2(x_2, n_2),\quad \ldots,\quad M_m(x_m, n_m)$$

або

$$M_1(x_1, W_1),\ M_2(x_2, W_2),\ \ldots,\ M_m(x_m, W_m).$$

Полігоном частот називають ламану, відрізки якої з'єднують точки$(x_1, n_1),\ (x_2,\ n_2),\ \ldots,\ (x_m,\ n_m)$.

Полігоном відносних частот називають ламану, відрізки якої з'єднують точки$(x_1,\ W_1),\ (x_2,\ W_2),\ \ldots,\ (x_m, W_m)$.

Полігон частот та частостей є аналогом щільності ймовірностей. Для побудови полігону частот на осі абсцис відкладають варіанти x_k ознаки *X*, а на осі ординат – відповідні їм частоти. Утворені точки з'єднують відрізками прямих і одержують полігон частот. Полігон, як правило, слугує для зображення дискретного варіаційного ряду.

Приклад. У результаті вибірки одержали такі ознаки: –3; 2; –1; –3; 5; –3 (табл. 2.3). Побудувати полігон частот цієї вибірки (рис. 2.1).

Таблиця 2.3

Вихідні дані

X	–3	–1	2	5
n_i	3	1	2	1

Важливість емпіричної функції розподілу вибірки в тому, що вона слугує для оцінки теоретичної функції розподілу генеральної сукупності.

Приклад. Побудувати емпіричну функцію за даними вибірки (табл. 2.4).

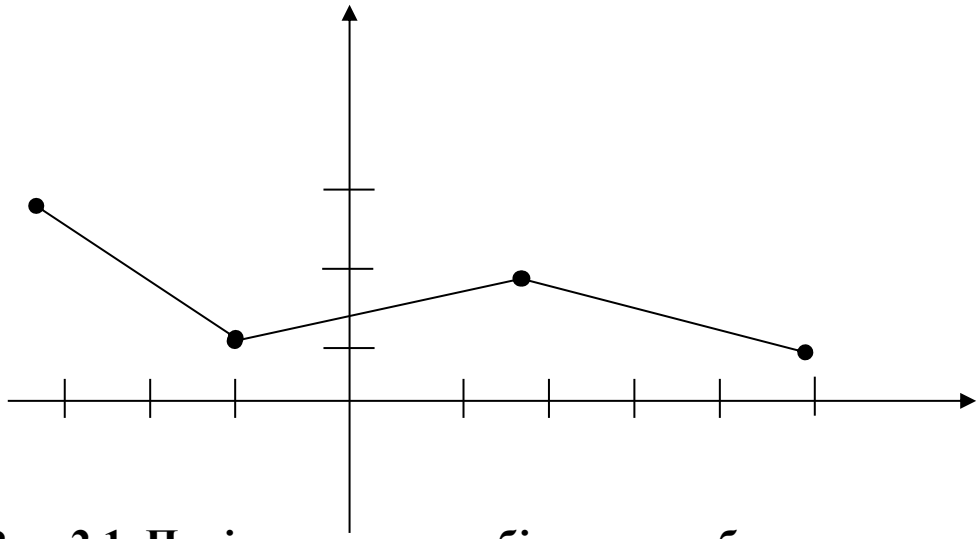


Рис. 2.1. Полігон частот вибіркового обстеження

Таблиця 2.4

Вихідні дані

Значення властивості, яка спостерігається	2	6	10
Частоти значень, які спостерігаються	12	18	30

Об'єм вибірки 12 + 18 + 30 = 60 = *n*.

Найменше зі значень 2, тому *F*(2)* = 0.

x < 6, а саме *x* = 2 спостерігалось 12 раз, тому $F^*(6)=\frac{12}{60}=0{,}2$.

Значення *x* = 10 спостерігалось 12 + 18 = 30 разів, тому $F^*(10)=\frac{30}{60}=0{,}5$.

Оскільки *x* > 10 найбільше зі значень, то $F^*(x>10)=1$ (рис. 2.2).

$$F^*(x)=\begin{cases}0, & при\ x\le 2;\\ 0{,}2, & при\ 2<x\le 6;\\ 0{,}5, & при\ 6<x\le 10;\\ 1, & при\ x>10.\end{cases}$$

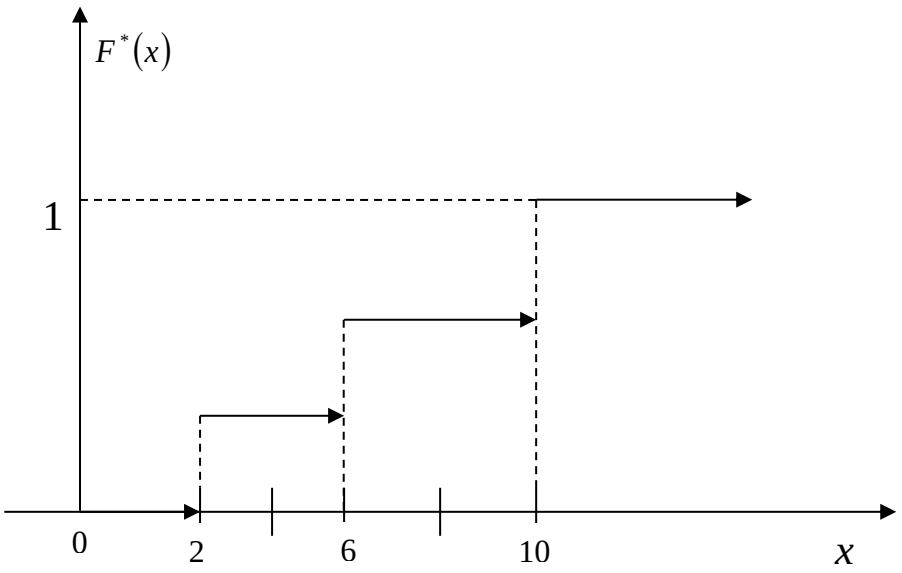


Рис. 2.2. Емпірична функція вибіркового обстеження

Гістограмою частот називають ступінчату фігуру, яка складається з прямокутників, основами яких є часткові інтервали варіант довжиною $h = x_k - x_{k-1}$, а висоти дорівнюють (n_k / h) частотам або частостям $n_r(W_i)$ інтервалів. **Гістограма** – це стовпчикова діаграма частот, а не даних!

Якщо з'єднати середини верхніх основ прямокутників відрізками прямої, то можна отримати полігон того ж розподілу.

Площа прямокутників гістограми частот дорівнює об'єму вибірки, а площа гістограми частостей – одиниці. Гістограма слугує тільки для зображення інтервальних варіаційних рядів.

Для побудови гістограми частот проміжок варіант [x_{min}; x_{max}], тобто від найменшого значення x_{min} до x_{max}, ділимо на декілька відрізків однакової довжини h. Потім відраховуємо суму частот тих значень варіант ознаки X, які належать кожному з одержаних відрізків.

Якщо в k-му відрізку кількість варіантів, що спостерігали з урахуванням їх частот, дорівнює n_k, то будують прямокутник $П_к$, основою якого буде k-ий відрізок довжиною h, а висотою – n_k/h.

Приклад. Маючи розподіл ознаки X вибірки (табл. 2.5), побудувати гістограму частот цього розподілу (рис. 2.3).

Таблиця 2.5

Вихідні дані

x_k	–2	0	1	2	3	5	7
n_k	4	5	7	8	6	2	1

$x_{\min} = -2$; $x_{\max} = 7$ $\left|\left[x_{\min}; x_{\max}\right]\right| = 9$ одиниць; $h = \frac{9}{4} = 2{,}25$

(табл. 2.6).

Таблиця 2.6

Поміжні розрахунки

Відрізки довжини $h = 2{,}25$	[–2; 0,25]	[0,25; 2,5]	[2,5; 4,75]	[4,75; 7]
Кількість значень n_k	9	15	6	9
Щільність частоти n_k/h	4	$\frac{20}{3}$	$\frac{8}{3}$	$\frac{4}{3}$

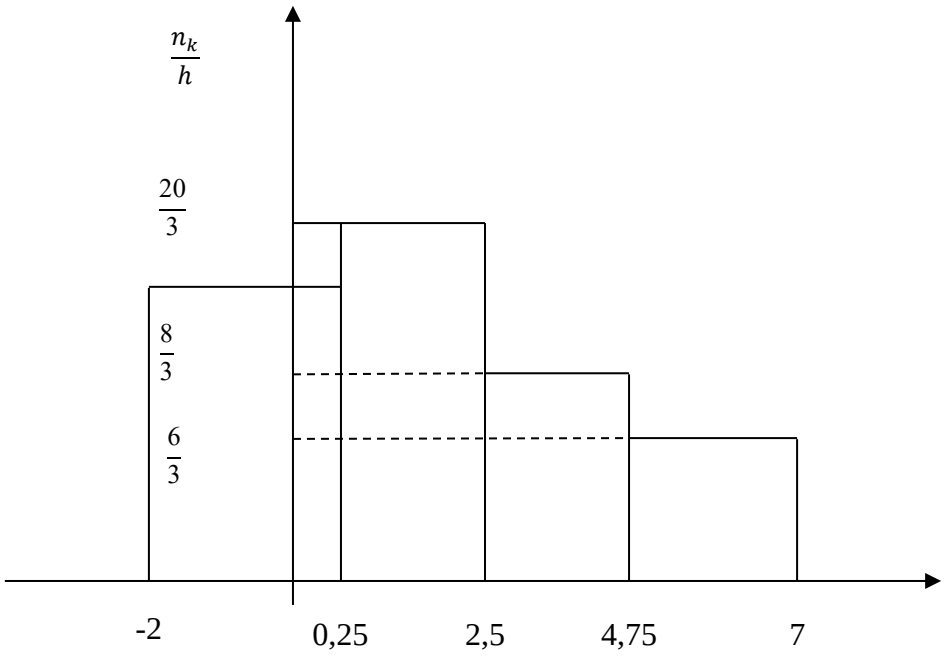


Рис. 2.3. Гістограма вибіркового обстеження

Приклад. Необхідно вивчити зміну продуктивності на одного робітника механічного цеху в звітному році порівняно з попереднім. Отримано такі дані про розподіл 100 робочих цеха за продуктивністю в звітному році (у процентах до попереднього).

97,8; 97,0; 101,7; 132,5;... ;142,0;104,2; 141; 122,1.

Розв'язання. Ранжируємо варіанти:

x_{min} = 97,0; 97,2;...; x_{max} = 142,0; n=100.

$$k = \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \lg n} = \frac{142 - 97}{1 + 3{,}322 \cdot 2} = \frac{45}{7{,}644} = 5{,}89\ \%;\ \ k = 6\ \%.$$

За початок першого рекомендують брати величину $x_{поч} = x_{min} - \frac{k}{2} = 97 - \frac{6}{2} = 94$ (%).

Згрупований ряд представимо у вигляді таблиці 2.7.

Таблиця 2.7

Проміжні розрахунки продуктивності праці одного робітника механічного цеху

№ з/п	Продуктивність у звітному році в процентах до попереднього x	Частота (кількість робітників) n_i	Частість (доля робітників) $W_i = \frac{n_i}{n}$	Накопичена частота $n_i^{нак}$	Накопичена частість $W_i^{нак} = \frac{n_i^{нак}}{n}$
1.	94–100	3	0,03	3	0,3
2.	100–106	7	0,07	10	0,10
3.	106–112	11	0,11	21	0,21
4.	112–118	20	0,2	41	0,41
5.	118–124	28	0,28	69	0,69
6.	124–130	19	0,19	88	0,88
7.	130–136	10	0,1	98	0,98
8.	136–142	2	0,02	100	1,00
	$\sum$	100	1		

Варіаційний ряд є статистичним аналогом (реалізацією) розподілу властивості (випадкової величини X). У цьому сенсі полігон (гістограма) аналогічний кривій розподілу, а емпірична функція розподілу – функції розподілу випадкової величини (рис. 2.4).

Кумулятивною кривою називається крива накопичених частот (частостей) (рис. 2.5).

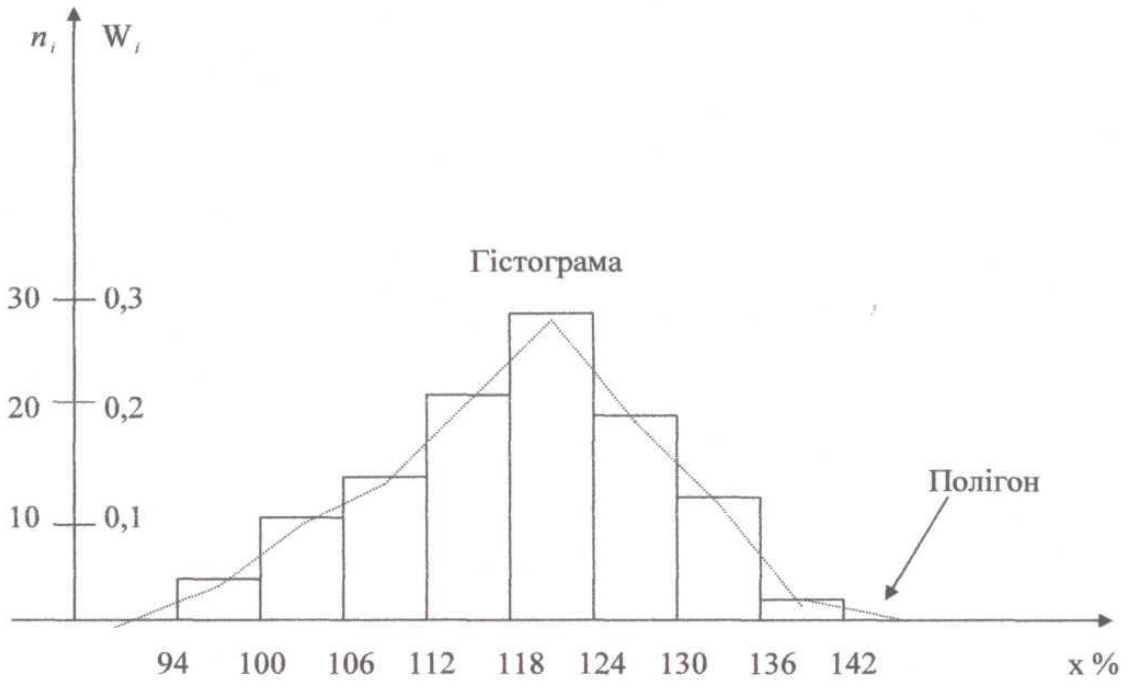


Рис. 2.4. Гістограма продуктивності праці одного робітника механічного цеху

Для дискретного ряду камулята має вигляд ламаної, яка з'єднує точки (x_i; $n_i^{нак}$) або (x_i; $W_i^{на}$), $i = 1, 2, \ldots. m$.

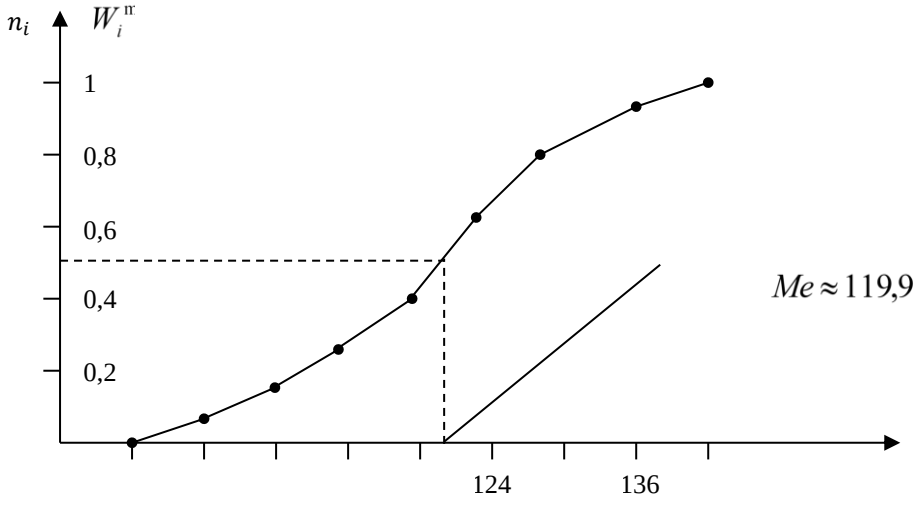


Рис. 2.5. Візуалізація кумулятивної частоти

Для інтервального варіаційного ряду ламана починається з точки, абсциса якої дорівнює початку першого інтервалу, а ордината – накопиченій частоті (частості), яка дорівнює нулю. Інші точки цієї ламаної відповідають кінцям інтервалів.

Основною метою розробки програмного продукту SPSS є обробка великих масивів даних, отриманих у процесі проведення анкетування. У таблиці наведений приклад застосування частот у різних шкалах вимірювання (табл. 2.8).

Таблиця 2.8

Види шкал

Номінальна шкала		**Порядкова шкала**		**Інтервальна шкала**	
Значення змінної					
Регіон проживання в Україні		Оцінка сьогодні матеріального стану сім'ї		Вікова шкала	
Категорія	Частота	Категорія	Частота	Категорія	Частота
Центр	622	Досить гарний	8	18–29	368
Південь	302	Гарний	100	30–39	342
Схід	386	Середній	706	40–49	301
Захід	491	Поганий	706	50–59	307
X	X	Значно поганий	224	60 и старше	482
X	X	Не можу відповісти	57	X	X

Звідси можна надати характеристику видам шкал:

– номінальна застосовується для вимірювання якісних ознак об'єкта дослідження. Для проведення розрахунків кожній властивості ознаки надається відповідний код, який дозволяє відрізняти одне значення від іншого;

– порядкова застосовується для оцінки властивостей ознаки, що міститься на різних рівнях відносно один одного. Для проведення розрахунків кодування проводиться за мірою зростання або спадання властивостей;

– інтервальна дозволяє проранжувати вимірювальну властивість і визначити, на скільки одиниць більше чи менше міститься в різних об'єктах.

Приклад. Побудований розподіл відпочиваючих на морі за кількістю днів (x) (рис. 2.6).

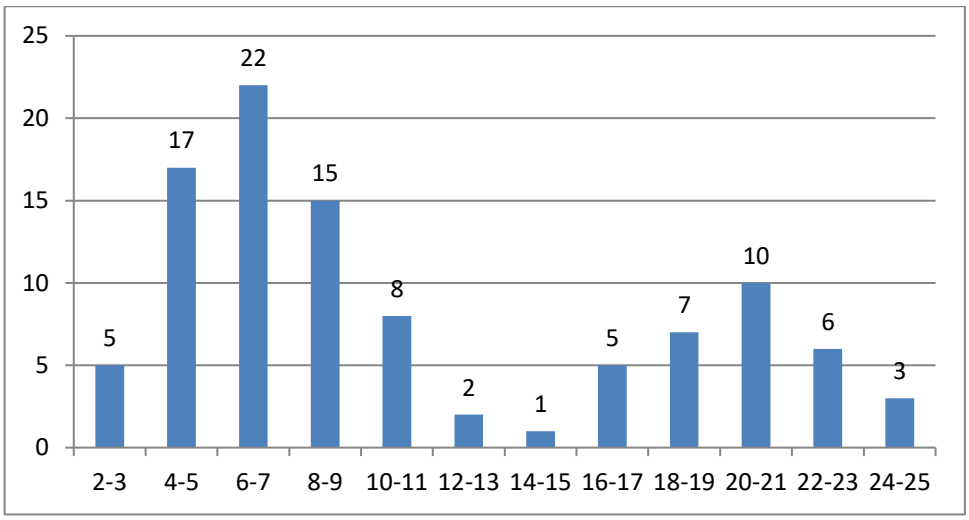


Рис. 2.6. Розподіл відпочиваючих на морі за кількістю днів

Висновок: розподіл має назву «двогорбий верблюд», потрібно розбити дані на дві групи (1 група – від 2 до 13, 2 група – від 14 до 25 днів) і аналізувати їх окремо.

Візуальне представлення розподілу за допомогою SPSS дає можливість отримати про нього якісну інформацію.

Приклад. Побудований розподіл респондентів за кількістю осіб, проживаючих разом (x) (рис. 2.7).

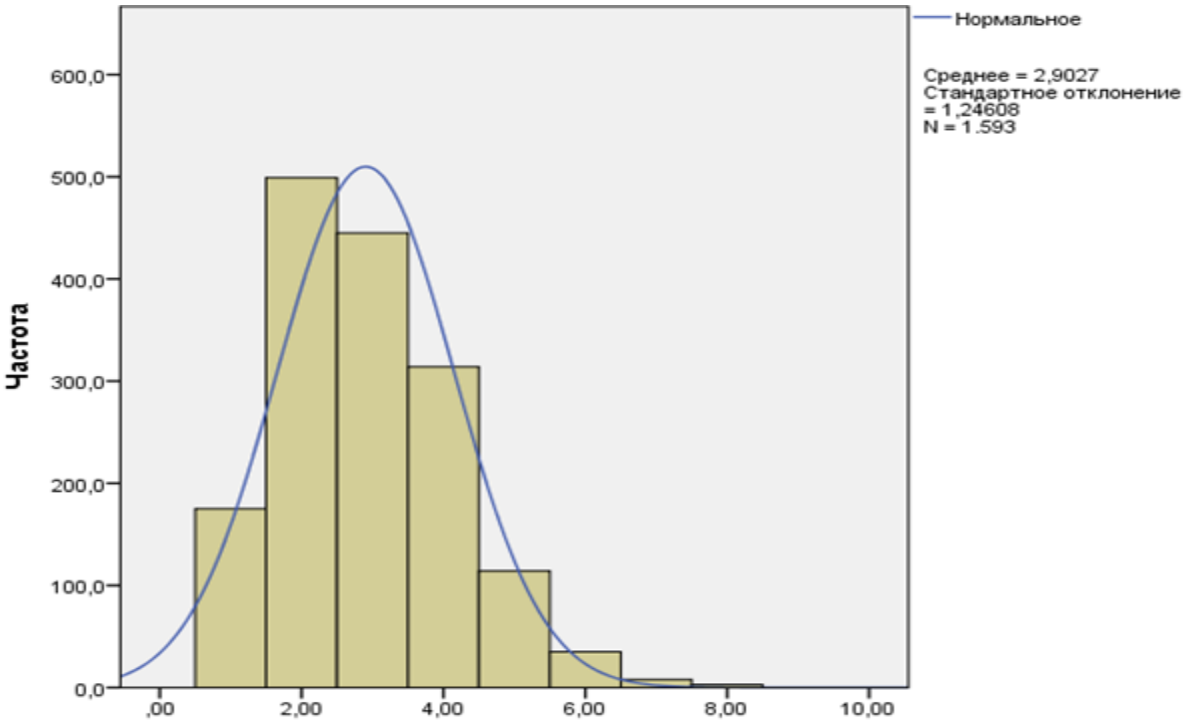


Рис. 2.7. Розподіл респондентів за кількістю осіб, які проживають разом

Висновок: з опитаних 1 593 респондентів у середньому проживає 3 особи. Найбільше з опитаних мають 2 членів родини, найменше – 8. Присутній нормальний розподіл з незначною правосторонньою асиметрією.

2.2. Показники центру розподілу

Середнє значення (mean) – рівне сумі всіх значень розподілу, поділене на їхню кількість.

Вибіркове середнє називається середнім арифметичним вибірки, тобто сума всіх значень вибірки, поділена на її кількість:

$$\bar{x} = \frac{\sum x_j}{n},$$

де $\sum x_j$ – сума всіх значень вибірки, n – об'єм вибірки.

Моду і медіану називають структурними (порядковими) середніми.

Медіана – значення ознаки, яке міститься посередині впорядкованого щодо зростання варіаційного ряду. Значення медіани для інтервального варіаційного ряду; це варіанта, яка припадає на середину упорядкованого ряду розподілу і ділить його на дві рівні за обсягом частини, обчислюють за формулою:

$$M_e = x_0 + h \cdot \frac{0{,}5\sum_1^m f_j - S_{f_{me}-1}}{f_{m_e}},$$

де M_e – медіана; x_0 – нижня межа медіанного інтервалу; h – ширина медіанного інтервалу; $\sum f$ – сума частот; $S_{f_{me}-1}$ – кумулятивна частота передмедіанного інтервалу; f_{me} – частота медіанного інтервалу.

Для визначення медіани використовують кумулятивні частоти або частки.

У дискретному ряді медіаною буде значення ознаки, для якої кумулятивна частота (сума нагромаджених частот) дорівнює або перевищує половину суми всіх частот.

Мода – значення ознаки, що найбільш часто зустрічається у ряді розподілу; найпоширеніше значення ознаки, тобто варіанта, який у ряді розподілу має найбільшу частоту, розраховується за такою формулою:

$$M_0 = x_0 + h \cdot \frac{f_{m_0} - f_{m_0-1}}{(f_{m_0} - f_{m_0-1}) + (f_{m_0} - f_{m_0+1})},$$

де M_0 – мода; x_0 – нижня межа модального інтервалу; $f_{m_0}, f_{m_0-1}, f_{m_0+1}$ – частоти (відносні частоти) модального, передмодального, післямодального інтервалів, відповідно; h – довжина інтервалу. У **дискретному** ряді розподілу **модальне значення** ознаки знаходять візуально за **найбільшою** частотою (**часткою**).

Залежно від виду даних можна розрахувати такі центри розподілів (табл. 2.9).

Таблиця 2.9

Застосування вимірів основної тенденції за видами змінних

Типове значення	Номінальні дані	Порядкові дані	Інтервальні дані
Мода	●	●	●
Медіана		●	●
Середнє			●

Для аналізу закономірностей розподілу використовують також квантилі, децилі тощо.

Приклад. Проведено розподіл підприємств за рівнем витрат на рекламу (табл. 2.10):

Таблиця 2.10

Проміжні розрахунки витрат підприємств

Групи підприємств за розміром витрат, тис. грн	**Кількість підприємств,** f_j	**Середина інтервалу,** x'_j	$x_j f_j$	**Кумулятивна частка,** S_f
до 5	17	4	68	17
5–7	39	6	234	56 (17 + 39)
7–9	**51**	8	408	107
9–11	42	10	420	149
11–13	29	12	348	178
13–15	15	14	210	193
15 і більше	7	16	112	200
Разом	200	х	1 800	х

$\bar{x} = 1800/200 = 9$ тис. грн у середньому одне підприємство витрачає на рекламу.

Найбільшу частоту має інтервал 7–9, $f_{m_0} = 51$, ширина інтервалу 2, нижня межа $x_0 = 7$, передмодальна частота $f_{m_0-1} = 39$, післямодальна частота $f_{m_0+1} = 42$.

$$M_0 = 7 + 2 \cdot \frac{51 - 39}{(51 - 39) + (51 - 42)} = 8{,}1 \text{ тис. грн}$$

(найбільш поширений рівень витрат серед підприємств на рекламу становить 8,1 тис. грн).

Половина обсягу сукупності $0{,}5\sum_1^m f_j = 0{,}5 \cdot 200 = 100$ припадає на інтервал 7–9 з частотою $f_{m_0} = 51$, кумулятивна частота перед медіанного інтервалу $S_{f_{me}-1} = 56$, тоді (половина підприємств витрачає до 8,7 тис. грн, інші менше ніж 8,7 тис. грн)

$M_e = 7 + 2 \cdot \frac{100 - 56}{51} = 8{,}7$ тис. грн.

Якщо найбільшу частоту мають два значення ознаки, вибірковий розподіл називається бімодальним.

2.3. Характеристика діапазону розподілу

Середня величина може бути однакова у різних значеннях (табл. 2.11).

Таблиця 2.11

Приклади варіаційних рядів

Показник	I	II	III
	99	90	1
	100	100	100
	101	110	199
Середнє значення	100	100	100

Висновок: усі три ряди мають однакове середнє значення, при цьому перший ряд має найменше відхилення значень від середнього (1), а третій найбільше (99).

Варіація – це мінливість (коливання) індивідуальних значень ознаки сукупності, тобто вимірювання ступеня коливання ознаки.

Характеристики варіації: інтенсивність структурних зрушень, щільність взаємозв'язків соціально-економічних явищ, точність результатів вибіркового обстеження.

Для вимірювання варіації використовують абсолютні та відносні показники варіації:

– абсолютні показники: розмах варіації, середнє лінійне та середнє квадратичне відхилення, дисперсія;

– відносними показниками варіації є коефіцієнт осциляції, лінійний і квадратичний коефіцієнти варіації.

Розмах варіації (R) **–** це різниця між максимальним і мінімальним значенням ознаки: $R = x_{\max} - x_{\min}$.

Середнє лінійне відхилення – середній з модулів відхилень:

– для генеральної сукупності:

$$\bar{l} = \frac{\sum_{1}^{m} \left| x_j - \bar{x} \right| \cdot f_j}{\sum_{1}^{m} f_j}, \text{ або } \bar{l} = \frac{\sum_{1}^{n} \left| x_j - \bar{x} \right|}{n} \text{ – незважена середня.}$$

– для вибірки: $\bar{l} = \frac{\sum_1^n \left| x_j - \bar{x} \right|}{n-1}$.

Стандартне відхилення (standard deviation) – корінь квадратний з дисперсії:

– для генеральної сукупності:

$\sigma = \sqrt{\frac{\sum_1^m \left(x_j - \bar{x}\right)^2 f_j}{\sum_1^m f_j}}$, або $\sigma = \sqrt{\frac{\sum_1^n \left(x_j - \bar{x}\right)^2}{n}}$ – незважена середня;

– для вибірки: $S = \sqrt{\frac{\sum_1^n \left(x_j - \bar{x}\right)^2}{n-1}}$.

Дисперсія – середній квадрат відхилень:

– для генеральної сукупності:

$\sigma^2 = \frac{\sum_1^m \left(x_j - \bar{x}\right)^2 f_j}{\sum_1^m f_j} = \overline{x^2} - \bar{x}^2$, або $\sigma^2 = \frac{\sum_1^m \left(x_j - \bar{x}\right)^2}{n}$;

– для вибірки: $S^2 = \frac{\sum_1^n (x_j - \bar{x})^2}{n-1}$.

Коефіцієнт осциляції – характеризує відносну варіацію крайніх значень ознаки навколо середньої: $V_R = \frac{R}{\bar{x}} \cdot 100\ \%$.

Коефіцієнти варіації:

– ***лінійний*** – характеризує співвідношення середнього лінійного відхилення ознаки та її середньої величини: $V_l = \frac{\bar{l}}{\bar{x}} \cdot 100\ \%$;

– ***квадратичний*** – характеризує співвідношення середнього квадратичного відхилення ознаки та її середньої величини:

$$V_\sigma = \frac{\sigma}{\bar{x}} \cdot 100\%.$$

Оцінка ступеня варіації:

$V < 15\ \%$ – слабка, V від 15 до 33 % – середня, $V > 33\ \%$ сильна.

У комерційній діяльності характеризує ризик, показує, наскільки невизначеною є ситуація.

Приклад. Визначити кандидатуру за віковою ознакою, яка може претендувати на керівну посаду в університеті. Вікова шкала дійсних завідувачів кафедр, декана і його заступників одного із факультетів наведена в таблиці 2.12.

Таблиця 2.12

Вікова шкала керівного складу факультету

Показники	**Повні роки**	$x_j - \bar{x}$	$\lvert x_j - \bar{x} \rvert$	$(x_j - \bar{x})^2$	**Варіанта, *x***	**Частота, *f***	**Кумулятивна частота, *S***
1	54	6,3	6,3	39,69	32	1	1
2	32	–15,7	15,7	246,49	39	1	2
3	43	–4,7	4,7	22,09	42	2	4
4	39	–8,7	8,7	75,69	43	1	5
5	45	–2,7	2,7	7,29	45	2	7
6	42	–5,7	5,7	32,49	54	1	8
7	42	–5,7	5,7	32,49	61	1	9
8	45	–2,7	2,7	7,29	74	1	10
9	74	26,3	26,3	691,69	x	x	x
10	61	13,3	13,3	176,89	x	x	x
Середнє значення	47,70	Усього	91,80	1 332,10	x	x	x

Мінімум (minimum) – найменше значення в розподілі (в нашому прикладі 32 роки).

Максимум (maximum) – найбільше значення в розподілі (у нашому прикладі 74 роки).

Розмах варіації: 74 – 32 = 42 роки.

Середнє лінійне відхилення:

$$\bar{l} = \frac{\sum_{1}^{n} \lvert x_j - \bar{x} \rvert}{n-1} = \frac{91{,}80}{10-1} = 10{,}20.$$

Дисперсія:

$$S^2 = \frac{\sum_{1}^{n} (x_j - \bar{x})^2}{n-1} = \frac{1332{,}10}{10-1} = 148{,}01.$$

Стандартне відхилення:

$S = \sqrt{S^2} = \sqrt{148{,}01} = 12{,}17$.

Коефіцієнт осциляції: $V_R = \frac{R}{\bar{x}} \cdot 100\ \% = \frac{42}{47{,}7} \cdot 100\ \% = 88{,}05\ \%$.

Коефіцієнти варіації:

– лінійний: $V_l = \frac{\bar{l}}{x} \cdot 100\% = \frac{10{,}20}{47{,}70} \cdot 100\% = 21{,}38\%$;

– квадратичний: $V_\sigma = \frac{S}{\bar{x}} \cdot 100\ \% = \frac{12{,}17}{47.7} \cdot 100\ \% = 25{,}51\ \%$.

Медіана: половина обсягу сукупності: $0{,}5\sum_1^m f_j = 0{,}5 \cdot 10 = 5$, припадає на значення 43 роки (у разі непарного числа ряду це була б медіана), але наш ряд складається з парного числа членів (8), тому обирається два середні значення варіанти $M_e = \frac{43 + 45}{2} = 44$ роки.

Прогнозне значення середнього визначаємо за допомогою довірчих інтервалів:

t-статистика Стьюдента для ймовірності 95 % та 10 – 1 = 9 ступенів свободи дорівнює 2,26:

$$E = t\frac{S}{\sqrt{n}} = 2{,}26\frac{12{,}17}{\sqrt{10}} = 8{,}698$$

$$\bar{x} - E < \mu < \bar{x} + E$$

$$47{,}7 - 8{,}698 < \mu < 47{,}7 + 8{,}698$$

$$39 < \mu < 56.$$

Висновок: середній вік керівного складу факультету становить 47,7 років, при цьому керівну посаду очолює і 32-річний науково-педагогічний персонал (мінімальне значення) і 74-річний (максимальне значення). Посаду може обійняти претендент у віці 39–56 років, який відповідає кваліфікаційним вимогам. Маємо бімодальний розподіл, найбільш вживаний вік керівного складу 45 і 42 роки. Половина керівного складу факультету має вік до 44 років (медіана), інші – старші. Розмах варіації становить 42 роки. Середня типова (сукупність однорідна) варіація становить 25,5 % (менше 33 %).

Приклад. Зробити прогноз у генеральній сукупності витрат підприємств та оцінити ризики (табл. 2.13).

Таблиця 2.13

Проміжні розрахунки витрат підприємства

Витрати, тис. грн, x_j	**Кількість підприємств, f_j**	$x_j - \bar{x}$	$\lvert x_j - \bar{x}\rvert \cdot f$	$(x_j - \bar{x})^2$	$(x_j - \bar{x})^2 \cdot f_j$
4	17	-5	85	25	425
6	39	-3	117	9	351
8	51	-1	51	1	51
10	42	1	42	1	42
12	29	3	87	9	261
14	15	5	75	25	375
16	7	7	49	49	343
Разом	**200**	х	**506**	х	**1 848**

Середнє значення:

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{4\cdot 17+6\cdot 39+8\cdot 51+10\cdot 42+12\cdot 29+14\cdot 15+16\cdot 7}{200} = 9 \text{ тис. грн,}$$

$$\bar{l} = \frac{\sum_1^m \lvert x_j - \bar{x}\rvert \cdot f_j}{\sum_1^m f_j} = \frac{506}{200} = 2{,}53 \text{тис. грн}$$

$$\sigma^2 = \frac{\sum_1^m (x_j - \bar{x})^2 f_j}{\sum_1^m f_j} = \frac{1848}{200} = 9{,}24 \text{ тис. грн,}$$

$$\sigma = \sqrt{\frac{\sum_1^m (x_j - \bar{x})^2 f_j}{\sum_1^m f_j}} = \sqrt{\sigma^2} = \sqrt{9{,}24} = 3{,}04 \text{ тис. грн.}$$

$$E = z\frac{\sigma}{\sqrt{n}} = 1{,}96\frac{3{,}04}{\sqrt{200}} = 0{,}42$$

$$\bar{x} - E < \mu < \bar{x} + E$$

$$9 - 0{,}42 < \mu < 9 + 0{,}42$$

$$8{,}58 < \mu < 9{,}42$$

$$V_\sigma = \frac{\sigma}{\bar{x}} \cdot 100\ \% = \frac{3{,}04}{9} \cdot 100\ \% = 33{,}8\ \%,$$

$$V_l = \frac{\bar{l}}{\bar{x}} \cdot 100\ \% = \frac{2{,}53}{9} \cdot 100\ \% = 28{,}11\ \%.$$

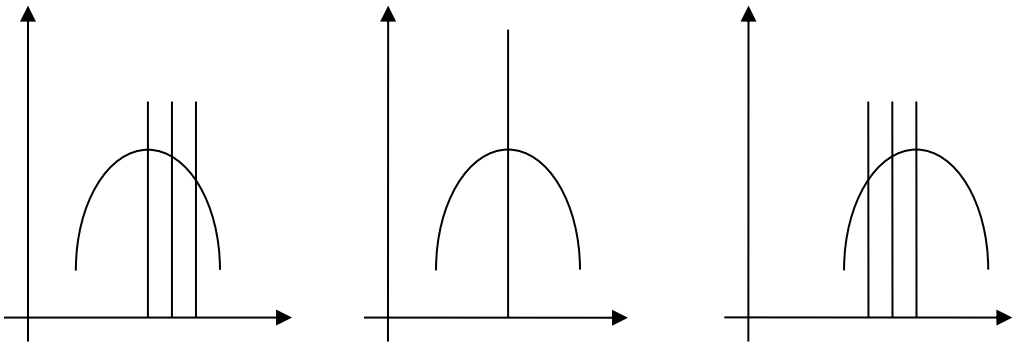
Висновок: у наступному періоді прогнозні витрати 200 підприємств становитимуть від 8,58 до 9,42 тис. грн. При цьому є ризики, що вони будуть більшими, присутня сильна ступінь варіації відносно середнього (33,8 %).

2.4. Різновиди та характеристики форм розподілів

За формою розподіляються на одно-, дво- і багатовершинні. Наявність двох і більше вершин свідчить про неоднорідність сукупності, поєднання у ній груп з різними рівнями ознаки. Розподіли якісно однорідних сукупностей, як правило, одновершинні.

Серед одновершинних розподілів є симетричні та асиметричні (скошені), гостро- і плосковершинні.

У симетричному розподілі рівновіддалені від центра значення ознаки мають однакові частоти, а в асиметричному – вершина розподілу зміщена. Напрям асиметрії протилежний напряму зміщення вершини. Якщо вершина зміщена ліворуч, то ця асиметрія правостороння, і навпаки. Асиметрія виникає внаслідок обмеженої варіації в одному напрямі або за умови домінування причини розвитку, яка веде до зміщення центра розподілу (рис. 2.8).



Правостороння асиметрія («хвіст» праворуч)	Симетричний розподіл	Лівостороння асиметрія («хвіст» ліворуч)
$A_S > 0$	$A_S = 0$	$A_S < 0$
$\bar{x} > M_e > M_0$	$\bar{x} = M_e = M_0$	$\bar{x} < M_e < M_0$.

Рис. 2.8. Візуалізація центрів розподілу (додаток Б)

Для відображення наближеності форми розподілу до нормального виду існує дві основні характеристики: асиметрія й ексцес.

Асиметрія (skewness) – відносне відхилення, яке характеризує напрям і міру скошеності в середині розподілу, тобто в який бік відносно середнього зміщені більшість значень розподілу. Нульове значення асиметрії вказує симетричність розподілу відносно середнього значення. Позитивна асиметрія вказує на здвиг розподілу в бік менших значень, а негативна – в бік більших значень. У більшості випадків за нормальний приймається розподіл з асиметрією, яка полягає в межах від –1 до +1. У дослідженнях, що не потребують високого рівня точності результатів, нормальним вважається розподіл з асиметрією, яка за модулем не перевищує 2.

Коефіцієнт асиметрії (за Пірсоном) – стандартне відхилення: $A_S = \dfrac{\bar{x} - M_0}{\sigma}$.

При $|A_S| < 0{,}25$ – асиметрія слабка;

при $0{,}25 < |A_S| < 0{,}5$ – асиметрія середня;

при $|A_S| > 0{,}5$ – асиметрія сильна.

Ексцес – відображає ступінь зосередженості елементів сукупності навколо центра розподілу. Якщо значення ексцесу близьке до нуля, то форма розподілу близька до нормального. Розраховується за формулою:

$$E_x = \frac{\mu_4}{\sigma^4}, \text{ де } \mu_4 = \frac{\sum_1^m \left(x_j - \bar{x}\right)^4 f_j}{\sum_1^m f_j}.$$

Ексцес більше 3 – гостровершинний, менше 3 – плосковершинний.

2.5. Ящикова діаграма, переваги застосування

Максимальну кількість інформації про ряд розподілу нам дає візуальне представлення результатів за допомогою ящикової діаграми (рис. 2.9).

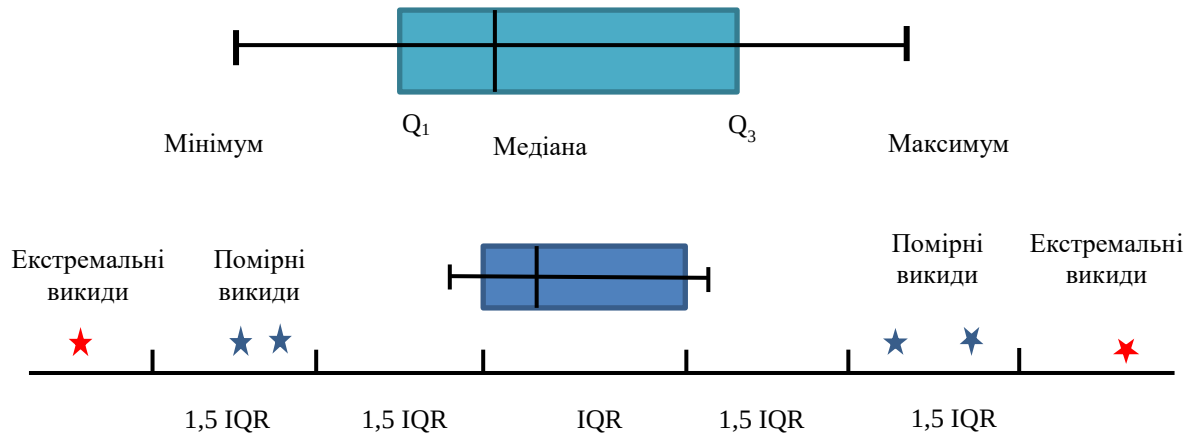


Рис. 2.9. Основні характеристики ящикової діаграми

Лівий і правий бік прямокутника (ящика) відповідають першому ($Q_1 = 25$ %) і третьому ($Q_3 = 75$ %) квантилям (значення, що виокремлюють ¼ і ¾ вибірки). Відстань між 1-м і 3-м квантилями – це міжквантильний розмах (або відстань):

$$IQR = Q_3 - Q_1.$$

Горизонтальні точки на кінці «вусів» – це максимальне і мінімальне значення розподілу. Окремі точки, які виходять за межі ящика, – це викиди, вони бувають помірні й екстремальні. Екстремальні викиди потрібно розглядати окремо. Під час порівняння вибірок, які поділені на категорії, цей вид діаграми найбільш наглядно характеризує розподіл.

Приклад. Знайти квантилі вибірки і міжквантильний розмах за даними 8, 9, 10, 11, 12, 13, **15**, 15, 18, 20, 20, 26, 53. З'ясувати, чи є екстремальні викиди.

Медіана дорівнює 15 (13 спостережень 7 значення є медіаною).

Медіана = Q_2= 15 (50 % вибірки).

$Q_1 = \frac{10+11}{2} = 10{,}5$ (25 % вибірки).

$Q_2 = \frac{20+20}{2} = 20$ (75 % вибірки).

Міжквантильний розмах:

$IQR = Q_3 - Q_1 = 20 - 10{,}5 = 9{,}5$ $1{,}5IQR = 1{,}5 \cdot 9{,}5 = 14{,}25$.

Визначаємо межі ящика:

– нижня межа:

$Q_1 - 1{,}5IQR = 10{,}5 - 14{,}25 = -3{,}75.$

Найнижче значення нашого ряду спостереження 8 (це – лівий «вус»);

– верхня межа:

$Q_3 + 1{,}5IQR = 20 + 14{,}25 = 34{,}25.$

До значення 34,25 крайнім є 26 (це – правий «вус»).

Розраховуємо межі помірних викидів (рис. 2.10):

– обчислюємо праву межу:

$3IQR = 3 \cdot 9{,}5 = 28{,}5$

$Q_3 + 3IQR = 20 + 28{,}5 = 48{,}5;$

– обчислюємо ліву межу:

$Q_1 - 3IQR = 10{,}5 - 28{,}5 = -18{,}0.$

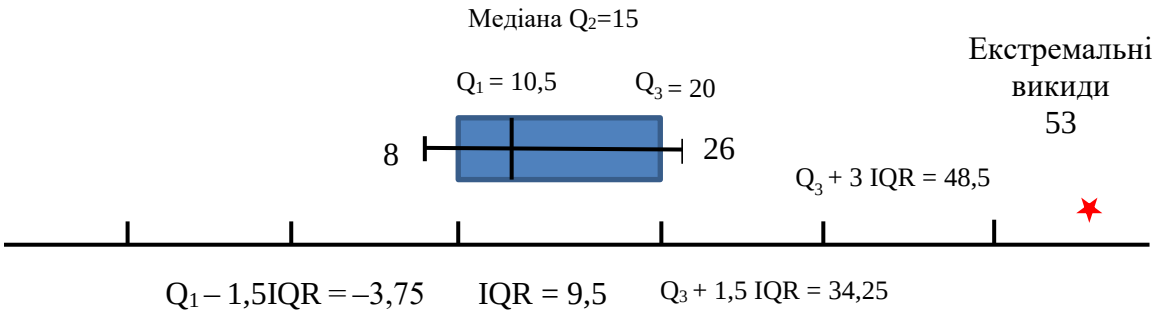


Рис. 2.10. Аналіз ряду спостереження за допомогою ящикової діаграми

У нашому дослідженні ліворуч викидів немає взагалі, межу можна було б і не розраховувати.

Висновок: усі спостереження, що виходять за межі 48,5, є екстремальними викидами, у нашому випадку це значення 53 тринадцяте спостереження.

2.6. Описова статистика в SPSS

Результати описової статистики можна отримати за допомогою трьох вмонтованих функцій (рис. 2.11–2.15):

1) описова: «**Аналіз**» – «**Описательные статистики**» – «**Описательные**»

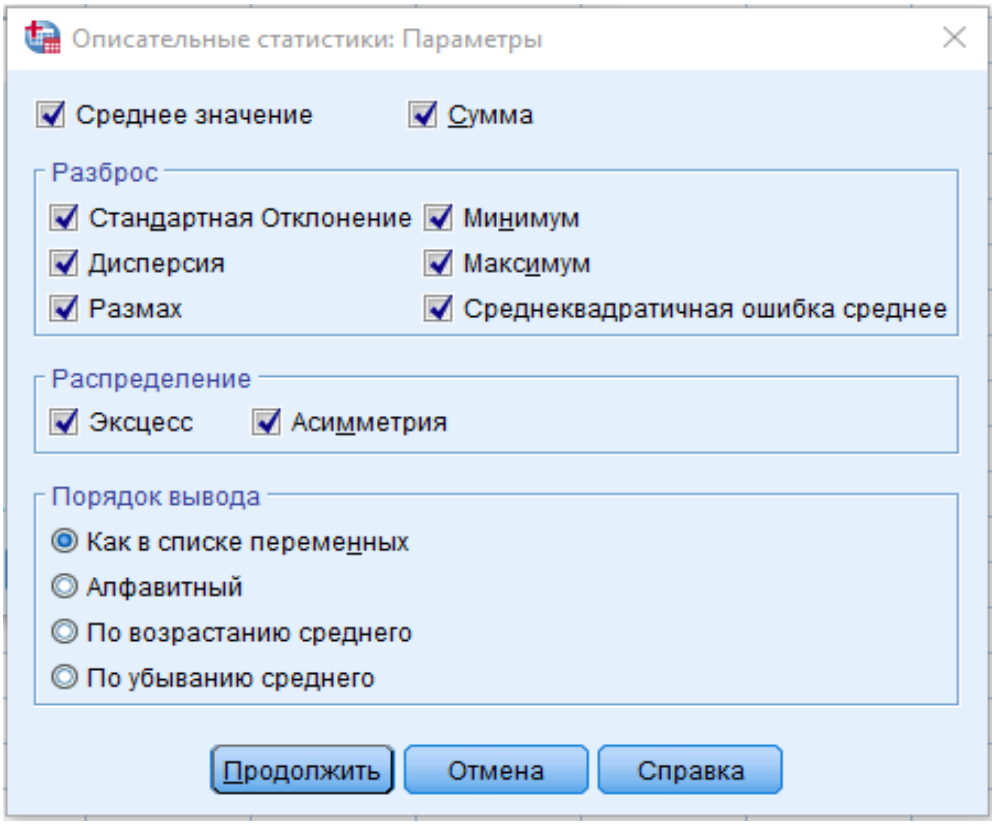


Рис. 2.11. Діалогове вікно SPSS

2) частоти: «**Анализ**» – «**Описательные статистики**» – «**Частоты**»

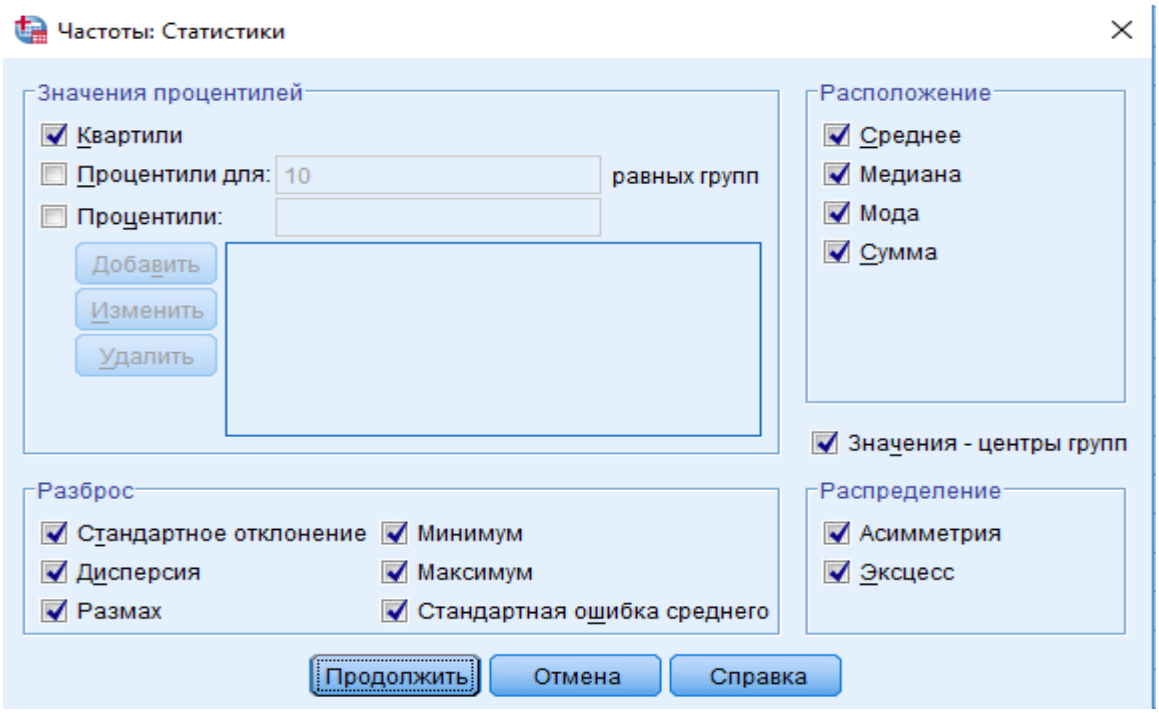


Рис. 2.12. Діалогове вікно SPSS

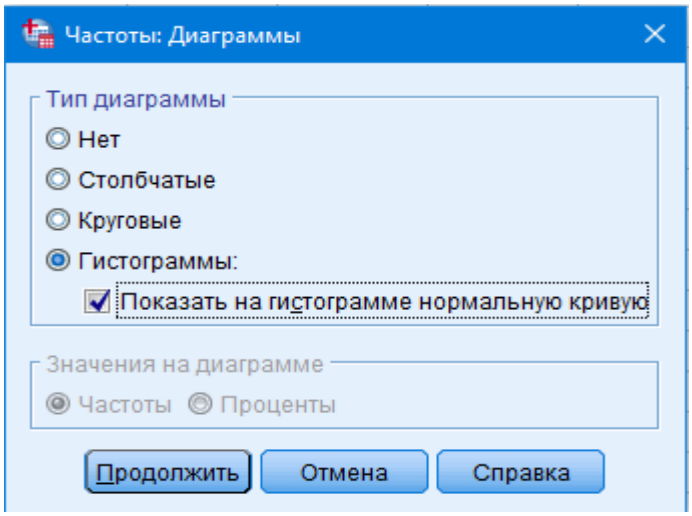


Рис. 2.13. Діалогове вікно SPSS

3) розвідувальний аналіз: «**Анализ» – «Описательные статистики» – «Разведочный анализ»**

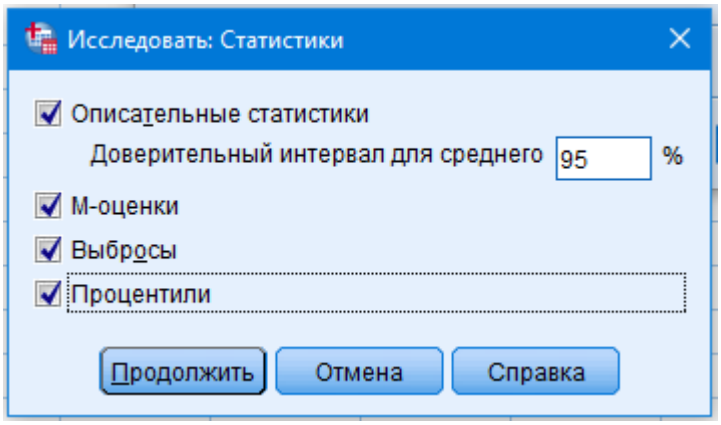


Рис. 2.14. Діалогове вікно SPSS

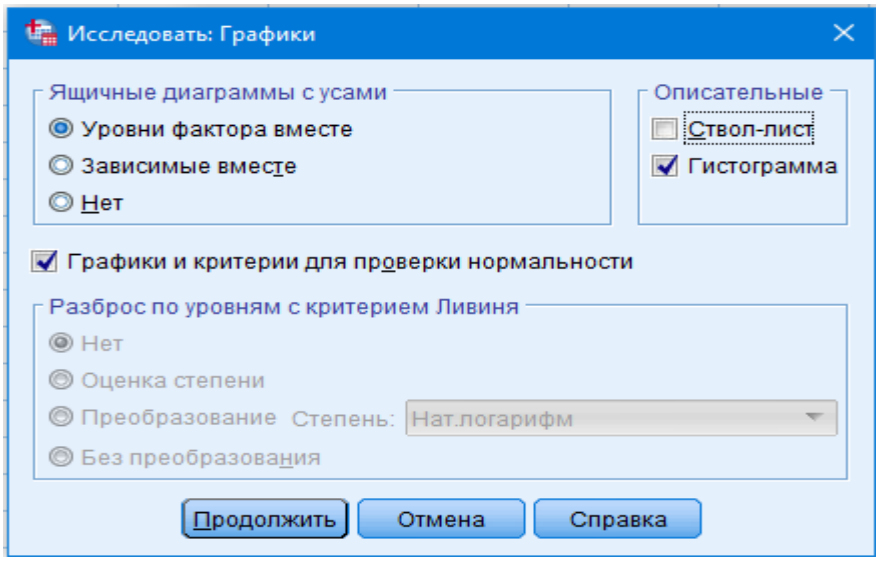


Рис. 2.15. Діалогове вікно SPSS

Приклад. Визначити кандидатуру за віковою ознакою, яка може претендувати на керівну посаду в університеті. Вікова шкала дійсних завідувачів кафедр, декана і його заступників одного із факультетів наведена: 54, 32, 43, 39, 45, 42, 42, 45, 74, 61.

Отримані за допомогою програми результати наведені в таблиці 2.14.

Таблиця 2.14

Результати описової статистики в SPSS

Описова статистика

	N	**Мінімум**	**Максимум**	**Середнє ($\bar{x}$)**	**Стд. відхилення (*S*)**
Повні роки	10	32	74	47,70	12,166
N валідних (цілим)	10				

Описова статистика

	N	**Размах (*R*)**	**Мінімум (min*x*)**	**Максимум (max*x*)**	**Середнє (*x*)**		**Стд. відхилення (*S*)**	**Диспе-рсія (S^2)**	**Асиметрія (A_s)**		**Ексцес (E_x)**	
	Статистика	**Статистика**	**Статистика**	**Статистика**	**Статистика**	**Стд. похибка**	**Статистика**	**Статистика**	**Статистика**	**Стд.похибка**	**Статистика**	**Стд. похибка**
Повні роки	10	42	32	74	47,70	3,847	12,166	148,011	1,215	0,687	1,444	1,334
N валід-них (цілим)	10											

Перевірка на нормальний розподіл:
ексцес/стандартна похибка ексцесу = 1,444 / 1,334 = 1,08
асиметрія/стандартна похибка асиметрії = 1,215 / 0,687 = 1,77.
Результат < 2 не порушує нормальний розподіл.

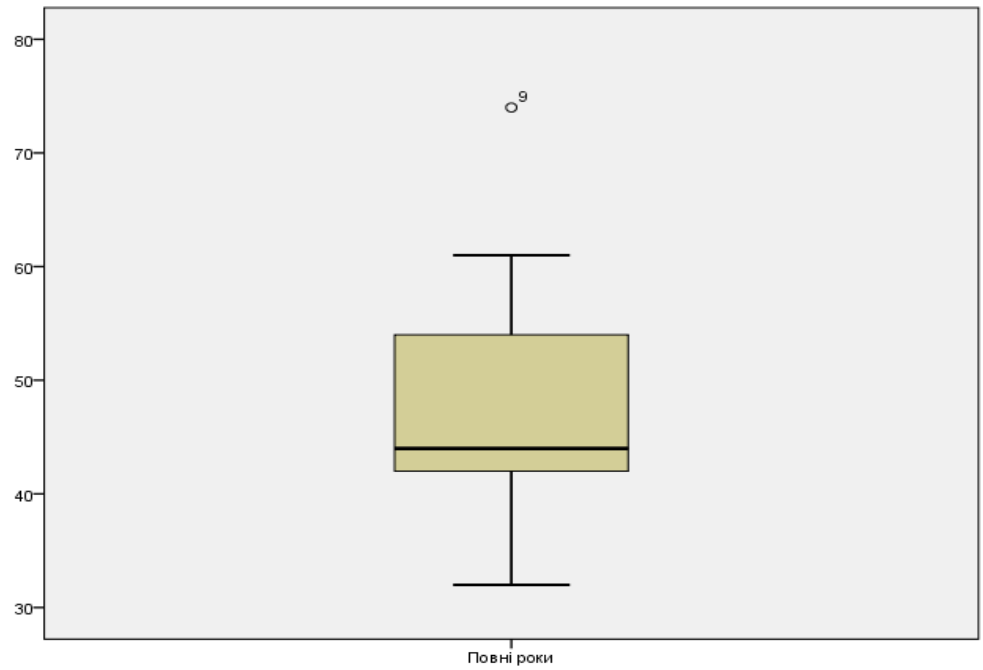


Рис. 2.16. Ящикова діаграма розподілу керівного складу

Висновок: середній вік керівного складу факультету становить 47,7 років, при цьому керівну посаду очолює і 32-річний науково-педагогічний персонал (мінімальне значення т. 2) і 74-річний (максимальне значення т. 9) (рис. 2.16). Посаду може обійняти претендент у віці 39–56 років, якій відповідає кваліфікаційним вимогам. Маємо бімодальний розподіл, найбільш вживаний вік керівного складу, – 45 і 42 роки. Половина керівного складу факультету має вік до 44 років (медіана), інші – старші. Розмах варіації становить 42 роки. Середня типова (сукупність однорідна) варіація становить 25,5 % (менше 33 %). 47,7 > 44 > 42 ($\bar{x} > M_e > M_0$) сильна правостороння асиметрія A_s = 1,215 (> 0,5), плосковершинний розподіл, ексцес 1,444 (менше 3-х), при цьому не порушено нормальний розподіл (доведено через аналіз співвідношення ексцесу й асиметрії).

Перелік питань для самоконтролю

1. Назвіть основні показники центру розподілу.
2. Назвіть відмінність між модою і медіаною.
3. Поясніть поняття «бімодальний» розподіл.
4. Назвіть характеристики діапазону розподілу.

5. Назвіть відмінність між середньоквадратичним і лінійним відхиленнями.

6. Назвіть основні різновиди форм розподілів.

7. Назвіть основні переваги застосування ящикової діаграми.

8. Поясніть поняття «гостровершинний» і «плосковершинний» ексцес.

9. Назвіть характеристику форми розподілу, що вказує міру скошеності в середині розподілу.

10. Назвіть особливість розрахунку зон помірних та екстремальних викидів.

11. Назвіть відмінність між стандартним відхиленням і дисперсією.

12. Якщо дослідник отримав бімодальний розподіл вибіркового спостереження, які його дії в оцінці статистичної сукупності?

13. Поясніть особливості застосування показників центру розподілу в номінальній, порядковій та інтервальній шкалах.

14. Поясніть структуру ящикової діаграми, вкажіть переваги і недоліки її застосування.

Тести

1. Описова статистика НЕ характеризує:

а) дисперсію;

б) форми розподілу;

в) центральну тенденцію;

г) коефіцієнт апроксимації.

2. Відносне відхилення, яке характеризує напрям і міру скошеності в середині розподілу, – це:

а) ексцес;

б) асиметрія;

в) середнє відхилення;

г) середнє квадратичне відхилення.

3. Нульове значення асиметрії вказує на:
а) симетричний розподіл
б) правостороннє відхилення
в) лівостороннє відхилення
г) однакове правостороннє та лівостороннє відхилення.

4. Позитивна асиметрія вказує на:
а) здвиг розподілу в бік більших значень;
б) здвиг розподілу праворуч;
в) здвиг розподілу ліворуч;
г) здвиг розподілу в бік менших значень.

5. Негативна асиметрія вказує на:
а) здвиг розподілу в сторону більших значень;
б) здвиг розподілу вправо;
в) здвиг розподілу вліво;
г) здвиг розподілу в сторону менших значень.

6. Слабка асиметрія |As| в межах:
а) до 0,25;
б) від 0,25 до 0,5;
в) від 0,5;
г) правильної відповіді немає.

7. Показник, який відображає ступінь зосередженості елементів сукупності навколо центру розподілу:
а) ексцес;
б) асиметрія;
в) середнє відхилення;
г) середнє квадратичне відхилення.

8. Якщо ексцес близький до 0, то розподіл:
а) нормальний;
б) близький до нормального;
в) симетричний;
г) асиметричний.

9. Якщо значення ексцесу більше чим 3, то розподіл:
а) має правосторонню асиметрію;
б) нормальний;
в) плосковершинний;
г) гостровершинний.

10. Якщо значення ексцесу менше ніж 3, то розподіл:
а) має правосторонню асиметрію;
б) нормальний;
в) плосковершинний;
г) гостровершинний.

11. Відсоток значень для кожної категорії за вирахуванням пропущених значень – це:
а) простий відсоток;
б) валідний відсоток;
в) кумулятивний відсоток;
г) складний відсоток.

12. Накопичений відсоток величини – це:
а) простий відсоток;
б) валідний відсоток;
в) кумулятивний відсоток;
г) правильної відповіді немає.

13. Кількість спостережень, в яких ознака (x) приймає певне значення або міститься в певному інтервалі, – це:
а) мода;
б) медіана;
в) частота;
г) середнє значення.

14. Валідні значення – це:
а) відсоток від загальної кількості з урахуванням пропусків;
б) правильної відповіді немає;
в) список значень змінних;
г) список факторів впливу.

15. Відсоток – це:

а) відсоток від загальної кількості з урахуванням пропусків;

б) кількість об’єктів, яка відповідає градації кожній змінній;

в) накопичений відсоток величини;

г) відсоток значень для кожної категорії за вирахуванням пропущених значень.

16. Графічне представлення даних – це:

а) кореляційний аналіз;

б) візуалізація даних;

в оптимізаційна модель;

г) бізнес-процес.

17. Найпоширеніші способи візуалізації даних:

а) графіка, схеми;

б) діаграми;

в) інфографіка, бізнес-аналітика;

г) усі відповіді правильні.

18. Види змінних поділяють на:

а) номінальні дані;

б) порядкові дані;

в) інтервальні (кількісні) дані;

г) усі відповіді правильні.

19. Види частот поділяють на:

а) номінальну шкалу;

б) порядкову шкалу;

в) інтервальні частоти;

г) усі відповіді правильні.

20. Номінальні дані – це:

а) якісна ознака;

б) кількісна ознака;

в) кумулятивна ознака;

г) правильної відповіді немає.

21. Оберіть порядкову шкалу – це:

а) гендерна ознака, регіон проживання;

б) сорт товару, рівень доходу;

в) кількісна ознака;

г) попередній відбір даних для аналізу.

22. Стандартне табличне подання з можливістю сортування, експорту та фільтрації даних – це:

а) статистика;

б) дані;

в) таблиця;

г) гістограма.

23. Гістограма – це:

а) спосіб графічного представлення, який складається з прямокутників;

б) графік у формі коробочки з вусиками;

в) графік у формі кола, який поділений на сегменти;

г) правильної відповіді немає.

24. Валідний відсоток – це:

а) правильної відповіді немає;

б) кількість об’єктів, яка відповідає градації кожної змінної;

в) накопичений відсоток величини;

г) відсоток значень для кожної категорії за вирахуванням пропущених значень.

25. Значення випадкової величини, що трапляється найчастіше в сукупності спостережень:

а) медіана;

б) мода;

в) вибіркове середнє;

г) правильної відповіді немає.

26. Медіана – це:

а) значення, в якому *x* набуває максимальної ознаки;

б) значення, в якому *x* набуває мінімальної ознаки;

в) величина ознаки, що розташована посередині ранжованого ряду вибірки;

г) величина ознаки, що не має відношення до вибірки.

27. Сума всіх значень вибірки поділена на їх кількість:

а) медіана;

б) мода;

в) вибіркове середнє;

г) середнє арифметичне.

28. Різниця між найбільшим та найменшим значеннями в ряду спостережень:

а) розмах R;

б) розмах P;

в) дисперсія;

г) частота.

29. Слугує для оцінки симетричності розподілу випадкової величини щодо середньої:

а) коефіцієнт кореляції;

б) коефіцієнт детермінації;

в) асиметрія;

г) правильної відповіді немає.

30. Вибіркове середнє:

а) середнє арифметичне всіх вибіркових значень;

б) сума всіх вибіркових значень;

в) добуток усіх вибіркових значень;

г) правильної відповіді немає.

31. Ексцес:

а) правильної відповіді немає;

б) лінійний або нелінійний;

в) гостровершинний або плосковершинний;

г) сильний або слабкий.

32. Бімодальний розподіл:

а) має одну точку;

б) має дві точки, а не одну;

в) не має точок;

г) правильної відповіді немає.

33. Позитивна асиметрія:

а) вказує на здвиг розподілу в бік більших значень;

б) вказує на здвиг розподілу в бік менших значень;

в) не вказує ні на що;

г) інша відповідь.

34. Форма розподілу близька до нормального, якщо ексцес:

а) близький до 1;

б) близький до 0;

в) міститься в межах від –1 до 1;

г) будь-яке значення.

35. У більшості випадків за нормальний розподіл приймається асиметрія, яка:

а) не перевищує значення |2|;

б) перевищує значення |2|;

в) більша 3;

г) правильної відповіді немає.

36. Відображає ступінь зосередженості елементів сукупності навколо центра розподілу:

а) критерій Фішера;

б) ексцес;

в) вибіркове середнє;

г) критерій Дарбіна-Уотсона.

Економічна інтерпретація статистичного аналізу

Приклад 1. Провести статистичний аналіз основних показників діяльності ПАТ КБ «ПриватБанк» за 2001–2017 роки за даними таблиці 1.

Таблиця 1

Основні показники діяльності ПАТ КБ «ПриватБанк» за 2001–2017 рр.

Рік	Фінансовий результат, млн грн	Активи, млн грн	Зобов'язання, млн грн	Депозити юр. і фіз. осіб
2001	33,48	7221,36	6 912,33	2 409,57
2002	35,07	10 259,71	9 837,99	4 336,65
2003	49,19	17 724,43	17 092,38	6 017,72
2004	289,8	14 671,25	13 308,96	10 317,32
2005	437,07	21 719,16	19 606,09	13 627,08
2006	382,52	30 652,74	27 493,53	20 220,28
2007	1 052,51	51 149,69	46 242,43	36 249,36
2008	990,5	78 410,04	70 515,63	55 244,57
2009	842,96	81 813,22	71 753,93	52 858,58
2010	1 239,78	109 752,52	98 001,05	77 139,56
2011	1 216,43	14 2236,7	12 5700,1	92 043,44
2012	1 410,42	169 570,39	151 391,27	106 275,74
2013	1 832,07	211 425,77	191 189	129 863,54
2014	652,63	201 471,24	180 045,37	144 343,78
2015	239,82	251 551,13	225 827,12	169 502,57
2016	628,51	269 032,37	240 100,58	191 603,55
2017	1 734,99	245 882,11	218 345,19	564 738,72

Джерело: побудовано на основі даних Офіційного сайту ПАТ КБ «ПриватБанк»[1].

У середньому ПАТ КБ «ПриватБанк» за сімнадцять років отримував 768,69 млн грн фінансового результату, 11 2620,22 млн грн активів та 100 786,05 млн грн зобов'язань за загальною сумою депозитів 98 634 млн грн (табл. 2). Стандартне відхилення, що показує розсіювання значень відносно її математичного сподівання, за фінансовими результатами становить 582,64, за активами – 9 499,27, за зобов'язаннями – 84 574,8 за депозитами – 134 703,4.

[1] Офіційний сайт ПАТ КБ «ПриватБанк» URL : https://privatbank.ua

Аналізуючи статистичні оцінки, необхідно зазначити, що фінансові результати, активи і зобов'язання мають відносно нормальний розподіл, а саме співвідношення асиметрії до її стандартної похибки й ексцесу, відповідно, вказує не на порушення умов нормальності (< 2), при цьому присутня середня правостороння асиметрія *As* < 0,5, плосковершинний розподіл (< 3).

Таблиця 2

Описова статистика показників ПАТ КБ «ПриватБанк» за 2001–2017 рр.

Показники	**Фінансовий результат, млн грн**	**Активи, млн грн**	**Зобов'язання, млн грн**	**Депозити юридичних і фізичних осіб**
Середнє	768,69	112 620,22	100 786,05	98 634,82
Мода	33,48	7 221,36	6 912,33	2 409,57
Медіана	652,63	81 813,22	7 153,93	55 244,57
Асиметрія	0,39	0,42	0,43	2,83
Ст. похибка асиметрії	0,55	0,55	0,55	0,55
Асиметрія / Ст. похибка асиметрії	0,71	0,76	0,78	5,15
Ексцес	–0,94	1,43	–1,43	9,54
Ст. похибка ексцесу	1,06	1,06	1,06	1,06
Ексцес / Ст. похибка ексцесу	–0,89	1,35	–1,35	9
Стандартне відхилення	582,64	94 992,27	84 758,8	134 703,68
Розподіл	$x_{ср}>M_e>M_o$	$x_{ср}>M_e>M_o$	$x_{ср}>M_e>M_o$	$x_{ср}>M_e>M_o$
Прогноз на 2018 рік				
Похибка (*E*)	299,59	48 843,95	43 582,02	69 263,11
Нижня / верхня межа до середнього	469,10/ 1 068,28	63 776,27/ 161 464,17	57 204,03/ 144 368,07	29 371,71/ 167 897,93
Трендовий аналіз /коефіцієнт апроксимації	1 658,92	276 886,78	247 120,26	428 566,8
	$R^2 = 0,72$	$R^2 = 0,94$	$R^2 = 0,94$	$R^2 = 0,76$

Джерело: побудовано автором.

Значно гірші показники розподілу за депозитами юридичних і фізичних осіб. Присутня значна правостороння асиметрія, гостровершинний розподіл, повністю відсутній нормальний розподіл. Зробимо прогноз двома способами: перший – відносно середнього значення, другий – аналіз часового ряду за допомогою рівняння тренду.

Прогнозне значення середнього визначаємо за допомогою довірчих інтервалів: *t*-статистика Стьюдента для ймовірності 95 % та 17 – 1 = 16 ступенів свободи дорівнює 2,12:

$$E = t\frac{S}{\sqrt{n}} = 2{,}12\frac{582{,}64}{\sqrt{17}} = 299{,}59$$

$$\bar{x} - E < \mu < \bar{x} + E$$

$$469{,}10 < \mu < 1068{,}28.$$

Фінансовий результат 2018 року становитиме 1,66 млрд грн, прогноз здійснено за допомогою степеневої функції. Активи комерційних банків 2018 року становитимуть 276,89 млрд грн, а зобов'язання – 247,12 млрд грн. За другим і третім показниками здійснено прогноз за допомогою лінійного тренду, а прогноз за депозитами юридичних осіб і фізичних осіб – за допомогою поліноміальної функції показав розмір у 428,57 млрд грн.

Аналізуючи динаміку фінансових результатів, отриманих ПриватБанком за цей період, однозначно видно тенденцію до нарощування його фінансових можливостей за виключенням декількох кризових років.

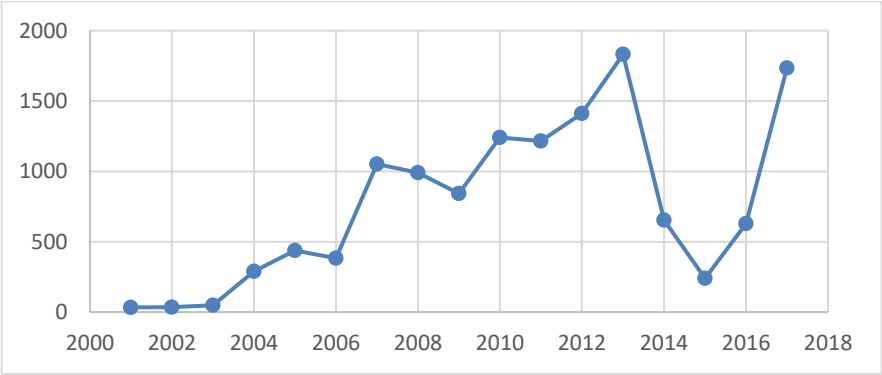


Рис. 1. Динаміка фінансового результату ПАТ КБ «ПриватБанк» за 2001–2017 роки

Джерело: побудовано автором.

Отже, згідно з графіком динаміки фінансових результатів банку (рис. 1) бачимо, що найменший результат банк мав у 2001–2003 роках (від 33 до 50 млн грн). Це явище було зумовлено нещодавнім створенням банку, низькою довірою населення до цього банку як нового на ринку, а також загальними економічними процесами, що відбувалися у пострадянській Україні. Також одним із кризових років, коли ПриватБанк мав малий фінансовий результат, а саме 239,8 млн грн, став 2015 рік. Це був рік з нестабільною політичною та економічною ситуацією в країні та процесом націоналізації самого банку.

Для усвідомлення причин таких змін у розмірах отриманих фінансових результатів банку варто дослідити основні його складові (рис. 2).

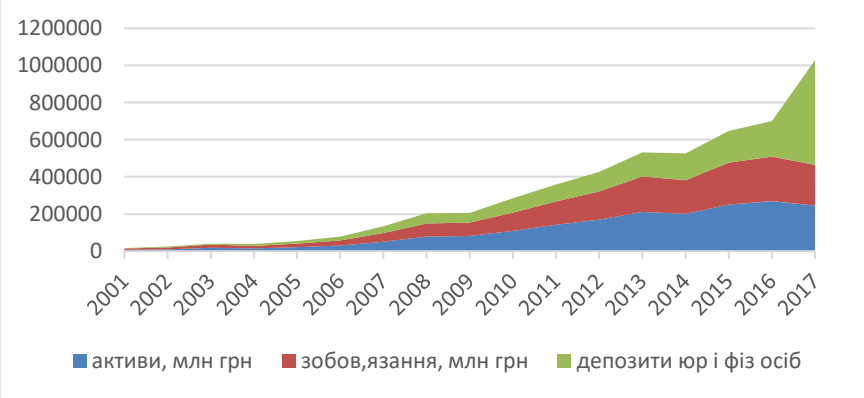


Рис. 2. Динаміка складових фінансового результату ПАТ КБ «ПриватБанк» за 2001–2017 роки

Джерело: побудовано автором.

Отже, найменша сума активів, зобов'язань та суми депозитів була 2001 року: 7 221,36, 6 912,33, 2 409,57 млн грн відповідно. Найбільшу суму активів банк мав 2016 року – 269 032 млн грн. Найбільшу суму зобов'язань 240 100 млн грн також 2016. Максимальну суму депозитів – 2017 року.

Висновок: аналізуючи тривалий період часу, необхідно враховувати зміну вартості грошей, а саме інфляційні процеси в країні. Депозити юридичних і фізичних осіб напряму залежать від таких змін. У подальшому необхідно, крім номінальних показників, дослідити реальні в цінах базового року. Цей захід зменшить коливання, що відбулися за рахунок збільшення курсу долара, покращить розподіл відносно середнього, тоді результати прогнозу будуть більш достовірнішими.

Загалом фінансове становище банку є стабільним. Поступово збільшується його надійність та фінансова стійкість. До речі, не може не втішати і той факт, що останнім часом до банку проявляють значний інтерес потенційні інвестори, зокрема міцні західні банківські структури, що хотіли б придбати акції ПриватБанку. Такий інтерес є яскравим доказом стабільного фінансового стану банку. Адже жодному інвестору навіть на думку не спаде вкладати кошти у структуру або бізнес, які мають проблеми та недостатній рівень надійності.

Приклад 2. Розглянуто особливості застосування програмного продукту IBM SPSS STATISTICS під час проведення статистичного аналізу собівартості продукції сільськогосподарського виробництва (табл. 3). Надано характеристику описовому, частотному і розвідувальному аналізам, які дають можливість дослідникові здійснити статистичний аналіз економічних показників.

Таблиця 3

Виробнича собівартість с/г продукції за регіонами України у 2017 р., млрд грн

Регіон	Виробнича собівартість продукції (робіт, послуг) с/г у підприємствах	Виробнича собівартість продукції (робіт, послуг) с/г у фермерських господарствах
Вінницька	25,069	4,292
Волинська	6,942	1,195
Дніпропетровська	19,913	4,635
Донецька	7,745	1,472
Житомирська	8,432	0,967
Закарпатська	,676	0,193
Запорізька	11,777	3,125
Івано-Франківська	5,125	0,771
Київська	29,973	2,326
Кіровоградська	14,202	5,062
Луганська	6,227	2,326
Львівська	7,200	1,487
Миколаївська	10,557	3,399
Одеська	14,687	4,925
Полтавська	24,702	3,360
Рівненська	5,447	00,538

Продовження таблиці 3

Сумська	14,010	2,039
Тернопільська	10,991	1,323
Харківська	17,879	3,174
Херсонська	10,549	2,904
Хмельницька	16,529	2,104
Черкаська	26,768	2,449
Чернівецька	1,849	0,402
Чернігівська	47,253	1,692
м. Київ	25,150	0,006

Джерело: Офіційний сайт Державної служби статистики України[2].

Результати дослідження. Можливості програмного продукту SPSS дозволяють провести статистичний аналіз, використовуючи його три основні види, а саме: описовий, частотний і розвідувальний аналізи. Описовий аналіз дає можливість проаналізувати вибірку за допомогою стандартних характеристик, а саме визначає валідність і загальну кількість спостережень, розмах варіації, мінімальне і максимальне значення, суму, стандартну похибку, дисперсію, при цьому середнє значення, асиметрія та ексцес розраховані зі стандартними похибками. Частотний, крім вищеперерахованих характеристик, додатково визначає: квантилі вибірки, частоту (відсоток) настання подій, розраховує їхнє накопичення та відображає на різних типах діаграм.

Розвідувальний аналіз найбільш повно характеризує середовище за допомогою нього і аналізує виробничу собівартість с/г продукції (робіт, послуг) у підприємствах і фермерських господарствах у розрізі регіонів за 2017 рік, а саме:

1. Дослідник отримує зведену інформацію про кількість наявних і пропущених спостережень, визначає їх відсотки. В обстежені залучено 24 регіони та місто Київ. Ця інформація набуває цінності під час обробки великих масивів статичних і динамічних даних.

[2] Офіційний сайт Державної служби статистики. URL : http://www.ukrstat.-gov.ua. (дата звернення 28.04.2019).

2. Аналізуючи описову статистику, можна стверджувати, що середня виробнича собівартість с/г продукції на підприємствах України становить 14,786 млрд грн, при цьому є регіони, в яких витрачається 0,676 млрд грн (мінімальне значення) і 47,253 млрд грн (максимальне значення). Половина регіонів України на підприємствах мають витрати до 11,777 млрд грн (медіана), інші – більші витрати. Розмах варіації становить 46,577 млрд грн. Середня нетипова (сукупність неоднорідна) варіація становить 71 % (квадратичний коефіцієнт варіації більше 33 %). Середнє значення більше за медіану (14,786 > 11,777), присутня сильна правостороння асиметрія (As = 1,303 > 0,5), плосковершинний розподіл (ексцес 2,299 менше 3-х). Середня виробнича собівартість с/г продукції у фермерських господарствах України становить 2,247 млрд грн, при цьому є регіони, в яких витрачається 0,006 млрд грн (мінімальне значення) і 5,062 млрд грн (максимальне значення). Половина регіонів України на фермерських господарствах мають витрати до 2,104 млрд грн (медіана), інші – більші витрати. Розмах варіації становить 5,056 млрд грн. Середня нетипова (сукупність неоднорідна) варіація становить 66 % (квадратичний коефіцієнт варіації більше 33 %). Присутня середня (As = 0,403 < 0,5) правостороння а (2,247 > 2,104) асиметрія, плосковершинний розподіл ексцес – 0,709 (менше 3-х) (табл. 4).

Таблиця 4

Описова статистика виробничої собівартості с/г продукції за 2017 р., млрд грн

Показники		С/в на підприємствах		С/в на фермерських господарствах	
		Статистика	Стандартна похибка	Статистика	Стандартна похибка
Середнє		14,786	2,109	2,247	0,296
95 % довірчий інтервал для середнього	Нижня границя	10,434		1,636	
	Верхня границя	19,138		2,857	
5 % усічене середнє		13,945		2,214	
Медіана		11,777		2,104	
Дисперсія		111,177		2,190	
Відхилення		10,544		1,480	

Мінімум	0,676		0,006	
Максимум	47,253		5,062	
Розмах	46,577		5,056	
Міжквантильний розмах	15,237		2,186	
Асиметрія	1,303	0,464	0,403	0,464
Ексцес	2,299	0,902	–0,709	0,902

Джерело: розраховано автором за даними[3].

Середня виробнича собівартість с/г продукції 2018 року на підприємствах становитиме від 10,434 до 19,138 млрд грн, а на фермерських господарствах – 1,636 – 2,858 млрд грн.

3. Квантильний розподіл визначає відсоток регіонів за обсягом витрат. Так, 75 % регіонів мають с/в на підприємствах у розмірі 22,491 млрд грн, а у фермерських господарствах – 3,267 млрд грн (табл. 5).

Таблиця 5

Квантилі, млрд грн

Показники		**Квантилі**						
		5	10	25	50	75	90	95
Зважене середнє	Виробнича с/в продукції с/г на підприємствах	1,028	3,815	7,071	11,777	22,491	28,610	42,069
	Виробнича с/в продукції с/г на фермерських господарствах	0,062	0,318	1,081	2,104	3,267	4,751	5,021

Джерело: розраховано автором.

4. Детальний аналіз екстремальних значень дає можливість визначити регіони, в яких витрати суттєво вирізняються від загальної сукупності, так, за підприємствами Чернігівської (47,253 млрд грн) і Київської (29,973 млрд грн) області мають максимальні витрати по Україні, а Закарпатської (0,676 млрд грн) і Чернівецька (1,849 млрд грн) області – мінімальні. За фермерськими господарствами вирізняються Кіровоградська (5,062 млрд грн), Одеська (4,925 млрд грн), Закарпатська (0,193 млрд грн) області і м. Київ (0,006 млрд грн) відповідно.

[3]Там само.

5. Побудовані графіки: ймовірнісний графік (очікування під час нормального розподілу за кількістю значень) і ймовірнісний графік з віддаленим трендом (кількість спостережень, що відхилені від нормального розподілу) дають можливість візуально зобразити усі спостереження та визначити їх відповідність умовам нормального розподілу.

6. Розрахований критерій Колмогорова-Смірнова вказує на порушення умов нормального розподілу, значимість показників більша за похибку (0,2 > 0,05 і 0,195 > 0,05). При цьому виробнича с/в продукції на підприємствах наближається до нормального, критерій Шапіро-Уілка значущості 0,021 < 0,05 (табл. 6).

Таблиця 6

Критерій нормальності

Показники	Колмогоров-Смірнов			Шапіро-Уілк		
	Статистика	Ст. св.	Значимість	Статистика	Ст. св.	Значимість
Виробнича собівартість продукції с/г у фермерських господарствах	0,096	25	0,200*	0,956	25	0,346
Виробнича собівартість продукції с/г у підприємствах	0,144	25	0,195	0,902	25	0,021

Джерело: розраховано автором.

Додатково перевіримо на відповідність нормального розподілу за допомогою ексцесу й асиметрії:

1) ексцес/стандартна похибка ексцесу;

2) асиметрія/стандартна похибка асиметрії.

За підприємствами розраховані значення 2,5 і 2,8, що підтверджує порушення умов нормального розподілу (> 2), а за фермерськими господарствами – 0,8 і 0,9 – результат менше 2-х, можна прийняти нормальний розподіл. Цінність цієї функції зростає, коли проводиться вибіркове дослідження і результати статистичного аналізу переносяться на всю генеральну сукупність.

7. Побудова гістограми дає можливість оцінити, яка кількість спостережень приймає певне значення або перебуває в певному інтервалі. Виробнича собівартість продукції на фермерських господарствах шести регіонів сягає до 1 млрд грн, ще шість перебуває в інтервалі від 2-х до 3-х млрд грн. Лише один регіон має витрати виробництва більше 5 млрд грн. На гістограмі присутня правостороння асиметрія (хвіст перебуває праворуч), наявний нормальний розподіл (рис. 3).

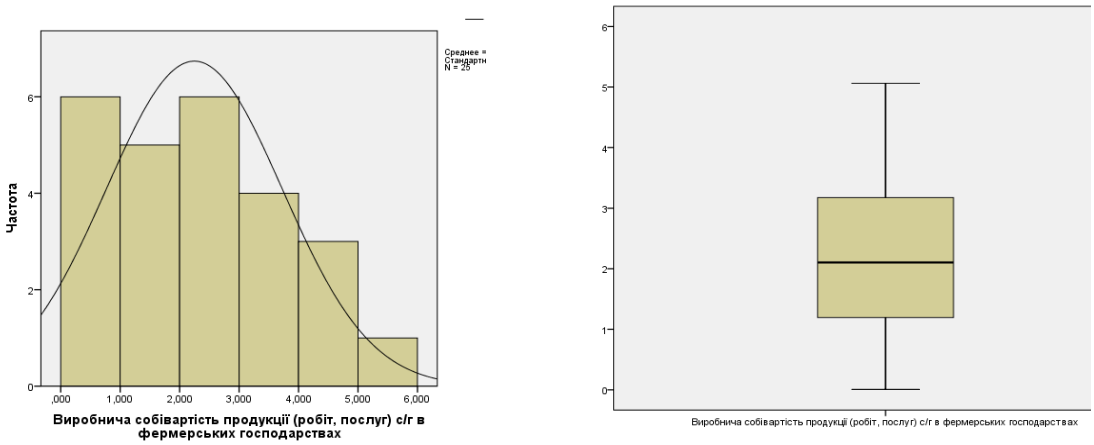


Рис. 3. Візуальне відображення собівартості продукції
Джерело: розраховано автором.

Візуальне зображення ряду спостереження за допомогою ящикової діаграми вказує на відсутність екстремальних викидів, дає можливість визначити медіану, мінімальне і максимальне значення. Нижній і верхній боки ящика відповідають першому (Q1 = 25 %) і третьому (Q3 = 75 %) квантилям (значення, що виокремлюють ¼ і ¾ вибірки). Ящикова діаграма зручна під час порівняння вибірок, які поділені на категорії, вона найбільш наглядно характеризує розподіл.

Висновки і перспективи для подальшого дослідження. Фермерське господарство в Україні лише розвивається, про що свідчать витрати с/в продукції. Середній розмір менший за витрати підприємств у 6,6 раза. Розвиток с/г в Україні нерівномірний. У подальшому необхідно проаналізувати інші економічні показники, щоб підтвердити ці припущення.

РОЗДІЛ 3
ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ

3.1. Принципи перевірки статистичних гіпотез

Одне з основних завдань статистики полягає в підтвердженні або спростуванні певного припущення (гіпотези) відносно генеральної сукупності, зробленої на основі вибіркової. Сама процедура зіставлення гіпотези з вибірковими даними називається перевіркою статистичної гіпотези. Зауважимо, що статистична гіпотеза підтверджується або спростовується але ніяк не доводиться. Гіпотеза може бути правильною або неправильною, під час її прийняття або спростування можлива помилка, тому у разі проведення статистичного тестування визначимо помилки 1 та 2 роду.

Статистична гіпотеза (гіпотеза) – це певне твердження (припущення) відносно генеральної сукупності, що перевіряється на основі вибірки. Гіпотеза називається параметричною, якщо перевіряється припущення про параметри розподілу певного виду, та непараметричною, якщо перевіряється гіпотеза про вид невідомого розподілу.

Перевірка статистичної гіпотези може бути проведена з використанням параметричних і непараметричних критеріїв.

На практиці найчастіше доводиться розв’язувати задачі двох типів за перевіркою гіпотез. Задачі першого типу пов’язані з перевіркою гіпотез про істотність відмінностей між параметрами статистичних сукупностей. Прикладом таких задач може бути оцінка вірогідності відмінностей між середніми, дисперсіями, коефіцієнтами кореляції, регресії та ін.

Задачі другого типу пов’язані з перевіркою гіпотез про істотність відмінностей законів розподілу. До них відносять задачі щодо визначення відповідності вибіркового розподілу теоретичному, частіше всього нормальному, оцінці близькості двох емпіричних розподілів, однорідності складу кількох сукупностей тощо.

Розглянемо деякі важливі передумови й особливості перевірки статистичних гіпотез задач першого типу, пов'язаних із застосуванням параметричних критеріїв і припущення нормального розподілу в генеральній сукупності. Вибір конкретних заходів з перевірки гіпотези залежить від таких умов:

1) обсяг вибіркової сукупності. Під час перевірки гіпотез за даними великих вибірок ($n > 30$) доцільно застосовувати χ^2-критерій нормального розподілу, а за даними малих вибірок ($n < 30$) – t -статистику Стьюдента;

2) рівність вибірок за чисельністю. Вибіркові сукупності за чисельністю можуть бути рівними і нерівними. Ці властивості необхідно враховувати під час практичної перевірки гіпотез про істотність відмінностей між середніми, зокрема під час розрахунку середньої помилки двох вибіркових середніх;

3) принципи формування вибірок. Прийоми перевірки статистичних гіпотез залежать від характеру формування вибіркових сукупностей. Якщо спостереження однієї вибірки не протиставляються спостереженням другої вибірки, то такі вибірки називають незалежними. Якщо ж спостереження однієї вибірки деякою мірою пов'язані зі спостереженнями другої вибірки, то такі вибірки називають залежними. Формування вибіркових сукупностей зумовлює різні прийоми оцінки вірогідності між середніми двох малих вибірок. Якщо вибірки незалежні, то статистичній оцінці підлягає різниця середніх, якщо залежні – середня різниця;

4) рівність дисперсій. Під час перевірки гіпотез щодо середніх можливі два випадки щодо вибіркових дисперсій: дисперсії рівні або нерівні. У зв'язку з цим виникає спеціальне завдання перевірки гіпотези про істотність відмінностей двох дисперсій. Для перевірки гіпотези про рівність двох дисперсій у генеральних сукупностях використовується критерій Фішера, який ґрунтується на співвідношенні двох вибіркових скоригованих дисперсій, що замінюють значення дисперсій у генеральних сукупностях. Критичні значення критерію Фішера знаходять за спеціальними таблицями за відповідною кількістю ступенів свободи і заданим рівнем значущості.

Залежно від того, рівні чи нерівні дисперсії в генеральних сукупностях, потрібно видозмінювати схему перевірки гіпотези.

Розв'язування задач другого типу пов'язано з перевіркою гіпотез відносно законів розподілу генеральних сукупностей. У практичних дослідженнях часто виникає необхідність встановити характер розподілу генеральних сукупностей. При цьому можуть виникати задачі трьох видів:

1) про узгодженість емпіричного (вибіркового) і теоретичного (генерального) розподілу;

2) про незалежність розподілу однієї ознаки від другої;

3) про однорідність двох та більше емпіричних розподілів.

Такі гіпотези, як і гіпотези відносно параметрів розподілу, перевіряють за допомогою спеціальних критеріїв згоди. Критерієм згоди називають критерій перевірки гіпотези щодо передбачуваного закону невідомого розподілу в генеральній сукупності. Ці критерії дозволяють встановити, узгоджуються чи не узгоджуються досліджувані розподіли з теоретичними розподілами, а також те, наскільки істотними є розбіжності між цими розподілами.

3.2. Основна й альтернативна гіпотези

Одна гіпотеза називається ***основною*** **(або** ***нульовою*****),** її позначають H_0, логічне заперечення основної гіпотези називається альтернативною та позначається H_1. Гіпотезу, що фіксує єдине значення параметра, називають ***простою,*** в протилежному випадку – ***складною***.

Правило, за яким приймається рішення про прийняття або відхилення гіпотези H_0 називають *статистичним критерієм перевірки* гіпотези H_0. Рішення приймають на основі вибірки $X_1, X_2, \dots X_n$, за якою формують спеціальну функцію вибірки $T_n = T(X_1, X_2, \dots X_n)$, цю функцію називають *статистикою критерію*.

Принцип перевірки гіпотез. Множина можливих значень статистики критерію T_n розділяється на дві підмножини, що не перетинаються: критичну сферу S, тобто сферу відхилення гіпотези

H_0 та сферу $\overline{S}$ прийняття цієї гіпотези. Якщо фактичне (спостережне, вирахуване на основі вибірки) значення критерію потрапляє в критичну сферу S, то основна гіпотеза H_0 відхиляється та приймається альтернативна H_1, у протилежному випадку приймається основна гіпотеза, тобто відхиляється альтернативна.

Як уже зауважувалось, під час перевірки статистичної гіпотези є певна ймовірність помилитися.

Помилка першого роду *полягає в тому, що відхиляється нульова гіпотеза, яка насправді є правильною.* Ймовірність цієї помилки позначається α та називається рівнем значущості критерію. За означенням $\alpha = P(H_1|H_0)$. Цю ймовірність задають, як правило, перед проведенням тесту, числове значення є стандартним $\alpha =$ 0,01; 0,05; 0,001; 0,005.

Помилка другого роду *полягає у відхиленні альтернативної гіпотези, що насправді правильна.* Ймовірність помилки другого роду позначається β. Відповідно до означення $\beta = P(H_0|H_1)$. Величина 1– β називається потужністю критерію. Мінімізувати ймовірність одночасно помилки 1-го та 2-го роду можливо лише за умови збільшення об'єму вибірки (табл. 3.1).

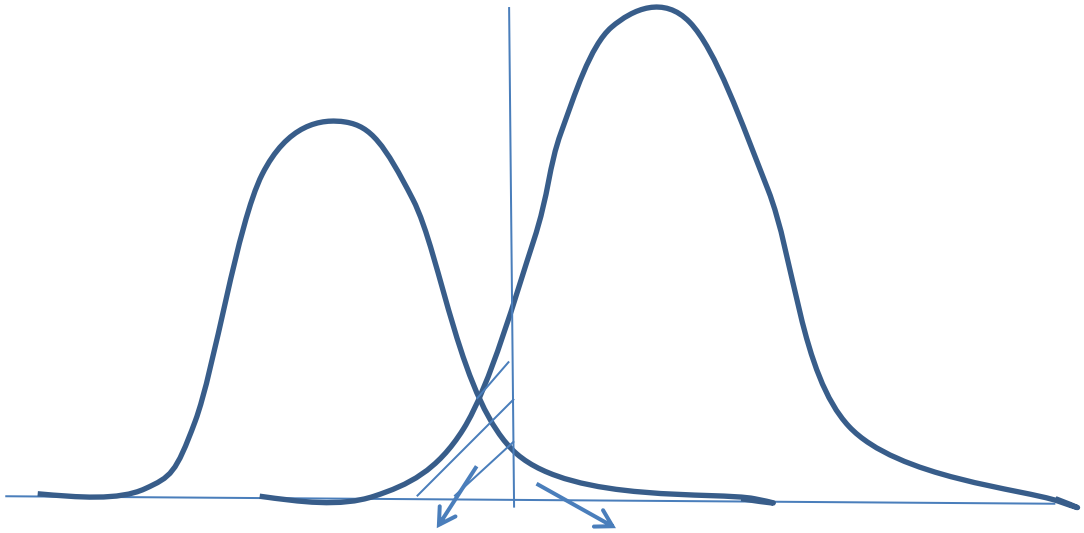
Таблиця 3.1

Гіпотеза H_0	**Відхиляється**	**Приймається**
Правильна	Помилка 1-го роду $\alpha = P(H_1\|H_0)$	Правильне рішення $1-\alpha = P(H_0\|H_0)$
Неправильна	Правильне рішення $1-\beta = P(H_1\|H_1)$	Помилка 2-го роду $\beta = P(H_0\|H_1)$

У радіолокації помилка першого роду – пропуск сигналу, другого роду – прийняття хибного сигналу; у юриспруденції α – ймовірність виправдовування винного, β – засудження невинного; у виробництві α – ризик постачальника, β – ризик споживача.

Рівень значимості гіпотези називають імовірність здійснити похибку першого роду, тобто відхилити правильно нульову гіпотезу (a = 10 % або 5 % або 1 %). Площа під нормальним розподілом **дорівнює 1.** Нам необхідно перевірити, чи це випадкове середнє або середнє іншого розподілу. Похибка характеризує, що середнє значення попало випадково, тоді відхиляється нульова гіпотеза.

Розширюється площа, знижується відсоток похибки, на рівні 0,05 відхиляємо, а за 0,01 не відхиляємо (рис. 3.1).



β – похибка II роду α – похибка I роду

1 – 0,95 = 0,05, 1 – 0,99 = 0,01.

Рис. 3.1. Візуалізація похибки першого і другого роду

Для візуалізації похибки застосуємо розподільчий калькулятор: https://gallery.shinyapps.io/dist_calc/ (рис. 3.2–3.5).

Distribution Calculator (розподільчий калькулятор)

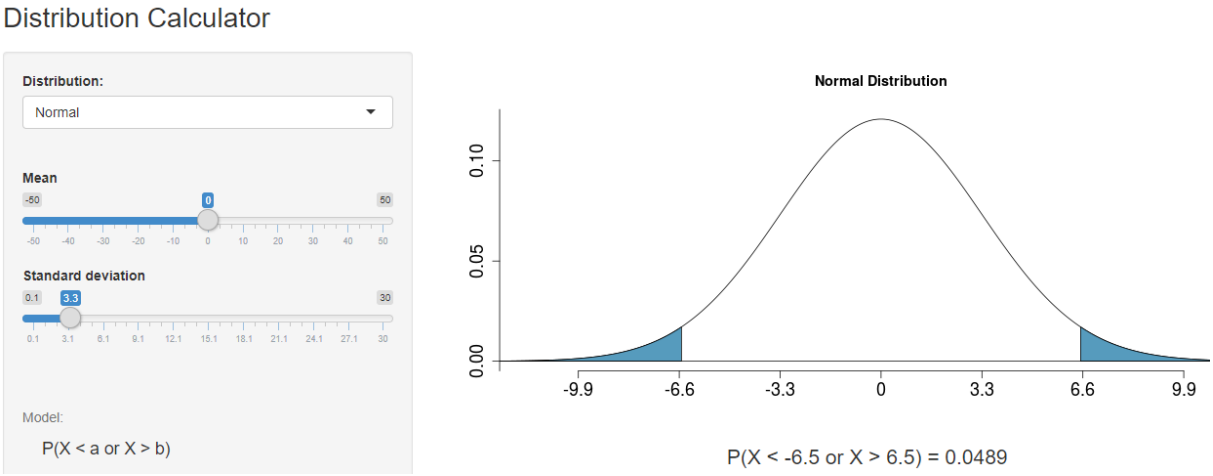


Рис. 3.2. Нормальний розподіл

standard deviation – середньоквадратичне відхилення
mean – середнє значення
Find area – знайдіть сферу:
both tails – обидва хвости lower tail – нижчий хвіст
upper tail – верхній хвіст middle tails – середні хвости

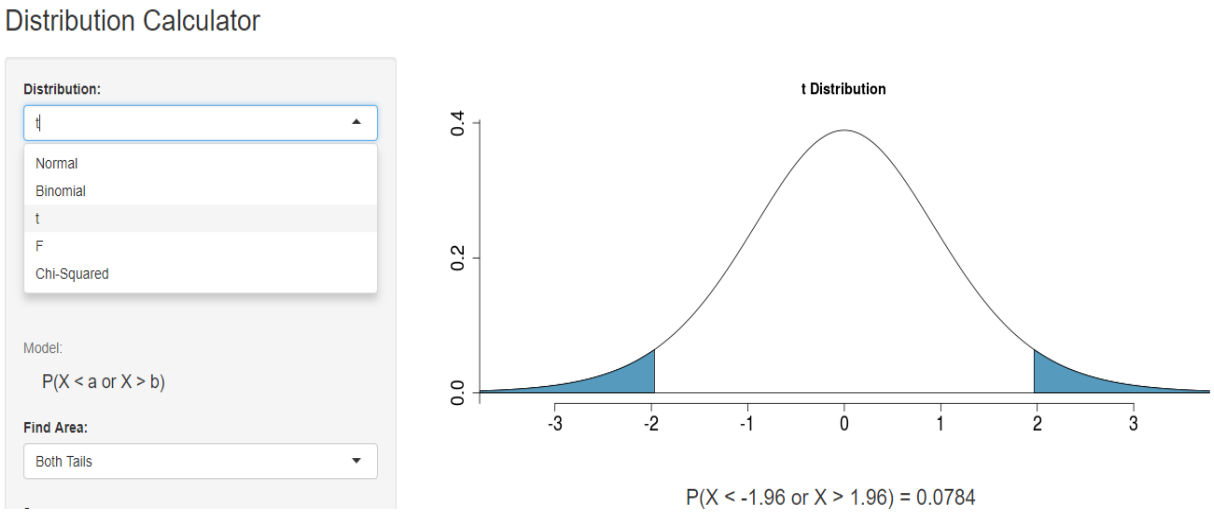


Рис. 3.3. *t*-розподіл

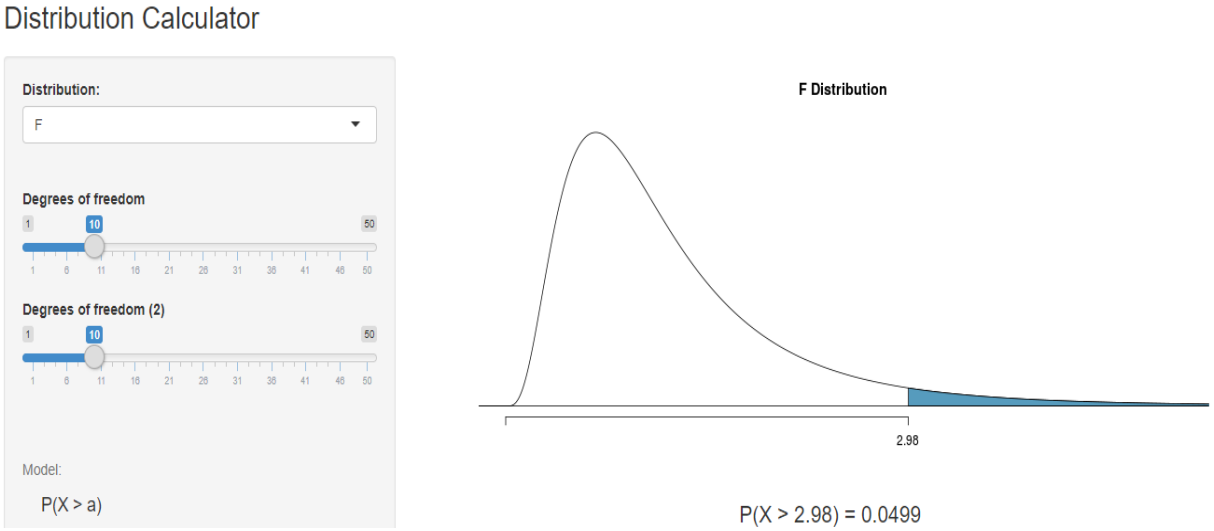


Рис. 3.4. *F*-розподіл

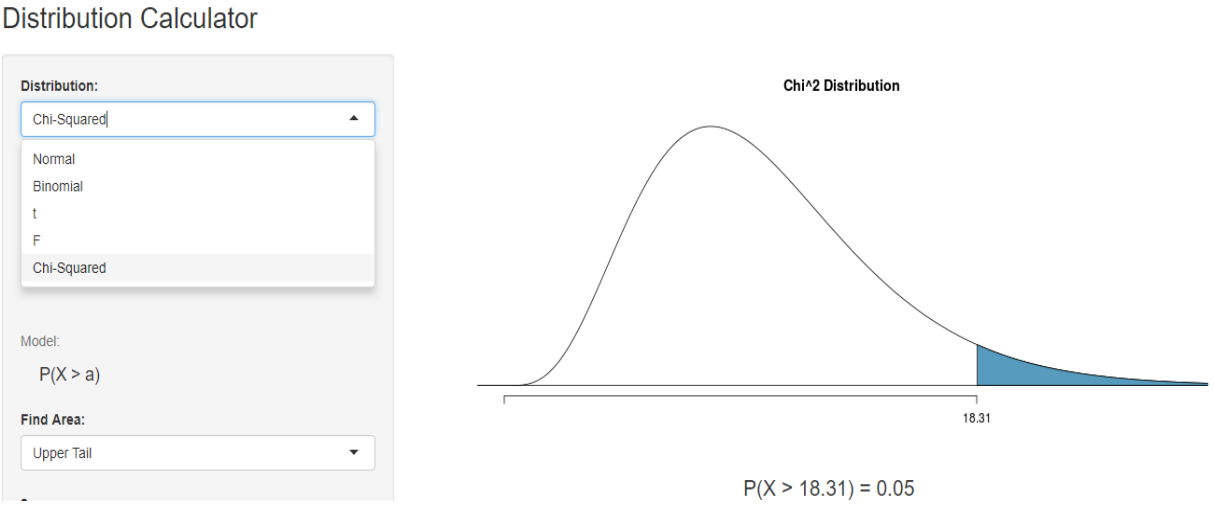


Рис.3.5. χ^2 – розподіл

3.3. Алгоритм проведення статистичного тестування

Перевірка статистичної гіпотези передбачає послідовне виконання таких етапів:

1. Оцінити інформацію та описати статистичну модель вибіркової сукупності.

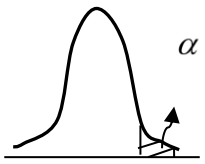
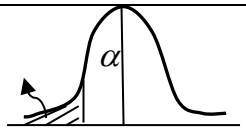
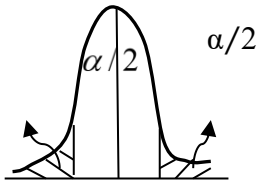
2. На основі вибірки $X_1, X_2, \dots X_n$ сформулювати нульову та альтернативну гіпотези.

3. Вибрати критерій перевірки. У кожному конкретному випадку підбирають статистику критерію $T_n = T(X_1, X_2, \dots X_n)$, зазвичай з нижче перерахованих: *U* – нормальний розподіл, розподіл *xi*-квадрат (додаток В); *t*-розподіл Стьюдента (додаток А), *F*-розподіл Фішера (додаток Г).

4. За статистикою критерію та рівнем значущості α у відповідній статистичній таблиці (у цьому підручнику ці таблиці подані в додатках) визначають критичне значення (критичну точку) та визначають критичну сферу *S*. Межі сфер визначаються, відповідно, із співвідношень для правосторонньої, лівосторонньої та двосторонньої сфер, що показано у таблиці 3.2.

Таблиця 3.2

Межі сфер статистичного критерію

Критичні сфери	Ймовірність помилки 1 роду	Візуалізація
Правостороння	$P(t_{\text{сп}} > t_{\text{кр}}) = \alpha$	
Лівостороння	$P(t_{\text{сп}} < t_{\text{кр}}) = \alpha$	
Двостороння	$P(t_{\text{сп}} < t_{\text{кр1}}) + P(t_{\text{сп}} > t_{\text{кр2}}) = \alpha$	
Симетрична двостороння	$P(t_{\text{сп}} < t_{\text{кр1}}) = P(t_{\text{сп}} > t_{\text{кр2}}) = \frac{\alpha}{2}$	

5. Для отриманої реалізації вибірки $x_1, x_2, \dots x_n$ розраховують фактичне значення критерію $T_{\text{спостер}} = T(x_1, x_2, \dots x_n) = t$.

6. Якщо розраховане значення t належить критичній сфері, то гіпотезу H_0 відхиляють; якщо ж $t \notin S$, тоді немає підстав для відхилення основної гіпотези, тобто H_0 приймають.

7. Формулювання висновків за результатами перевірки нульової гіпотези.

Необхідно зауважити, що у разі відсутності статистичних таблиць із критичним значенням можна скористатися значимістю критерію (p-значення).

Приклад. Побудувати гістограму за даними вибірки та зробити припущення про закон розподілу.

Таблиця 3.3

Вихідні дані

$x_i; x_{i+1}$	0–4	4–8	8–12	12–16	16–20	20–24
n_i	40	24	16	12	8	4
$\frac{n_i}{h}$	10	6	4	3	2	1

$h = 4 - 0 = 4$.

Сполучивши середини прямокутників гістограми, за виглядом отриманої кривої можемо зробити припущення, тобто висунути гіпотезу про існування експоненціального розподілу. Кожну гіпотезу варто перевірити, а для цього потрібні емпіричні та теоретичні частоти. Емпіричні – це частоти знайдені за даними вибірки, теоретичні частоти – обчислюємо за формулами.

У випадку *дискретного закону розподілу* теоретичні частоти шукають за формулою: $n_i' = n \cdot p_i$, де n – об'єм вибірки; p_i – ймовірність спостережуваного значення $X = x_i$, яка обчислюється за умови, що X взяте за припущеним законом.

У більшості випадків закон розподілу досліджуваної випадкової величини невідомий, але є певні підстави для припущення, що він є, наприклад, нормальним, показниковим, рівномірним, гіпергеометричним тощо.

Нехай необхідно перевірити гіпотезу H_0 про те, що випадкова величина розподілена за певним законом, що задається функцією розподілу $F_0(x)$, тобто основна гіпотеза має вигляд H_0: $F_X(x) = F_0(x)$.

Альтернативна гіпотеза суперечить основній, тобто H_1: $F_X(x) \neq F_0(x)$. Для перевірки гіпотези про розподіл випадкової величини X проводимо вибірку і представляємо її у формі таблиці 3.4 статистичного розподілу.

Таблиця 3.4

Вибірка статистичного розподілу

x_i	x_1	x_2	…	x_m
n_i	n_1	n_2	…	n_m

З об'ємом вибірки $n = \sum_{i=1}^{m} n_i$.

3.4. Види розподілів

Дискретна випадкова величина X має біноміальний закон розподілу з параметрами $n \in N$, $p \in [0;1]$, якщо вона набуває значення 0, 1, 2, ..., m, … n з імовірностями, що обчислюються за формулою Бернуллі: $P(X = m) = C_n^m p^m q^{n-m}$.

Той факт, що випадкова величина X, розподілена за біноміальним законом з параметрами $n \in N$, $p \in [0;1]$, позначається таким чином X~B (n; p) (табл. 3.5).

Таблиця 3.5

Вибірка біноміального розподілу

x_i	0	1	2	…	M	…	n
p_i	q^n	$C_n^1 p^1 q^{n-1}$	$C_n^2 p^2 q^{n-2}$	…	$C_n^m p^m q^{n-m}$	….	p^n

Біноміальний закон розподілу широко **використовується** в теорії і практиці статистичного контролю якості продукції, під час описування функціонування систем масового обслуговування та в інших сферах.

Математичне сподівання випадкової величини X, розподіленої за біноміальним законом, обчислюється за формулою M(X) = np.

Дисперсія випадкової величини X, розподіленої за біноміальним законом, обчислюється за формулою D(X) = npq.

Дискретна випадкова величина X має розподіл Пуассона з параметром $\lambda = np,\ 0 < \lambda < \infty,\ n \in N,\ p \in (0;1)$, якщо вона приймає значення 0, 1, 2, ..., *m*, … – це кількість успіхів у *n* незалежних випробуваннях Бернуллі, за умови якщо $n \to +\infty,\ p \to 0$ з ймовірностями, що шукають за формулою (табл. 3.6): $P(X = m) = \frac{\lambda^m e^{-\lambda}}{m!}$.

Таблиця 3.6

Вибірка розподілу Пуассона

x_i	0	1	2	…	m	…
p_i	$e^{-\lambda}$	$\lambda e^{-\lambda}$	$\frac{\lambda^2}{2!}e^{-\lambda}$	…	$\frac{\lambda^m e^{-\lambda}}{m!}$	….

Розподіл Пуассона за параметром λ позначається таким чином: PS(λ).

За допомогою цього закону **досліджується** кількість збоїв на автоматичній лінії, відмов складної системи, що працює в нормальному режимі, вимог на обслуговування, які поступають за одиницю часу в системах масового обслуговування тощо.

Якщо випадкова величина X має розподіл Пуассона з параметром, тоді: математичне сподівання $M(X) = \lambda$; дисперсія $D(X) = \lambda$.

Приклад. Кожен зі 150-ти знаків повідомлення, незалежно від інших знаків, може бути спотвореним під час проходження лініями зв'язку з ймовірністю 0,003. Обчислити ймовірність того, що отримане повідомлення:

1) не містить жодного спотвореного знака;

2) містить не менше ніж два спотворених знаки.

Нехай випадкова величина *X* – кількість спотворених знаків, серед 150 знаків, що містяться у повідомленні. Так, задана випадкова величина розподілена за законом Пуассона, параметр розподілу $\lambda = np = 150 \cdot 0{,}003 = 0{,}45$. Використовуючи твердження 7,5 для $k = 0{,}1$, маємо:

1) $P(X=0)=e^{-0,45}\frac{0,45^1}{1!}\approx 0,2869$; 2) $P(X\geq 2)=1-P(X<2)=1-\left(P(X=0)+P(X=1)\right)=1-\left(e^{-0,45}\cdot\frac{0,45^0}{0!}+e^{-0,45}\cdot\frac{0,45^1}{1!}\right)\approx 1-0,9246=0,0754.$

Припустимо, що існує множина, що містить N елементів, серед яких рівно M елементів володіють певною властивістю. З основної множини відбирають *n* елементів. Випадкова величина X – кількість елементів *m* з властивістю, серед відібраних має **гіпергеометричний закон розподілу,** якщо вона приймає значення 1, 2, ... min (n, M) з ймовірностями $P(X=m)=\frac{C_M^m\cdot C_{N-M}^{n-m}}{C_N^n}$.

Гіпергеометричний розподіл з параметрами *N, M, n* позначається таким чином: X~ $H(N,M,n)$ (табл. 3.7).

Таблиця 3.7

Вибірка гіпергеометричного розподілу

x_i	0	1	2	…	m
p_i	$\frac{C_M^0\cdot C_{N-M}^n}{C_N^n}$	$\frac{C_M^1\cdot C_{N-M}^{n-1}}{C_N^n}$	$\frac{C_M^2\cdot C_{N-M}^{n-2}}{C_N^n}$	…	$\frac{C_M^m\cdot C_{N-M}^{n-m}}{C_N^n}$

Гіпергеометричний закон розподілу широко **використовується** в практиці статистичного приймального контролю якості промислової продукції, в задачах, пов'язаних з організацією вибіркових досліджень, та в інших сферах.

Якщо випадкова величина X має гіпергеометричний закон розподілу з параметрами *N, M, n* , тоді:

1) математичне сподівання $M(X)=n\frac{M}{N}$;

2) дисперсія $D(X)=n\frac{M}{N-1}\left(1-\frac{M}{N}\right)\left(1-\frac{n}{N}\right)$.

Приклад. У лотереї «Спортлото 6 із 45» грошові призи отримують учасники, які вигравали 3, 4, 5 і 6 видів спорту із відібраних випадково шести видів із 45. Розмір призу збільшується зі збільшенням кількості вгаданих видів спорту. Знайти закон розподілу

випадкової величини X – число вгаданих видів спорту серед випадково відібраних шести. Яка ймовірність отримати грошовий приз? Знайти математичне сподівання та дисперсію цієї випадкової величини.

Розв'язання

Таблиця 3.8

Вихідні дані

x_i	0	1	2	3	4	5	6
p_i	0,40056	0,42413	0,15147	0,02244	0,00137	0,00003	0,0000001

$$M(X) = 6 \cdot \frac{6}{45} = 0{,}8;\ D(X) = 6 \cdot \frac{6}{45-1} \cdot \left(1 - \frac{6}{45}\right) \cdot \left(1 - \frac{6}{45}\right) = 0{,}6145;$$

$$P(3 \le X \le 6) = 0{,}0244 + 0{,}00137 + 0{,}00003 + 0{,}0000001 = 0{,}02383 \approx 0{,}24.$$

Усі твердження зведемо в таблицю 3.9.

Таблиця 3.9

Характеристика видів розподілів

Розподіл	**Параметри**	**Ймовірність $P\{X = k\}$**	**Математичне сподівання**	**Дисперсія**
Бернуллі	$p \in [0,1]$	p для k=1; 1-p для k=0	p	$p(1-p) = pq$
Біноміальний	$n \in N, p \in [0;1$	$P_n(k) = C_n^k p^k q^{n-k}$; $k = 0,1,2,\dots n$	np	npq
Геометричний	$p \in (0,1)$	$\mathrm{P}(X = k) = pq^{k-1}$; $k = 1,2,3,\dots$	$1/p$	$\frac{q}{p^2}$
Гіпергеомет-ричний	N,M,n	$P(X = k) = \frac{C_M^k C_{N-M}^{n-k}}{C_N^n}$ k = 0, 1, 2, 3, … min(n,M)	$\frac{M \cdot n}{N}$	$n\frac{M}{N-1}\left(1-\frac{M}{N}\right)\left(1-\frac{n}{N}\right.$
Пуассона	$\lambda \in (0,\infty)$	$P(X = k) = \frac{\lambda^k}{k!}\ell^{-\lambda}$ $k = 1,2,3,\dots n$	λ	λ

Випадкова величина X розподілена рівномірно на інтервалі (a, b), якщо щільність розподілу є сталою для всіх можливих значень X, тобто:

$$f(x) = \begin{cases} \frac{1}{\mathrm{b}-\mathrm{a}}, \mathrm{x} \in (\mathrm{a}, \mathrm{b}), \\ 0, \mathrm{x} \notin (\mathrm{a}, \mathrm{b}). \end{cases}$$

Той факт, що випадкова величина X розподілена рівномірно, на інтервалі (a, b) позначається таким чином: $X \sim U[a,b]$.

Якщо випадкова величина X рівномірно розподілена на інтервалі (a, b), то:

1) функція розподілу ймовірностей випадкової величини X має вигляд:

$$F_X(x) = \begin{cases} 0, \text{якщо } x \le a, \\ \frac{x-a}{b-a}, \text{якщо } a < x \le b, \\ 1, \text{якщо } x > b; \end{cases}$$

2) математичне сподівання $M(X) = \frac{a+b}{2}$;

3) дисперсія $D(X) = \frac{(\mathrm{b-a})^2}{12}$;

4) початковий момент $\mu_r = \frac{b^{r+1} - a^{r+1}}{(b-a)\cdot(r+1)}$;

5) ймовірність влучення X в інтервал $(\alpha;\beta)$ дорівнює $P(\alpha \le X \le \beta) = \frac{\beta - \alpha}{b-a}$ за умови, що кінці інтервалу належать відрізку, на якому рівномірно розподілено випадкову величину.

Приклад. Електропотяги метрополітену йдуть регулярно з інтервалом 2 хв. Пасажир виходить на платформу у випадковий момент часу. Яка ймовірність того, що чекати пасажирові прийдеться не більше півхвилини. Знайти інтегральну та диференціальну функції розподілу ймовірностей, що описують рух електропотягів метрополітену, та побудувати їх графіки. Знайти математичне сподівання та середньо квадратичне відхилення випадкової величини X – часу очікування електропотягу.

Випадкова величина X – час очікування електропотягу на часовому відрізку [0; 2] має рівномірний закон розподілу з параметрами: $a = 0$; $b = 2$. Тому функції розподілу ймовірностей та щільності матимуть вигляд:

$$F(x)=\begin{cases}0, & x\le 0\\ \frac{x}{2}, 0<x\le 2\\ 1, x>2\end{cases};\quad f(x)=\begin{cases}0, & x\le 0\\ \frac{1}{2}, & 0<x\le 2\\ 1, & x>2\end{cases};$$

Числові характеристики дорівнюють:

$$M(X)=\frac{0+2}{2}=1;\ D(X)=\frac{(2-0)^2}{12}=\frac{1}{3};\ \sigma(X)=\sqrt{\frac{1}{3}}\approx 0{,}58.$$

Ймовірність того, що пасажиру прийдеться чекати не більше пів хвилини, дорівнює: $P(0\le X\le 0{,}5)=\int_0^{0,5}0{,}5dx=0{,}5x\Big|_0^{0,5}=0{,}25-0=0{,}25$ (або використовуючи геометричне означення ймовірності, матимемо відношення площі квадрата зі стороною 0,5 до площі прямокутника зі сторонами 2 і 0,5).

Випадкова величина X розподілена експоненційно з параметром λ (або за показниковим законом), якщо щільність розподілу ймовірностей задається формулою:

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

Той факт, що випадкова величина X розподілена експоненційно з параметром λ позначається таким чином $X\sim Ex(\lambda)$.

Якщо випадкова величина X розподілена за показниковим законом з параметром λ , то:

1) функція розподілу ймовірностей випадкової величини X має вигляд $F_\lambda(x)=\begin{cases}0, & x\le 0\\ 1-e^{-\lambda x}, & x>0\end{cases}$;

2) математичне сподівання $M(X)=\frac{1}{\lambda}$;

3) дисперсія $D(X)=\frac{1}{\lambda^2}$;

4) ймовірність влучення X в інтервал $(\alpha;\beta)$ дорівнює $P(\alpha\le X\le\beta)=\ell^{-\alpha\lambda}-\ell^{-\beta\lambda}$.

Показниковий розподіл відіграє досить важливу роль у теорії систем масового обслуговування (СМО) та теорії надійності. У теорії СМО λ – середня кількість подій, що в середньому можуть

відбутися за одиницю часу, відповідно, довжина часового проміжку між двома сусідніми подіями розподілена за показниковим законом.

Приклад. Встановлено, що термін ремонту телевізора є випадкова величина *X,* що розподілена за показниковим законом розподілу. Визначити ймовірність того, що на ремонт телевізора необхідно не менше 20 днів, якщо відомо, що середній термін ремонту телевізорів становить 15 діб. Знайти щільність ймовірностей, інтегральну функцію розподілу ймовірностей і середньоквадратичне відхилення випадкової величини *X.*

За умовою відомо, що випадкова величина *X* розподілена за показниковим законом розподілу і має математичне сподівання, що дорівнює 15, тому: $M(X)=\frac{1}{\lambda}=15$; $\lambda=\frac{1}{15}$; $\sigma(X)=M(X)=15$ (днів).

Функції щільності і розподілу ймовірностей матимуть вигляд:

$$f(x)=\begin{cases}0, & x<0\\ \frac{1}{15}e^{-\frac{1}{15}x}, & x\geq 0;\end{cases} \quad F(x)=\begin{cases}0, & x<0\\ 1-e^{-\frac{1}{15}x}, & x\geq 0.\end{cases}$$

Ймовірність того, що на ремонт телевізора необхідно не менше 20 днів знайдемо двома способами:

$$P(X\geq 20)=P(20\leq X<\infty)=\int_{20}^{\infty}\frac{1}{15}e^{-\frac{1}{15}x}dx$$

або

$$P(X\geq 20)=1-P(X<20)=1-F(20)=1-(1-e^{-\frac{20}{15}})=0{,}264.$$

Випадкова величина X розподілена за нормальним законом з параметрами a,σ, якщо щільність розподілу задається формулою:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Той факт, що випадкова величина *X* розподілена нормально з параметрами a,σ, позначається таким чином: $X\sim N(a,\ \sigma^2)$.

Якщо випадкова величина X розподілена нормально з параметрами $a,\ \sigma^2$, то

1) функція розподілу ймовірностей випадкової величини X має вигляд:

$$F_x(x) = \Phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{-\frac{(t-a)^2}{2\sigma^2}}\, dt;$$

2) математичне сподівання, мода та медіана $M(X) = Mo(X) = Me(X) = a$;

3) дисперсія: $D(X) = \sigma^2$;

4) асиметрія: $As = 0$;

5) ексцес: $Es = 0$;

6) ймовірність влучення X в інтервал $(\alpha;\beta)$ дорівнює

$$P(\alpha \le X \le \beta) = \Phi\left[\frac{\beta - a}{\sigma}\right] - \Phi\left[\frac{\alpha - a}{\sigma}\right].$$

Нормальний розподіл **називають стандартним (або нормованим)**, якщо математичне сподівання $\mathrm{M(X)} = 0$, дисперсія $\mathrm{D(X)} = 1$.

Нормальний розподіл **називають загальним**, якщо математичне сподівання M(X) і середнє квадратичне відхилення довільні.

За умови $\mathrm{X} \sim \mathrm{N}(a,\ \sigma^2)$, ймовірність того, що X:

– прийме значення менші (або не більші) за α обчислюється за формулою:

$$\mathrm{P(X<\alpha)=P(X\le\alpha)=\Phi(\alpha)=0{,}5+\Phi_0\left(\frac{\alpha\text{-}a}{\sigma}\right)};$$

– прийме значення більші (або не менші) за α обчислюється за формулою:

$$\mathrm{P(X>\alpha)=P(X\ge\alpha)=1\text{-}\Phi(\alpha)=0{,}5\text{-}\Phi_0\left(\frac{\alpha\text{-}a}{\sigma}\right)};$$

– прийме значення з інтервалу (α; β) обчислюється за формулою:

$$\mathrm{P(\alpha<X<\beta)=\Phi(\beta)\text{-}\Phi(\alpha)=\Phi_0\left(\frac{\beta\text{-}a}{\sigma}\right)\text{-}\Phi_0\left(\frac{\alpha\text{-}a}{\sigma}\right)};$$

– ймовірність того, що X відхилиться від математичного сподівання $\mathrm{M(X)} = a$ на величину не більшу ніж $\delta > 0$, обчислюється за формулою $\mathrm{P(|X\text{-}a|<\delta)=2\Phi_0\left(\frac{\delta}{\sigma}\right)}$.

Приклад. Вважається, що зріст чоловіка певної вікової групи є нормально розподілена випадкова величина X з параметрами $\sigma^2(X)=36; M(X)=a=173$. Знайти: щільність ймовірностей та інтегральну функцію розподілу ймовірностей; долю костюмів 4-го росту (176–182 см) і 3-го росту (170–176 см), які потрібно передбачити в загальному об'ємі виробництва для цієї вікової групи.

Функції щільності і розподілу ймовірностей матимуть вигляд:

$$f(x)=\frac{1}{6\sqrt{2\pi}}e^{-\frac{(x-173)^2}{2\cdot 36}};$$

$$F(x)=\frac{1}{6\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{(x-173)^2}{2\cdot 36}}dx=\frac{1}{2}+\frac{1}{2}\Phi\left(\frac{x-173}{6}\right).$$

Доля костюмів 4-го росту (176-182см), які потрібно передбачити в загальному обсязі виробництва для цієї вікової групи, дорівнює:

$$P(176\le X\le 182)=\Phi\left(\frac{182-173}{6}\right)-\Phi\left(\frac{176-173}{6}\right)=\Phi(1,5)-\Phi(0,5)=$$

0,4332 – 0,1915 = 0,2417.

Доля костюмів 3-го росту (170–176 см), які потрібно передбачити в загальному обсязі виробництва для цієї вікової групи, дорівнює:

$$P(170\le X\le 176)=\Phi\left(\frac{176-173}{6}\right)-\Phi\left(\frac{170-173}{6}\right)=\Phi(0,5)+\Phi(0,5)=$$

2·0,1915=0,3830.

Функція Лапласа $\Phi_0(t)=\frac{1}{\sqrt{2\pi}}\cdot\int_0^x e^{-\frac{t^2}{2}}dt$ табульована, відповідні значення функції представлені в додатку.

Приклад. Обчислити ймовірність того, що нормально розподілена величина X відхилиться від математичного сподівання $M(X)=a$ на величину $\sigma, 2\sigma, 3\sigma$.

Використовуючи останню формулу твердження, проведемо необхідні розрахунки:

$$P(|X-a|<\sigma)=2\Phi_0\left(\frac{\sigma}{\sigma}\right)=2\Phi_0(1)=2\cdot 0,3413=0,6826;$$

$$P(|X-a|<2\sigma)=2\Phi_0\left(\frac{2\sigma}{\sigma}\right)=2\Phi_0(2)=2\cdot 0,4772=0,9544;$$

$P(|X-a|<3\sigma)=2\Phi_0\left(\frac{3\sigma}{\sigma}\right)=2\Phi_0(3)=2\cdot 0{,}4965=0{,}9973.$

З останнього прикладу випливає **правило трьох сигм** ($\mathbf{3\sigma}$): якщо випадкова величина X має нормальний розподіл, то відхилення цієї випадкової величини від її математичного сподівання за абсолютною величиною не перевищує потроєне середнє квадратичне відхилення.

Якщо $X \sim N(a, \sigma^2)$, то стандартизація випадкової величини відбувається за формулою: $Z = \frac{X-a}{\sigma} \sim N(0,1)$.

Нормальний розподіл має важливе практичне значення, це підтверджують такі факти:

✓ важливі закони розподілу дискретних випадкових величин при $n \to \infty$ апроксимуються нормальним (наприклад, Біноміальний, Пуассона);

✓ важливі статистичні критерії і оцінки, що використовуються в математичній статистиці, базуються на законах розподілу випадкових величин, що являють собою суму нормально розподілених випадкових величин (наприклад, розподіл Стьюдента, Фішера, хі-квадрат розподіл);

✓ нормально розподілені випадкові величини використовуються у разі доведення Закону великих чисел та центральної граничної теореми;

✓ значна кількість випадкових величин, з якими стикаємося в житті, «розподілені нормально», тобто підкоряються нормальному закону розподілу.

Неперервна випадкова величина X має гамма-розподіл, якщо щільність розподілу задається формулою:

$$f(x)=\begin{cases} 0, & x<0, \\ \frac{\lambda^{\alpha}}{\Gamma(\alpha)}\cdot x^{\alpha-1}\cdot e^{-\lambda x}, & x\geq 0. \end{cases}$$

Той факт, що випадкова величина X має гамма-розподіл з параметрами α, λ позначається таким чином: $X \sim G(\alpha, \lambda)$.

Тут $\Gamma(x)=\int_0^{+\infty} t^{x-1}e^{-t}dt$ – гамма-функція Ейлера, якщо $x > 0$, для цілих значень аргументу $\Gamma(n+1) = n!$

Якщо випадкова величина X має гамма-розподіл з параметрами α, λ, то

1) функція розподілу ймовірностей задається формулою:

$$F(x) = \begin{cases} 0, x < 0, \\ \dfrac{\lambda^{\alpha}}{\Gamma(\alpha)} \cdot \int\limits_{0}^{x} x^{\alpha-1} \cdot e^{-\lambda x} dx \, , x \geq 0; \end{cases}$$

2) математичне сподівання $M(X) = \frac{\alpha}{\lambda}$;

3) дисперсія $D(X) = \frac{\alpha}{\lambda^2}$.

Якщо неперервна випадкова величина X має гамма-розподіл і параметр α(k) набуває лише додатних цілих значень (параметр позначають k), то гамма розподіл перетворюється в розподіл Ерланга k-того порядку, функція, щільності набуває вигляду:

$$f(x) = \begin{cases} 0, x<0 \\ \dfrac{\lambda}{(k-1)!} \cdot (\lambda x)^{k-1} \cdot e^{-\lambda x}, x \geq 0, k=1,2,\dots . \end{cases}$$

Той факт, що випадкова величина X, розподілена за законом Ерланга з параметрами $\alpha, \lambda,$ позначається таким чином: $X \sim ERL(k, \lambda)$.

Якщо випадкова величина X має розподіл Ерланга k-го порядку, то:

1) функція розподілу ймовірностей задається формулою:

$$F(x) = \begin{cases} 0, x<0, \\ \dfrac{\lambda}{(k-1)!} \cdot \int\limits_{0}^{x} (\lambda x)^{k-1} \cdot e^{-\lambda x} dx \, , x \geq 0; \end{cases}$$

2) математичне сподівання $M(X) = \frac{k}{\lambda}$;

3) дисперсія $D(X) = \frac{k}{\lambda^2}$.

Неперервна випадкова X має логарифмічно-нормальний (логнормальний) розподіл, якщо її логарифм задовольняє нормальний закон:

$$F(x) = P(X \prec x) = P(\ln X \prec \ln x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \int\limits_{-\infty}^{\ln x} e^{-\frac{(t-\ln a)^2}{2\sigma^2}} dt;$$

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}x}e^{-\left(\frac{\ln x-\ln a}{2\sigma^2}\right)^2};$$

$$M(x)=ae^{\frac{\sigma^2}{2}};D(x)=a^2e^{\sigma^2}(e^{\sigma^2}-1);M_0(x)=ae^{-\sigma^2};M_e(x)=a.$$

Нехай $X_1, X_2, \ldots, X_n$ – випадкові величини, незалежні, однаково розподілені за нормованим нормальним законом, тоді сума квадратів цих випадкових величин $\chi^2_n = \sum_{i=1}^{n} X_i^2$ має розподіл x^2 з $v = n$ ступенями свободи.

Якщо випадкова величина X має розподіл χ^2 з n ступенями свободи, то:

1) щільність розподілу ймовірностей задається формулою:

$$f(x)=\begin{cases}0, & x\le 0,\\ \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}\cdot x^{\frac{n}{2}-1}\cdot e^{-\frac{x}{2}}, & x>0;\end{cases}$$

2) математичне сподівання $M(X) = n$;

3) дисперсія $D(X) = 2n$.

Нехай Y – випадкова величина, що має нормований нормальний розподіл, випадкова величина X задається **функцією щільності розподілу**:

$$f(x)=\begin{cases}0, & x\le 0,\\ \frac{n^{\frac{n}{2}}}{2^{\frac{n}{2}-1}\Gamma\left(\frac{n}{2}\right)}\cdot x^{n-1}\cdot e^{-\frac{nx^2}{2}}, & x>0;\end{cases}$$

тоді випадкова величина $Z=\frac{Y}{X}$ має розподіл Стьюдента з n ступенями свободи.

Якщо випадкова величина Z має розподіл Стьюдента n ступенями свободи, то

1) щільність розподілу ймовірностей задається формулою:

$$f(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\cdot\left(1+\frac{z^2}{n}\right)^{\frac{n+1}{2}},\ -\infty<z<+\infty$$

2) математичне сподівання, мода та медіана M(Z) = Mo(Z) = Me(Z) = 0;

3) дисперсія $D(Z)=\frac{n}{n-2}$.

Нехай X та Y – випадкові величини, що мають закони розподілу:

$$f(\mathrm{x})=\begin{cases} 0, \mathrm{x}\le 0 \\ \dfrac{(\mathrm{n}_1)^{\frac{\mathrm{n}_1}{2}}}{2^{\frac{\mathrm{n}_1}{2}}\Gamma\left(\frac{\mathrm{n}_1}{2}\right)}\cdot \mathrm{x}^{\frac{\mathrm{n}_1}{2}-1}\cdot \mathrm{e}^{-\frac{\mathrm{n}_1\mathrm{x}^2}{2}}, \mathrm{x}>0, \end{cases}$$

$$f(y)=\begin{cases} 0, \mathrm{y}\le 0 \\ \dfrac{(\mathrm{n}_2)^{\frac{\mathrm{n}_2}{2}}}{2^{\frac{\mathrm{n}_2}{2}}\Gamma\left(\frac{\mathrm{n}_2}{2}\right)}\cdot \mathrm{x}^{\frac{\mathrm{n}_2}{2}-1}\cdot \mathrm{e}^{-\frac{\mathrm{n}_2\mathrm{y}^2}{2}}, \mathrm{y}>0, \end{cases}$$

тоді випадкова величина $Z=\frac{Y}{X}$ має розподіл Фішера-Снедекора з кількістю ступенів свободи, що дорівнює n_1, n_2.

Якщо випадкова величина Z має розподіл Стьюдента з n ступенями свободи, то

1) математичне сподівання $M(Z)=\frac{n_1}{n_2-2}$;

2) дисперсія $D(Z)=\frac{2n_1^2(n_2+n_1-2)}{n_2(n_1-2)^2(n_1-4)}$.

Узагальнені характеристики розподілів наведені в таблиці 3.10.

Таблиця 3.10

Характеристика видів розподілів

Розподіл	**Параметри**	**Щільність розподілу ймовірностей**	**M(X)**	**D(X)**
Рівномірний на (a,b)	$a<b$; $a,b\in R$	$f(x)=\begin{cases}\frac{1}{b-a}, & x\in(a,b)\\ 0, & x\notin(a,b)\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Показниковий розподіл	$\lambda\in(0,\infty)$	$f(x)=\begin{cases}\lambda e^{-\lambda x}, & x\ge 0,\\ 0, & x<0\end{cases}$	1/λ	1/λ
Нормований (стандартний) нормальний розподіл	–	$f(x)=\frac{1}{\sqrt{2\pi}}e^{\frac{x^2}{2}}$	0	1
Нормальний розподіл	$a\in R$, $\sigma^2>0$	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(x-a)^2}{2\sigma^2}}$	a	σ^2
Гамма розподіл	$\alpha>0$, $\lambda>0$	$f(x)=\begin{cases}0, x<0\\ \frac{\lambda^\alpha}{\Gamma(\alpha)}\cdot x^{\alpha-1}\cdot e^{-\lambda x}, x\ge 0.\end{cases}$	$\frac{\alpha}{\lambda}$	α/λ^2

Розподіл Ерланга	$k = 1, 2, \ldots, \lambda > 0$	$f(x) = \begin{cases} 0, x < 0 \\ \frac{\lambda}{(k-1)!} \cdot (\lambda x)^{k-1} \cdot e^{-\lambda x}, x \geq 0, k = 1,2, \ldots . \end{cases}$	$\frac{k}{\lambda}$	k/λ^2
χ^2 – розподіл	$n \in N$	$f(x) = \begin{cases} 0, x \leq 0 \\ \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \cdot x^{\frac{n}{2}-1} \cdot e^{-\frac{x}{2}}, x > 0. \end{cases}$	n	$2n$
Розподіл Стьюдента (t-розподіл)	$n \in N$	$f(z) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \cdot \left(1 + \frac{z^2}{n}\right)^{-\frac{n+1}{2}}, -\infty < z < +\infty$	0	$\frac{n}{n-2}$
Розподіл Фішера-Снедекора (F-розподіл)	$k \in N$, $k \in N$	$f_F(x, k_1, k_2) = C_{k_1 k_2} \cdot x^{\frac{k_1}{k_2}-1} \left(1 + \frac{k_1}{k_2} x\right)^{\frac{k_1+k_2}{2}}$, $x>0$	$\frac{n}{n-2}$	$k_1 = n-1$ $k_2 = m-1$

3.5. Перевірка умови нормальності розподілу

У ході дослідження необхідно зробити висновок про те, чи узгоджуються результати спостережень із зробленим припущенням. Для цього використовується спеціально підібрана величина – критерій згоди.

Критерієм згоди називають статистичний критерій перевірки гіпотези про можливий закон невідомого розподілу. У математичній статистиці є багато критеріїв згоди, що названі, як правило, на честь учених, які ці критерії сформулювали та довели: Пірсона, Колмогорова, Фішера, Смирнова й інші. Найбільш поширеними і вживаними є критерій Пірсона та критерій Колмогорова.

Критерій x^2 Пірсона

Для перевірки основної гіпотези про закон розподілу випадкової величини необхідно всю сферу значень випадкової величини X розбити на m інтервалів $x_0 - x_1;\ x_1 - x_2; \ldots; x_{m-1} - x_m$ та підрахувати ймовірності p_i; $i=\overline{1,m}$ попадання випадкової величини X у відповідний інтервал $x_{i-1} - x_i$. З цією метою використовують

формулу $P\{\alpha \le X \le \beta\} = F^*(\beta) - F^*(\alpha)$. Тоді теоретична кількість значень випадкової величини *X*, що має потрапити у *i*-тий інтервал, може бути розрахована за формулою $n \cdot p_i$. Отримуємо у результаті статистичний розподіл випадкової величини X (табл. 3.11) та теоретичний ряд розподілу.

Таблиця 3.11

Δ_i	Δ_1	Δ_2	…	Δ_m
n_i	n_1	n_2		n_m
n'_i	$n'_1 = np_1$	$n'_2 = np_2$	…	$n'_m = np_m$

Якщо емпіричні частоти (n_i) суттєво відрізняються від теоретичних ($n'_i = np_i$), то основну гіпотезу H_0 необхідно відхилити, в протилежному випадку – прийняти. Критерієм, що характеризує ступінь різниці між теоретичними та емпіричними частотами, є величина, запропонована К. Пірсоном, розраховується вона за формулою:

$$\chi^2_{cn} = \sum_{i=1}^{m} \frac{(n_i - n_i')^2}{n_i'}.$$

Відповідно, теоремі Пірсона за умови $n \to \infty$ статистика має розподіл χ^2 з $k = m - r - 1$ ступенями свободи, тут m – кількість (інтервалів) вибірки, r – кількість параметрів можливого розподілу.

Алгоритм використання критерію згоди Пірсона

1. Обчислити основні числові характеристики вибірки.

2. Обчислити теоретичні частоти для варіант вибірки за формулою: $n_i' = \frac{hn}{\overline{\sigma}}\varphi(t_i)$, де *n* – об'єм вибірки; *h* – крок (різниця між сусідніми варіантами, які є рівновіддаленими); $\overline{\sigma}$ – вибіркове середньоквадратичне відхилення;

$\varphi(t_i)$ – функція Лапласа з аргументом: $t_i = \frac{x_i - \overline{x}}{\overline{\sigma}}$.

3. Вирахувати $\chi^2_{\text{спост}}$ – вибіркове значення статистики критерію: $\chi^2_{сп} = \sum_{i=1}^{m} \frac{(n_i - n_i')^2}{n_i'}$.

4. Знайти ступінь вільності за формулою $k = m - r - 1$, де m – кількість інтервалів, r – кількість параметрів розподілу.

5. Знайти за таблицею критичну точку $\chi^2_{\alpha,k}$, яка відповідає заданому рівню значущості α відповідно до кількості ступенів свободи за таблицею χ^2–розподілу.

6. Порівняти $\chi^2_{\text{спост}}$ – вибіркове значення статистики та критичну точку $\chi^2_{\alpha,k}$. Якщо $\chi^2_{\text{спост}} \leq \chi^2_{\alpha,k}$, то гіпотеза H_0 не протирічить емпіричним даним і її приймають; якщо $\chi^2_{\text{спост}} \geq \chi^2_{\alpha,k}$, то основна гіпотеза відхиляється.

Важливим для використання критерію Пірсона є виконання необхідної умови: в кожному з інтервалів має бути не менше 5 спостережень (тобто $n_i \geq 5$). Якщо в окремих інтервалах кількість спостережень менше 5, необхідно провести об'єднання відповідних інтервалів, тобто провести укрупнення.

Критерій Колмогорова

У випадку перевірки простої гіпотези критерій Колмогорова є найпростішим критерієм перевірки гіпотези про закон розподілу випадкової величини. Цей критерій пов'язує емпіричну функцію розподілу $F_n^*(x)$ з функцією розподілу $F(x)$ неперервної випадкової величини X.

Нехай маємо конкретну вибірку $x_1;\ x_2; \dots x_n$ з невідомою функцією розподілу $F(x)$ та емпіричною функцією розподілу $F_n^*(x)$. Висувається пара гіпотез: основна H_0: $F(x) = F_0(x)$ та альтернативна H_A: $F(x) \neq F_0(x)$.

Сутність критерію проста: дослідити спеціальну функцію, яка являє собою максимум відхилення емпіричної функції розподілу від теоретичної. Колмогоров довів, що для функції $D_n = \max\limits_{-\infty<x<\infty} |F_n^*(x) - F_0(x)|$ має місце границя:

$$P\{\sqrt{n}D_n < x\} \to K(x), n \to \infty.$$

Іншими словами, випадкова величина $\sqrt{n}D_n$ має розподіл, що незалежно від розподілу випадкової величини X наближається до

розподілу Колмогорова. Тут $K(x)$ – функція розподілу Колмогорова, для якої складена таблиця 3.12, яку можна використовувати для розрахунків за умови, що кількість спостережень $n \geq 20$.

Таблиця 3.12

Вихідні дані

α	0,2	0,1	0,05	0,02	0,01	0,001
x_0	1,073	1,224	1,358	1,520	1,627	1,950

Алгоритм використання критерію згоди Колмогорова

1. Будується емпірична і теоретична функції розподілу.
2. $\lambda \leq \lambda_\alpha$ Визначається міра розбіжності між теоретичним і емпіричним розподілом за формулою $D = \max\left|F^*(x) - F(x)\right|$ та обчислюється $\lambda = D\sqrt{n}$.
3. Маючи рівень значущості a з відношення $P(\lambda_\alpha) = \alpha$ та використовуючи відповідні таблиці, знаходимо відповідне критичне значення λ_a.
4. Якщо знайдене значення λ буде більше критичного λ_a, визначеного на рівні значущості α, то нульова гіпотеза про те, що випадкова величина X має заданий закон розподілу, відкидається. Якщо, то вважають, що гіпотеза не протирічить дослідним даним.

Критерій Колмогорова досить часто використовується на практиці завдяки своїй простоті, але його використання можливе лише тоді, коли теоретична функція розподілу задана повністю. На жаль, це буває досить нечасто.

3.6. Перевірка статистичних гіпотез у SPSS

Для перевірки нормальності розподілу в SPSS застосовують:

Критерій Колмогорова-Смірнова дозволяє знайти точку, в якій сума накопичених розбіжностей між двома розподілами є найбільшою, і оцінити достовірність цієї розбіжності.

Критерій Шапіро-Уілка – особливістю застосування цього критерію є те, що він дозволяє перевірити нормальність розподілу

малих вибірок, при цьому отримані оцінки критерію нормальності більш точніші, що позитивно вирізняє його серед інших альтернативних критеріїв.

Розрахуємо ці критерії на прикладі.

Приклад. Визначити кандидатуру за віковою ознакою, яка може претендувати на керівну посаду в університеті. Вікова шкала дійсних завідувачів кафедр, декана і його заступників одного із факультетів наведена: 54, 32, 43, 39, 45, 42, 42, 45, 74, 61.

Застосуємо можливості SPSS для оцінки розподілу: «**Анализ» – «Описательные статистики» – «Разведочный анализ».**

Обирається: вкладка «Графики» (рис. 3.5), спосіб візуалізації «Гистограмма». Ставиться відмітка «Графики и критерии для проверки нормальности», яка містить функції критеріїв Колмогорова-Смірнова і Шапіро-Уілка.

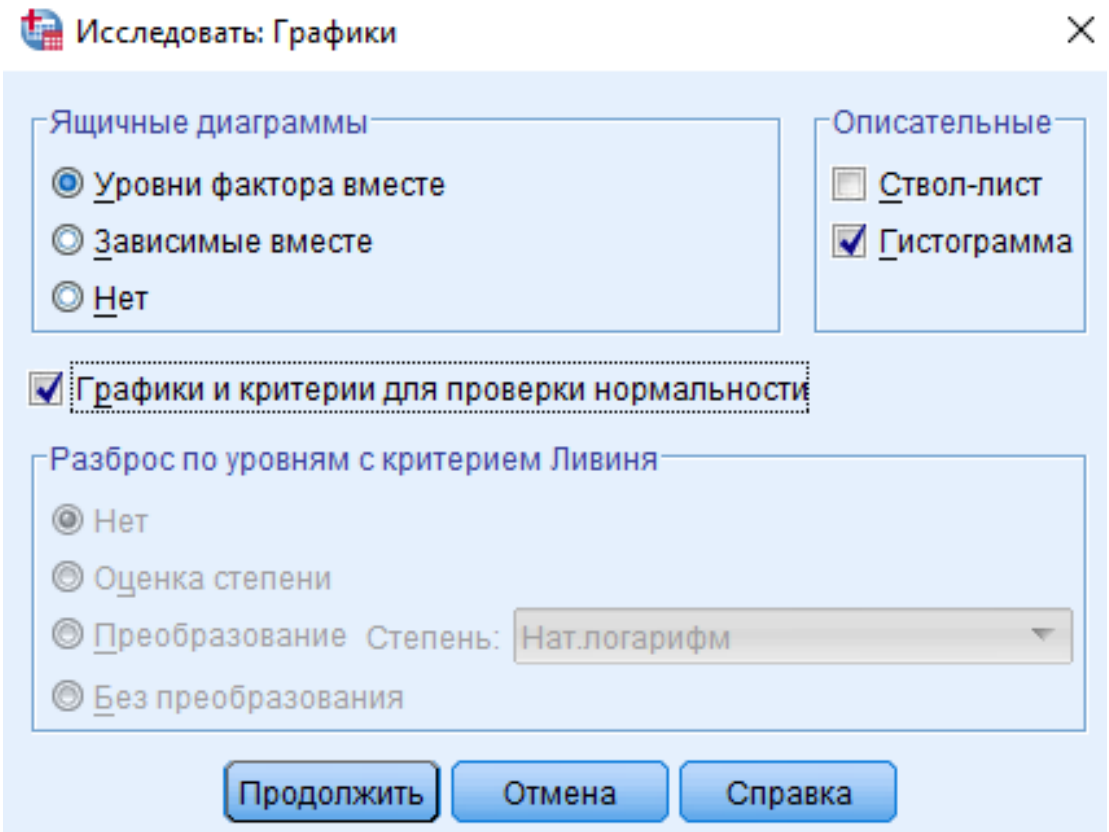


Рис. 3.5. Діалогове вікно SPSS

Натиснувши клавішу «Продолжить», отримаємо результати (табл. 3.13).

Таблиця 3.13

Критерії перевірки нормальності

	Колмогоров-Смірнов[a]			Шапіро-Уілк		
	Статистика	**ст. св.**	**Значимість**	**Статистика**	**ст. св.**	**Значимість**
Повні роки	0,288	10	0,019	0,886	10	0,153

a. Поправка значимості Лільефорса.

Нормальний розподіл (приймається гіпотеза H_0), якщо значимість показників < 0,05.

Висновок: за критерієм Колмогорова-Смірнова приймається H_0 гіпотеза (0,019 < 0,05), розподіл нормальний, при цьому критерій Шапіро-Уілка цю гіпотезу відхиляє (0,153 > 0,05).

Програмний продукт SPSS для оцінки розподілу дає можливість застосувати такі методи порівняння середніх величин (табл. 3.14):

Таблиця 3.14

Основні методи порівняння середніх величин

Методи порівняння середніх величин					
Параметричні методи перевірки			**Непараметричні методи**		
Одна вибірка: *t*-тест, *z*-тест	Дві вибірки		Одна вибірка: χ^2-квадрат, критерій Колмогорова-Смірнова	Дві вибірки	
	Незалежні вибірки: двогруповий *t*-тест, *z*-тест	Парні вибірки: парний *t*-тест		Незалежні вибірки: критерії Манна-Уітні, Крускала-Уоллеса, Медіани, Мозеса, Вальда-Вольфовіца, Джонкхіра-Терпстри, Колмогорова-Смірнова	Парні вибірки: критерії знаків, Уілкоксона, Фрідмана, Макнемара, Маргінальної однорідності

Параметричні критерії порівняння середніх величин

Розглянемо приклади застосування основних параметричних методів, що застосовують для перевірки гіпотез статистичних досліджень:

***t*-критерій** – це одновимірний метод перевірки гіпотез, застосовується для невеликих вибірок, коли невідоме стандартне відхилення. Результат порівняння середніх значень із застосуванням *t*-критерію оцінюється за достовірне значущості (*p*-значимість),

що є мірою достовірності обчислених результатів. Якщо розраховане $p < 0{,}05$, це означає, що вихідна гіпотеза (H_0) може бути відхилена з ймовірністю помилки менше 5 %, при цьому розраховується довірчий інтервал, який дорівнює 95 %.

Приклад. Перевіримо гіпотезу про те, що абітурієнти одного із факультетів у середньому поставили перший пріоритет (додаток Д).

Більш детально розглянемо застосування *t*-критерію (рис. 3.6) за допомогою SPSS: «**Анализ**» – «**Сравнение средних**» – «**Одновыборочный *t*-критерий**».

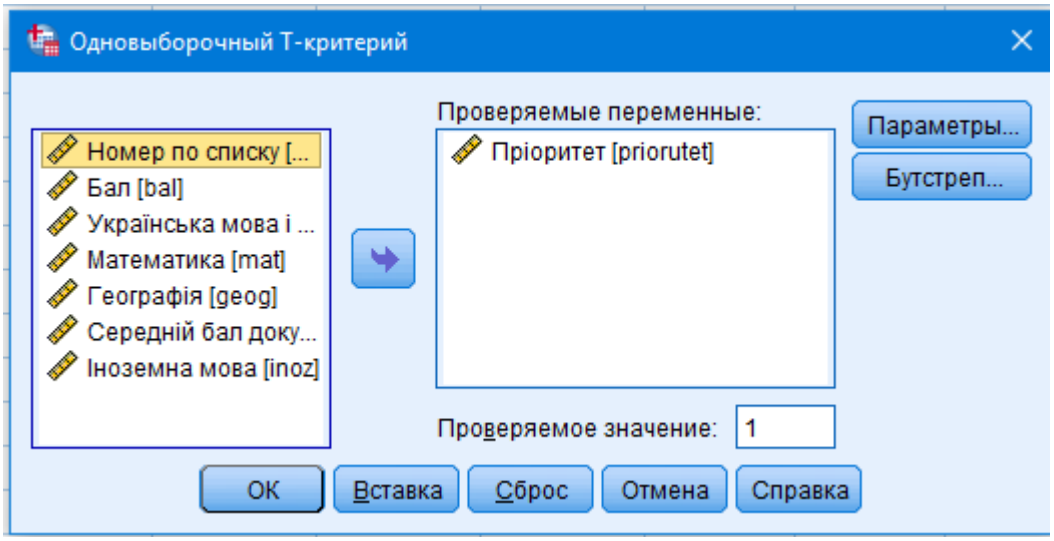


Рис. 3.6. Діалогове вікно SPSS

Обирається змінна, за якою перевіряється гіпотеза. Для цього необхідно виділити в лівому діалоговому вікні змінну і натиснути стрілку, що дасть змогу перенести змінну в праве вікно. Обов'язково проставляється значення гіпотези (в нашому випадку перший пріоритет) у комірці «Проверяемое значение». Натискається клавіша ОК. Результати дослідження відображено в таблицях 3.15–3.16.

Таблиця 3.15

Одновибіркова статистика

	N	Середнє	Середньоквадратичне відхилення	Середньоквадратична похибка середнього
Пріоритет	74	2,28	1,724	0,200

Таблиця 3.16

Одновибірковий критерій

	Значення критерію = 1					
	т	Ступені свободи	Значимість (двосторонння)	Середня різниця	95 % довірчий інтервал для різниці	
					Нижня	Верхня
Пріоритет	6,405	73	0,000	1,284	0,88	1,68

Висновок: H_0 гіпотеза відхиляється 0,00 < 0,05. Абітурієнти в середньому проставляють пріоритет 2,28, а не 1, при цьому це значення не потрапляє в довірчий інтервал:

$2{,}28 - 0{,}88 \leq 2{,}28 \leq 2{,}28 + 1{,}68$

$1{,}4 \leq 2{,}28 \leq 3{,}96$.

***t*-критерій для незалежних вибірок** визначає вплив однієї незалежної змінної на іншу залежну. Незалежна змінна є дихотомічною або бінарною.

Перевіряються 2 тести:

1) рівність середніх;
2) рівність дисперсій.

Приклад. Дослідити, чи впливає стать абітурієнта одного із факультетів на якість навчання в школі (загальний рейтинговий бал), обравши довірчий інтервал 95 % (додаток Д).

Функція в SPSS: «**Анализ» – «Сравнение средних» – «*t*-критерий для независимых выборок».**

Обирається змінна «Бал», яка характеризує загальний рейтинговий бал абітурієнтів. Вказується критерій, за яким будуть групувати дані – стать (рис. 3.7), за допомогою стрілок.

Групується за бінарною змінною стать. Натиснути клавішу «Задать групы» і проставити значення груп: 0 – жінка, 1 – чоловік (рис. 3.8).

У вкладці «Параметры» проставляємо рівень довірчого інтервалу (рис. 3.9).

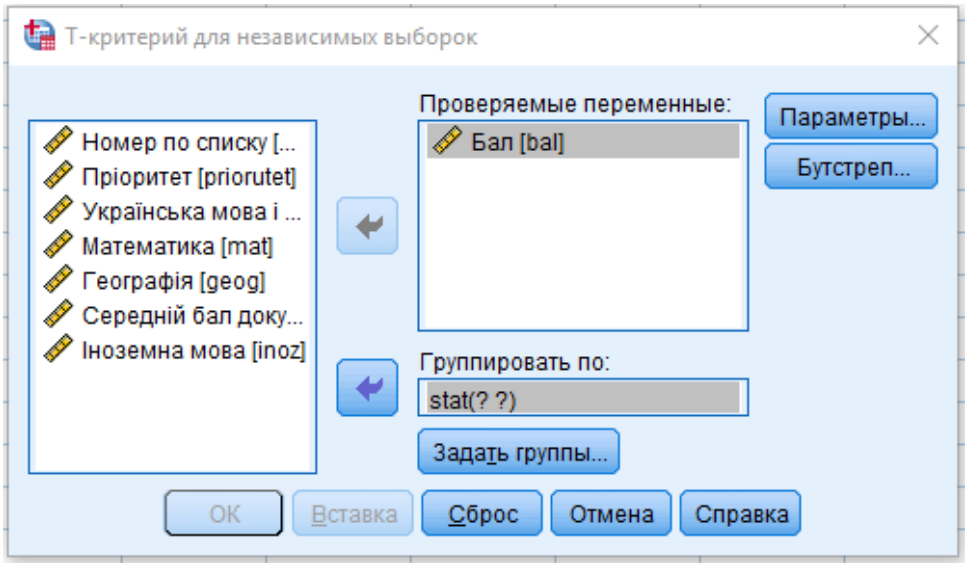


Рис. 3.7. Діалогове вікно SPSS

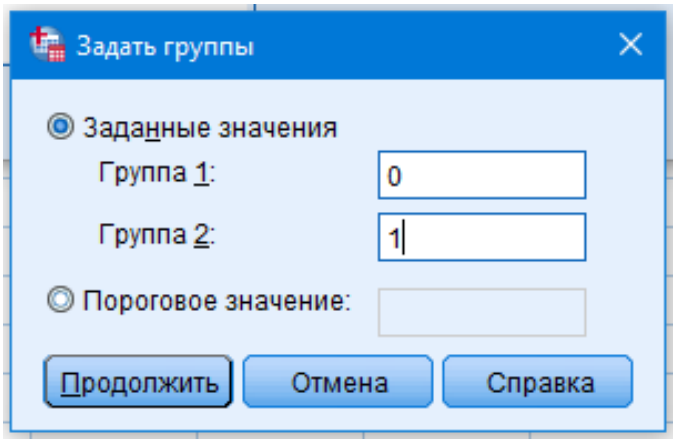


Рис. 3.8. Діалогове вікно SPSS

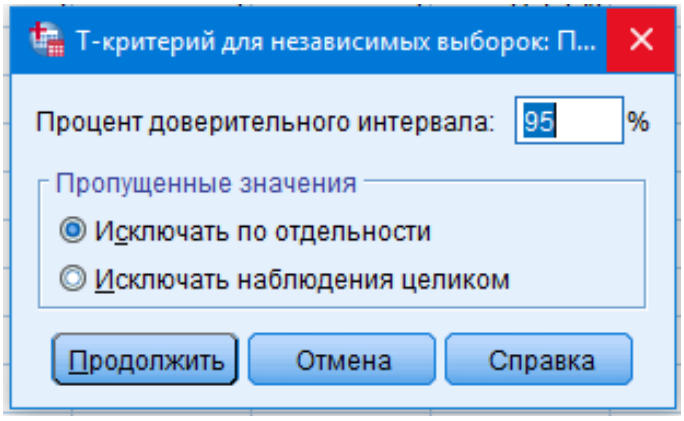


Рис. 3.9. Діалогове вікно SPSS

Натискаємо клавіші «Продолжить» і «ОК». Результати дослідження відображено в таблицях 3.17–3.18.

Таблиця 3.17

Статистика групи

	Стать	N	Середнє	Середньоквадратичне відхилення	Середньоквадратична похибка середнього
Бал	жінка	55	159,973	12,565	1,694
	чоловік	23	149,987	13,807	2,879

У вибірку увійшло 55 жінок і 23 чоловіка, які вступили на один із факультетів.

Таблиця 3.18

Критерій для незалежних вибірок

	Критерій рівності дисперсії Лівіня		*t*-критерій для рівності середніх						
								95 % довірчий інтервал для різниці	
Бал	F	Значимість	т	ст.св.	Знач. (двостороння)	Середня різниця	Середньоквадратична похибка різниці	Нижня	Верхня
Передбачається рівність дисперсій	0,017	0,898	3,109	76	0,003	9,986	3,212	3,588	16,384
Не передбачається рівність дисперсій			2,989	38,019	0,005	9,986	3,341	3,223	16,749

Правильність висунутої гіпотези можна перевірити за допомогою величини «значимість» за критерієм Лівіня (табл. 3.18), яка дорівнює 0,898 (*F*-критерій має ймовірність більшу за 0,05). Отже, гіпотезу про рівність дисперсій не відхиляємо з імовірністю помилки 0 %, що вище порогового значення 5 %. Нульова гіпотеза (Но) приймається, аналізується верхній рядок. Якщо б нульова гіпотеза відхилялася, то аналізували б результати за нижнім рядком.

Висновок: у цьому випадку використовується *t*-критерій, відповідний твердженням «передбачається рівність дисперсій». У таблиці можна бачити, що *t* дорівнює 3,109, ступенів свободи – 76, двостороння значимість дорівнює 0,003, яке менше допустимого рівня, що дорівнює 0,05. Отже, гіпотезу про те, що чоловіки і

жінки мають однаковий бал, відхиляємо. Оскільки середній бал жінок 159,973, а чоловіків – 149,987, що на десять балів менше.

***t*-критерій для залежних вибірок** застосовують під час аналізу однієї групи респондентів для різних цілей.

Приклад. Дослідити вплив на загальний рейтинговий бал абітурієнта, одного із факультетів, оцінки ЗНО із математики й української мови та літератури (додаток Д).

Функція в SPSS: «**Анализ» – «Сравнение средних» – «*t*-критерий для парных выборок»** (рис. 3.10).

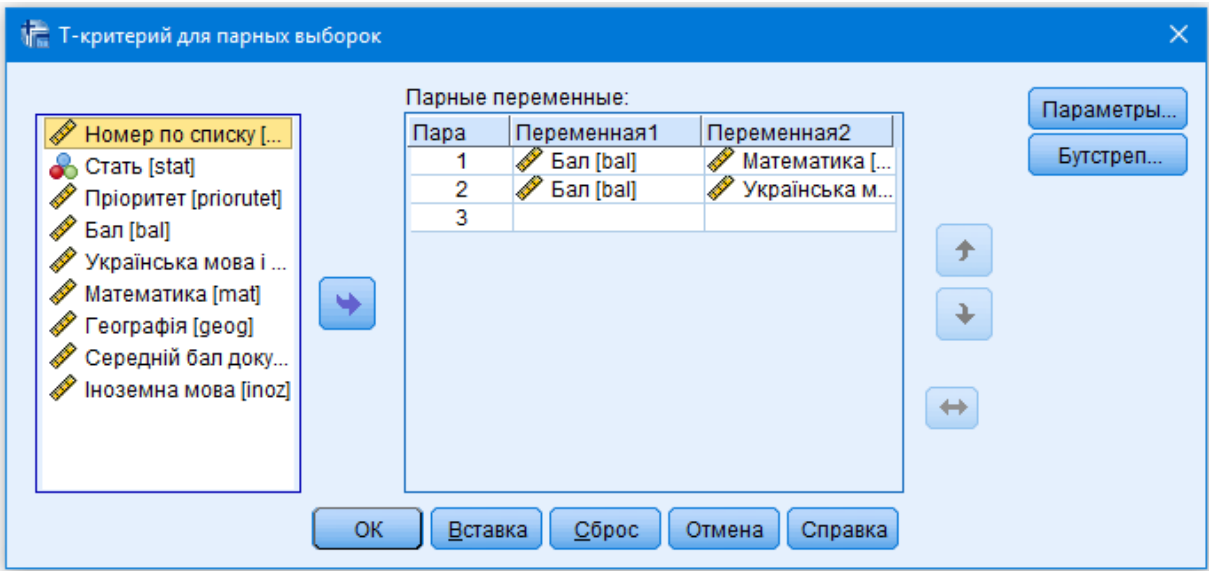


Рис. 3.10. Діалогове вікно SPSS

За чергою обирається перша і друга пари змінних, що досліджуються. Натискаємо клавішу «ОК». У таблицях 3.19–3.21 відображені результати дослідження.

Таблиця 3.19

Статистика парних вибірок

		Середнє	№	Середньоквадратичне відхилення	Середньоквадратична похибка середнього
Пара 1	Бал	157,029	78	13,645	1,545
	Математика	142,795	78	20,207	2,288
Пара 2	Бал	157,0293	78	13,645	1,545
	Українська мова і література	161,590	78	18,592	2,105

Таблиця 3.20

Кореляції парних вибірок

		№	Кореляція	Значимість
Пара 1	Бал & Математика	78	0,769	0,000
Пара 2	Бал & Українська мова і література	78	0,812	0,000

Таблиця 3.21

Критерій парних вибірок

		Парні різності					т	ст.	Знач. (двосторонн.)
		Середнє	Середньокв. відхилення	Середньокв. похибка середнього	95 % довірчий інтервал для різниць				
					Нижня	Верхня			
Пара 1	Бал – Математика	14,234	13,062	1,479	11,289	17,179	9,624	77	0,000
Пара 2	Бал – Українська мова і література	–4,561	10,948	1,240	–7,029	–2,093	–3,679	77	0,000

Висновок: загальна середня рейтингова оцінка абітурієнтів – 157,029. Абітурієнти краще знають українську мову і літературу (середнє значення 161,59), ніж математику (середнє значення 142,795). Вплив результатів ЗНО з української мови на загальний рейтинговий бал більший (кореляція 0,812), ніж з математики (кореляція 0,769), при цьому обидва показники вказують на значний зв’язок.

Непараметричні критерії порівняння середніх величин

Непараметричні методи застосовуються для перевірки гіпотез про параметри генеральної сукупності у випадку, коли змінна не розподілена нормально або дослідник застосовує номінальні та порядкові дані. Результати непараметричних методів мають меншу інформативність і не такі точні, ніж параметричні критерії. Розглянемо критерії, що найчастіше застосовують у прикладних дослідженнях.

Критерій Манна-Уітні застосовують для порівняння двох незалежних вибірок (аналог *t*-критерію для незалежних вибірок). Спочатку ранжують усі об'єкти без врахування належності до груп порівняння, а потім розраховують середні ранги для кожної із двох груп. Після знаходження середніх рангів визначається *p*-значимість. Тест можна не застосовувати у разі наявного нормального розподілу.

Приклад. Дослідити вплив статті на загальний рейтинговий бал абітурієнта одного із факультетів (додаток Д).

Функція в SPSS: «**Анализ**» – «**Непараметрические критерии**» – «**Устаревшие диалоговые окна**» – «**Для двух независимых выборок**» (рис. 3.11).

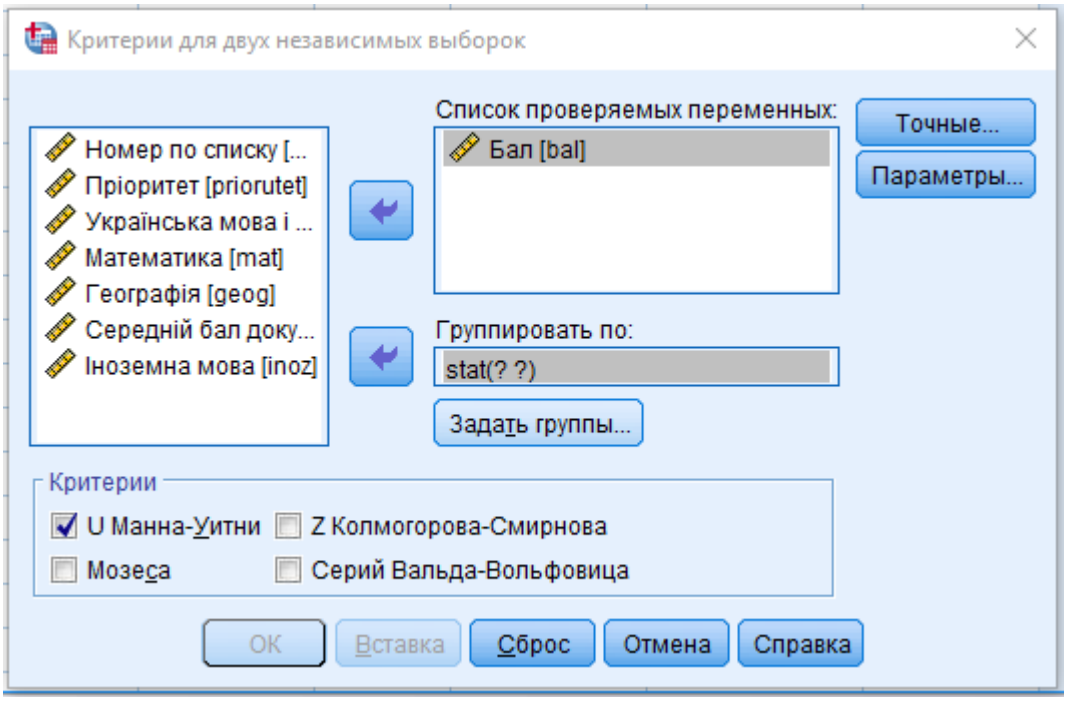


Рис. 3.11. Діалогове вікно SPSS

З лівого діалогового вікна обираємо загальний рейтинговий бал абітурієнтів, за допомогою стрілки переміщуємо в праве діалогове вікно. Групуються змінні по бінарній змінній стать. Для цього необхідно обрати вкладку «Задать группы» і вказати стать: 0 – жінка, 1 – чоловік (такий шифр мають первинні дані в таблиці). Натискаємо клавішу «Продолжить» (рис. 3.12).

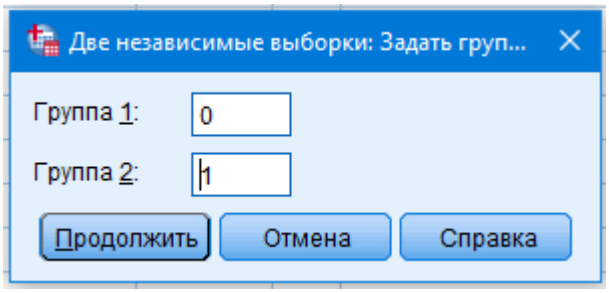


Рис. 3.12. Діалогове вікно SPSS

Для отримання більшої інформація про змінні обирається вкладка «Параметры», проставляються мітки описової статистики (рис. 3.13).

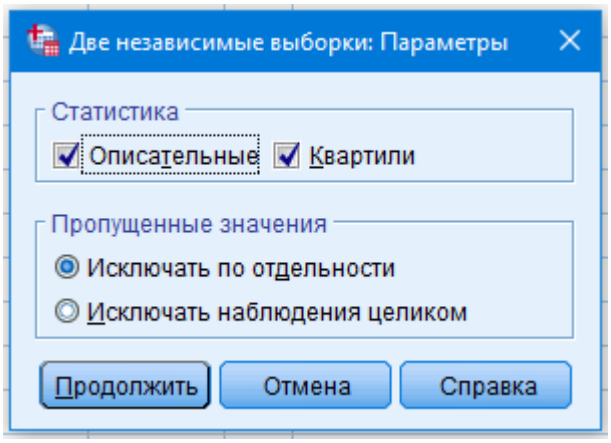


Рис. 3.1.3. Діалогове вікно SPSS

Натиснувши клавіші «Продолжить» і «ОК», отримаємо результати аналізу (табл. 3.22–3.24).

Таблиця 3.22

Описова статистика

	№	Середнє	Середньокв. відхилення	Мінімум	Максимум	Процентилі		
						25	50-я (медіана)	75-я
Бал	78	157,09	13,645	118,728	182,631	149,945	159,044	166,439

Таблиця 3.23

Ранги

	Стать	N	Середній ранг	Сума рангів
Бал	жінка	55	44,18	2430,00
	чоловік	23	28,30	651,00
	Усього	78		

Таблиця 3.24

Статистичні критерії

	Бал
U Манна-Уітні	375,000
W Вілкоксона	651,000
Z	–2,822
Асимптот. значимість (2-стороння)	0,005

а. Згруповано за змінною: стать.

Висновок: середній рейтинговий бал жінок – 44,18, а для чоловіків – 28,30. Оскільки розмір рівня значимості $p = 0{,}005 < 0{,}05$, можна стверджувати, що успішність жінок, вища за успішність чоловіків, є статистично значимою. Аналізуючи описову статистику, необхідно зазначити, що лише 25 % абітурієнтів мають рейтинговий бал до 149,945, 50 % – до 159,044, усі інші вищий, що вказує на високий рівень підготовки студентів першого курсу.

Критерій знаків. Під час застосування методу спочатку для кожного об’єкта визначається знак різниць значень. Підраховується кількість додатних, від’ємних і нульових різниць, а потім розраховується *p*-значимість.

Приклад. Порівняємо результати тестування школярів зі знання числових рядів і рівня обізнаності (додаток Е).

Функція в SPSS: «**Анализ**» – «**Непараметрические критерии**» – «**Устаревшие диалоговые окна**» – «**Для двух связаных выборок**» (рис. 3.14).

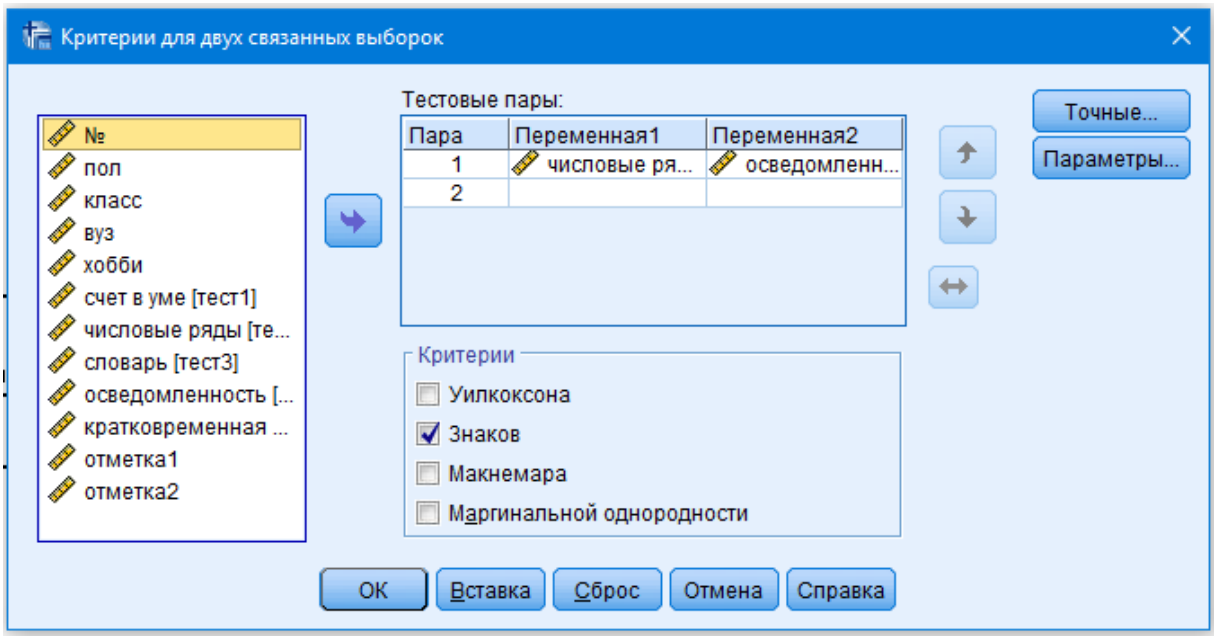


Рис. 3.14. Діалогове вікно SPSS

Обираємо змінні, що порівнюються. Переміщуємо їх у праве діалогове вікно. Обираємо критерій перевірки «Знаков». У вкладці «Точные» ставимо мітку «Только асимптотически». Натискаємо клавішу «Продолжить» (рис. 3.15).

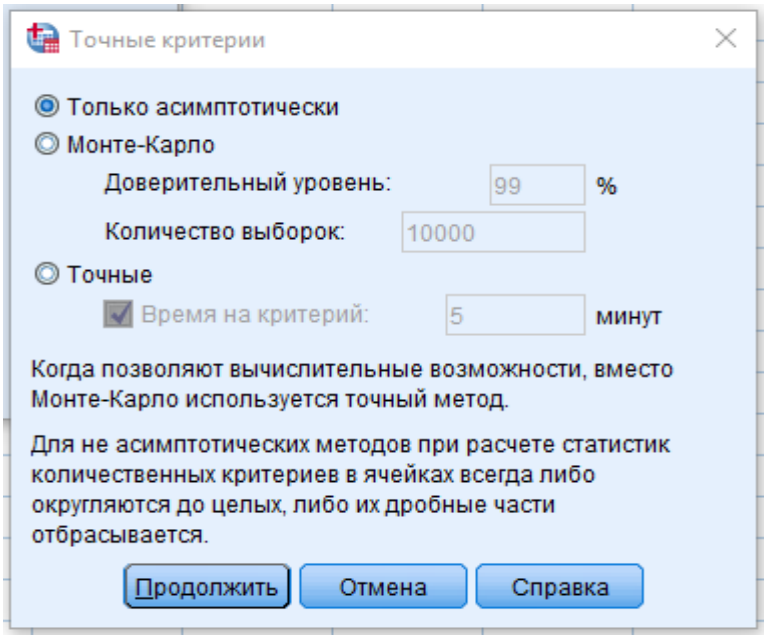


Рис. 3.15. Діалогове вікно SPSS

Для отримання додаткової інформація про змінні обирається вкладка «Параметры», проставляються мітки описової статистики (рис. 3.16).

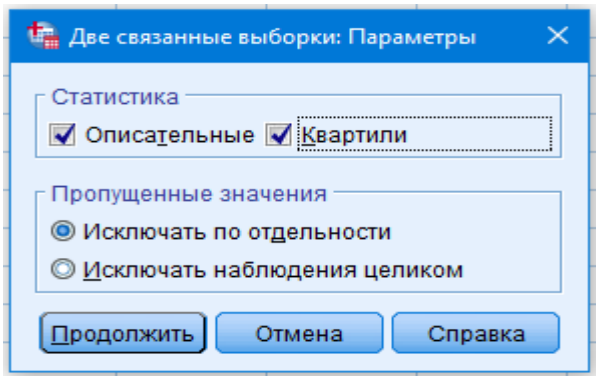


Рис. 3.16. Діалогове вікно SPSS

Натискаємо клавіші «Продолжить» і «ОК». Результати наведені в таблицях 3.25–3.27.

Таблиця 3.25

Описова статистика

	№	Середнє	Середньокв. відхилення	Мінімум	Максимум	Процентилі		
						25	50-я (медіана)	75-я
Числові ряди	100	10,35	2,768	4	17	8,25	10,50	12,00
Обізнаність	100	11,51	3,040	4	21	10,00	12,00	13,00

Таблиця 3.26

Частоти

		N
Обізнаність – числові ряди	Від'ємні різниці[a]	39
	Додатні різниці[b]	57
	Співпадаючі спостереження[c]	4
	Усього	100

a. Обізнаність < числові ряди.
b. Обізнаність > числові ряди.
c. Обізнаність = числові ряди.

Таблиця 3.27

Статистичні критерії

	Обізнаність – числові ряди
Z	–1,735
Асимптот. значимість (2-стороння)	0,083

a. Критерій знаків.

Висновок: у 39 випадках значення змінної тесту числові ряди виявилися меншими за значення тесту обізнаність, а в 57 випадках значення змінної тесту числові ряди перевищили значення змінної тест обізнаність, до того ж 4 рази встановлено рівність значень обох змінних. При цьому $p = 0{,}083 > 0{,}05$ можна стверджувати, що відмінність між результатами тестів є статистично незначимою.

Критерій Уілкоксона ґрунтується на основі підрахунку абсолютних різниць між параметрами значень з подальшим їх ранжуванням. Потім розраховується середнє значення рангів для додатних і від'ємних різниць.

Приклад. Порівняємо результати тестування школярів зі знання числових рядів і рівня обізнаності (додаток Е).

Функція в SPSS: «**Анализ**» – «**Непараметрические критерии**» – «**Устаревшие диалоговые окна**» – «**Для двух связаных выборок**» (рис. 3.17).

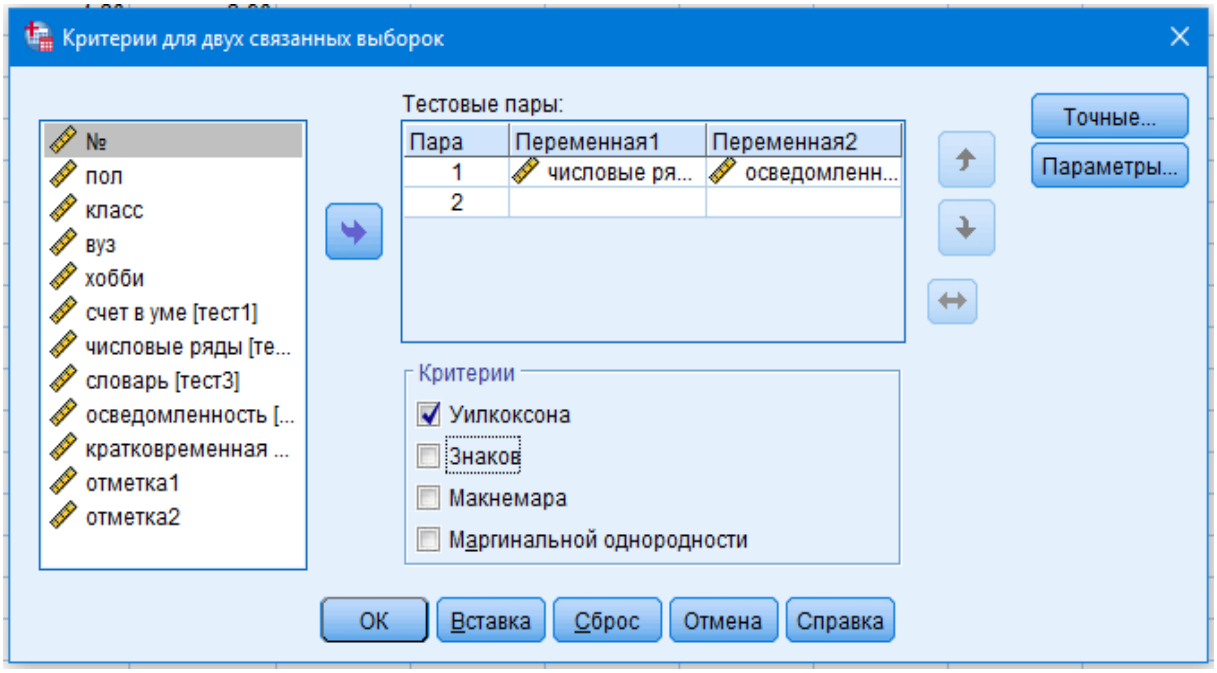


Рис. 3.17. Діалогове вікно SPSS

Аналогічно обираємо змінні, що порівнюються. Переміщуємо їх у праве діалогове вікно. Обираємо критерій перевірки «Уилкоксона». Натискаємо клавішу «ОК» і отримуємо результати аналізу (табл. 3.28–3.29).

Таблиця 3.28

Ранги

		№	Середній ранг	Сума рангів
Обізнаність – числові ряди	Від'ємні ранги	39[a]	42,26	1 648,00
	Додатні ранги	57[b]	52,77	3 008,00
	Співпадаючі спостереження	4[c]		
	Всього	100		

a. Обізнаність < числові ряди.

b. Обізнаність > числові ряди.

c. Обізнаність = числові ряди.

Таблиця 3.29

Статистичні критерії

	Обізнаність – числові ряди
Z	–2,493[b]
Асимптот. значимість (2-стороння)	0,013

a. Критерій знакових рангів Вілкоксона.

b. На основі від'ємних рангів.

Висновок: рівень значимості $p = 0{,}013 < 0{,}05$ можна стверджувати, що відмінність між результатами тестів є статистично значимою. Критерій Вілкоксона є більш потужним, ніж критерій знаків.

Критерій Крускала-Уоллеса (порівняння К незалежних вибірок)

Цей непараметричний критерій є альтернативою одновимірному (межгруповому) дисперсійному аналізу. Сутність критерію полягає в представлені всіх значень вибірок, що порівнюються, у вигляді однієї загальної послідовності рангів, з подальшим розрахунком середнього рангу для кожної вибірки. Якщо виконується статистична гіпотеза про відсутність відмінностей, то стверджують, що всі середні ранги приблизно рівні і наближені до загального середнього рангу. Нульова гіпотеза – це відмінності відсутні.

Приклад. Перевіримо гіпотезу про те, чи існує зв'язок між успішністю з математики і хобі школярів (додаток Е).

Функція в SPSS: «**Анализ**» – «**Непараметрические критерии**» – «**Устаревшие диалоговые окна**» – «**Для К независимых выборок**» (рис. 3.18).

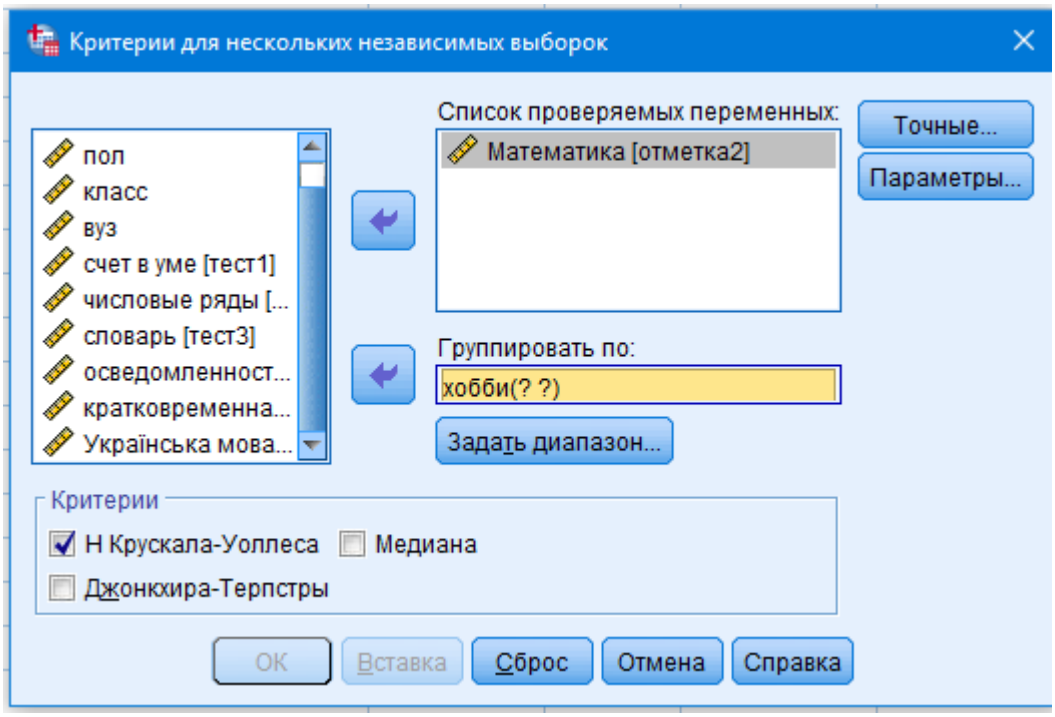


Рис. 3.18. Діалогове вікно SPSS

Обираємо змінну, що досліджують, у верхнє діалогове вікно, і змінну, за якою групують, у праве нижнє. Задаємо діапазон порівняння, а саме види хобі. Для цього натискають вкладку «Задать диапазон» (рис. 3.19).

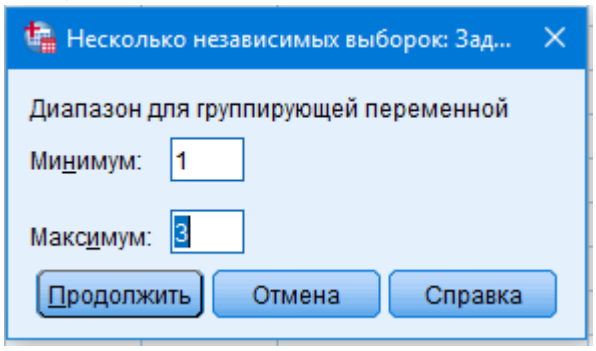


Рис. 3.19. Діалогове вікно SPSS

У вкладці «Точные» ставимо мітку «Только асимптотически». Натискаємо клавішу «Продолжить» (рис. 3.20).

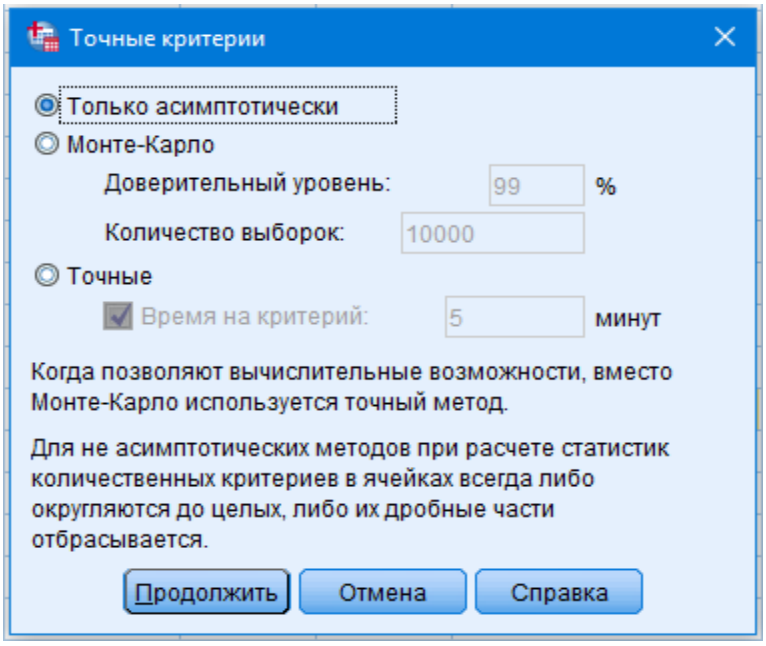


Рис. 3.20. Діалогове вікно SPSS

Натиснувши клавіші «Продолжить» і «ОК», отримаємо результати аналізу (табл. 3.30–3.31).

Таблиця 3.30

Ранги

	Хобі	N	Середній ранг
Математика	Спорт	33	38,18
	Комп'ютер	37	54,28
	Мистецтво	30	59,38
	Усього	100	

Таблиця 3.31

Статистичні критерії[a,b]

	Математика
Хі-квадрат	9,437
Ст. св.	2
Асимптот. значимість	0,009

a. Критерій Крускала-Уоллеса.

b. Групуюча змінна: хобі.

Отже, рівень значимості $p = 0{,}009 < 0{,}05$, можна стверджувати, що існує статистично достовірний зв'язок позашкільних захоплень школярів з успішністю з математики. Вплив є статистично значимим.

Щоб визначити силу зв'язку, доцільно попарно порівняти різні співвідношення хобі. Для цього використаємо непараметричний критерій для двох незалежних вибірок (Манна-Уітні):

Функція в SPSS: «**Анализ**» – «**Непараметрические критерии**» – «**Устаревшие диалоговые окна**» – «**Для двух независимых выборок**» (рис. 3.21).

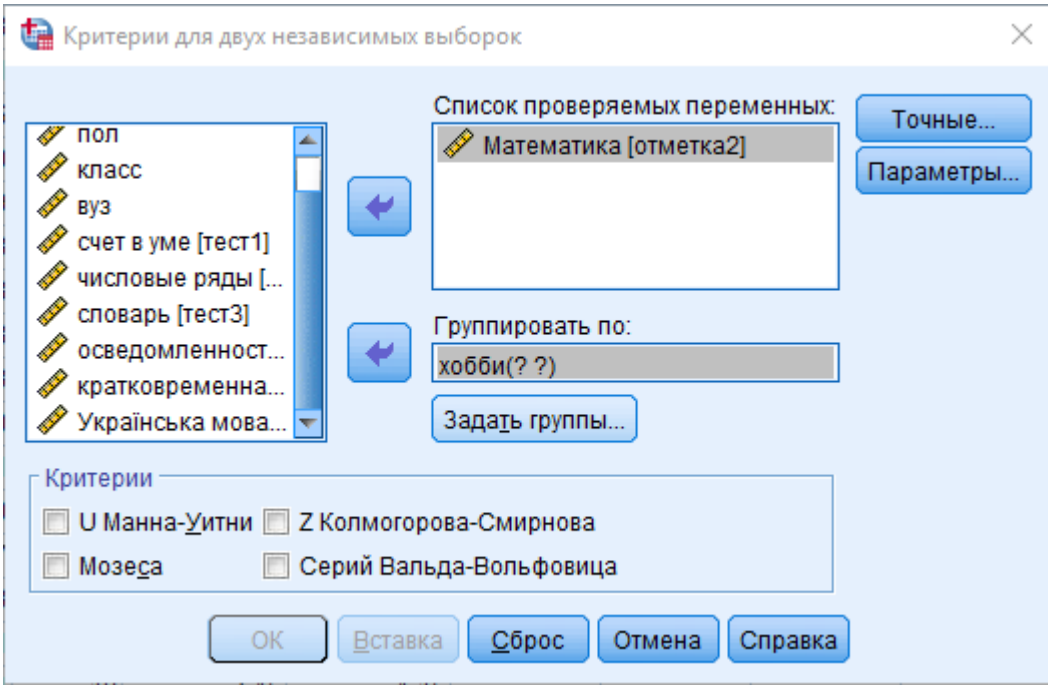


Рис. 3.21. Діалогове вікно SPSS

Натиснути клавішу «Задать группы» і порівняти по черзі хобі: спорт (1) і комп'ютер (3), спорт (1) і мистецтво (2), комп'ютер (2) і мистецтво (3). Встановити позначку на критерію «Манна-Уитни» (рис. 3.22–3.25).

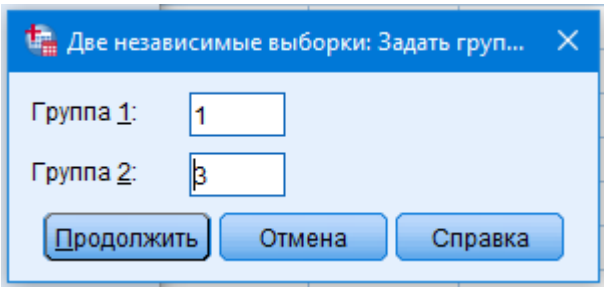


Рис. 3.22. Діалогове вікно SPSS

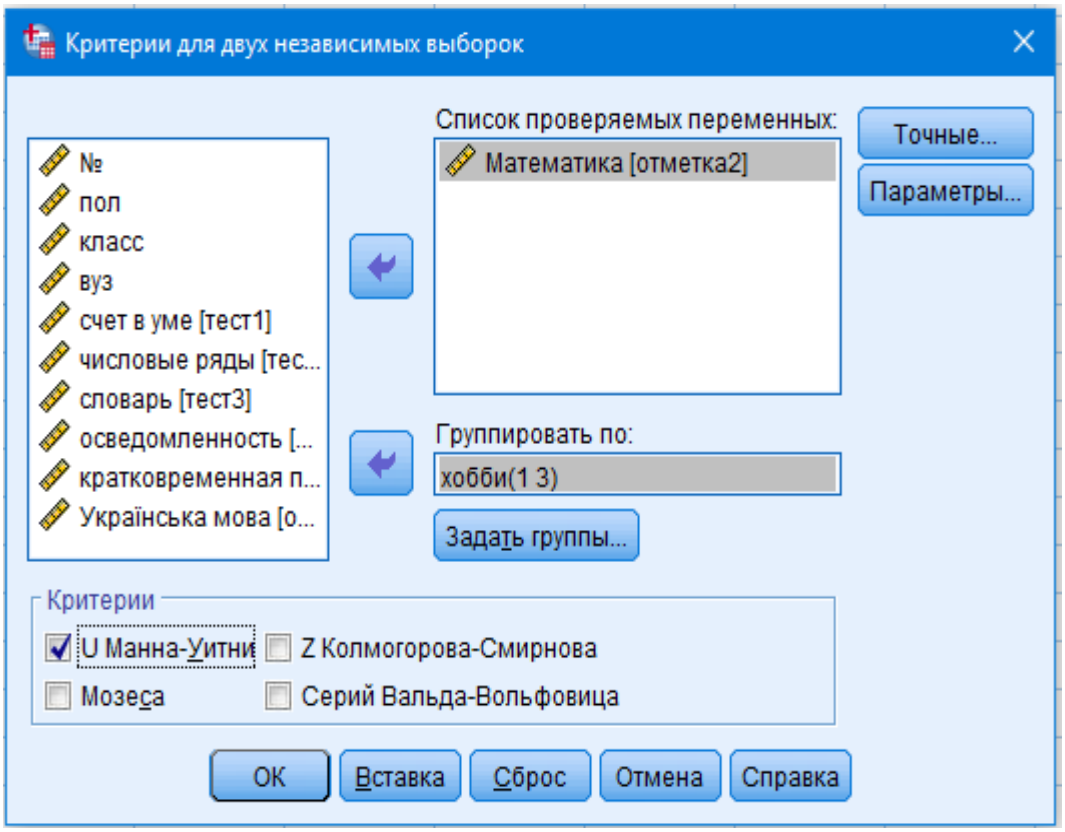


Рис. 3.23. Діалогове вікно SPSS

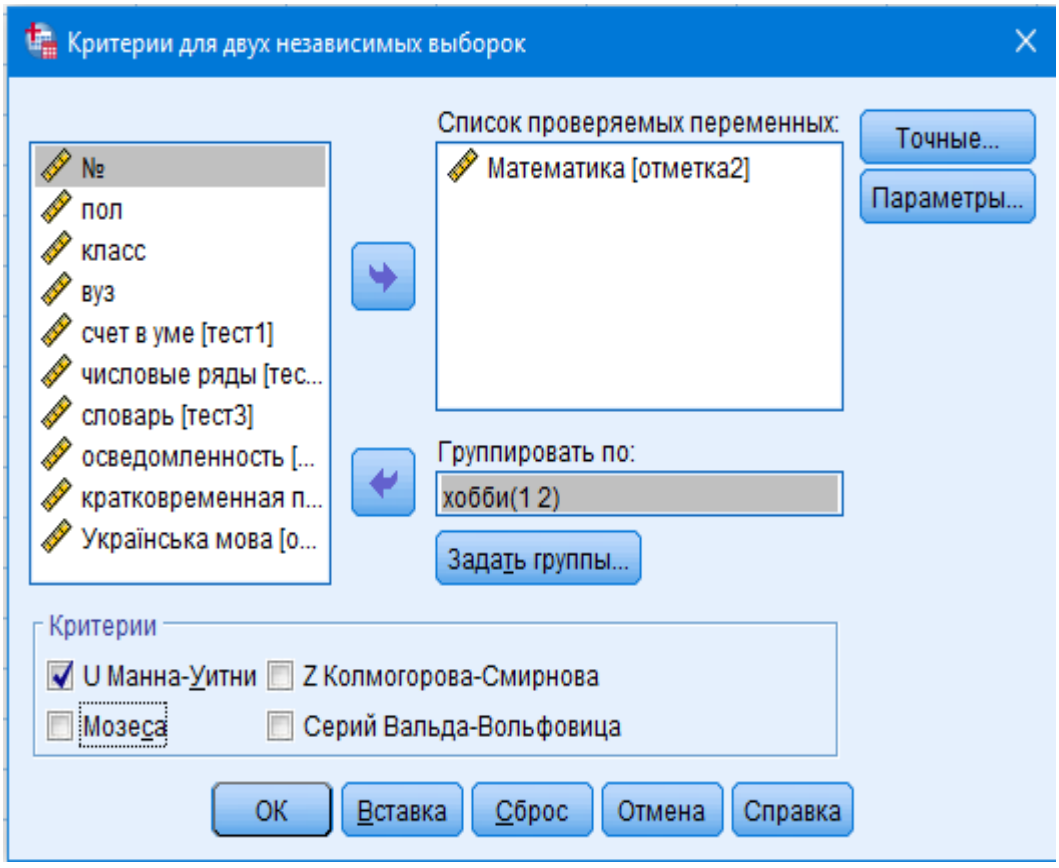


Рис. 3.24. Діалогове вікно SPSS

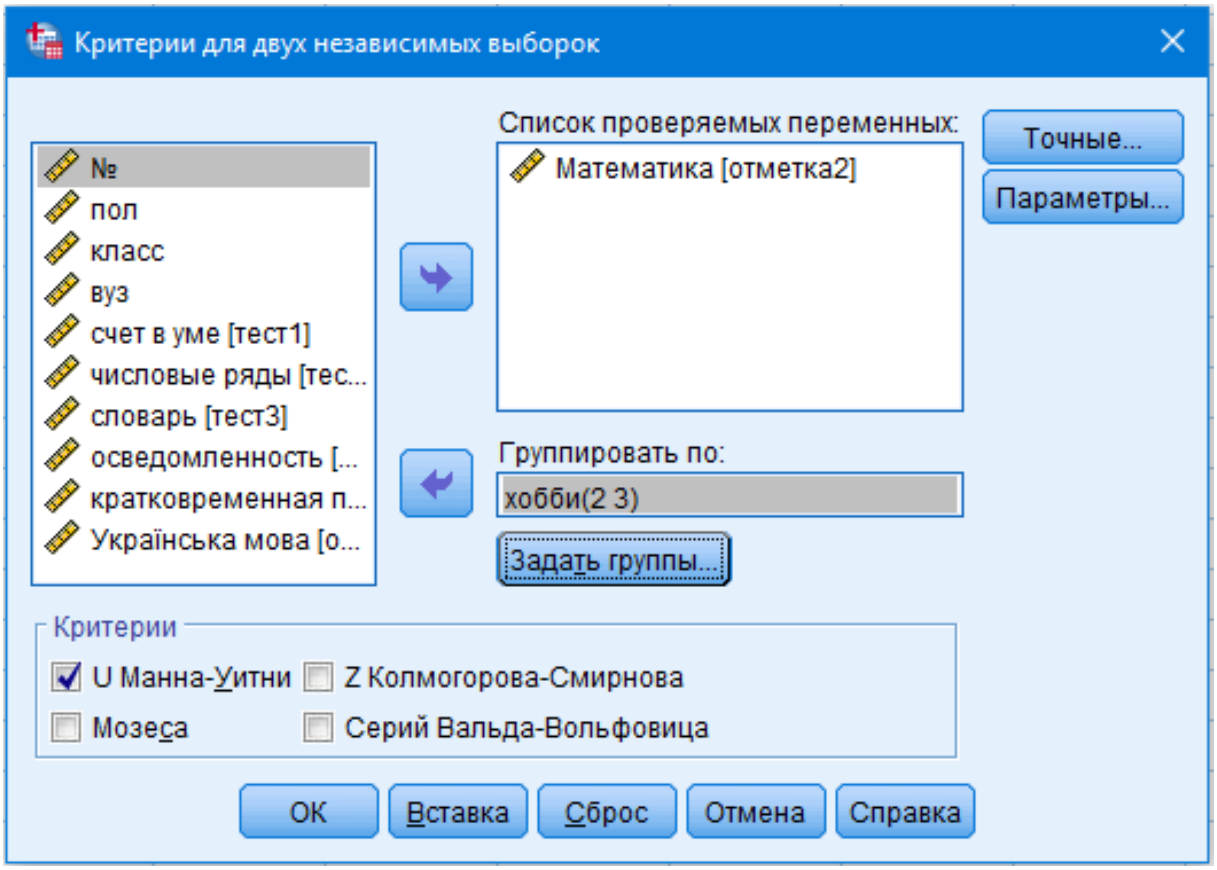


Рис. 3.25. Діалогове вікно SPSS

Натиснути клавішу «ОК». У таблицях 3.32–3.37 поданий детальний аналіз.

Таблиця 3.32

Ранги

	Хобі	N	Середній ранг	Сума рангів
Математика	Спорт	33	29,15	962,00
	Комп'ютер	37	41,16	1 523,00
	Усього	70		

Таблиця 3.33

Статистичні критерії[a]

	Математика
U Манна-Уітні	401,000
W Вілкоксона	962,000
Z	–2,472
Асимптот. значимість (2-стороння)	0,013

а. Групуюча змінна: хобі.

Таблиця 3.34

Ранги

	Хобі	N	Середній ранг	Сума рангів
Математика	Спорт	33	26,03	859,00
	Мистецтво	30	38,57	1 157,00
	Усього	63		

Таблиця 3.35

Статистичні критерії[a]

	Математика
U Манна-Уітні	298,000
W Вілкоксона	859,000
Z	–2,718
Асимптот. значимість (2-стороння)	0,007

а. Групуюча змінна: хобі.

Таблиця 3.36

Ранги

	Хобі	N	Середній ранг	Сума рангів
Математика	Комп'ютер	37	32,12	1 188,50
	Мистецтво	30	36,32	1 089,50
	Усього	67		

Таблиця 3.37

Статистичні критерії[a]

	Математика
U Манна-Уітні	485,500
W Вілкоксона	1 188,500
Z	–0,880
Асимптот. значимість (2-стороння)	0,379

а. Групуюча змінна: хобі.

Висновок: діти, у яких хобі комп'ютер (середній ранг 41,16) мають кращі оцінки з математики, ніж ті, що відвідують спортивні секції (29,15), оцінка значуща $p = 0{,}013 < 0{,}05$. Якщо порівнювати дітей, захоплених мистецтвом (середній ранг 38,57), то вони мають кращі оцінки з математики, ніж ті, що відвідують спортивні секції (26,03), оцінка значуща $p = 0{,}007 < 0{,}05$. У третьому поєднані характеризувати недоцільно, оскільки оцінка незначуща $p = 0{,}379 > 0{,}05$.

Критерій Фрідмана. Застосовують для порівняння парних залежних вибірок. Він ґрунтується на ранжуванні ряду повторних вимірювань для кожного об'єкта вибірки.

Приклад. Перевірити значимість результатів тестування школярів (додаток Е).

Функція в SPSS: «**Анализ**» – «**Непараметрические критерии**» – «**Устаревшие диалоговые окна**» – «**Для К связаных выборок**» (рис. 3.26):

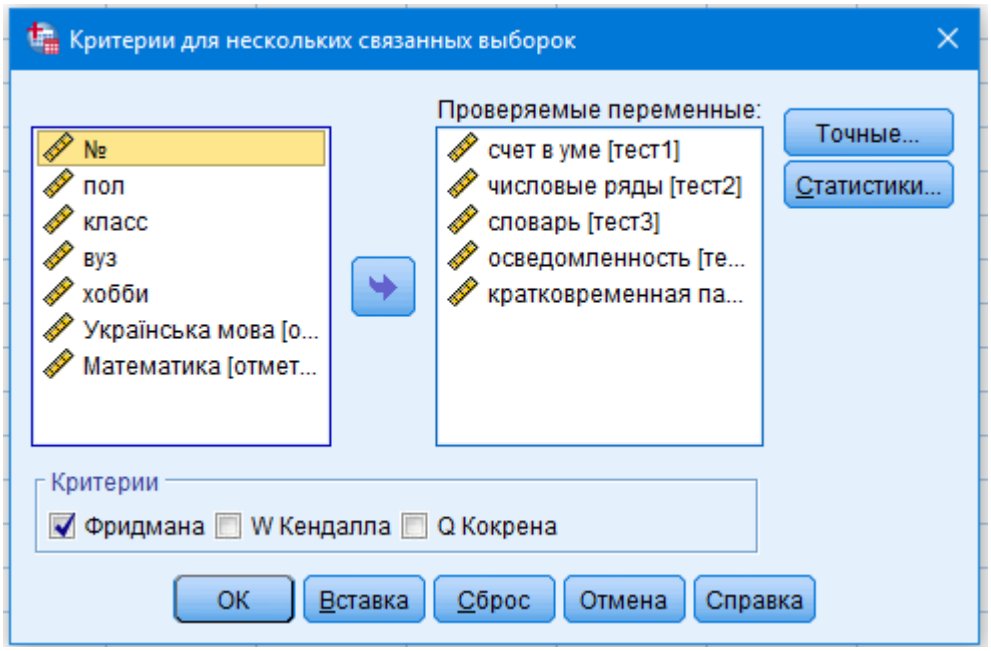


Рис. 3.26. Діалогове вікно SPSS

Обрати з правого діалогового вікна значення для перевірки. Встановити позначку на критерію «Фридмана». Натиснути клавішу «ОК». Результати аналізу наведені в таблицях 3.38–3.39.

Таблиця 3.38

Ранги

	Середній ранг
Рахунок в умі	2,50
Числові ряди	2,65
Словник	3,41
Обізнаність	3,08
Короткочасна пам'ять	3,37

Таблиця 3.39

Статистичні критерії[a]

N	100
Хі-квадрат	29,696
ст. св.	4
Асимптот. значимість	0,000

а. Критерій Фрідмана.

Висновок: рівень значимості $p = 0{,}000 < 0{,}05$ можна стверджувати про статистичну достовірність різниці між п'ятьма результатами тестування школярів. При цьому найкращі результати вони показали зі знання термінів – тест-словник (середній ранг 3,41), найгірші отримали за рахунок в умі (середній ранг 2,5).

Перелік питань для самоконтролю

1. Поясніть сутність статистичної гіпотези.
2. Назвіть основні умови застосування параметричних методів.
3. Поясніть принципи перевірки гіпотез.
4. Поясніть відмінність між помилкою I і II родів.
5. Назвіть етапи перевірки статистичної гіпотези.
6. Назвіть види розподілів, основні характеристики.
7. Поясність значимість критеріїв згоди, назвіть їх.
8. Назвіть параметричні методи перевірки статистичних гіпотез, що застосовуються у програмному продукті SPSS.
9. Назвіть види непараметричних методів перевірки статистичних гіпотез, що застосовуються у програмному продукті SPSS.
10. Охарактеризуйте непараметричні методи перевірки статистичних гіпотез, що застосовують для незалежних вибірок.
11. Охарактеризуйте непараметричні методи перевірки статистичних гіпотез, що застосовують для парних вибірок.
12. Поясність особливість застосування критеріїв перевірки нормальності розподілу в SPSS.

Тести

1. Певне твердження щодо генеральної сукупності, що перевіряється на основі вибірки – це:

а) вибіркове спостереження;

б) статистична гіпотеза;

в) рівність дисперсії;

г) правильної відповіді немає.

2. Якщо відхиляється нульова гіпотеза, яка насправді є правильною, то це:

а) помилка першого роду;

б) помилка другого роду;

в) помилка першого і другого родів;

г) правильної відповіді немає.

3. Якщо відхиляється альтернативна гіпотеза, яка насправді є правильною, то це:

а) помилка першого роду;

б) помилка другого роду;

в) помилка першого і другого родів;

г) правильної відповіді немає.

4. Ймовірність здійснити похибку першого роду (тобто відхилити вірну нульову гіпотезу) називають:

а) рівнем значимості середнього значення;

б) рівнем значимості дисперсії;

в) рівнем значимості гіпотези;

г) правильної відповіді немає.

5. Чому дорівнює площа нормального розподілу:

а) 0;

б) 4;

в) 1;

г) правильної відповіді немає.

6. За формулою Бернуллі обчислюється дискретна випадкова величина, що має:

а) біноміальний закон розподілу;

б) гіпергеометричний розподіл;

в) геометричний розподіл;

г) рівномірний розподіл.

7. Параметр λ застосовується під час розрахунку розподілу:

а) експоненційного;

б) гіпергеометричного;

в) геометричного;

г) рівномірного.

8. Статистичний критерій перевірки гіпотези про можливий закон невідомого розподілу називається:

а) критерієм згоди;

б) дискретним законом;

в) статистичним твердженням;

г) правильної відповіді немає.

9. До параметричних методів порівняння середніх величин відносять:

а) критерії знаків, Фрідмана й Уілкоксона;

б) t-тест, z-тест;

в) критерії Манна-Уітні і Крускала-Уоллеса

г) правильної відповіді немає.

10. До непараметричних методів порівняння середніх величин, що застосовують для незалежної вибірки, відносять:

а) критерії знаків, Фрідмана й Уілкоксона;

б) t-тест, z-тест;

в) критерії Манна-Уітні і Крускала-Уоллеса;

г) правильної відповіді немає.

11. До непараметричних методів порівняння середніх величин, що застосовують для парної вибірки, відносять:

а) критерії знаків, Фрідмана і Уілкоксона;

б) t-тест, z-тест;

в) критерії Манна-Уітні і Крускала-Уоллеса;

г) правильної відповіді немає.

12. Критерій Шапіро-Уілка доцільно застосовувати для перевірки:

а) великих вибірок;

б) малих вибірок;

в) розмір вибірки не має значення;

г) правильної відповіді немає.

13. До непараметричних методів порівняння середніх величин, що застосовують для незалежної вибірки, відносять:

а) критерії Макнемара і Маргінальної однорідності;

б) t-тест, z-тест;

в) критерії Медіани, Джонкхіра-Терпстри і Мозеса;

г) правильної відповіді немає.

14. До непараметричних методів порівняння середніх величин, що застосовують для парної вибірки, відносять:

а) критерії Макнемара і Маргінальної однорідності;

б) t-тест, z-тест;

в) критерії Медіани, Джонкхіра-Терпстри і Мозеса;

г) правильної відповіді немає.

Економічна інтерпретація статистичних гіпотез

Приклад. Провести статистичний аналіз 78 абітурієнтів 2019 року, які вступили в навчально-науковий інститут обліку, аналізу та аудиту (далі – ННІОАА). Для аналізу взяти такі статистичні показники: стать вступника, загальний бал, оцінка ЗНО з української мови і літератури, математики, географії, іноземної

мови, середній бал документа про освіту та пріоритет, який абітурієнт ставив для Університету ДФС України (додаток Д[4]).

Зменшення кількості абітурієнтів за останні роки, введення зовнішнього незалежного оцінювання (далі – ЗНО), встановлення бар'єру із ЗНО допустимого для вступу, загальна тенденція зменшення населення і народжуваності негативно впливають на діяльність закладів вищої освіти (далі – ЗВО). Введення безвізового режиму спрощує абітурієнтам вступити на навчання у ЗВО ближнього зарубіжжя, де вартість оплати нижча за контрактом, ніж в Україні, а рівень життя і перспективи значно вищі. Стає актуальним проведення оцінки знань абітурієнтів з метою регулювання переліку дисциплін, які необхідно встановлювати під час вступу та відношення їх до Університету ДФС України як місця навчання.

Результати показали (табл. 1, 2), що 2019 року абітурієнти в середньому проставляють пріоритет 2,28 і довірчий інтервал:

$$2{,}28–0{,}88 \leq 2{,}28 \leq 2{,}28+1{,}68;$$

$$1{,}4 \leq 2{,}28 \leq 3{,}96.$$

Таблиця 1

Одновибіркова статистика

	N	**Середнє**	**Середньоквадратичне відхилення**		**Середньоквадратична похибка середнього**
Пріоритет	74	2,28	1,724		0,200

Джерело: розраховано авторами.

Таблиця 2

Одновибірковий критерій

	Значення критерію = 1					
	т	Ступені свободи	Значимість (двостороння)	Середня різниця	95 % довірчий інтервал для різниці	
					Нижня	Верхня
Пріоритет	6,405	73	0,000	1,284	0,88	1,68

Джерело: розраховано авторами.

[4] Вступ.ОСВІТА.UA. URL: https://vstup.osvita.ua/r11/3457/

У вибірку ввійшло 55 жінок і 23 чоловіка, які обрали ННІОАА. За допомогою критерію Лівіня знайдено (табл. 3, 4), що середній бал жінок 159,97, які вступили на навчання, вищий за чоловіків майже на десять балів і становить 149,99.

Таблиця 3

Статистика групи

	Стать	N	Середнє*	Середньоквадратичне відхилення	Середньо-квадратична похибка середнього
Бал	Жінка	55	159,973	12,565	1,694
	Чоловік	23	149,987	13,807	2,879

*Загальний бал вступника.

Джерело: розраховано авторами.

Таблиця 4

Критерій для незалежних вибірок

Бал	Критерій Лівіня		t-критерій для рівності середніх						
	F	Значимість	t	ст.св.	Знач. (двостороння)	Середня різниця	Середньоквадратична похибка різниці	95 % довірчий інтервал	
								Нижня	Верхня
Передбачається рівність дисперсій	0,017	0,898	3,109	76	0,003	9,986	3,212	3,588	16,384
Не передбачається			2,989	38,019	0,005	9,986	3,341	3,223	16,749

Джерело: розраховано авторами.

Оцінимо вплив статті на загальний рейтинговий бал абітурієнта за допомогою статистичних критеріїв Манна-Уітні і Вілкоксона (табл. 5, 6, 7).

Таблиця 5

Описова статистика

	N	Середнє	Середньокв. відхилення	Мінімум	Максимум	Процентилі		
						25	50-я	75-я
Бал	78	157,09	13,645	118,728	182,631	149,945	159,044	166,439

Джерело: розраховано авторами.

Таблиця 6

Ранги

	Стать	N	Середній ранг	Сума рангів
Бал	Жінка	55	44,18	2430,00
	Чоловік	23	28,30	651,00
	Усього	78		

Джерело: розраховано авторами.

Таблиця 7

Статистичні критерії[a]

Показник	Бал
U Манна-Уітні	375,000
W Вілкоксона	651,000
Z	-2,822
Асимптот. значимість (2-стороння)	0,005

а. Згруповано за змінною: стать.

Джерело: розраховано авторами.

Результати аналізу показують, що середній рейтинговий бал жінок 44,18, а чоловіків – 28,30. Оскільки розмір рівня значимості $p = 0{,}005 < 0{,}05$, можна стверджувати, що успішність жінок вища за успішності чоловіків і є статистично значимою. Аналізуючи описову статистику, необхідно зазначити, що лише 25 % абітурієнтів мають рейтинговий бал до 149,945, 50 % – до 159,044, всі інші – вищий, що вказує на високий рівень підготовки студентів першого курсу.

Проведено оцінку впливу на загальний рейтинговий бал абітурієнта оцінки ЗНО з математики і української мови та літератури (табл. 8, 9, 10).

Таблиця 8

Статистика парних вибірок

		Середнє	N	Середньоквадратичне відхилення	Середньоквадратична похибка середнього
Пара 1	Бал	157,029	78	13,645	1,545
	Математика	142,795	78	20,207	2,288
Пара 2	Бал	157,029	78	13,645	1,545
	Українська мова і література	161,590	78	18,592	2,105

Джерело: розраховано авторами.

Таблиця 9

Кореляції парних вибірок

		N	Кореляція	Значимість
Пара 1	Бал & Математика	78	0,769	0,000
Пара 2	Бал & Українська мова і література	78	0,812	0,000

Джерело: розраховано авторами.

Аналіз показав, що абітурієнти краще знають українську мову і літературу (середнє значення 161,59), ніж математику (середнє значення 142,795). Вплив результатів ЗНО з української мови на загальний рейтинговий бал більший (кореляція 0,812), ніж з математики (кореляція 0,769), при цьому обидва показники вказують на значний зв’язок.

Проведено оцінку впливу на загальний рейтинговий бал абітурієнта оцінки ЗНО з іноземної мови і географії (табл. 11, 12, 13).

Таблиця 10

Критерій парних вибірок

		Парні різності					т	ст. св.	Знач. (двосторонній)
		Середнє	Середньокв. відхилення	Середньокв. похибка середнього	95 % довірчий інтервал для різниць				
					Нижня	Верхня			
Пара 1	Бал – Математика	14,234	13,062	1,479	11,289	17,179	9,624	77	0,000
Пара 2	Бал – Українська мова і література	–4,561	10,948	1,240	–7,029	–2,093	–3,679	77	0,000

Джерело: розраховано авторами.

Таблиця 11

Статистика парних вибірок

Пара		Середнє	N	Середньоквадратичне відхилення	Середньоквадратична похибка середнього
1	Бал	163,887	17	11,676	2,832
	Іноземна мова	143,588	17	18,252	4,427
2	Бал	155,1175	61	13,622	1,744
	Географія	151,951	61	19,910	2,549

Джерело: розраховано авторами.

Таблиця 12

Кореляції парних вибірок

		N	Кореляція	Значимість
Пара 1	Бал & Іноземна мова	17	0,483	0,050
Пара 2	Бал & Географія	61	0,703	0,000

Джерело: розраховано авторами.

Таблиця 13

Критерій парних вибірок

		Парні різності					т	ст. св.	Знач. (двосторонн)
		Середнє	Середньокв. відхилення	Середньокв. похибка середнього	95 % довірчий інтервал для різниць				
					нижня	верхня			
Пара 1	Бал – Іноземна мова	20,299	16,242	3,939	11,948	28,650	5,15	16	,000
Пара 2	Бал – Географія	3,167	14,163	1,813	–0,461	6,794	1,75	60	,086

Джерело: розраховано авторами.

Абітурієнти краще знають географію (151,951), ніж іноземну мову (143,588). Вплив результатів ЗНО з географії на загальний рейтинговий бал сильний (0,703), а з іноземної мови слабкий (кореляція 0,483).

Отже, на перший курс ННІОАА вступили студенти з високим рівнем підготовки. При цьому жіночої половини майже в три рази більше і рівень їхньої підготовки вищий, ніж у чоловіків. Обираючи цей інститут, абітурієнти ішли цілеспрямовано, виставляючи середній пріоритет 2,28, на нашу думку, на їхнє рішення вплинула велика кількість бюджетних місць і якісна профорієнтаційна робота колективу інституту. Студенти краще знають українську мову, ніж математику, і краще географію, ніж іноземну мову.

РОЗДІЛ 4
ДИСПЕРСІЙНИЙ АНАЛІЗ У SPSS

4.1. Умови застосування однофакторного дисперсійного аналізу

Дисперсійний аналіз (ANOVA? Analysis of Variance) – це метод перевірки гіпотез про рівність трьох і більше середніх. Визначається вплив однієї (незалежної) змінної на іншу (залежну). При цьому незалежна змінна повинна бути номінальною або порядковою, і залежна – метричною.

У практичній діяльності застосування дисперсійного аналізу дає можливість відповісти на такі запитання:

1. Чи залежать вподобання споживачів до торгової марки від рівня їхнього доходу?

2. Чи залежать вподобання споживачів до торгової марки від виду рекламних заходів?

3. Чи залежать вподобання груп споживачів від місць придбання товару?

4. Чи впливає рівень освіти респондентів на рівень їхнього доходу?

5. Чи розрізняються географічні сегменти за товарними перевагами споживачів?

Особливості застосування основних критеріїв перевірки гіпотез в залежно від характеристики групи:

1) передбачається рівність дисперсії:

– якщо групи мають однаковий розмір, застосовуються QP-Э-Г-У або Тьюки (потужний критерій, коли багато груп);

– Шеффе слабший за Тьюки, застосовують для визначення різниці групах різного розміру;

– якщо важливим є контроль похибки I типу, коли у групах до 5 N – Бонферроні;

– якщо розмір груп слабо вирізняється – Габріель;

– якщо розмір груп сильно вирізняється – GT2 Гохберга;

– для порівняння з контрольною групою – Даннетт;

2) не передбачається рівність дисперсії:

– універсальним є метод Т2 Тамхейна;

– якщо важливим є контроль похибки І типу, застосовують Т3 Даннетт;

– якщо є сумніви у рівності дисперсій або за наявності груп з різною кількістю спостережень, застосовують критерій Геймс-Хоуелла.

Приклад. За даними 30 магазинів проаналізуємо взаємозв'язок між обсягами продаж і витратами на рекламу (табл. 4.1).

Нульова гіпотеза – в торговельних точках з різним рівнем витрат на рекламу обсяг продажу в середньому однаковий. Тобто не існує зв'язку між рівнем витрат на рекламу й обсягом продажу в магазині.

Таблиця 4.1

Рівень витрат на рекламу в магазинах

	Рівень витрат на рекламу (факторна ознака)		
№	**низький (1)**	**середній (2)**	високий (3)
Обсяг продажу, тис. грн (результативна ознака)			
1.	5	8	10
2.	7	8	9
3.	6	7	10
4.	4	9	8
5.	5	6	9
6.	2	4	8
7.	3	5	9
8.	2	5	7
9.	1	6	7
10.	2	4	6

Функція в SPSS: «**Анализ**» – «**Сравнение средних**» – «**Однофакторный дисперсионный анализ**».

Обираються залежна і незалежні змінні, які переміщуються в праві діалогові вікна за допомогою стрілки (рис. 4.1).

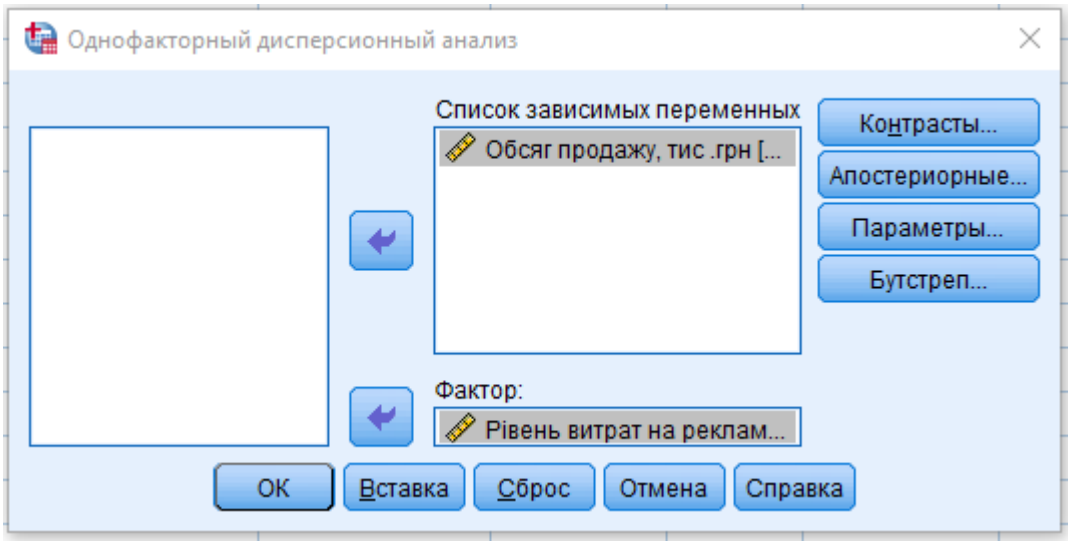


Рис. 4.1. Діалогове вікно SPSS

Обирається вкладка «Параметры» і ставляться мітки описова статистика, перевірка однорідності дисперсії, графік середніх (рис. 4.2).

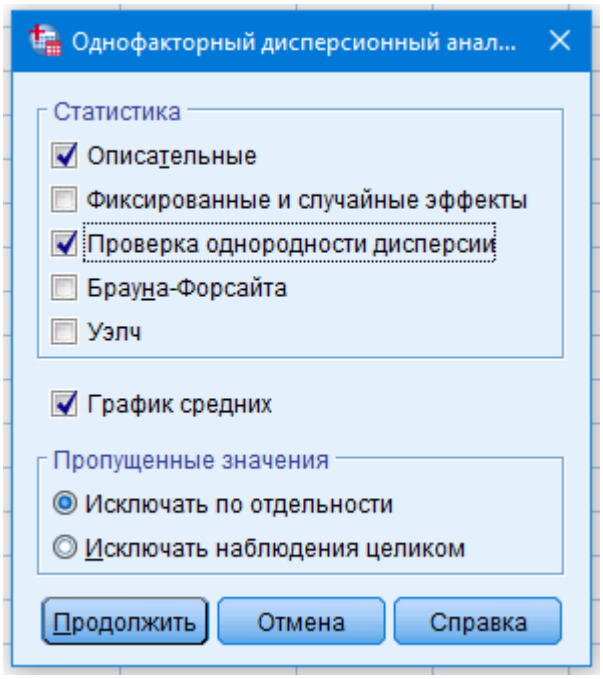


Рис. 4.2. Діалогове вікно SPSS

Натиснувши вкладку «Апостериорные», обираємо критерії перевірки гіпотез для двох випадків: у разі наявності (Тьюкі) або відсутності (Т2 Тамхейна) рівності дисперсії (рис. 4.3).

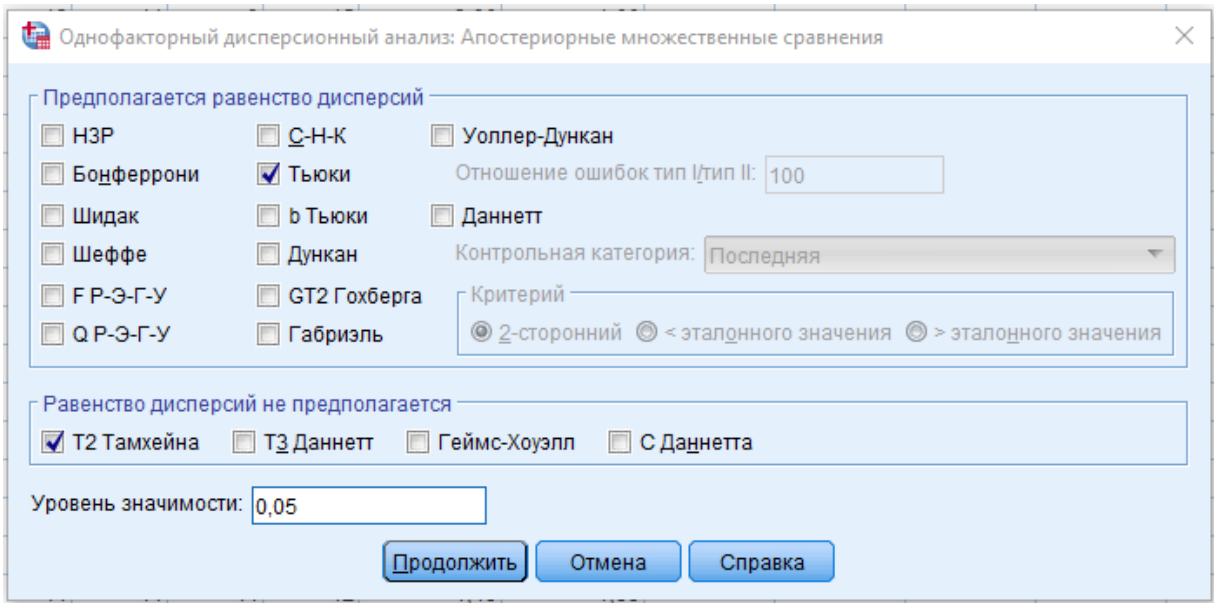


Рис. 4.3. Діалогове вікно SPSS

Натискаються клавіші «Продолжить» і «ОК».

Функція в SPSS: «**Анализ**» – «**Сравнение средних**» – «**Однофакторный дисперсионный анализ**».

Описова статистика застосовується для перевірки значущості сформованих груп. Кількість об'єктів дослідження в кожній групі повинна бути більше 2. Якщо у процесі дослідження є сформована група з однією відповіддю респондента, вона вважається незначущою і її вилучають (табл. 4.2).

Таблиця 4.2

Описова статистика обсягу продажу, тис. грн

	N	Середнє	Середньокв. відхилення	Стандартна похибка	95 % довірчий інтервал для середнього значення		Мінімум	Максимум
					Нижня межа	Верхня межа		
Низький	10	3,70	2,003	,633	2,27	5,13	1	7
Середній	10	6,20	1,751	,554	4,95	7,45	4	9
Високий	10	8,30	1,337	,423	7,34	9,26	6	10
Всього	30	6,07	2,532	,462	5,12	7,01	1	10

Усі групи мають практичну значимість, кількість респондентів у кожній групі 10.

Таблиця 4.3

Критерій однорідності дисперсій обсягу продажу, тис. грн

Статистика Лівіня	ст.св.1	ст.св.2	Значимість
1,353	2	27	0,275

Статистика Лівіня дає змогу перевірити гіпотезу «Дисперсії у порівнюваних групах рівні» (табл. 4.3). Значимість 0,275 означає, що гіпотеза може бути відхилена з ймовірністю похибки 27,5 %. Отже, гіпотеза не відхиляється, і це означає, що дисперсії рівні. Якщо значимість менше 0,05, гіпотеза може бути відхилена, тобто дисперсії нерівні.

Перевірка правильності нульової гіпотези проводиться за алгоритмом, наведеним у табл. 4.4, у магазинах з різним рівнем витрат на рекламу обсяг продажу в середньому однаковий. Не існує зв'язку між рівнем витрат на рекламу й обсягом продажу в магазині.

Таблиця 4.4

ANOVA

	Сума квадратів	**ст. св.**	**Середній квадрат**	**F**	**Значимість**
Міжгрупами	SS_B	k-1	$SS_B/k\text{-}1 = MS_B$	MS_B/MSw	
Всередині груп	SSw	n-k	SSw/n-k = MSw		
Усього	SS_B+SS_W	n-1			

Таблиця 4.5

Обсяг продажу, тис. грн

	Сума квадратів	**ст. св.**	**Середній квадрат**	**F**	**Значимість**
Міжгрупами	106,067	2	53,033	17,944	0,000
Всередині груп	79,800	27	2,956		
Усього	185,867	29			

Значимість F-критерію 0,000 вказує, що гіпотеза неправильна і повинна бути відхилена з ймовірністю 0 %. Зміна рівня витрат на рекламу впливає на зміну обсягу продаж у магазинах (табл. 4.5).

Множинні порівняння дають змогу визначити групи, в яких відмінності найбільш значні. У випадку, коли дисперсії рівні і групи мають однаковий розмір, аналізують дані за допомогою тесту Тьюки. У разі нерівності дисперсій застосовують результати тесту Т2 Тамхейна (табл. 4.6).

Таблиця 4.6

Множинні порівняння

Залежна змінна: Обсяг продажу, тис. грн

	(I) Рівень витрат на рекламу	(J) Рівень витрат на рекламу	Середня різниця (I-J)	Стандартна похибка	Значимість	95 % довірчий інтервал	
						Нижня межа	Верхня межа
Тьюки HSD	низький	середній	–2,500*	0,769	0,008	–4,41	–0,59
		високий	–4,600*	0,769	0,000	–6,51	–2,69
	середній	низький	2,500*	0,769	0,008	0,59	4,41
		високий	–2,100*	0,769	0,029	–4,01	–0,19
	високий	низький	4,600*	0,769	0,000	2,69	6,51
		середній	2,100*	0,769	0,029	0,19	4,01
Тамхейна	низький	середній	–2,500*	0,841	0,025	–4,72	–0,28
		високий	–4,600*	0,762	0,000	–6,63	–2,57
	середній	низький	2,500*	0,841	0,025	0,28	4,72
		високий	–2,100*	0,697	0,023	–3,95	–0,25
	високий	низький	4,600*	0,762	0,000	2,57	6,63
		середній	2,100*	0,697	0,023	0,25	3,95

** Середня різниця значима на рівні 0,05.*

У таблиці пари, які характеризуються значною різницею середніх, позначаються зірочкою. Кожна пара, позначена зірочкою, – це означає, що різниця між рівнем витрат на рекламу в магазині значна для кожної з груп. Результат достовірний, адже *p*-значення < 0,05.

Графік середніх величин дає змогу візуально побачити значимість різниці рівнів обсягу продажу в магазинах з різним рівнем витрат на рекламу (рис. 4.4).

Отже, будь-який рівень витрат на рекламу впливає на обсяги продажу в магазинах.

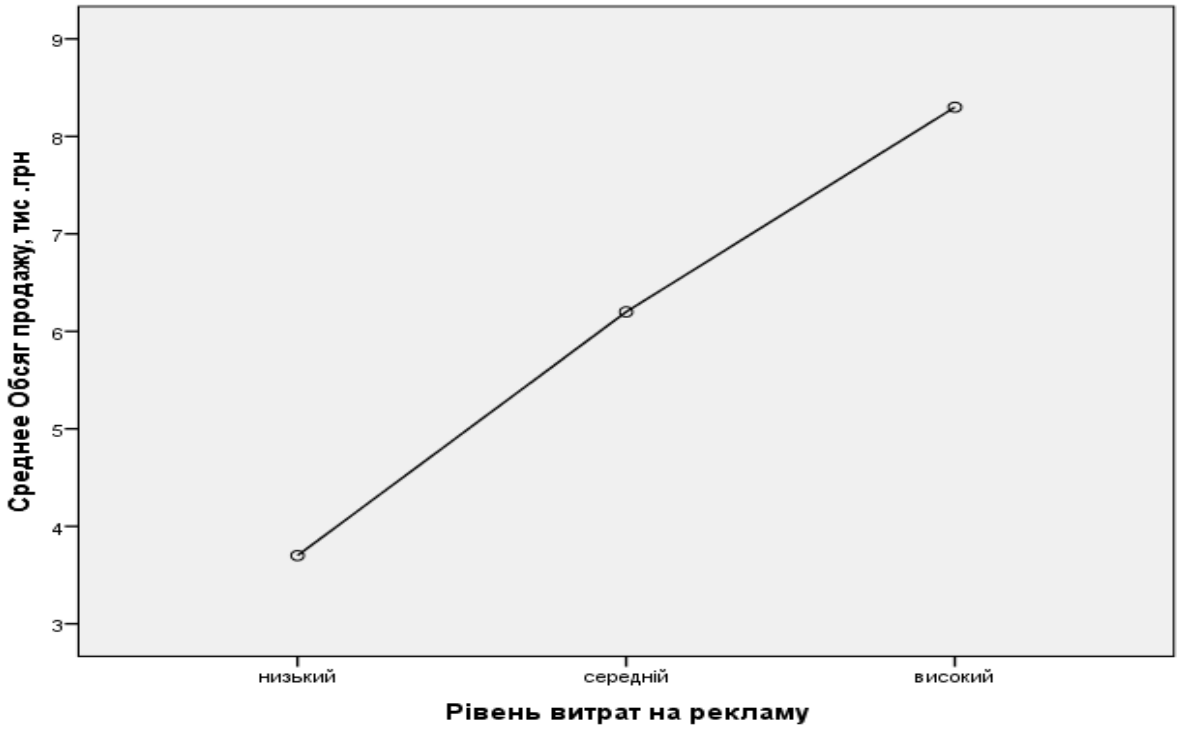


Рис. 4.4. Графік середніх величин значимості обсягу продажу в магазинах з різним рівнем витрат на рекламу

Приклад. Визначити залежність вміння школярів рахувати в умі від їх хобі (додаток Е).

Функція в SPSS: «**Анализ**» – «**Сравнение средних**» – «**Однофакторный дисперсионный анализ**» (рис. 4.5).

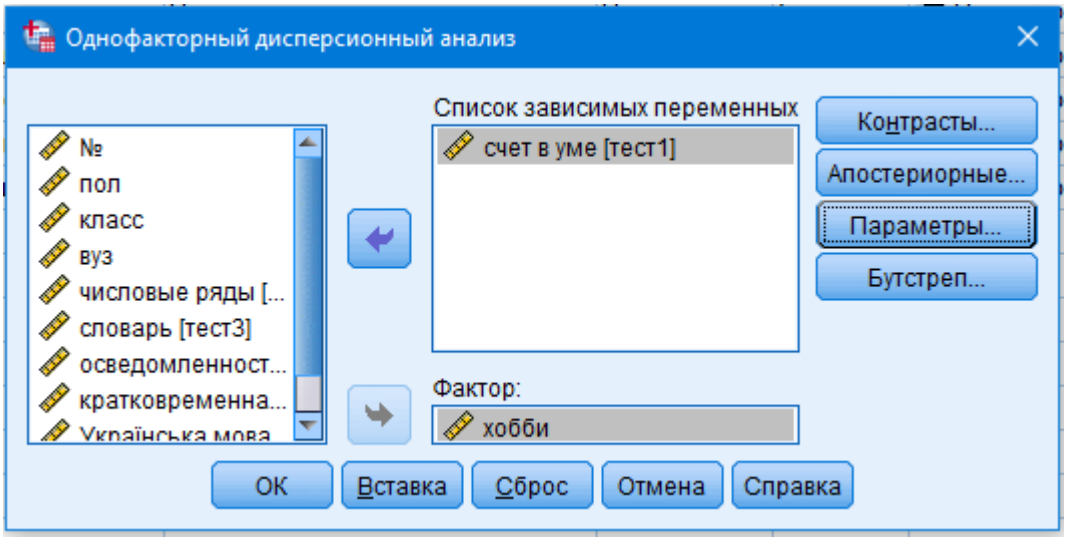


Рис. 4.5. Діалогове вікно SPSS

Розраховується аналогічно як вищеописаний, але оскільки групи мають різну кількість спостережень для оцінки рівності дисперсій обирається критерій Шеффе (4.7–4.11).

Таблиця 4.7

Описова статистика рахунку в умі у розрізі хобі

	N	**Середнє**	**Середньокв. відхилення**	**Стандартна похибка**	**95 % довірчий інтервал для середнього значення**		**Мінімум**	**Максимум**
					Нижня межа	**Верхня межа**		
Спорт	33	9,67	2,582	0,449	8,75	10,58	4	14
Комп’ютер	37	11,43	2,102	0,346	10,73	12,13	7	17
Мистецтво	30	9,43	2,885	0,527	8,36	10,51	5	15
Усього	100	10,25	2,653	0,265	9,72	10,78	4	17

Таблиця 4.8

Критерій однорідності дисперсій

Рахунок в умі			
Статистика Лівіня	ст.св.1	ст.св.2	Значимість
1,858	2	97	0,161

Таблиця 4.9

ANOVA

Рахунок в умі					
	Сума квадратів	**ст.св.**	**Середній квадрат**	**F**	**Значимість**
Між групами	82,969	2	41,484	6,556	0,002
Всередині груп	613,781	97	6,328		
Всього	696,750	99			

Таблиця 4.10

Множинні порівняння

Залежна змінна: Рахунок в умі

	(I) Хобі	(J) Хобі	Середня різниця (I-J)	Стандартна похибка	Значимість	95 % довірчий інтервал	
						Нижня межа	Верхня межа
Шеффе	Спорт	комп’ютер	–1,766*	0,602	0,016	–3,26	–0,27
		мистецтво	,233	0,635	0,935	–1,34	1,81

	Комп’ютер	спорт	1,766*	0,602	0,016	0,27	3,26
		мистецтво	1,999*	0,618	0,007	0,46	3,54
	Мистецтво	спорт	–0,233	0,635	0,935	–1,81	1,34
		комп’ютер	–1,999*	0,618	0,007	–3,54	-0,46
Тамхейн	Спорт	комп’ютер	–1,766*	0,567	0,008	–3,16	-0,37
		мистецтво	,233	0,692	0,982	–1,47	1,94
	Комп’ютер	спорт	1,766*	0,567	0,008	0,37	3,16
		мистецтво	1,999*	0,630	0,008	0,44	3,55
	Мистецтво	спорт	–0,233	0,692	0,982	–1,94	1,47
		комп’ютер	–1,999*	0,630	0,008	–3,55	–0,44

** Середня різниця значима на рівні 0,05.*

Таблиця 4.11

Рахунок в умі

			Підмножина для альфа = 0,05	
	Хобі	N	1	2
Шеффе[a,b]	Мистецтво	30	9,43	
	Спорт	33	9,67	
	Комп’ютер	37		11,43
	Значимість		0,931	1,000

Захоплення школярів комп’ютером (11,43) суттєво вирізняється серед інших двох захоплень (рис. 4.6).

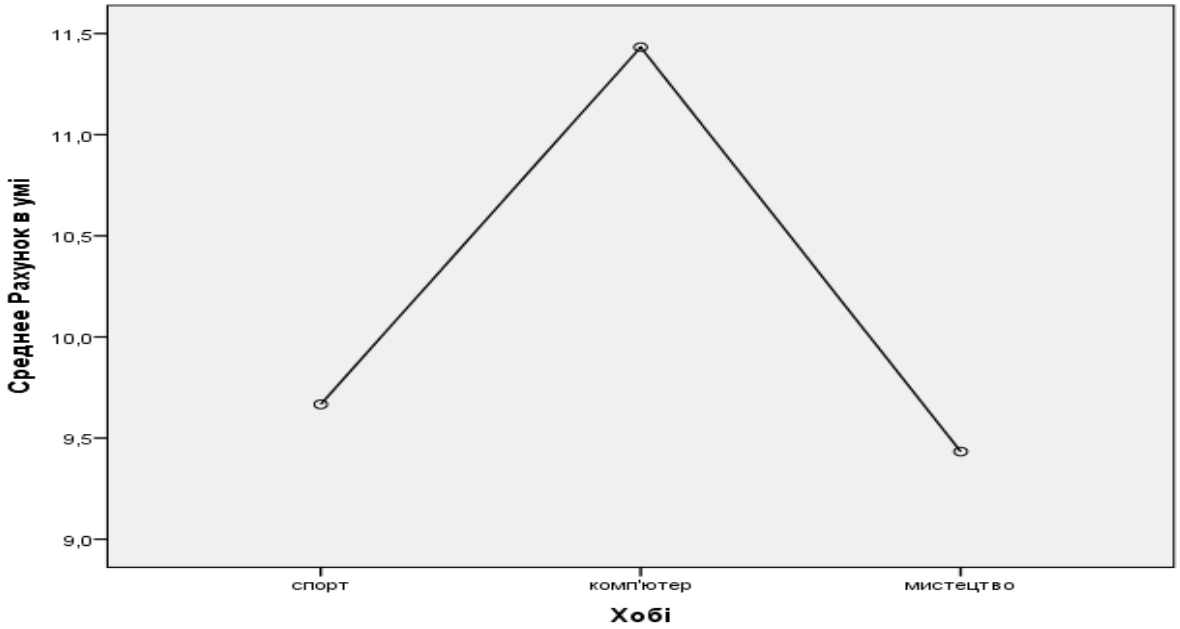


Рис. 4.6. Графік середніх величин хобі з рівнем рахувати в умі

Висновок: статистика Лівіня вказує, що дисперсії рівні (*p*-значимість 0,161 > 0,05). Значимість F-критерія 0,002 вказує, що гіпотеза не правильна і повинна бути відхилена з ймовірністю 0,2 %. Хобі школярів впливає на їхню здатність рахувати в умі. Результати розрахунку школярів в умі, для тих, хто захоплюється комп'ютером, статистично більші, ніж для тих, хто захоплюється спортом і мистецтвом. При цьому ті, хто захоплюється спортом і мистецтвом за результатами розрахунку в умі статистично не вирізняються. Зірочкою відмічені пари вибірок, для яких різниця середніх значень статистично достовірна (*p* – значимість < 0,05).

4.2. Двофакторний дисперсійний аналіз

Приклад. У ході маркетингових заходів розроблено два типи рекламного ролику: гумористичний і стандартний. Реклама розміщувалася у вихідні і робочі дні. Для дослідження обрано 16 потенційних клієнтів, які були випадково розподілені на 4 групи (1 група дивилася гумористичний ролик у робочий день, 2 – гумористичний ролик у вихідний день, 3 – стандартний ролик у робочий день, 4 – стандартний ролик у вихідний день) (табл. 4.12). Кожний респондент переглядав ролик, і оцінювали ефективність реклами за 20-бальною системою. Необхідно оцінити ефективність рекламних заходів з використанням двостороннього дисперсійного аналізу при α = 0,01.

Таблиця 4.12

Оцінка респондентами реклами

	День тижня:	
Тип ролику:	1) робочий	2) вихідний
1) гумористичний	6, 10, 11, 9	15, 18, 14, 16
2) стандартний	8, 13, 12, 10	19, 20, 13, 17

Цей двофакторний дисперсійний аналіз (табл. 4.13) має декілька нульових гіпотез: одна для кожної незалежної змінної та одна для взаємозв'язку.

H_0 – тип ролика і день не мають ефекту взаємодії на ефективність реклами.

H_1 – тип ролика і день мають ефект взаємодії на ефективність реклами.

H_0 – ефективність реклами не залежить від типу ролика.

H_1 – ефективність реклами залежить від типу ролика.

H_0 – ефективність реклами не залежить від дня тижня.

H_1 – ефективність реклами залежить від дня тижня.

Таблиця 4.13

Алгоритм двофакторного дисперсійного аналізу

	Сума квадратів	**Ступені свободи**	**Середнє квадратичне відхилення**	**F**
Фактор А	SS_A	a-1	MS_A	F_A
Фактор В	SS_B	b-1	MS_B	F_B
Взаємодія, АхВ	SS_{AxB}	(a-1)(b-1)	MS_{AxB}	F_{AxB}
Похибка	SS_e	ab(n-1)	MS_e	X
Всього			X	X

SS_A – сума квадратів для фактора А;

SS_B – сума квадратів для фактора В;

SS_{AxB} – сума квадратів для взаємодії факторів;

SS_e – сума квадратів для похибки;

a – кількість рівнів фактора А;

b – кількість рівнів фактора В;

n – кількість об’єктів у кожній групі.

Функція в SPSS: «**Анализ**» – «**Общая линейная модель**» – «**ОЛМ–одномерная**».

Обираються залежна і фіксовані фактори з лівого діалогового вікна і за допомогою стрілок переноситься в праві (рис. 4.7).

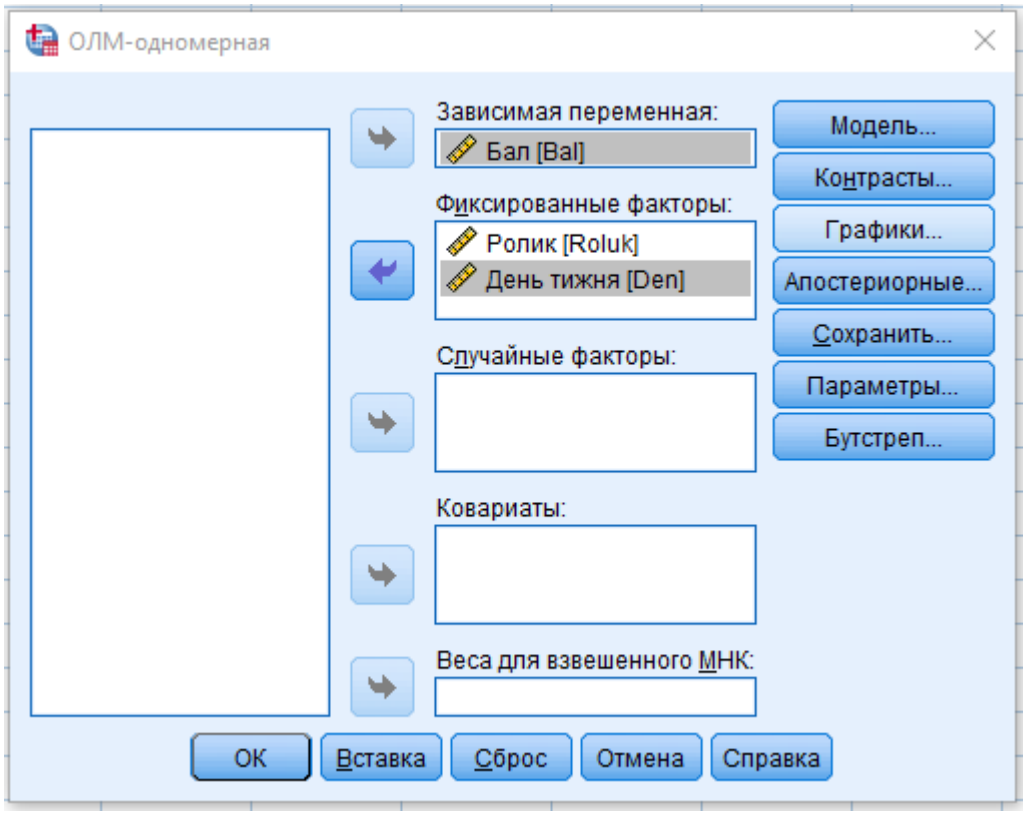


Рис. 4.7. Діалогове вікно SPSS

У вкладці «Графіки» обирають характеристики візуального відображення (рис. 4.8, 4.9).

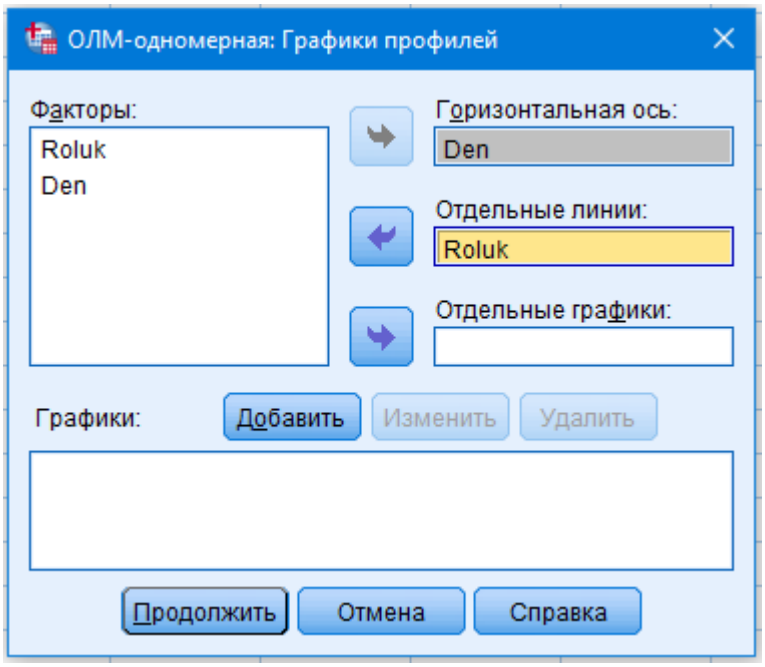


Рис. 4.8. Діалогове вікно SPSS

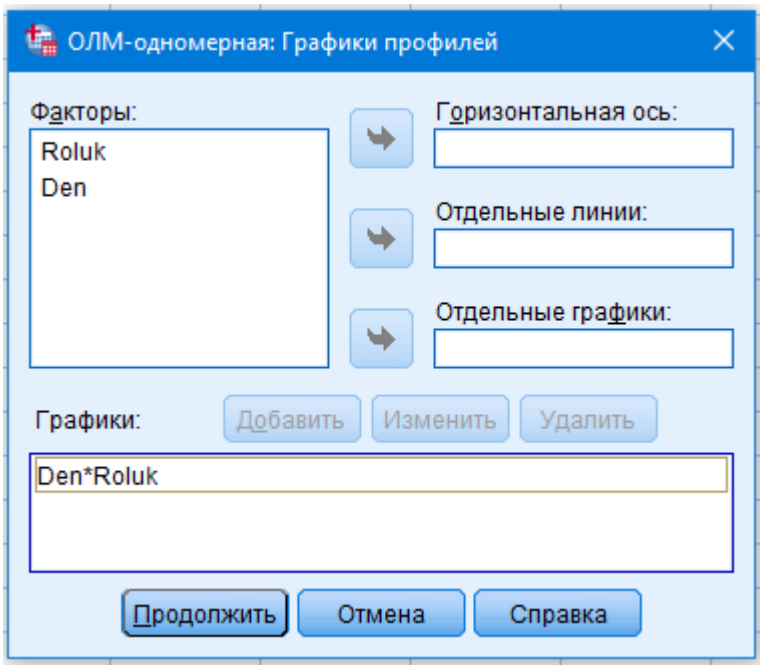


Рис. 4.9. Діалогове вікно SPSS

Натискають вкладку «Параметри» і переміщують у праве діалогове вікно OVERALL (уся модель), ставлять мітки за основними характеристиками, а саме: описова статистика, оцінки розмірів ефекту, критерії однорідності. Відповідно до умови проставляють рівень значимості α = 0,01 (рис. 4.10).

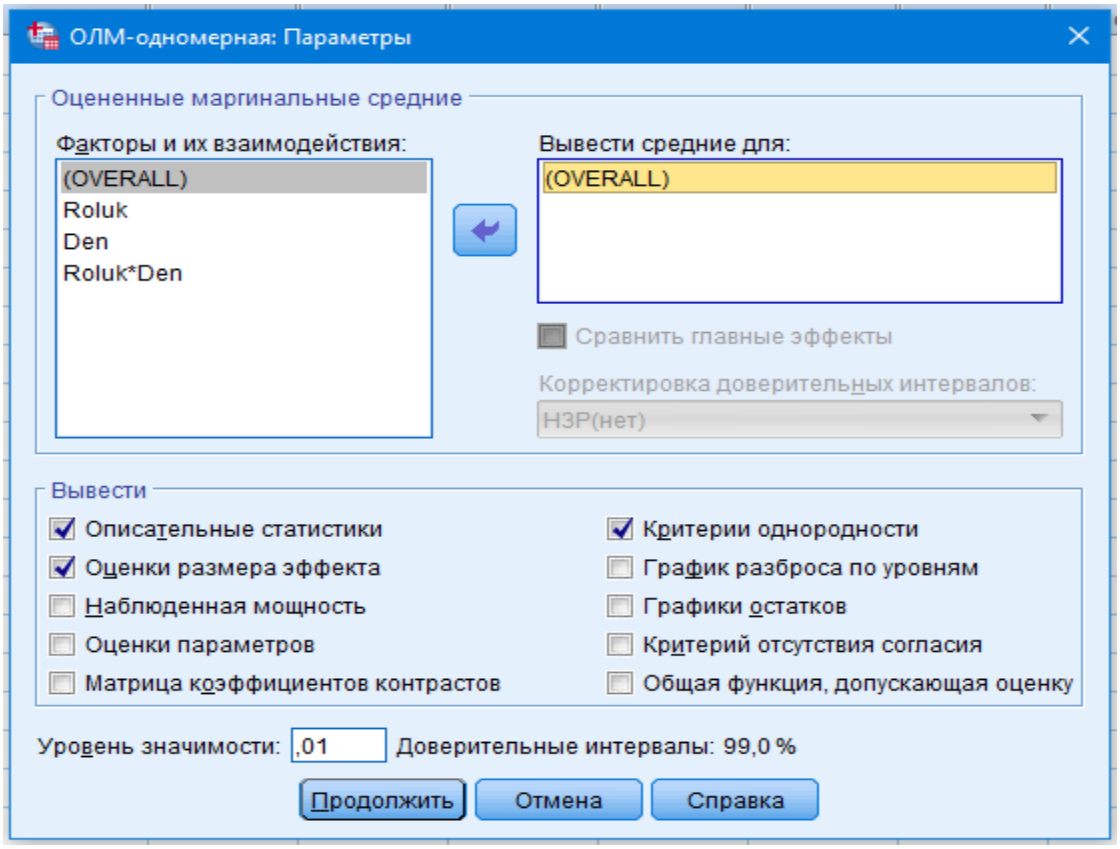


Рис. 4.10. Діалогове вікно SPSS

Натискають клавіші «Продолжить» і «ОК». Результати аналізу наведенні в таблицях 4.14–4.18.

Таблиця 4.14

Міжгрупові фактори

		Мітка значення	N
Ролик	1	Гумористичний	8
	2	Стандартний	8
День тижня	1	Робочий	8
	2	Вихідний	8

Таблиця 4.15

Описова статистика

Залежна змінна: Бал

Ролик	День тижня	Середнє	Стандартна похибка	N
гумористичний	Робочий	9,00	2,160	4
	Вихідний	15,75	1,708	4
	Всього	12,38	4,033	8
стандартний	Робочий	10,75	2,217	4
	Вихідний	17,25	3,096	4
	Всього	14,00	4,276	8
Всього	Робочий	9,88	2,232	8
	Вихідний	16,50	2,449	8
	Усього	13,19	4,102	16

Таблиця 4.16

Критерій рівності дисперсій похибок Лівіня[a]

Залежна змінна: Бал

F	ст.св.1	ст.св.2	Значимість
0,473	3	12	0,707

Перевіряє нульову гіпотезу, що дисперсія похибок залежної змінної однакова за групами.

a. Структура: Вільний член + Roluk + Den + Roluk * Den.

Таблиця 4.17

Критерії міжгрупових ефектів

Залежна змінна: Бал

Джерело	Сума квадратів типу III	ст. св.	Середній квадрат	F	Значи-мість	Часткова ета-квад-рат
Скоригована модель	186,188[a]	3	62,063	11,242	0,001	0,738
Вільний член	2 782,563	1	2 782,563	504,011	0,000	0,977
Roluk	10,563	1	10,563	1,913	0,192	0,138
Den	175,563	1	175,563	31,800	0,000	0,726
Roluk * Den	0,063	1	0,063	0,011	0,917	0,001
Похибка	66,250	12	5,521			
Усього	3 035,000	16				
Скорегований підсумок	252,438	15				

a. R-квадрат = ,738 (Скоригований R-квадрат = ,672).

Таблиця 4.18

Загальне середнє

Залежна змінна: Бал

Середнє	Стандартна похибка	99 % довірчий інтервал	
		Нижня межа	Верхня межа
13,188	0,587	11,393	14,982

Висновок: статистика Лівіня вказує, що дисперсії рівні (*p*-значимість 0,707 > 0,05). Дослідження гіпотез:

H_0 – тип ролика і день не мають ефекту взаємодії на ефективність реклами.

H_1 – тип ролика і день мають ефект взаємодії на ефективність реклами.

Значимість F-критерія 0,917 вказує, що H_0 гіпотеза правильна і не відхиляється з імовірністю 91,7 %. Тип ролика і день не мають ефекту взаємодії на ефективність реклами.

H_0 – ефективність реклами не залежить від типу ролика.

H_1 – ефективність реклами залежить від типу ролика.

Значимість F-критерія 0,192 вказує, що H_0 гіпотеза правильна і не відхиляється з імовірністю 19,2 %. Тип ролика не має впливу на ефективність реклами.

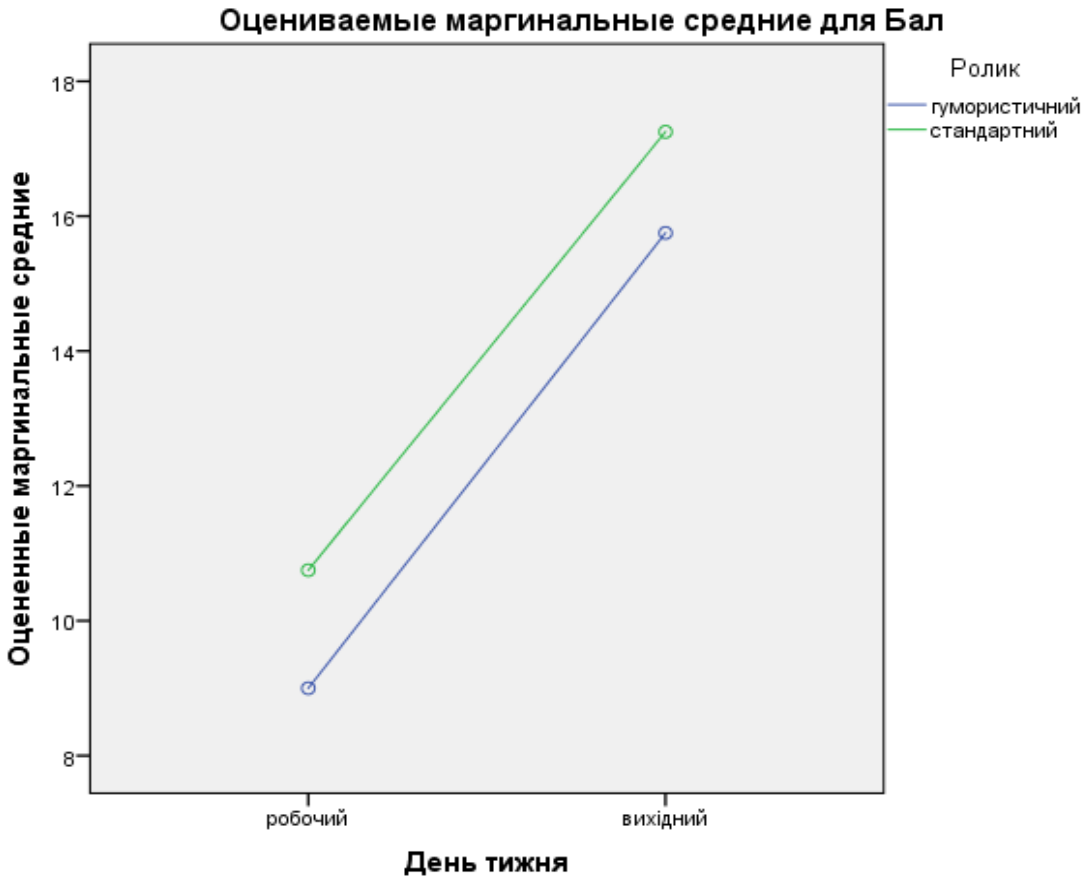


Рис. 4.11. Маргінальні середні в оцінці ефективності реклами

H_0 – ефективність реклами не залежить від дня тижня.

H_1 – ефективність реклами залежить від дня тижня.

Значимість F-критерія 0,000 вказує, що H_0 гіпотеза неправильна і ми її відхиляємо з імовірністю 0 %. Приймаємо гіпотезу H_1 ефективність реклами залежить від дня тижня. Аналізуючи маргінальні середні для ефективності реклами, можемо стверджувати, що стандартний ролик оцінюється вище (рис. 4.11).

4.3. Дисперсійний аналіз з трьома факторами

Приклад. Дослідити ступінь впливу факторних змінних *x*: стать, хобі на результативну ознаку *y*: знання з математики (додаток Е). Виявити:

– чи існує різниця між знаннями дівчат і хлопців;

– чи різняться оцінки з математики між групами захоплень школярів;

– чи існує вплив факторних ознак між собою.

Функція в SPSS: «**Анализ**» – «**Общая линейная модель**» – «**ОМЛ-одномерная**» (рис. 4.12–4.14).

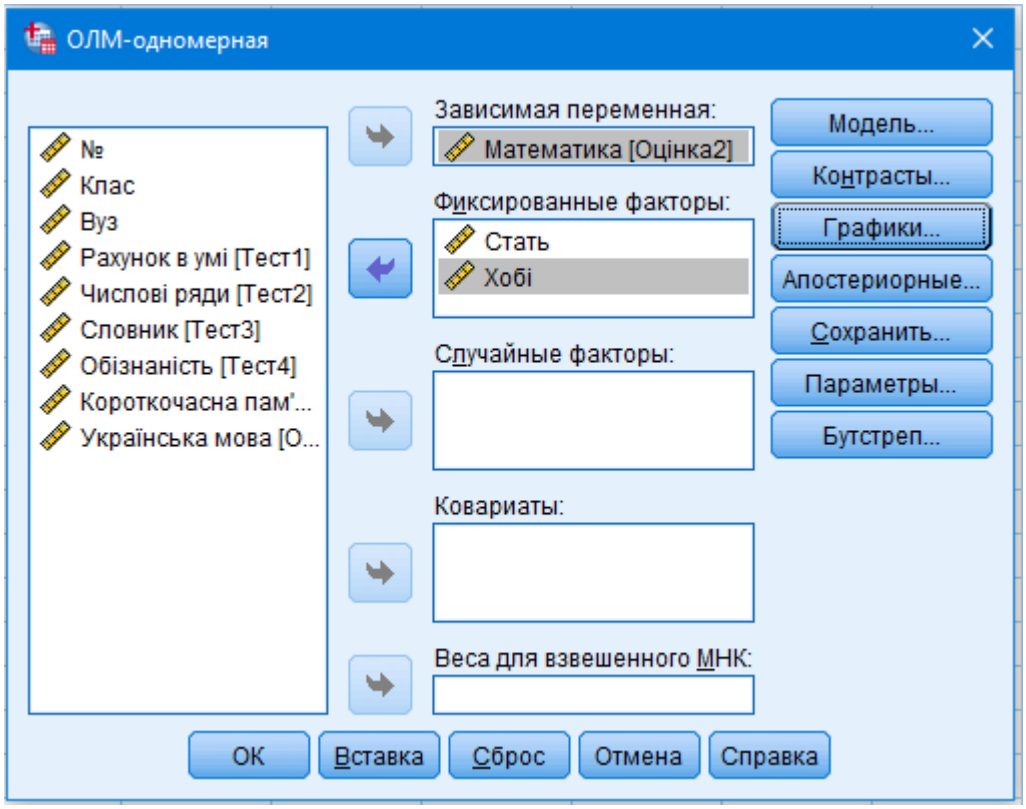


Рис. 4.12. Діалогове вікно SPSS

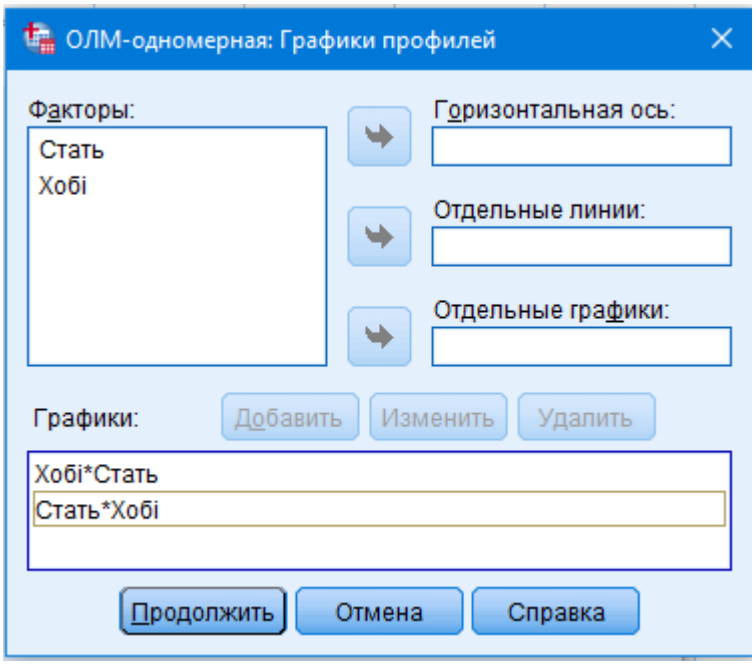


Рис. 4.13. Діалогове вікно SPSS

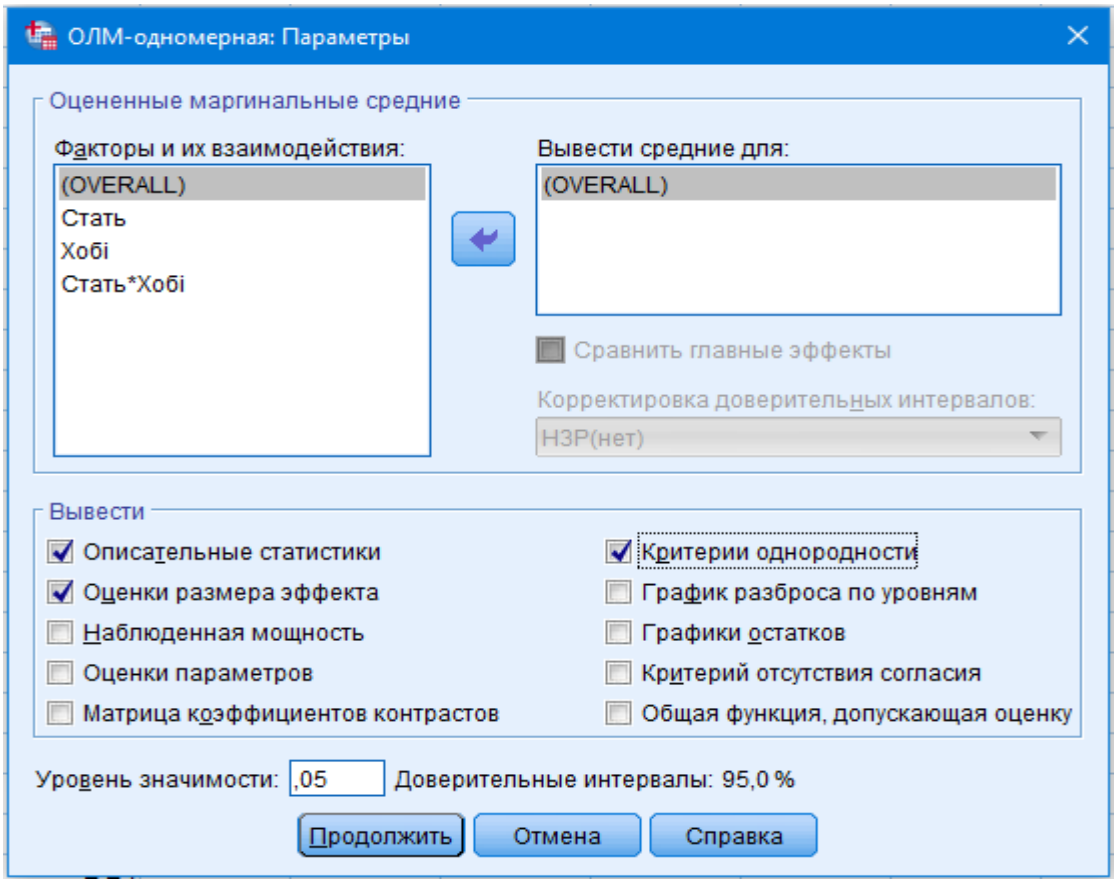


Рис. 4.14. Діалогове вікно SPSS

Натиснувши клавіші «Продолжить» і «ОК», отримаємо результати дослідження в таблицях 4.19–4.22.

Таблиця 4.19

Міжгрупові фактори

		Мітка значення	N
Стать	1	Жінка	61
	2	Чоловік	39
Хобі	1	Спорт	33
	2	Комп'ютер	37
	3	Мистецтво	30

Таблиця 4.20

Описова статистика

Залежна змінна: Математика

Стать	Хобі	Середнє	Стандартне відхилення	N
ЖІН.	Спорт	4,2267	0,24339	15
	Комп'ютер	4,2474	0,28552	19
	Мистецтво	4,3278	0,27080	27
	Всього	4,2779	0,26856	61

ЧОЛ.	Спорт	4,0000	0,25896	18
	Комп'ютер	4,2667	0,22361	18
	Мистецтво	4,1000	0,17321	3
	Всього	4,1308	0,26622	39
Усього	Спорт	4,1030	0,27327	33
	Комп'ютер	4,2568	0,25390	37
	Мистецтво	4,3050	0,26953	30
	Усього	4,2205	0,27589	100

Таблиця 4.21

Критерій рівності дисперсій похибок Лівіня[a]

Залежна змінна: Математика

F	ст.св.1	ст.св.2	Значимість
0,572	5	94	0,721

Перевіряє нульову гіпотезу, що дисперсія похибок залежної змінної дорівнює за групами.

а. Структура: Вільний член + Стать + Хобі + Стать * Хобі.

Таблиця 4.22

Результати дисперсійного аналізу

Джерело	Сума квадратів типу III	ст. св.	Середній квадрат	F	Значи-мість
Скорегована модель	1,282[a]	5	0,256	3,854	0,003
Вільний член	1 054,384	1	1 054,384	1 5849,389	0,000
Стать	0,315	1	0,315	4,738	0,032
Хобі	0,364	2	0,182	2,735	0,070
Стать * Хобі	0,303	2	0,152	2,278	0,108
Похибка	6,253	94	0,067		
Усього	1 788,798	100			
Скорегований підсумок	7,535	99			

$R^2 = 0,170$ (Скорегований $R^2 = 0,126$).

Висновок: змінна стать має статистично достовірний вплив на знання школярів з математики (середнє значення для чоловіків 4,13, а жінок – 2,28). Критерій Фішера 4,74 значимий ($p = 0,032 < 0,05$). Змінна хобі не впливає на результати з математики школярів,

оскільки $p = 0{,}07 > 0{,}05$. Також не виявлено статистично достовірного взаємозв'язку між незалежними змінними статті і хобі ($p = 0{,}108 < 0{,}05$). Коефіцієнт детермінації $R^2 = 0{,}170$ вказує на ті результати з математики школярів, які лише на 17 % залежать від хобі і статті учнів, і на 83 % від інших чинників, включаючи стохастичну змінну. Середній бал оцінок для дівчат майже не залежить від захоплень. При цьому хлопці, що захоплюються комп'ютером, мають вищий бал з математики (рис. 4.15).

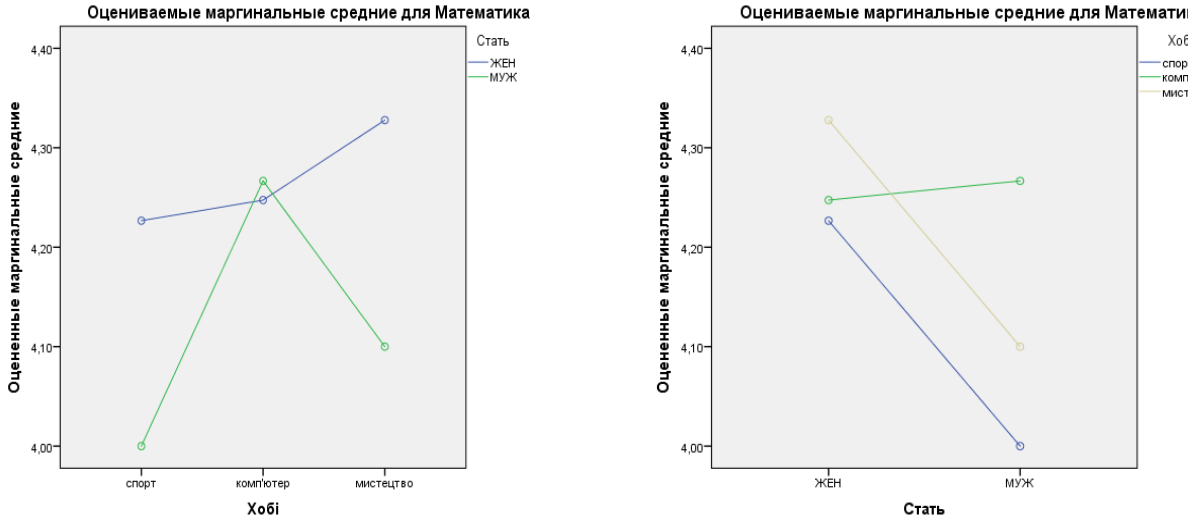


Рис. 4.15. Маргінальні середні величини для математики за статтю і хобі

4.4. Асоціативний аналіз

У випадку аналізу незалежної (факторної) і залежної (результативної) змінних, які представлені за допомогою номінальних шкал, застосовують асоціативний аналіз. Первинну інформацію групують у таблицю зв'язаності. Як правило, вона складається для двох і більше ознак і містить частоти для кожного набору значень. Під час проведення аналізу дослідник має дати відповідь на запитання:

1. Чи існує взаємозв'язок між номінальними змінними?
2. Якщо існує, то яка його сила?
3. Який характер і направлення зв'язку?

Приклад. Дослідити, чи існує залежність між вподобаннями школярів і їх статтю. Якщо існує, визначити її силу і напрям. Опитано 100 школярів, які мають такі вподобання: спорт, комп'ютер і мистецтво (додаток Е).

Функція в SPSS: «**Анализ» – «Описательная статистика» – «Таблицы сопряженности».**

Обираються дані згідно з умовою задачі (рис. 4.16).

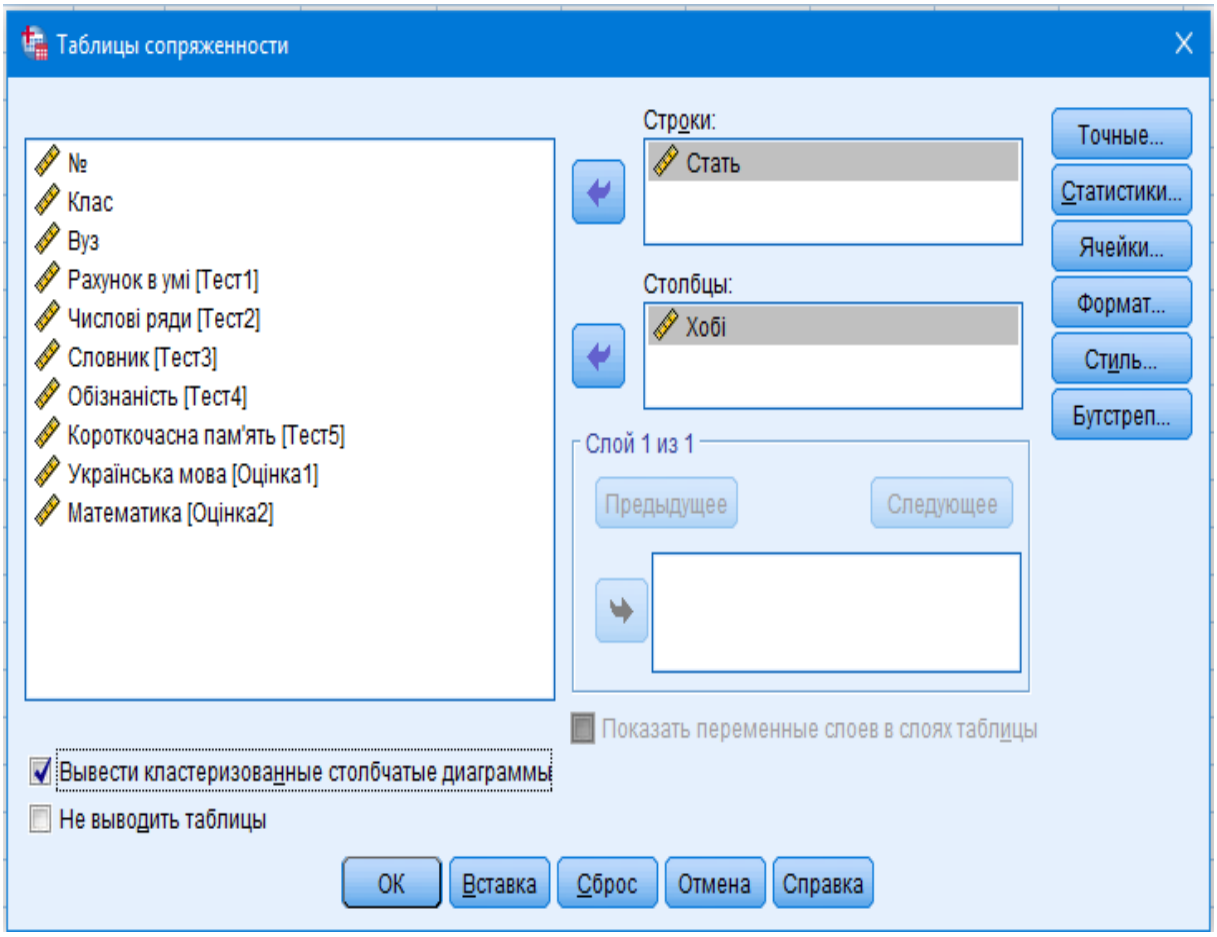


Рис. 4.16. Діалогове вікно SPSS

У вкладці «Статистики» ставляться мітки *Хі*-квадрат, *Фі* і V Крамера, кореляції. Натискається клавіша «Продолжить» (рис. 4.17).

У вкладці «Ячейки» обираються характеристики таблиці пов'язаності, кількість спостережень, очікуванні значення, відсотки та нестандартні залишки (рис. 4.18).

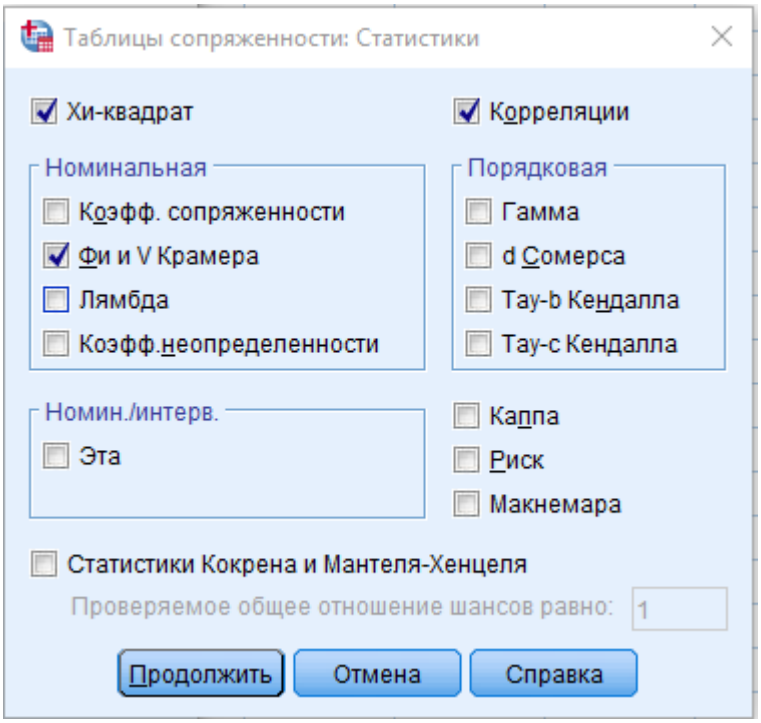


Рис. 4.17. Діалогове вікно SPSS

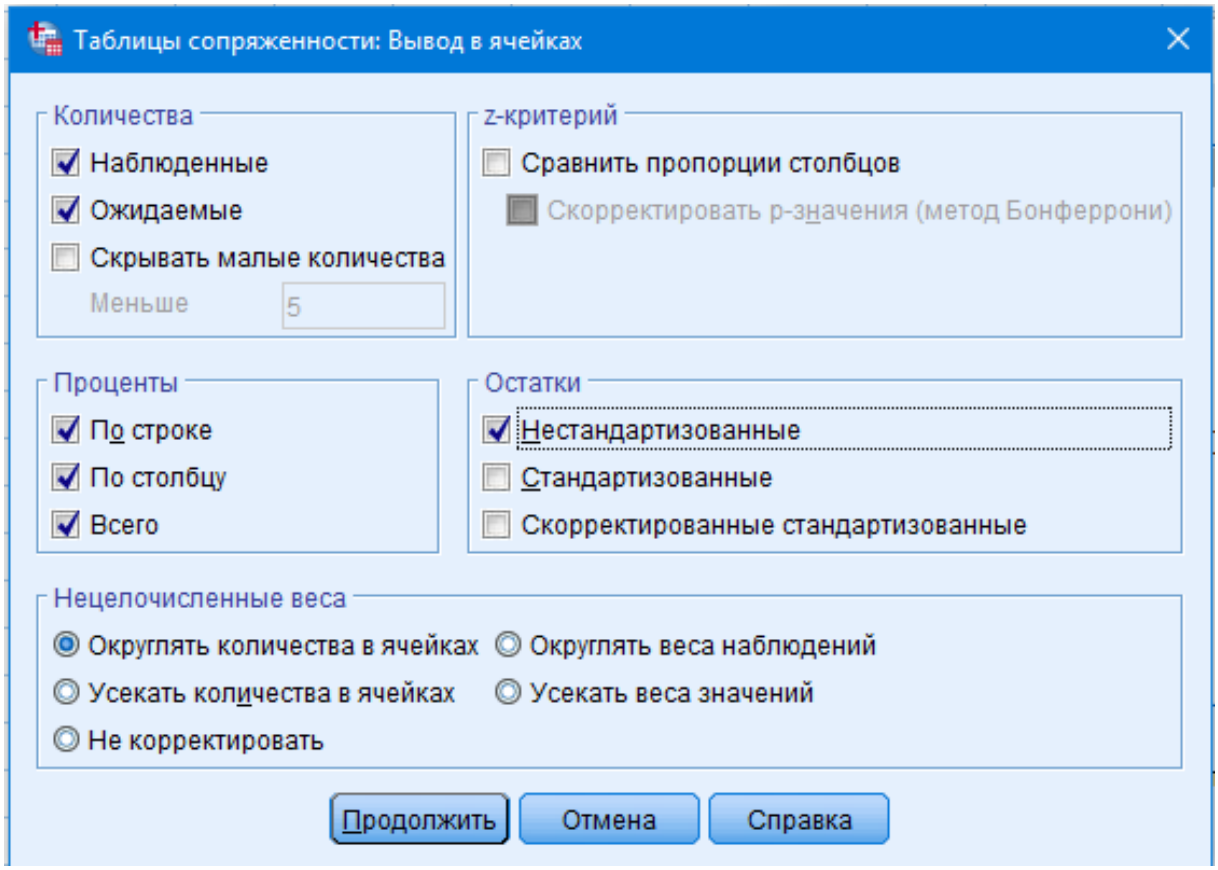


Рис. 4.18. Діалогове вікно SPSS

Натискаються клавіші «Продолжить» і «ОК».

Таблиця 4.23

Таблиця пов'язаності

Факторна ознака	Хобі (результативна ознака)			Усього
	Спорт	Комп'ютер	Мистецтво	
Кількість жінок	15	19	27	61
Кількість чоловіків	18	18	3	39
Усього	33	37	30	100

Функція в SPSS: «**Анализ» – «Описательная статистика» – «Таблицы сопряженности»** (табл. 4.24–4.25).

На першому етапі перевіряється валідність результатів анкетування (кількість наданих відповідей) (табл. 4.24).

Таблиця 4.24

Зведений звіт за спостереженнями

	Спостереження					
	Валідні		Пропущені		Усього	
	N	Проценти	N	Проценти	N	Проценти
Стать * Хобі	100	100,0 %	0	0,0 %	100	100,0 %

Таблиця 4.25

Комбінаційна таблиця Стать*Хобі

			Хобі			Усього
			Спорт	Комп'ютер	Мистецтво	
Стать	ЖІН.	Кількість	15	19	27	61
		Очікувана кількість	20,1	22,6	18,3	61,0
		% в Стать	24,6 %	31,1 %	44,3 %	100,0 %
		% в Хобі	45,5 %	51,4 %	90,0 %	61,0 %
		% загального підсумку	15,0 %	19,0 %	27,0 %	61,0 %
		Залишок	–5,1	-3,6	8,7	
	ЧОЛ.	Кількість	18	18	3	39
		Очікувана кількість	12,9	14,4	11,7	39,0
		% в Стать	46,2 %	46,2 %	7,7 %	100,0 %
		% в Хобі	54,5 %	48,6 %	10,0 %	39,0 %
		% загального підсумку	18,0 %	18,0 %	3,0 %	39,0 %
		Залишок	5,1	3,6	-8,7	
Усього		Кількість	33	37	30	100
		Очікувана кількість	33,0	37,0	30,0	100,0

	% в Стать	33,0 %	37,0 %	30,0 %	100,0 %
	% в Хобі	100,0 %	100,0 %	100,0 %	100,0 %
	% загального підсумку	33,0 %	37,0 %	30,0 %	100,0 %

Таблиця 4.26

Критерії хі-квадрат

	Значення	**ст.св.**	**Асимптотична значимість (2-сторонння)**
Хі-квадрат Пірсона	15,405[a]	2	0,000
Відносини правдоподібності	17,504	2	0,000
Лінійно-лінійний зв'язок	12,652	1	0,000
Кількість допустимих спостережень	100		

a. Для кількості комірок 0 (0,0 %) пропонується значення, менше 5. Мінімальне очікуване число дорівнює 11,70.

Таблиця 4.27

Симетричні міри

		Значення	**Асимптотична середньоквадратична похибка[a]**	**Приблизна T[b]**	**Приблизна значимість**
Номінал	Фи	0,392			0,000
	V Крамера	0,392			0,000
	Коефіцієнти зв'язаності	0,365			0,000
Інтервал	R Пірсона	–0,357	0,085	–3,789	0,000[c]
Порядковий	Кореляція Спірмена	–0,355	0,087	–3,763	0,000[c]
Кількість допустимих спостережень		100			

a. Не передбачає нульової гіпотези.
b. Використання асимптотичної середньоквадратичної похибки в припущенні нульової гіпотези.
c. Обґрунтовано на нормальній апроксимації.

Для перевірки статистичних гіпотез застосовують непарамет¬ричні критерії, а саме $\chi 2$. Умови застосування якого є: кількість спостережень повинна бути більша 5, а вибірка випадковою.

Окремо його можна розрахувати, застосувавши функцію в SPSS: «Анализ» – «Непараметрические критерии» – «Устаревшие диалоговые окна» – «Хи-квадрта». Аналіз таблиці 4.26 пов’язаності також дозволяє його отримати.

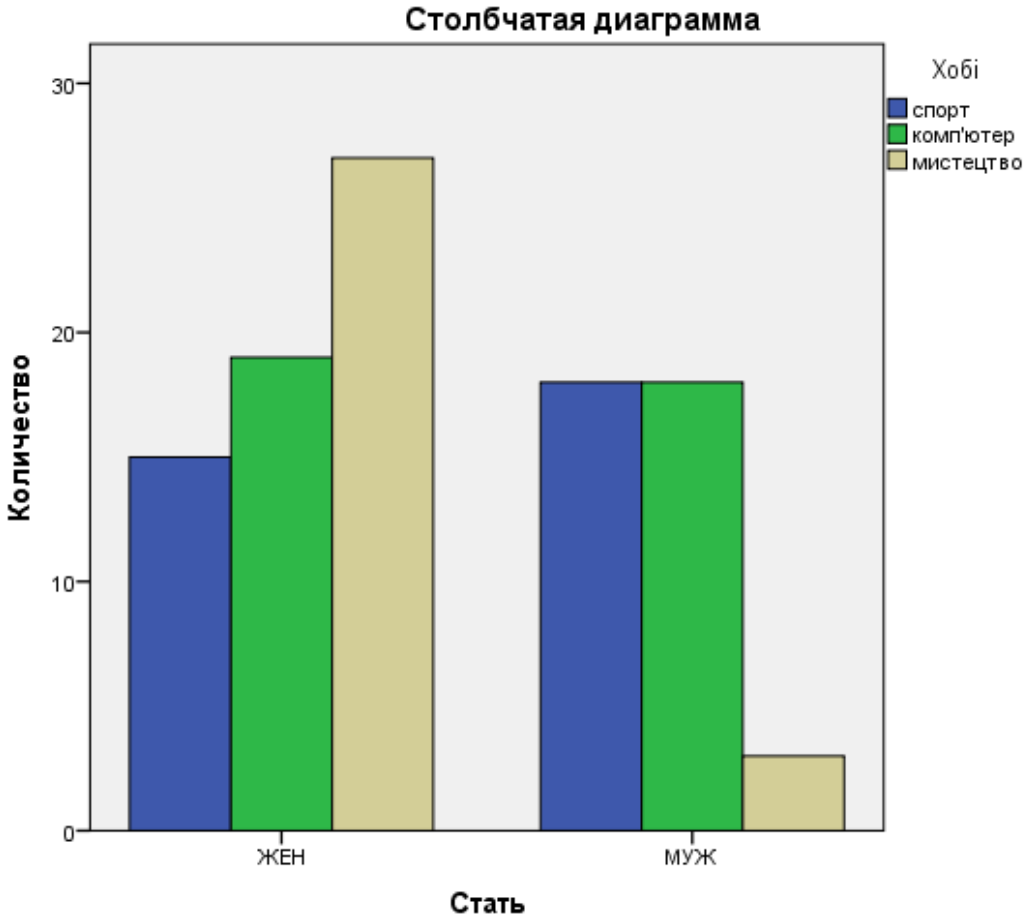


Рис. 4.19. Візуальне відображення хобі за статтю

Висновок: відмінність спостережень і очікуваних величин вказує на те, що зміні стать і хобі пов’язані між собою. Це підтверджує значення χ^2 (15,405 p = 0,00 < 0,05). Серед жінок переважає захоплення мистецтвом (рис. 4.19), на інші види хобі гендерна ознака майже не впливає. При цьому аналіз симетричних мір вказує, що даний зв’язок слабкий (від 0–0,3 відсутній, від 0,3–0,5 слабкий, від 0,5–0,7 середній, від 0,7 до 1 сильний, значення порівнюються за моделями), а саме: значення Фи, V Крамера, коефіцієнти пов’язаності, R Пірсона і кореляція Спірмена мають значення більші ніж 0,36, але менші ніж 0,5 (табл. 4.27).

Перелік питань для самоконтролю

1. Назвіть методи статистичного аналізу, які ґрунтуються на порівняно середніх величинах.

2. Поясніть сутність нульової гіпотези, що перевіряє дослідник у ході порівняння середніх величин.

3. Назвіть показник, який застосовують для перевірки нульової гіпотези.

4. Наведіть приклади застосування дисперсійного аналізу.

5. Поясніть сутність дисперсійного аналізу.

6. Назвіть, які типи шкал застосовують під час проведення дисперсійного аналізу, особливості їх застосування.

7. Назвіть основні види Т-тестів і дисперсійного аналізу, поясніть відмінність між ними.

8. Поясніть, для чого застосовується тест Лівіні.

9. Назвіть основні кроки однофакторного дисперсійного аналізу під час проведення перевірки практичної значущості результатів опитування респондентів.

10. Поясніть відмінність між однофакторним і двофакторним дисерсійними аналізами.

11. Поясніть, як впливають результати тесту Лівіні на подальший хід проведення дисперсійного аналізу.

12. Поясніть випадки застосування апостеріорних тестів у дисперсійному аналізі.

Тести

1. Метод перевірки гіпотез про рівність трьох і більше середніх – це:

а) вибіркове спостереження;

б) статистична гіпотеза;

в) дисперсійний аналіз;

г) правильної відповіді немає.

2. У разі рівності дисперсії застосовують критерій:

а) QP-Э-Г-У;

б) Т2 Тамхейна;

в) Т3 Даннетт;
г) Геймс-Хоуелл.

3. У разі рівності дисперсії застосовують критерій:
а) Геймс-Хоуелла;
б) Т2 Тамхейна;
в) Т3 Даннетт;
г) Даннетт.

4. У разі рівності дисперсії застосовують критерій:
а) Геймс-Хоуелла;
б) Шеффе;
в) Т2 Тамхейна;
г) Т3 Даннетт.

5. У разі рівності дисперсії застосовують критерій:
а) Геймс-Хоуелла;
б) Тьюки;
в) Т2 Тамхейна;
г) Т3 Даннетт.

6. У разі рівності дисперсії застосовують критерій:
а) Геймс-Хоуелла;
б) Т2 Тамхейна;
в) Т3 Даннетт;
г) Бонферроні.

7. У разі рівності дисперсії застосовують критерій:
а) Габріель;
б) Т2 Тамхейна;
в) Т3 Даннетт;
г) Геймс-Хоуелл.

8. У разі рівності дисперсії застосовують критерій:
а) Т3 Даннетт;
б) Т2 Тамхейна;

в) GT2 Гохберга;
г) Геймс-Хоуелл.

9. У разі відсутності рівності дисперсії застосовують критерій:
а) Даннетт;
б) Т2 Тамхейна;
в) GT2 Гохберга;
г) Шеффе.

10. У разі відсутності рівності дисперсії застосовують критерій:
а) Т3 Даннетт;
б) Габріель;
в) GT2 Гохберга;
г) Бонферроні.

11. У разі відсутності рівності дисперсії застосовують критерій:
а) Геймс-Хоуелла;
б) Габріель;
в) Даннетт;
г) Бонферроні.

12. Оберіть критерій, який застосовують для дисперсійного аналізу, коли обсяг групи до 5:
а) Тьюки;
б) Габріель;
в) Даннетт;
г) Бонферроні.

13. Оберіть критерій для дисперсійного аналізу, якщо передбачається рівність дисперсій і групи мають однаковий розмір:
а) QP-Э-Г-У і Тьюки;
б) Габріель і Даннетт;
в) Шеффе і Бонферроні;
г) Геймс-Хоуелла.

14. Оберіть критерій для дисперсійного аналізу, якщо передбачається рівність дисперсій і групи мають різний розмір:

а) Тьюки;

б) Т3 Даннетт;

в) Шеффе;

г) Геймс-Хоуелла.

15. Оберіть критерій для дисперсійного аналізу, якщо не передбачається рівність дисперсій і групи мають різну кількість спостережень:

а) Тьюки;

б) Даннетт;

в) Шеффе;

г) Геймс-Хоуелла.

16. Оберіть критерій, який перевіряє гіпотезу про рівність дисперсій у порівнюваних групах:

а) Тьюки;

б) Даннетт;

в) Шеффе;

г) Лівіня.

17. Якщо вибіркові спостереження мають номінальну школу як за залежними, так і незалежними змінними, то застосовують:

а) асоціативний аналіз;

б) однофакторний дисперсійний аналіз;

в) двофакторний дисперсійний аналіз;

г) правильної відповіді немає.

18. Оберіть критерій для дисперсійного аналізу, якщо передбачається рівність дисперсій і розмір груп сильно вирізняється:

а) Тьюки;

б) Даннетт;

в) Шеффе;

г) GT2 Гохберга.

19. Оберіть критерій для дисперсійного аналізу, якщо передбачається рівність дисперсій і необхідно порівняти з контрольною групою:

а) Тьюки;

б) Даннетт;

в) Шеффе;

г) GT2 Гохберга.

20. Оберіть універсальний критерій для дисперсійного аналізу, якщо не передбачається рівність дисперсій:

а) Т2 Тамхейна;

б) Даннетт;

в) Шеффе;

г) GT2 Гохберга.

РОЗДІЛ 5
КОРЕЛЯЦІЙНИЙ АНАЛІЗ

5.1. Основні поняття кореляційного аналізу

Завдяки англійському біологу і статисту Френсісу Гальтону наприкінці XIX століття в статистику був уведений термін «кореляція» (*correlation*). Карл Пірсон (1857–1936) і Джордж Юла (1871–1951) у своїх працях розробили і ввели в прикладні статистичні дослідження термін «парний коефіцієнт кореляції», який і сьогодні застосовують для вивчення взаємозв'язків змінних.

Кореляція – це статистична залежність між випадковими величинами, що має імовірнісний характер.

Кореляційний аналіз – статистичний метод, який дозволяє визначити, чи є залежність між випадковими змінними, який вона має напрям і наскільки сильна.

Кореляційний аналіз дозволяє досліднику:

1) вибрати з урахуванням специфіки і природи аналізованих змінних відповідний інструмент для вимірювання тісноти статистичного зв'язку (коефіцієнт кореляції, кореляційне відношення, ранговий коефіцієнт кореляції тощо);

2) оцінити знайдене за наявними вибірковими даними, його числове значення за допомогою точкової та інтервальних оцінок;

3) перевірити гіпотезу про те, що отримане значення аналізованого вимірювача зв'язку дійсно свідчить про наявність статистичного зв'язку, тобто перевірити досліджувану кореляційну характеристику на статистично значущу відміну від нуля;

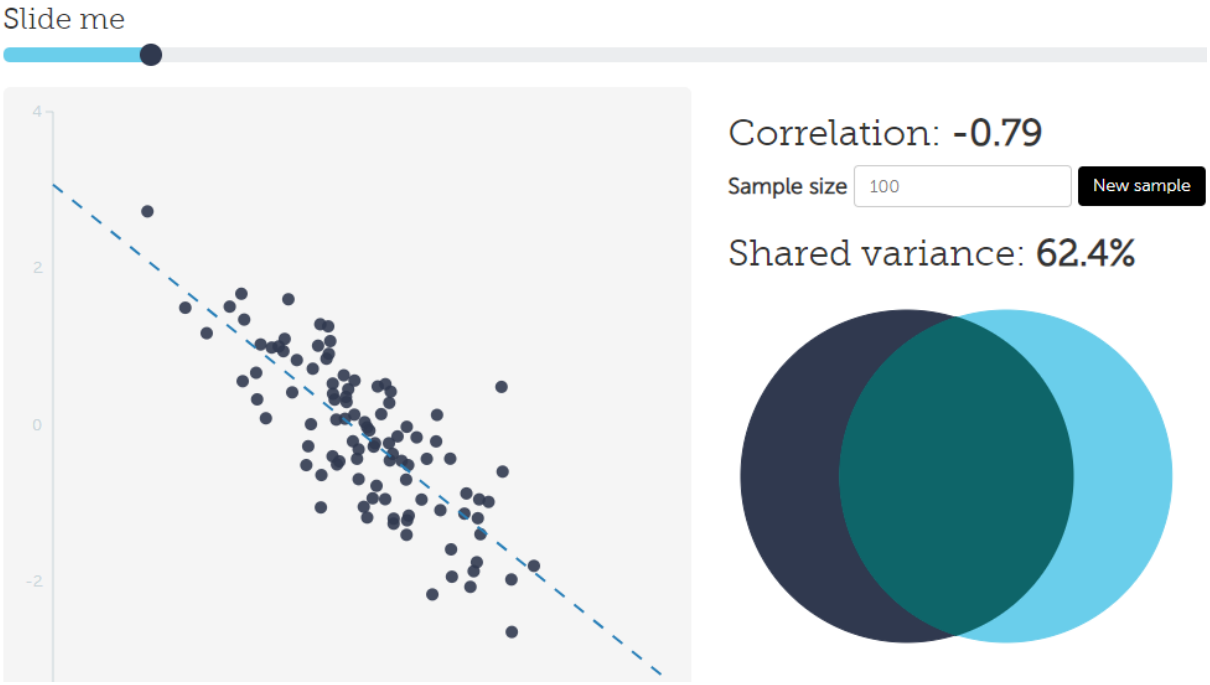
4) визначити структуру зв'язків між компонентами досліджувані багатовимірної ознаки, зіставивши кожній парі відповідь: зв'язок є чи немає.

Розрізняють два види залежності між економічними явищами: функціональну або жорстко детерміновану і статистичну або стохастичну, вірогідну.

Розглянемо приклади діаграми розсіювання емпіричних значень змінних X і У за допомогою сучасних онлайн-технологій:

Interpreting Correlations an interactive visualization (візуалізація кореляційних зв'язків): http://rpsychologist.3com/d3/correlation.

Негативна лінійна залежність (негативна кореляція або зворотній зв'язок) – у разі зростання значень однієї змінної зменшуються значення іншої, розраховане значення існує в межах $-1 < r < 0$ (рис. 5.1).



Slide me – рухайте мене. Sample size – об'єм вибірки.
New Sample – нова вибірка. Shared variance – спільна варіація.

Рис. 5.1. Візуалізація сильної негативної кореляції

Чітка негативна кореляція дорівнює –1, коли прослідковується 100 % спільна варіація двох змінних, при цьому випадкові змінні повністю накладаються на пряму (яка прямує з нижнього правого кута – у верхній лівий), відсутня зона розсіювання (рис. 5.2).

Позитивна лінійна залежність ($0 < r < 1$) – зростання рівня значень однієї змінної супроводжується підвищенням значення іншої змінної (рис. 5.3).

Чітка позитивна кореляція дорівнює +1, коли прослідковується 100 % спільна варіація двох змінних, при цьому напрям функції з лівого нижнього кута – в правий верхній (рис. 5.4).

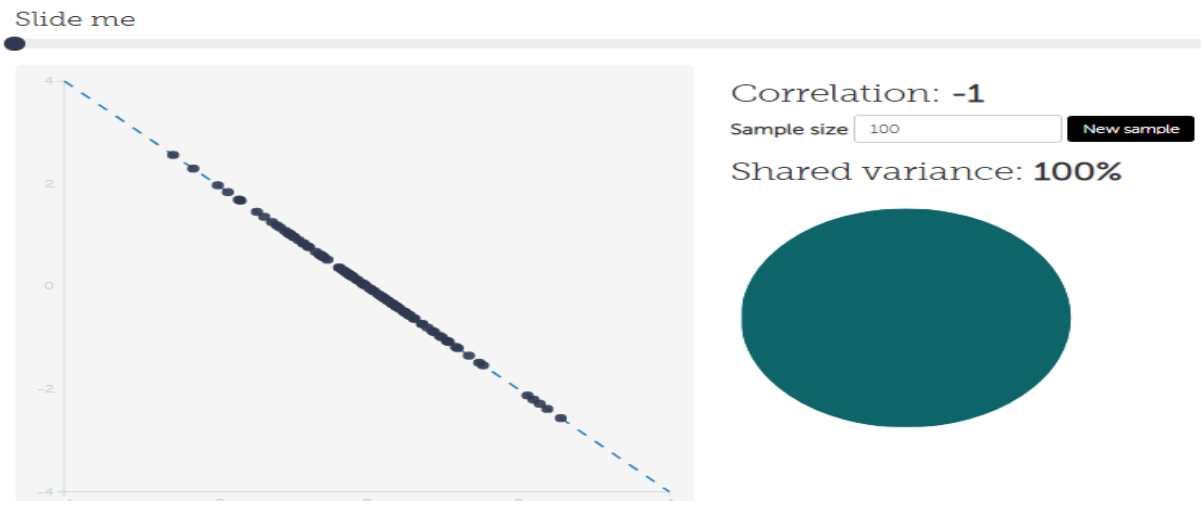


Рис. 5.2. Візуалізація чіткої негативної кореляції

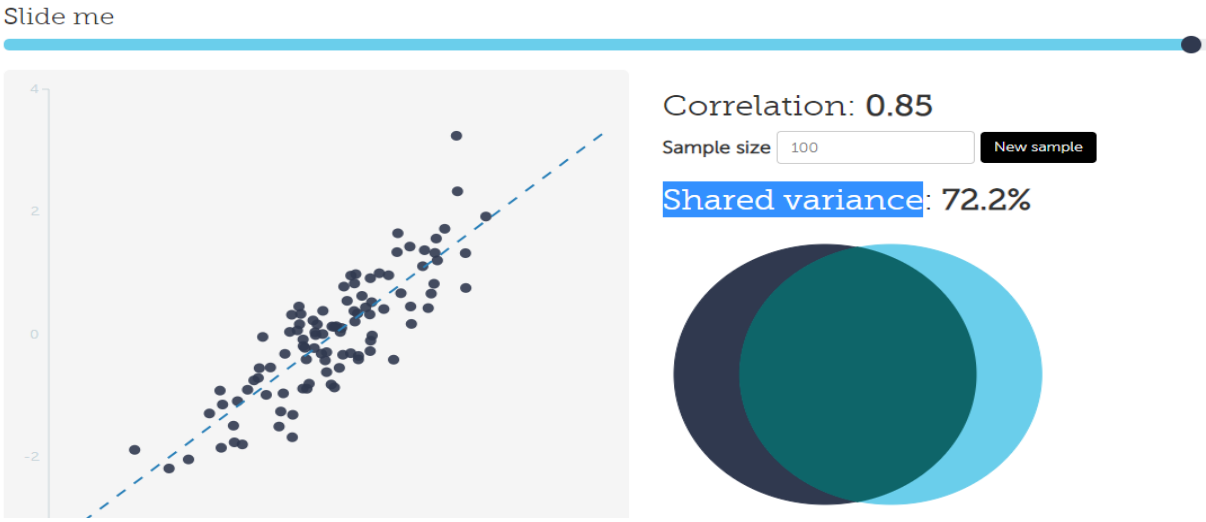


Рис. 5.3. Візуалізація сильної позитивної кореляції

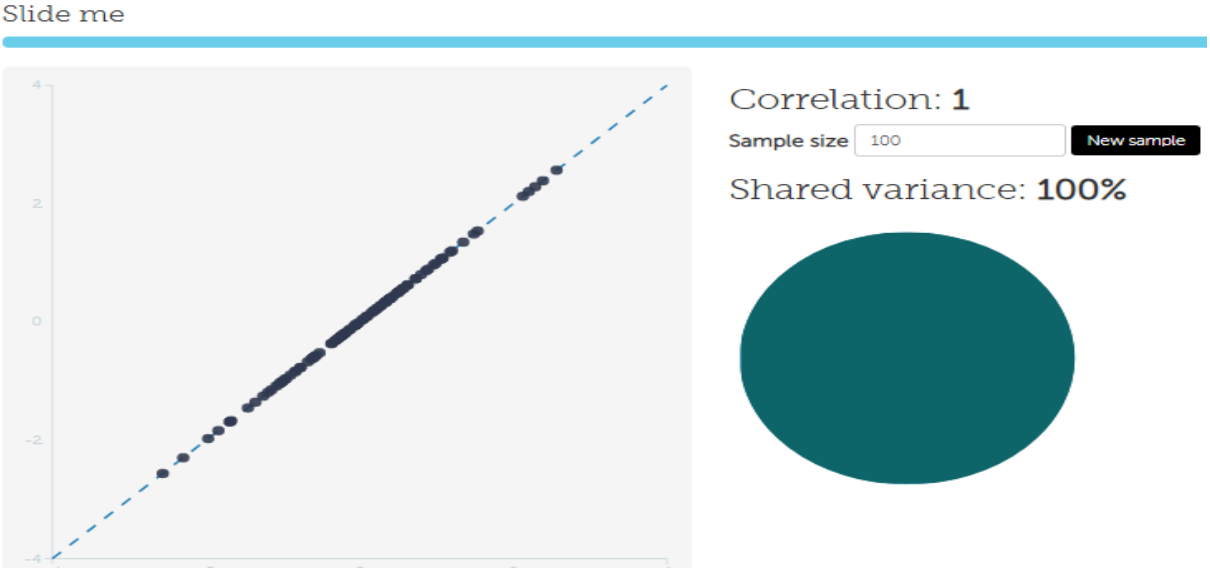


Рис. 5.4. Візуалізація чіткої позитивної кореляції

Зауважимо, що строга кореляція є математичною абстракцією і не зустрічається в реальних дослідженнях.

За відсутності зв'язку між змінними йдеться про **нульову кореляцію**. При цьому нульова загальна кореляція може свідчити про відсутність лише лінійної залежності, а не про відсутність будь-якого статистичного зв'язку (рис. 5.5).

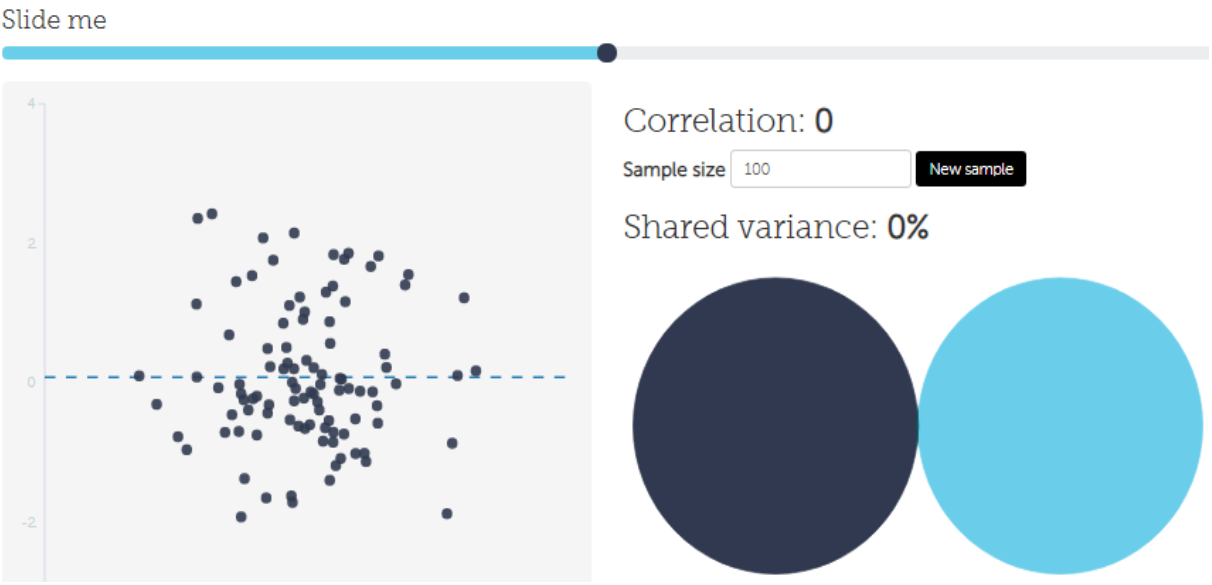


Рис. 5.5. Візуалізація нульової кореляції

З вищенаведеного можна описати межі і надати характеристику коефіцієнту кореляції. Кількісна міра кореляційного зв'язку оцінюється за значенням у межах від –1 до +1. Від'ємні значення коефіцієнта кореляції вказують на зворотний зв'язок, додатні – на прямий. Якщо оцінювати інтенсивність зв'язку, то він розглядається за модулем за такими межами:

від 0 до 0,3 відсутній;

від 0,3 до 0,5 слабкий;

від 0,5 до 0,7 середній;

від 0,7 до 1 сильний.

Чим ближче коефіцієнт кореляції наближається до 1, тим сильніший зв'язок, і навпаки, якщо розраховане значення коефіцієнта кореляції наближається до 0, тим слабша тенденція.

5.2. Основні методи вимірювання кореляційного взаємозв'язку

Методи вимірювання взаємозв'язку тісно пов'язані із вимірювальними шкалами результативної і факторними ознаками.

Загальноприйнятий коефіцієнт кореляції Пірсона (r) застосовують для оцінки зв'язку між двома змінними, які вимірюються в метричній шкалі (кількісна ознака) і мають нормальний розподіл:

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{S_x S_y},$$

де $\bar{x}$ – вибіркове середнє величини X, $\bar{y}$ – вибіркове середнє величини Y, $\overline{xy}$ – вибіркове середнє величини XY, S_x – вибіркове середнє квадратичне відхилення величини X, S_y – вибіркове середнє квадратичне відхилення величини Y. При цьому вибіркові середні і середньоквадратичні відхилення існують за формулами:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i\, n_i,\ \bar{y} = \frac{1}{n}\sum_{j=1}^{m} y_i\, n_j,\ \overline{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{m} x_i\, y_j n_{ij},$$

$$S_x = \sqrt{\frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - (\frac{1}{n}\sum_{i=1}^{k} x_i\, n_i)^2},$$

$$S_y = \sqrt{\frac{1}{n}\sum_{j=1}^{m} y_j^2 n_j - (\frac{1}{n}\sum_{j=1}^{m} y_j\, n_j)^2}.$$

У разі застосовування незгрупованих даних, наявності лінійного зв'язку між результативною та факторними ознаками використовується спрощений лінійний коефіцієнт кореляції Пірсона, який характеризує тісноту зв'язку і його напрям:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - ((\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i)/n)}{\sqrt{(\sum_{i=1}^{n} x_i^2 - ((\sum_{i=1}^{n} x_i)^2/n))(\sum_{i=1}^{n} y_i^2 - ((\sum_{i=1}^{n} y_i)^2/n))}},$$

де n – кількість спостережень, $y_{\text{і}}$ – індивідуальні значення результативної ознаки, x_i – індивідуальні значення факторної ознаки.

$$t_i = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}} = \frac{r\sqrt{n-k}}{\sqrt{1-r^2}}.$$

Для перевірки значимості коефіцієнта кореляції використовують t – статистику Сьюдента.

Порядок перевірки значимості коефіцієнта за *t*-статистикою Стьюдента:

1. Обираємо рівень значимості α (1 або 5 %).
2. Розраховуємо кількість ступенів свободи ($n-k$).
3. За таблицею розподілу Стьюдента визначаємо критичне значення *t* для *a*, *n* – *k*, де *k* = *m* + *1*, а *m* – кількість факторів.

Якщо розраховане значення *t-статистики* за модулем (за абсолютною величиною) більше табличного значення, то розрахований коефіцієнт кореляції є істотним для обраного рівня значимості *a*. В іншому випадку коефіцієнти неістотні.

Приклади. На підприємстві проводять дослідження ефективного використання фонду оплати праці, тис. грн (факторна змінна *x*), визначають його вплив на результати діяльності суб'єкта господарювання. За результативну ознаку взято чистий дохід від реалізації продукції (тис. грн) за десять років. Дані наведені в таблиці 5.1. Знайти коефіцієнт кореляції Пірсона і перевірити його значимість.

Таблиця 5.1

Проміжні розрахунки за даними підприємства

№	x_i	y_i	x_i^2	$x_i y_i$	y_i^2
1.	33,8362	125,5460	1 144,8884	4 247,9996	1 5761,7981
2.	36,5619	164,2310	1 336,7725	6 004,5974	26 971,8214
3.	39,2876	198,2530	1 543,5155	7 788,8846	39 304,2520
4.	42,0133	253,0230	1 765,1174	10 630,3312	64 020,6385
5.	44,7390	287,3650	2 001,5781	12 856,4227	82 578,6432
6.	47,4647	295,1130	2 252,8977	14 007,4500	87 091,6828
7.	50,1904	289,1530	2 519,0763	14 512,7047	83 609,4574
8.	52,9161	296,1230	2 800,1136	15 669,6743	87 688,8311
9.	55,6418	315,4650	3 096,0099	17 553,0404	99 518,1662
10.	58,3675	324,5640	3 406,7651	18 943,9893	105 341,7901
Σ	**461,0185**	**2 548,8360**	**21 866,7346**	**122 215,0942**	**691 887,0809**

З вихідних даних можна припустити, що між даними є лінійна залежність, тобто їх можна апроксимувати прямою лінією.

$$r = \frac{122215{,}0942-((461{,}0185*2548{,}8360)/10)}{\sqrt{\left(21866{,}7346-((461{,}0185)^2/10)\right)\left(691887{,}0809-((2548{,}8360)^2/10)\right)}} = 0{,}9256,$$

$$t = \frac{0{,}9256*\sqrt{10\text{-}2}}{1\text{-}0{,}9256^2} = \frac{2{,}61799}{0{,}37855} = 6{,}916,$$

$$t_{\text{таб}} = (\alpha = 0{,}05;\ \gamma = 10\text{–}2 = 8) = 2{,}306,$$

$$t_{\text{розр}} = 6{,}916 > 2{,}306 = t_{\text{табл}}$$

Висновок: на підприємстві існує прямий сильний лінійний зв'язок між фондом оплати праці і чистим доходом підприємства. *t*-статистика розрахована значно більша за критичне табличне значення, модель адекватна.

Необхідно зауважити, що у випадку інших форм розподілів і навіть коли одна із величин є дискретною, критерій кореляції Пірсона дає відносно точні результати. При цьому у разі відсутності нормального розподілу і наявності нелінійного зв'язку між змінними все ж таки рекомендується користуватися коефіцієнтом парної рангової кореляції Спірмена:

$$p_{AB} = 1 - \frac{6\sum_{i=1}^{n}\left(R_{Ai} - R_{Bi}\right)^2}{n\left(n^2 - 1\right)},$$

де R_{Ai} , R_{Bi} – ранги оцінок *i*-го напряму.

У разі наявності рангів, які повторюються, застосовують таку формулу:

$$p_{AB} = 1 - \frac{\sum_{i=1}^{n}\left(R_{Ai} - R_{Bi}\right)^2}{\frac{1}{6}\left(n^3 - n\right) - \frac{1}{12}\left(T_A + T_B\right)},$$

де T_A, T_B – показники зв'язаних рангів *A* і *B*, що визначаються таким чином:

$$\sum_{l=1}^{L} T_l = \sum_{l=1}^{L}\left(t_l^3 - t_l\right),$$

l – кількість груп зв'язаних (однакових) рангів,t_l – кількість зв'язаних рангів у *l*-ій групі.

Коефіцієнт парної рангової кореляції може приймати значення від –1 до 1.

Приклад. Експерти оцінюють Державну податкову адміністрацію (далі – ДПА) однієї з областей України за 10 напрямами роботи щодо поповнення державного бюджету (табл. 5.2). За ре-

зультатами оцінки побудуйте матрицю рангів. Розрахуйте узгодженість експерта з думками кожного з решти експертів. Результати подайте у вигляді матриці коефіцієнтів парної рангової кореляції.

Таблиця 5.2

Оцінка експертів напрямів роботи ДПА

№	Напрями	Експерти			
		1	2	3	4
1.	Податок з доходів фізичних осіб	100	100	90	80
2.	Податок на прибуток підприємств	90	100	80	100
3.	Плата за землю	90	80	100	90
4.	Податок на додану вартість	90	70	70	70
5.	Бюджетне відшкодування ПДВ	70	90	50	60
6.	Акцизний збір із вироблених в Україні товарів	80	60	60	50
7.	Акцизний збір із ввезених на територію України товарів	50	60	40	50
8.	Єдиний податок для суб'єктів малого бізнесу	50	60	30	40
9.	Частина прибутку (доходів) державних підприємств, яка утримується в бюджет	40	50	20	0
10.	Рентні платежі	60	50	10	20

Матрицю балів перетворюємо на матрицю рангів табл. 5.3 (на прикладі другого експерта розглянуто у таблицях 5.3, 5.4).

Таблиця 5.3

Приклад ранжування початкових даних

Бали	№	Розрахунок	Ранжований ряд
100	1	$\frac{1+2}{2}=1{,}5$	1,5
100	2		1,5
80	4	-	4
70	5	-	5
90	3	-	3
60	6	$\frac{6+7+8}{3}=7$	7
60	7		7
60	8		7
50	9.	$\frac{9+10}{2}=9{,}5$	9,5
50	10.		9,5

Таблиця 5.4

Ранжовані вихідні дані

№	Напрями	Експерти			
		1	**2**	**3**	**4**
1.	Податок з доходів фізичних осіб	1	1,5	2	3
2.	Податок на прибуток підприємств	3	1,5	3	1
3.	Плата за землю	3	4	1	2
4.	Податок на додану вартість	3	5	4	4
5.	Бюджетне відшкодування ПДВ	6	3	6	5
6.	Акцизний збір із вироблених в Україні товарів	5	7	5	6,5
7.	Акцизний збір із ввезених на територію України товарів	8,5	7	7	6,5
8.	Єдиний податок для суб'єктів малого бізнесу	8,5	7	8	8
9.	Частина прибутку (доходів) державних підприємств, яка утримується в бюджет	10	9,5	9	10
10.	Рентні платежі	7	9,5	10	9

Обчислимо узгодженість перших двох експертів (для інших пар експертів пропонується виконати розрахунки самостійно). Для цього розрахуємо коефіцієнт парної рангової кореляції:

$$p_{AB} = 1 - \frac{\sum_{i=1}^{n} (R_{Ai} - R_{Bi})^2}{\frac{1}{6}(n^3 - n) - \frac{1}{12}(T_A + T_B)}.$$

Скористаємося розрахунковою таблицею 5.5.

Таблиця 5.5

Проміжні розрахунки

№	Напрями	Ранги експертів		$R_{1i} - R_{2i}$	$(R_{1i} - R_{2i})^2$
		R_{1i}	R_{2i}		
1.	Податок з доходів фізичних осіб	1	1,5	–0,5	0,25
2.	Податок на прибуток підприємств	3	1,5	1,5	2,25
3.	Плата за землю	3	4	–1	1,00
4.	Податок на додану вартість	3	5	–2	4,00
5.	Бюджетне відшкодування ПДВ	6	3	–3	9,00
6.	Акцизний збір із вироблених в Україні товарів	5	7	–2	4,00

7.	Акцизний збір із ввезених на територію України товарів	8,5	7	1,5	2,25
8.	Єдиний податок для суб'єктів малого бізнесу	8,5	7	1,5	2,25
9.	Частина прибутку (доходів) державних підприємств, яка утримується в бюджеті	10	9,5	0,5	0,25
10.	Рентні платежі	7	9,5	–2,5	6,25
Разом		X	X	X	31,5

Зв'язані ранги, відповідно, дорівнюють:

Кількість груп зв'язаних рангів за першим експертом (3; 3; 3), (8,5; 8,5), другим експертом (1,5; 1,5), (7; 7; 7), (9,5; 9,5).

Кількість зв'язаних рангів за першим експертом становить: $t_1 = 3$, $t_2 = 2$, за другим експертом $t_3 = 2$, $t_4 = 3$, $t_5 = 2$, $t_6 = 2$. Звідси,

$$T_1 = (3^3 - 3) + (2^3 - 2) = 30;$$

$$T_2 = (2^3 - 2) + (3^3 - 3) + (2^3 - 2) = 36.$$

Маємо: $$p_{1,2} = 1 - \frac{31{,}5}{\frac{1}{6}(10^3 - 10) - \frac{1}{12}(30 + 36)} = 0{,}802.$$

$$t = \frac{0{,}802\sqrt{10\text{-}2}}{1\text{-}0{,}802^2} = \frac{2{,}2684}{0{,}3568} = 6{,}358$$

$$t_{\text{таб}} = (\alpha = 0{,}05;\ \gamma = 10 - 2 = 8) = 2{,}306.$$

Розрахований коефіцієнт парної рангової кореляції (0,802) значимий ($t_{\text{розр}} = 6{,}358 > 2{,}306 = t_{\text{табл}}$).

Матриця коефіцієнтів парної рангової кореляції аналогічно розрахованих за всіма чотирма експертами наводиться в таблиці 5.6.

Таблиця 5.6

Матриця парних коефіцієнтів кореляції Спірмена

Експерти	**Експерти**			
	1	**2**	**3**	**4**
1	1	0,803	0,886	0,867
2		1	0,833	0,917
3			1	0,930
4				1

Висновки. Аналізуючи матрицю парних коефіцієнтів рангової кореляції, можна стверджувати про досить велику ступінь збігу думок експертів, особливо між другим, третім та четвертим експертами.

Для малих вибірок застосовують рангові коефіцієнти кореляції Спірмена або Кендала. Коли обидві змінні вимірюються в номінальних шкалах, застосовують критерій асоціації Фі (покроковий аналіз дано в темі 4.4. Асоціативний аналіз змінних).

5.3. Кореляційний аналіз у SPSS

Приклад. Визначити взаємозв'язок між результативною ознакою: обсяг промислової продукції, млн грн (Y) і факторними: активи комерційних банків, млн грн (X_1), обсяг операцій фондового ринку, млн грн (X_2), обсяг операцій валютного ринку, млн грн (X_3) (рис. 5.6). Побудувати матрицю парних коефіцієнтів кореляції, провести аналіз за 16 років.

Файл Правка Вид Данные Преобразовать Анализ Прямой маркетинг Графика

	Rik	PP	AKB	OFR	OVR	пер
1	2001	210,84	70,27	68,50	13,48	
2	2002	229,63	125,43	108,60	20,44	
3	2003	289,12	120,99	203,00	32,91	
4	2004	400,76	196,95	321,30	35,01	
5	2005	468,56	316,97	403,80	40,90	
6	2006	551,73	561,13	492,80	55,72	
7	2007	717,08	599,40	754,30	82,24	
8	2008	917,04	926,09	883,40	99,31	
9	2009	806,55	880,30	1067,30	121,13	
10	2010	1043,11	942,09	1537,80	201,41	
11	2011	1305,31	1054,28	2147,50	298,02	
12	2012	1367,93	1127,19	2506,46	370,16	
13	2013	1322,41	1278,10	1676,97	638,34	
14	2014	1428,84	1316,85	2331,94	165,80	
15	2015	1776,60	1254,39	2171,59	303,79	
16	2016	2158,03	1256,30	2127,55	5090,33	
17						

Рис. 5.6. Вихідні дані моделі

Функція в SPSS: «**Анализ**» – «**Корреляции**» – «**Парные**».

Із лівого діалогового вікна обираємо змінні і за допомогою стрілки переміщуємо праворуч. Для перевірки зв'язку застосуємо коефіцієнт кореляції Пірсона, оскільки у нас усі вхідні дані мають кількісні ознаки (рис. 5.7).

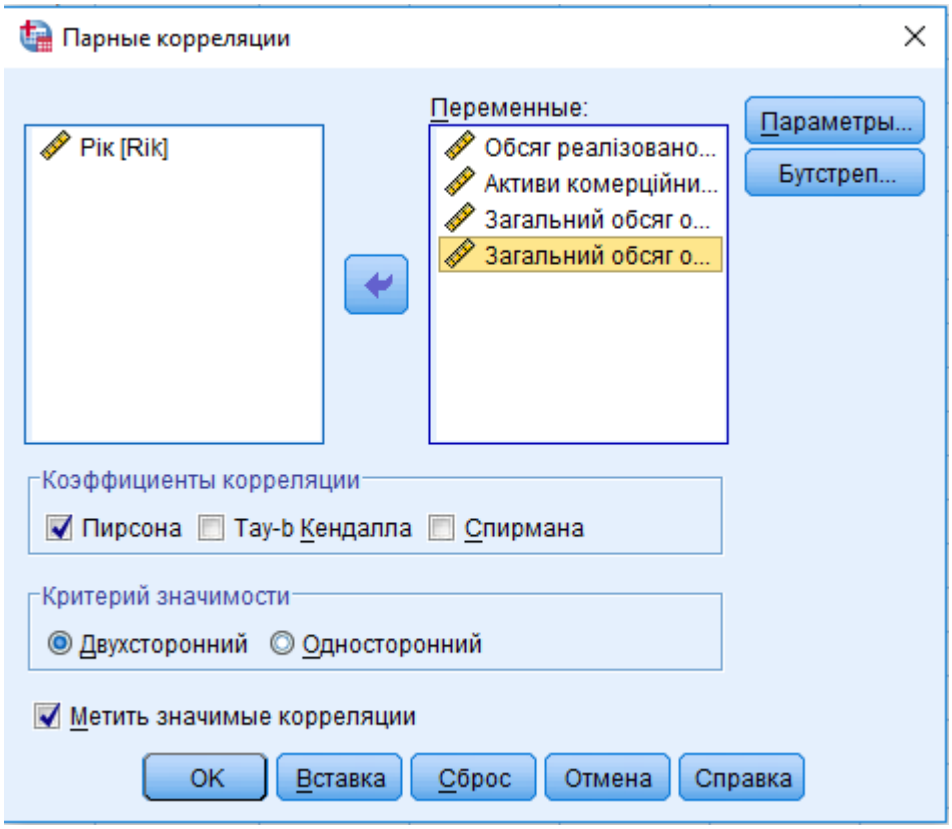


Рис. 5.7. Парний коефіцієнт кореляції Пірсона

Після натискання клавіші «ОК» отримаємо вивід результатів (табл. 5.7).

Таблиця 5.7

Матриця парних коефіцієнтів кореляції Пірсона

Показники		**Обсяг реалізованої промислової продукції (Y)**	**Активи комерційних банків (X1)**	**Обсяг операцій фондового ринку (X2)**	**Обсяг операцій валютного ринку (X3)**
Обсяг реалізованої промислової продукції , млрд грн (Y)	Кореляція Пірсона	1	0,924**	0,922**	0,640**
	Знач. (двостороння)		0,000	0,000	0,008
	Кореляція Пірсона	0,924**	1	0,933**	0,387

Активи комерційних банків, млрд грн (X1)	Знач. (двостороння)	0,000		0,000	0,139
Обсяг операцій фондового ринку, млрд грн (X2)	Кореляція Пірсона	0,922**	0,933**	1	0,381
	Знач. (двостороння)	0,000	0,000		0,146
Обсяг операцій валютного ринку, млрд грн (X3)	Кореляція Пірсона	0,640**	,387	0,381	1
	Знач. (двостороння)	0,008	00,139	0,146	

** Кореляція значима на рівні 0,01 (двостороння).

Автоматично в програмному продукті SPSS закладена функція, яка спрощує досліднику інтерпретувати результати, а саме відзначення різною кількістю зірочок результатів дослідження (табл. 5.8).

Таблиця 5.8

Інтерпретація рівнів значущості

Рівень статистичної значимості, *p*	**Статистична інтерпретація**	**Позначення в SPSS**
$p < 0{,}001$	Максимально значимий зв'язок	***
$0{,}001 \le p \le 0{,}01$	Досить значимий зв'язок	**
$0{,}01 < p \le 0{,}05$	Значимий зв'язок	*
$0{,}05 < p \le 0{,}10$	Слабозначимий	
$p > 0{,}10$	Незначимий	

У практичній діяльності застосовують критичне значення лінійного коефіцієнта кореляції для визначення порогового значення для порівняння факторів з метою включення в регресійну модель:

$$r_{кр} = \left[\frac{t^2_{табл}}{t^2_{табл} + (n-2)} \right]^{1/2},$$

де n – кількість спостережень, табличне значення *t-статистики* для $\gamma = (n - 2)$ ступенів свободи та $a = 0{,}05$. Усі факторні ознаки, що мають розрахований коефіцієнт кореляції вище критичного

значення, можна застосувати для побудови регресійної моделі, адже кореляційний аналіз є першим кроком у регресійному аналізі.

Критичне значення *t*-статистика Стьюдента обирається із таблиці для $a = 0{,}05$ і $\gamma = (n - 2) = 16 - 2 = 14$ ступенів свободи його значення дорівнює 2,145. Звертаємо увагу, незважаючи, що розглядається три факторні ознаки, порівнюють два парних коефіцієнти кореляції, тому завжди для визначення рівня ступенів свободи *(n – 2)*. Визначаємо критичне значення лінійного коефіцієнта кореляції для матриці парних коефіцієнтів кореляції:

$$r_{кр} = \sqrt{\frac{2{,}145^2}{2{,}145^2 + (16-2)}} = 0{,}4973.$$

Висновок: між результативною (Y) і факторними ознаками X_1 (0,924) і X_2 (0,922) сильний прямий лінійний зв'язок. Обсяг операцій валютного ринку має середній прямий лінійний вплив на обсяг реалізованої промислової продукції (0,64). Усі розраховані значення парних коефіцієнтів кореляції Пірсона значно вищі за його критичне значення. За побудови множинної регресії необхідно враховувати, що активи комерційних банків (X_1) мають тісний лінійний зв'язок з іншою факторною ознакою обсягом операцій фондового ринку (X_2) ($r_{x1x2} = 0{,}933$), що показує наявність між ними мультиколінеарності. Необхідно провести додаткові дослідження.

Приклад. Експерти оцінюють ДПА однієї з областей України за 10 напрямами роботи щодо поповнення державного бюджету (табл. 5.9). Розрахуйте узгодженість експертів за допомогою матриці парних коефіцієнтів рангової кореляції.

Таблиця 5.9

Оцінка експертів напрямів роботи ДПА

№	Напрями	Експерти			
		1	2	3	4
1.	Податок з доходів фізичних осіб	100	100	90	80
2.	Податок на прибуток підприємств	90	100	80	100
3.	Плата за землю	90	80	100	90
4.	Податок на додану вартість	90	70	70	70
5.	Бюджетне відшкодування ПДВ	70	90	50	60

6.	Акцизний збір із вироблених в Україні товарів	80	60	60	50
7.	Акцизний збір із ввезених на територію України товарів	50	60	40	50
8.	Єдиний податок для суб'єктів малого бізнесу	50	60	30	40
9.	Частина прибутку (доходів) державних підприємств, яка утримується в бюджеті	40	50	20	0
10.	Рентні платежі	60	50	10	20

Функція в SPSS: «**Анализ**» – «**Корреляции**» – «**Парные**».

Обираємо оцінки всіх чотирьох експертів. Відмічаємо рангові парні коефіцієнти кореляції Спірмена і Кендалла (рис. 5.8).

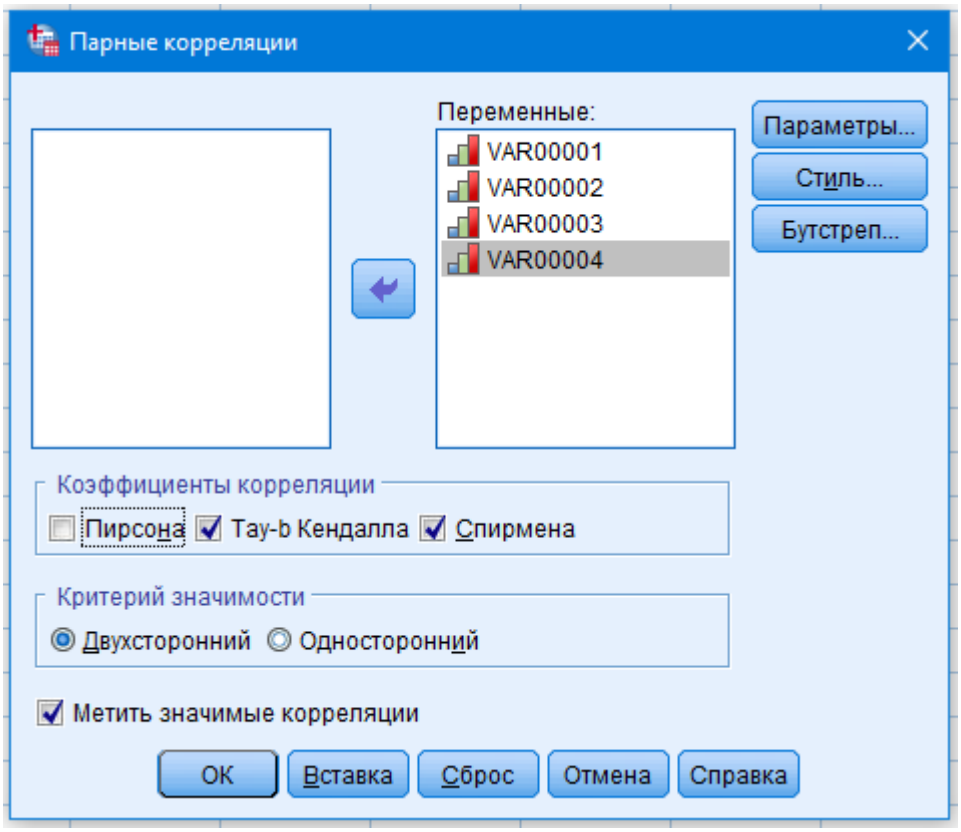


Рис. 5.8. Діалогове вікно SPSS

Натискаємо клавішу «ОК» і отримаємо такі результати (табл. 5.10):

Таблиця 5.10

Матриця рангових парних коефіцієнтів кореляції

			VAR001	VAR002	VAR003	VAR004
Тау-b Кендалла	VAR001	Коефіцієнт кореляції	1,000	0,667*	0,768**	0,706**
		Знач. (двостороння)	0.	0,012	0,003	0,006
	VAR002	Коефіцієнт кореляції	0,667*	1,000	0,707**	0,810**
		Знач. (двостороння)	0,012	.	0,006	0,002
	VAR003	Коефіцієнт кореляції	0,768**	0,707**	1,000	0,809**
		Знач. (двостороння)	0,003	0,006	0.	0,001
	VAR004	Коефіцієнт кореляції	0,706**	0,810**	0,809**	1,000
		Знач. (двостороння)	0,006	0,002	0,001	0.
Ро Спірмена	VAR001	Коефіцієнт кореляції	1,000	0,803**	0,886**	0,867**
		Знач. (двостороння)	0.	0,005	0,001	0,001
	VAR002	Коефіцієнт кореляції	0,803**	1,000	0,833**	0,917**
		Знач. (двостороння)	0,005	0.	0,003	0,000
	VAR003	Коефіцієнт кореляції	0,886**	0,833**	1,000	0,930**
		Знач. (двостороння)	0,001	0,003	0.	0,000
	VAR004	Коефіцієнт кореляції	0,867**	0,917**	0,930**	1,000
		Знач. (двостороння)	0,001	0,000	0,000	0.

* Кореляція значима на рівні 0,05 (двостороння).

** Кореляція значима на рівні 0,01 (двостороння).

Висновок: порівнюючи два методи оцінки, необхідно відмітити, що показники Кендалла мають менші значення, але вони значимі. Рівень коефіцієнтів рангових кореляцій Спірмена розрахований на рівні значимості 0,01. Усі оцінки парних коефіцієнтів кореляції Спірмена вказують на прямий, лінійний, сильний зв'язок. Найбільша узгодженість оцінки ДПА між четвертим експертом і другим, четвертим і третім. Найслабша – між першим і другим.

Частинний коефіцієнт кореляції. Якщо дві змінні корелюють, завжди можна припустити, що ця кореляція обумовлена впливом третьої змінної як загальної причини спільної змінності перших двох змінних. Для перевірки достатньо виключити вплив цієї третьої змінної і розрахувати кореляцію двох перших без врахування впливу третьої змінної (її значення фіксується). Така кореляція називається частинною кореляцією.

Приклад. Кореляційний аналіз взаємозв'язку між обсягом промислової продукції, млн грн (Y) і факторними ознаками показав лінійну залежність між активами комерційних банків, млн грн (X_1) і обсягом операцій фондового ринку, млн грн (X_2). Провести частинний кореляційний аналіз.

Функція в SPSS: «**Анализ**» – «**Корреляции**» – «**Частные**».

1. Спочатку проаналізуємо залежність результативної ознаки обсягу реалізованої промислової продукції (Y) з факторною змінною обсягом операцій фондового ринку (X_2). При цьому зафіксуємо змінну (X_1) активи комерційних банків (рис. 5.9).

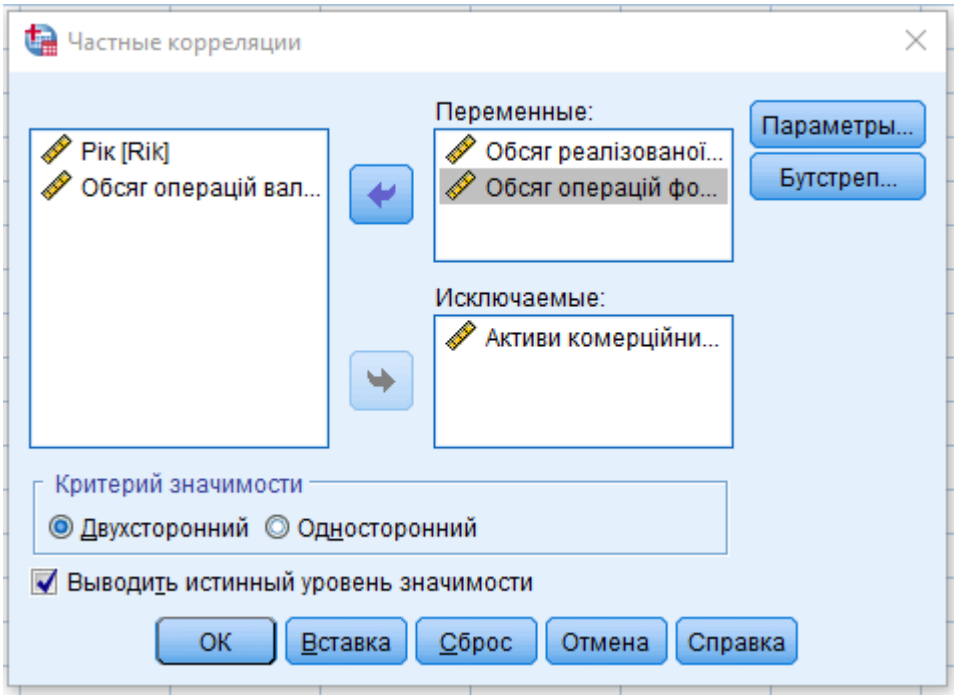


Рис. 5.9. Діалогове вікно SPSS

Натискаємо клавішу «ОК» і отримуємо результати дослідження (табл. 5.11).

Таблиця 5.11

Частинна кореляція

Змінні впливу			**Обсяг реалізованої промислової продукції, млрд грн (Y)**	**Обсяг операцій фондового ринку, млрд грн (X2)**
Активи комерційних банків, млрд грн (X1)	Обсяг реалізованої промислової продукції, млрд грн(Y)	Кореляція	1,000	0,434
		Значимість (двостороння)	0.	0,106
		ст. св.	0	13
	Обсяг операцій фондового ринку, млрд грн (X2)	Кореляція	0,434	1,000
		Значимість (двостороння)	0,106	0.
		ст. св.	13	0

Висновок: обсяг операцій фондового ринку слабо впливає на обсяг реалізованої продукції, коефіцієнт частинної кореляції 0,434, він вказує на прямий слабкий лінійний зв'язок. При цьому за результатами матриці парних коефіцієнтів кореляції Пірсона $r_{yx2} = 0{,}922$.

2. На другому етапі проаналізуємо залежність результативної ознаки обсягу реалізованої промислової продукції (Y) з факторною змінною активами комерційних банків (X_1), а обсяг операцій фондового ринку (X_2) зафіксуємо.

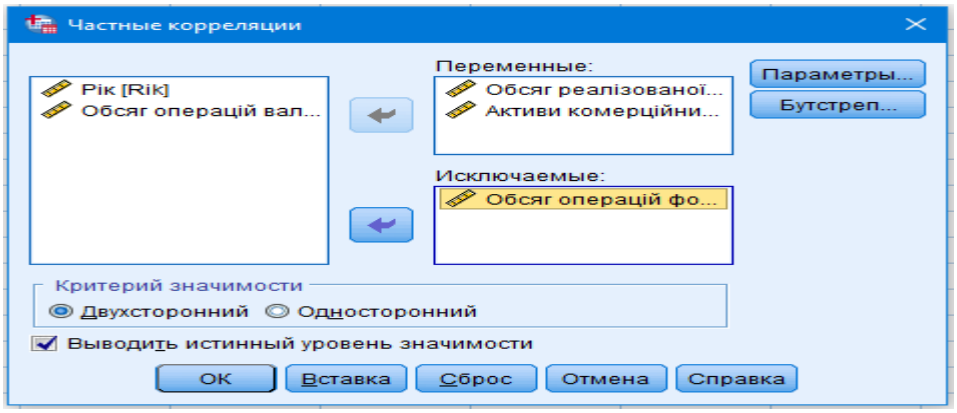


Рис. 5.10. Діалогове вікно SPSS

Натискаємо клавішу «ОК» і отримуємо результати 5.12.

Таблиця 5.12

Частинна кореляція

Змінні впливу			Обсяг реалізованої промислової продукції, млрд грн (Y)	Активи комерційних банків, млрд грн (X1)
Обсяг операцій фондового ринку, млрд грн (X2)	Обсяг реалізованої промислової продукції, млрд грн (Y)	Кореляція	1,000	0,461
		Значимість (двосторонння)	0.	0,083
		ст. св.	0	13
	Активи комерційних банків, млрд грн (X1)	Кореляція	0,461	1,000
		Значимість (двосторонння)	0,083	0.
		ст. св.	13	0

Висновок: активи комерційних банків мають також слабкий прямий лінійний вплив на обсяг реалізованої промислової продукції, коефіцієнт частинної кореляції 0,461, хоча дещо вище за вплив операцій фондового ринку. При цьому за результатами матриці парних коефіцієнтів кореляції Пірсона r_{yx2} = 0,924. Звідси випливає, що обидві змінні, доповнюючи одна одну дали високі оцінки парним коефіцієнтам кореляції Пірсона. Маємо наявний негативний вплив лінійної залежності між факторними ознаками, а саме мультиколінеарності.

Перелік питань для самоконтролю

1. Поясніть сутність кореляційного аналізу.
2. Чим відрізняється негативна лінійна залежність від позитивної.
3. Поясніть у чому особливість чіткої і нульової кореляції.
4. Назвіть основні методи вимірювання кореляційного зв'язку.
5. Назвіть, який метод кореляційного аналізу застосовують для метричних даних.
6. Назвіть, який метод кореляційного аналізу застосовують для порядкових даних.
7. Назвіть, який метод кореляційного аналізу застосовують для номінальних даних.
8. У чому особливість методу парної кореляції Спірмена, поясність, коли його застосовують.
9. У чому особливість методу парної кореляції Пірсона, поясність, коли його застосовують.
10. Поясніть, для чого застосовується критичне значення парного коефіцієнта кореляції.

Тести

1. Статистичний метод, який дозволяє визначити, чи є залежність між випадковими змінними, і виміряти її щільність та напрям – це:

а) кореляційний аналіз;

б) дисперсійний аналіз;
в) описова статистика;
г) правильна відповідь відсутня.

2. Яка залежність між величинами, якщо коефіцієнт кореляції буде наближений до 0:
а) відсутня;
б) пропорційна;
в) обернена;
г) слабка.

3. Як називається кореляційний зв'язок, при якому значення результативної ознаки змінюється у протилежному напрямі щодо факторної:
а) криволінійний;
б) обернений;
в) прямий;
г) прямолінійний.

4. В яких межах міститься коефіцієнт кореляції:
а) [0;1];
б) [–1; +1];
в) [–1; 0];
г) [0; 4].

5. Для вимірювання тісноти взаємозв'язку між двома ознаками, що включені у модель, визначають:
а) правильна відповідь відсутня;
б) парні коефіцієнти кореляції;
в) парні коефіцієнти асиметрії;
г) парні коефіцієнти детермінації.

6. Розрахований коефіцієнт кореляції «–0,45» – це означає:
а) слабкий обернений зв'язок;
б) відсутній зв'язок;
в) середній обернений зв'язок;

г) слабкий прямий зв’язок.

7. В яких межах полягає значення коефіцієнта кореляції:

а) від 0 до 4;

б) від 0 до 1;

в) від –4 до +4;

г) від –1 до +1.

8. Коефіцієнт кореляції 0,95 – це означає:

а) факторна змінна на 95 % залежить від результативної і на 5 % від інших величин, неврахованих у моделі;

б) змінна на 95 % залежить від факторних і на 5 % від інших величин, неврахованих у моделі;

в) тісний прямий лінійний зв’язок;

г) тісний обернений лінійний зв’язок.

9. Аналіз матриці парних коефіцієнтів кореляції визначає наявність:

а) автокореляції;

б) гетероскедастичності;

в) гомоскедастичності;

г) мультиколінеарності.

10. Зв’язки, в яких існують невраховані фактори, мають назву:

а) функціональних;

б) детермінованих;

в) стохастичних;

г) змішаних.

11. *p*-значення моделі, що вказує на значущість оцінок повинно бути:

а) < 0,05;

б) > 0,05;

в) = 0,05;

г) більшим за табличне значення.

12. Якщо під час зростання значень однієї змінної зменшується значення іншої – це:

а) позитивна лінійна залежність;

б) консервативна лінійна залежність;

в) негативна лінійна залежність;

г) правильна відповідь відсутня.

13.Чітка негативна кореляція дорівнює:

а) –100;

б) –10;

в) –1;

г) правильна відповідь відсутня.

14. Строга позитивна кореляція дорівнює:

а) 100;

б) 10;

в) 1;

г) правильна відповідь відсутня.

15. Відсутність зв’язку між змінними – це:

а) чітка позитивна кореляція;

б) чітка негативна кореляція;

в) нульова кореляція;

г) правильна відповідь відсутня.

16. Для визначення кореляційного зв’язку між номінальними даними використовується критерій:

а) Пірсона;

б) Спірмена;

в) Кендала;

г) асоціації Фі.

17. Для визначення кореляційного зв’язку між порядковими даними використовується критерій:

а) Пірсона;

б) Спірмена;

в) Кендала;

г) асоціації Фі.

18. Зв’язки, в яких значення результативної ознаки на 100 % відповідає факторній:

а) функціональні;

б) кореляційні;

в) стохастичні;

г) змішані.

19. Модель, у якій використовується випадкова величина (похибка) – це модель:

а) функціональна;

б) детермінована;

в) стохастична;

г) змішана.

20. Для визначення кореляційного зв’язку між метричними даними використовується критерій:

а) Пірсона;

б) Спірмена;

в) Кендала;

г) асоціації Фі.

Економічна інтерпретація кореляційного аналізу

Приклад. 1. Провести кореляційний аналіз фінансових показників банку та макроекономічних важелів за даними таблиці 1.

Таблиця 1

Фінансові результати ПАТ КБ «ПриватБанк» за 2001–2018 рр.

Рік	Активи, млн грн	Власний капітал, млн грн	Зобов'язання, млн грн	Депозити, млн грн	Фінансовий результат, млн грн	Курс долара	Інфляція*	ВВП, млн грн	Доходи населення, млн грн
2001	7221,36	303,71	6912,33	2409,57	33,48	5,43	106,1	204587	108835
2002	10259,71	430,36	9837,99	4336,65	35,07	5,29	99,4	225810	191946
2003	17724,43	726,46	17092,38	6017,72	49,19	5,33	108,2	267344	211922
2004	14671,25	1451,99	13308,96	10317,32	289,80	5,34	112,3	345113	264382
2005	21719,16	2165,39	19606,09	13627,08	437,07	5,35	110,3	441452	381404
2006	30652,74	3957,79	27493,53	20220,28	382,52	5,05	111,6	544153	472061
2007	51149,69	5377,56	46242,43	36249,36	1052,51	5,05	116,6	720731	623289
2008	78410,04	8445,14	70515,63	55244,57	990,50	5,45	122,3	948056	845641
2009	81813,22	10790,49	71753,93	52858,58	842,96	7,79	112,3	913345	894286
2010	109752,50	8860,20	98001,05	77139,56	1239,78	7,93	109,1	1082569	1101175
2011	142236,70	13545,17	125700,10	92043,44	1216,43	7,95	104,6	1316600	1266753
2012	169570,40	14897,55	151391,30	106275,70	1410,42	7,99	99,6	1408889	1457864
2013	211425,80	16352,07	191189,00	129863,50	1832,07	7,99	100,5	1454931	1548733
2014	201471,20	18100,74	180045,40	144343,80	652,63	11,88	124,9	1566728	1516768
2015	251551,10	21256,46	225827,10	169502,60	239,82	21,86	143,3	1979458	1772016
2016	269032,40	26348,00	240100,60	191603,60	628,51	25,55	112,4	2383182	2051331
2017	245882,10	23619,00	218345,20	564738,70	1734,99	26,62	113,7	2982920	2652082
2018	278048,00	31464,00	246534,00	231055,00	18289,00	28,15	109,8	3558706	3219518

Джерело: побудовано на основі джерел[5].

* Індексів споживчих цін.

[5]Офіційний сайт Державної служби статистики України. URL : http://www.ukrstat.gov.ua/operativ/gdn/dvn/arh_dvn2001.html
Офіційний сайт ПАТ КБ «ПриватБанк». URL : https://privatbank.ua/about/finansovaja-otchetnost.
URL : https://finance.ua/ua/org/-/ua/banks/privatbank/finres/2001/06
URL : https://tables.finance.ua/ru/currency/official/-/1/2001/01/03

Аналізуючи динаміку фінансових результатів, отриманих ПриватБанком за 2001–2018 рр., однозначно видно тенденцію до нарощування його фінансових можливостей за виключенням декількох кризових років. Проаналізуємо вплив наведених у таблиці 1 факторів на динаміку фінансових результатів ПАТ КБ «ПриватБанк» за допомогою кореляційного аналізу в програмному продукті SPSS.

На власний капітал ПАТ КБ «ПриватБанку» переважна більшість показників має істотний вплив, крім інфляції (r = 0,27). Згідно з нашим дослідженням індекс споживчих цін (інфляція) взагалі не впливає на діяльність банку, обрані показники мають парний коефіцієнт кореляції від 0,08 до 0,29 (табл. 2). При цьому прослідковується слабкий лінійний взаємозв'язок її з курсом долара 0,34, тому що у разі збільшення інфляційних процесів у країні курс долара теж зростає. У свою чергу, на такий важливий показник, як власний капітал банку, що має прямий зв'язок із його фінансовими результатами, чинять істотний вплив: активи (r = 0,98), депозити (r = 0,76), зобов'язання (r = 0,98) банку, а також курс долара (0,90), ВВП (r = 0,97) та доходи населення (r = 0,98).

Таблиця 2

Матриця парних коефіцієнтів кореляції

Показники	Активи	Власний капітал	Зобов'язання	Депозити	Курс долара	Інфляція	ВВП	Доходи населення	Фінансовий результат
Активи	1	,980**	1,000**	,756**	,860**	,290	,932**	,946**	,454
		,000	,000	,000	,000	,242	,000	,000	,059
Власний капітал	,980**	1	,978**	,757**	,897**	,273	,971**	,976**	,569*
	,000		,000	,000	,000	,272	,000	,000	,014
Зобов'язання	1,000**	,978**	1	,753**	,857**	,292	,929**	,944**	,449
	,000	,000		,000	,000	,239	,000	,000	,061
Депозити	,756**	,757**	,753**	1	,816**	,190	,837**	,825**	,303
	,000	,000	,000		,000	,451	,000	,000	,221
Курс долара	,860**	,897**	,857**	,816**	1	,340	,932**	,899**	,539*
	,000	,000	,000	,000		,167	,000	,000	,021
Інфляція	,290	,273	,292	,190	,340	1	,231	,197	-,082
	,242	,272	,239	,451	,167		,357	,432	,747
ВВП	,932**	,971**	,929**	,837**	,932**	,231	1	,995**	,655**
	,000	,000	,000	,000	,000	,357		,000	,003
Доходи населення	,946**	,976**	,944**	,825**	,899**	,197	,995**	1	,652**
	,000	,000	,000	,000	,000	,432	,000		,003
Фінансовий результат	,454	,569*	,449	,303	,539*	-,082	,655**	,652**	1
	,059	,014	,061	,221	,021	,747	,003	,003	

Джерело: власні розрахунки.

Аналізуючи матрицю парних коефіцієнтів кореляції Пірсона, необхідно зазначити, що на фінансові результати банка його активи (r = 0,45), зобов'язання (r = 0,45), розмір депозитів (r = 0,30) та інфляція (r = –0,08) не чинять впливу, адже коефіцієнти кореляції Пірсона за ними є надто низькими, що вказує на слабкий зв'язок (від 0,3 до 0,5) між даними показниками, а за інфляцією взагалі відсутній (від 0 до 0,3). Найістотніший зв'язок фінансові результати банку мають з ВВП (r = 0,66) та розміром доходів населення (r = 0,65) за рік. Економічне середовище країни має середній прямий лінійний вплив на ендогенну зміну. Ця тенденція є логічною, адже чим більше виготовлено в країні товарів, тим більшим є його ВВП, а отже, і можливість збільшувати обсяг вільних коштів на руках у населення, що приводить до затребуваності банківських послуг, таких як: відкриття депозитів, збільшення валютно-обмінних операцій та інших послуг. Так само і у разі зменшення ВВП та доходів населення люди все частіше звертаються до банку за кредитами та розстрочками, що, в свою чергу, також позитивно впливає на фінансові результати банку, хоча в цьому випадку присутній ризик їхнього непогашення. Вплив власного капіталу (r = 0,57) та курсу національної валюти до долара США (r = 0,54) майже на одному рівні.

У практичній діяльності застосовують критичне значення лінійного коефіцієнта кореляції (порогове значення), яке використовують для порівняння з розрахованими парними коефіцієнтами. Він враховує зміну сили зв'язку від кількості спостережень у ряду динаміки. Під час здійснення кореляційного аналізу доцільно характеризувати ознаки, які мають розрахований коефіцієнт кореляції більший за критичний[6]:

$$r_{кр} = \left[\frac{t^2_{табл}}{t^2_{табл} + (n-2)} \right]^{1/2}, \qquad (1)$$

[6]Паянок Т. М., Лаговський В. В., Краєвський В. М. та ін. Аналітика та прогнозування соціально-економічних процесів і податкових надходжень : монографія К.: ЦП «Компринт», 2019. 426 с.

де n – кількість спостережень, табличне значення t-статистики для $\gamma = (n-2)$ ступенів свободи та $\alpha = 0{,}05$. Розрахуємо критичне значення для нашої моделі $\gamma = (18-2) = 16$ ступенів свободи та $\alpha = 0{,}05$:

$$r_{кр} = \sqrt{\frac{2{,}1199^2}{2{,}1199^2+16}} = 0{,}468. \quad (2)$$

Продовжимо аналізувати фактори, які задовольняють цю вимогу, а саме: власний капітал (x_1), курс національної валюти до долара США (x_2), ВВП (x_3) та доходи населення (x_4). Між деякими екзогенними змінними присутня лінійна залежність (мультиколінеарність), тобто показники доповнюють один одного, що не дає можливості розмежувати вплив кожного на ендогенну змінну (табл. 3).

Таблиця 3

Лінійна залежність між факторними змінними

X1 – власний капітал X3 – ВВП	X1 – власний капітал X4 – доходи населення	X2 – Курс долара X3 – ВВП	X3 – ВВП X4 – Доходи населення
$r_{x1x3} = 0{,}971$	$r_{x1x4} = 0{,}976$	$r_{x2x3} = 0{,}932$	$r_{x3x4} = 0{,}995$

Джерело: власні розрахунки.

Щоб оцінити вплив кожного фактора на фінансові результати, доцільно розрахувати частинну парну кореляцію. Для перевірки достатньо виключити вплив другої змінної і розрахувати кореляцію між ендогенною змінною і однією екзогенною (інша екзогенна фіксується). Доходи населення і ВВП є складовими один одного, тому частинну кореляцію за ними розраховувати не доцільно. З таблиці 4 видно, що ВВП має більший вплив на фінансовий результат ($r = 0{,}523$), ніж власний капітал ($r = -0{,}372$).

Аналізуючи наступну пару факторних ознак, доцільно відмітити вплив доходів населення на фінансові результати ($r = 0{,}537$), який перевищує вплив власного капіталу ($r = -0{,}405$) (табл. 5).

Порівнюючи ВВП ($r = 0{,}502$) з курсом долара ($r = -0{,}264$), необхідно заначити, що перший показник має сильніший вплив (табл. 6).

Таблиця 4

Частинна кореляція між фінансовим результатом, власним капіталом і ВВП

Змінні управління			**Фінансовий результат**	**Власний капітал**
ВВП	Фінансовий результат	Кореляція	1,000	–0,372
		Значимість ст.св.		0,141
	Власний капітал	Кореляція	–0,372	1,000
		Значимість ст.св.	0,141	
Змінні управління			Фінансовий результат	ВВП
Власний капітал	Фінансо-вий резуль-тат	Кореляція	1,000	0,523
		Значимість ст.св.		0,031
	ВВП	Кореляція	0,523	1,000
		Значимість ст.св.	0,031	

Джерело: власні розрахунки.

Таблиця 5

Частинна кореляція між фінансовим результатом, власним капіталом і доходами населення

Змінні управління			**Фінансовий результат**	**Власний капітал**
Доходи населення	Фінансовий результат	Кореляція	1,000	–0,405
		Значимість ст.св.		0,107
	Власний капітал	Кореляція	-0,405	1,000
		Значимість ст.св.	0,107	
Змінні управління			**Фінансовий результат**	**Доходи населення**
Власний капітал	Фінансовий результат	Кореляція	1,000	0,537
		Значимість ст.св.		0,026
	Доходи населення	Кореляція	0,537	1,000
		Значимість ст.св.	0,026	

Джерело: власні розрахунки.

Таблиця 6

Частина кореляція між фінансовим результатом, курсом долара і ВВП

Змінні управління			**Фінансовий результат**	**ВВП**
Курс долара	Фінансовий результат	Кореляція	1,000	0,502
		Значимість ст.св.		0,040
	ВВП	Кореляція	0,502	1,000
		Значимість ст.св.	0,040	
Змінні управління			**Фінансовий результат**	**Курс долара**
ВВП	Фінансо-вий резуль-тат	Кореляція	1,000	-0,264
		Значимість ст.св.		0,306
	Курс до-лара	Кореляція	-0,264	1,000
		Значимість ст.св.	0,306	

Джерело: власні розрахунки.

Проведений кореляційний аналіз вказує, що на фінансові результати ПАТ КБ «ПриватБанк» мають суттєвий вплив ВВП і доходи населення України. Загалом фінансове становище банку є стабільним. Поступово збільшується його надійність та фінансова стійкість. До речі, не може не втішати і той факт, що останнім часом до банку проявляють значний інтерес потенційні інвестори, зокрема міцні західні банківські структури, що хотіли б придбати акції ПриватБанку. Такий інтерес є яскравим доказом стабільного фінансового стану банку. Адже жодному інвестору навіть на думку не спаде вкладати кошти у структуру або бізнес, які мають проблеми та недостатній рівень надійності.

Висновки: ПАТ КБ «ПриватБанк» займає лідируючі позиції за всіма фінансовими показниками в галузі та довірою населення, адже обслуговує третину вкладів населення країни і понад 24 % населення України. Кореляційний аналіз показав, що на власний капітал ПАТ КБ «ПриватБанку» переважна більшість аналізованих показників має значний вплив, крім інфляції. Активи, зобов’язання, розмір депозитів та індекс споживчих цін на фінансові

результати банку не чинять впливу. Найістотніший зв’язок фінансові результати банку мають з ВВП та розміром доходів населення за рік. Вплив власного капіталу та курсу національної валюти до долара США має середній вплив.

Хоча інфляція ніяк не впливає на фінансові показники банку, вважаємо за необхідність подальшого дослідження реальної тенденції фінансових результатів у цінах базового року. Цей захід зменшить коливання, що відбулися за рахунок збільшення курсу долара, і покращить розподіл відносно середнього, що забезпечить достовірніший прогноз.

РОЗДІЛ 6
МНОЖИННИЙ РЕГРЕСІЙНИЙ АНАЛІЗ

6.1. Сутність та види регресійного аналізу

Регресія – це вимірювання одностороннього стохастичного зв'язку між факторними і результативними ознаками. Регресійний аналіз дає математичний опис залежності між змінними, побудована модель дозволяє отримувати прогнози й оцінювати вплив факторних ознак на результативну.

У разі побудови регресійної моделі необхідно чітко розмежовувати характеристики вхідних даних моделі. Для цього більш детально розглянемо класифікацію змінних величин у регресійних моделях:

1. **Ендогенна змінна** (результативна змінна) – ознака, яка характеризує наслідок дії фактора або факторів, залежна змінна (Y).

2. **Екзогенні змінні** (факторні змінні) – причини та умови, які необхідні для виникнення певного наслідку (x), екзогенні зміни визначають ендогенні, але самі не перебувають під їх впливом.

Таким чином, між ендогенними та екзогенними змінними існують тільки односторонні стохастичні причинні відносини.

3. **Наперед визначенна змінна (лагова).** Під лаговою розуміють змінну, значення якої відстає на один або декілька періодів. Якщо $x_{t,k}$ – значення звичайної змінної x_k, зафіксовані в цей момент часу t, то $x_{t-1,\,k}$ – її лагові значення зміщенні на один період.

4. **Спільно залежні змінні** – це звичайні ендогенні змінні, які пояснюються регресійною моделлю в момент часу t. Між ними існують багатосторонні зв'язки і визначаються не одним рівнянням, а одночасними рівняннями моделі. Приклад: грошовий обіг y_1 і оборотність грошей y_2, між ними існують односторонні співвідношення.

5. **Збурення або латентні змінні** – це економічні величини, які не входять у рівняння регресійної моделі, але впливають на спільно залежні змінні. Вони формуються завдяки випадковим впливам, помилкам і припущенням. Збурення є стохастичними

змінними, їх значення знаходять як залишки за окремими рівняннями після оцінки невідомих параметрів моделі.

6. **Фіктивні змінні** (дихотомічні) – це категорійні змінні, які мають якісні характеристики, ці змінні бінарного типу, тобто кожна змінна може приймати всього два значення – одиниця і нуль.

Класифікація видів регресії:

0. За числом змінних: проста (парна) і множинна (багатофакторна).

1. Відносно форми залежності, моделі поділяють на: лінійну та нелінійну:

1) лінійна регресія (зміна x призводить до відносно рівномірної зміни y):$Y = a+\beta x + u$, де

величина y складається з двох складових:

невипадкові складові $a+\beta x$, де $x = \{x_1, x_2, \ldots, x_n\}$, а a і β – параметри рівняння, знаходяться методом найменших квадратів:

$$\beta = \frac{n \cdot \sum xy - \sum x \cdot \sum y}{n \cdot \sum x^2 - \left(\sum x\right)^2},\ a = \bar{y} - b \cdot \bar{x};$$

випадкова складова (збурення, помилки) $u = \{u_1, u_2, \ldots, u_n\}$.

2) нелінійна регресія:

– степенева (з прискорення – вплив обсягу грошової маси (x) на рівень інфляції (y)): $Y = ax^{\beta} + u$;

– гіпербола (з уповільненням – залежність продуктивності праці робітників (y) від рівня зарплати (x)): $Y = \alpha + \frac{\beta}{x} + u$;

– парабола (зі зміною напряму зв'язку – залежність урожайності зернових (y) від кількості опадів (x)): $Y = \alpha + \beta x + cx^2 + u$.

6.2. Етапи регресійного аналізу в SPSS

Проаналізуємо на конкретному прикладі етапи регресійного аналізу.

Приклад. Побудувати множину регресійну модель, визначити вплив основних фінансових показників на обсяг реалізації промислової продукції (табл. 6.1).

Таблиця 6.1

Вихідні дані для регресійного аналізу

Рік	Обсяг реалізованої промислової продукції (Y)	Активи комерційних банків (X1)	Обсяг операцій фондового ринку (X2)	Обсяг операцій валютного ринку (X3)
2001	210,84	70,27	68,50	13,48
2002	229,63	125,43	108,60	20,44
2003	289,12	120,99	203,00	32,91
2004	400,76	196,95	321,30	35,01
2005	468,56	316,97	403,80	40,90
2006	551,73	561,13	492,80	55,72
2007	717,08	599,40	754,30	82,24
2008	917,04	926,09	883,40	99,31
2009	806,55	880,30	1067,30	121,13
2010	1043,11	942,09	1537,80	201,41
2011	1305,31	1054,28	2147,50	298,02
2012	1367,93	1127,19	2506,46	370,16
2013	1322,41	1278,10	1676,97	638,34
2014	1428,84	1316,85	2331,94	165,80
2015	1776,60	1254,39	2171,59	303,79
2016	2158,03	1256,30	2127,55	5090,33

1. **Постановчий етап** – визначення об’єкта дослідження кінцевої мети моделювання, сукупності факторів і показників, які найбільше характеризують модель, їх роль та місце в дослідженні економічних процесів.

Метою дослідження є визначення впливу фінансового сектору України на промисловість. В Україні відсутня статистика за вказаний період за обсягами виробництва, тому обрано обсяг реалізованої продукції, який впливає на кінцеву мету діяльності суб’єкта господарювання – отримання прибутку.

2. **Інформаційний етап** – ретроспективний аналіз економічної сутності досліджуваного процесу, формулювання та формалізація інформаційної бази моделі щодо достатньої кількості одиниць у сукупності (у 8 разів більше, ніж число факторів, при цьому 1 факторна не менша ніж 8 точок, а множинна регресія не менше ніж n = 3*кількість факторів).

Розрахуємо мінімальний обсяг ряду спостереження: 8 < 3*3 = 9 < 16.

Шістнадцять років достатня кількість спостережень.

3. **Формалізаційний етап** – прийняття гіпотези взаємозв'язку, вибір загального виду моделі, специфікація зв'язків у рівняннях.

На першому етапі аналізують інформаційний ряд на наявність впливових викидів. Усі впливові спостереження є викидами, при цьому не всі викиди є впливовими. Викиди, які суттєво виріз-няються від значень ряду спостереження, але візуально їх можна розмістити на побудовану лінійну апроксимацію, є невпливовими.

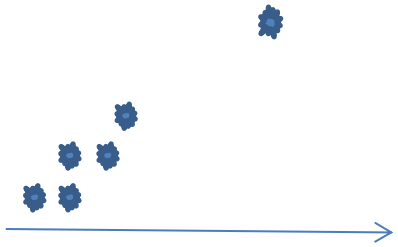


Рис. 6.1. Викид невпливовий

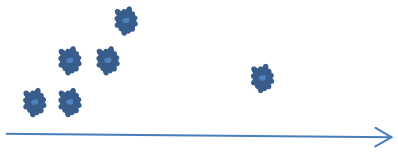


Рис. 6.2. Викид впливовий

Впливові викиди змінюють кут нахилу регресійної прямої, що призводить до зміни коефіцієнта детермінації. Якщо аналізу-ють результати анкетування респондентів, то під час аналізу викидів, у першу чергу, перевіряють наявність арифметичної помилки, а потім аналізують типовість значення: функція в SPSS: «**Данные**» – **«Поиск необычных наблюдений»**.

Зі списка «Переменные» обирають вхідні змінні і переносять у праве діалогове вікно за допомогою стрілки. Натискають «ОК».

У нашому дослідження нетиповим спостереженням є точка 10, оскільки у нас динамічний ряд, цей показник із ряду спостереження вилучати не можна. Візуально відобразимо викиди за усіма вхідними даними моделі за допомогою функції в SPSS: «**Графіка**» – «**Мастер диаграмм**».

З переліку «Галерея» обирається «Рассеяния/Точки», після чого в нижньому діалоговому вікні з'являються їхні види. Фіксуємо ліву кнопку мишки на «Матрица диаграмм рассеяния» і перетягуємо у верхнє діалогове вікно. Після цього зі списку «Переменные» обираємо вхідні дані і переносимо по черзі у верхнє діалогове вікно в нижній прямокутник.

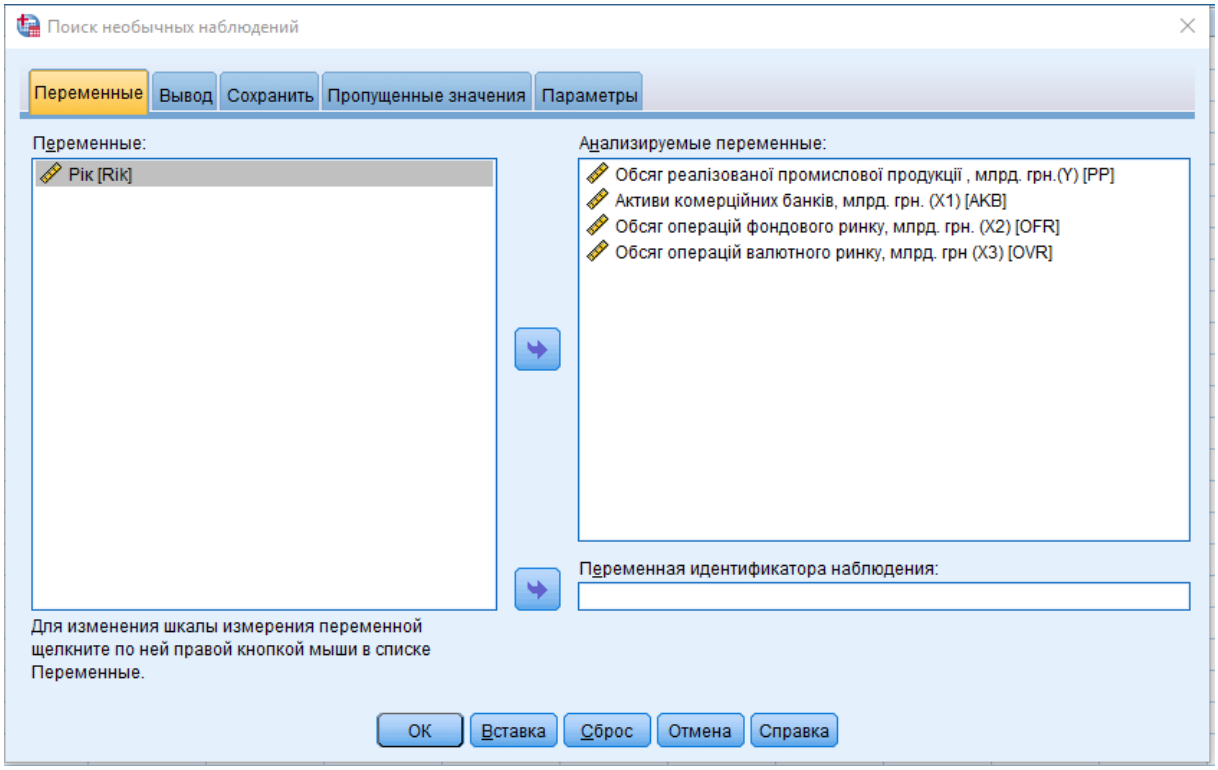


Рис. 6.3. Виявлення нетипових спостережень

Отриманий результат не містить достатньої інформації. Для збільшення його інформативності здійснимо кілька кроків:

1. Побудуємо тренд для всіх функцій. Натиснемо на малюнок двічі лівою клавішою мишки. Активізується малюнок і з'явиться нова панель управління ним «Редактор диаграмм». На верхній панелі обираємо «Элементы» – «Линия апроксимации для итога».

2. По діагоналі побудуємо діаграми. На верхній панелі обираємо «Параметры» – «Показать диаграммы на диагонали».

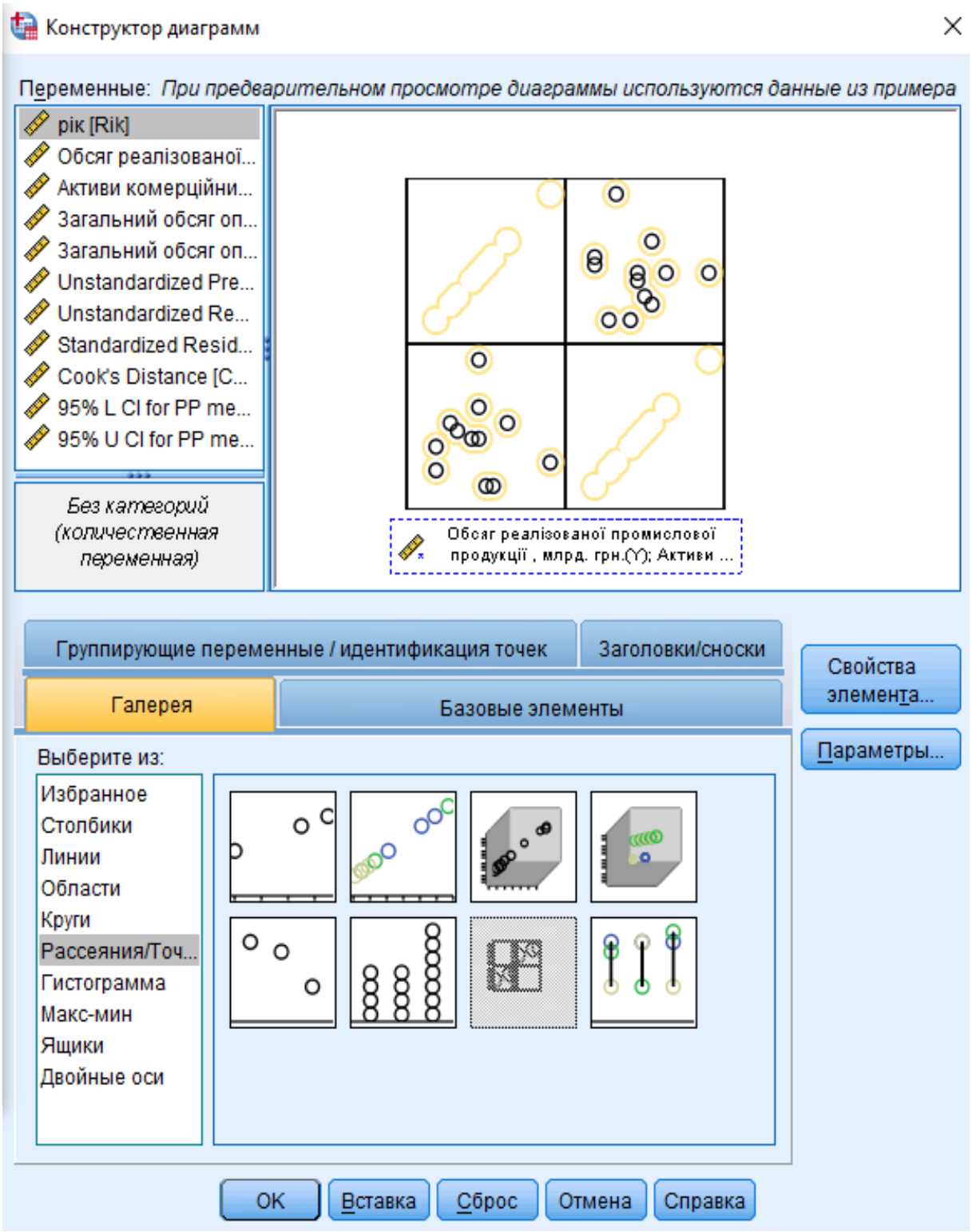


Рис. 6.4. Побудова діаграми розсіювання за всіма показниками

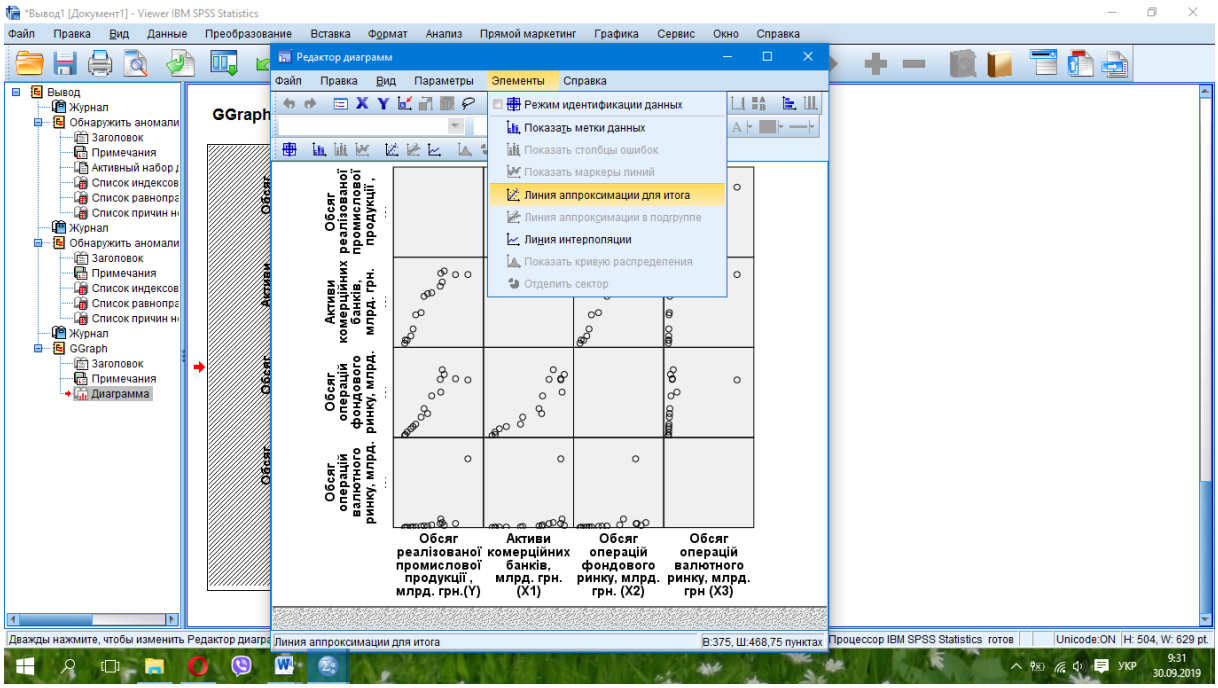


Рис. 6.5. Побудова лінійної апроксимації

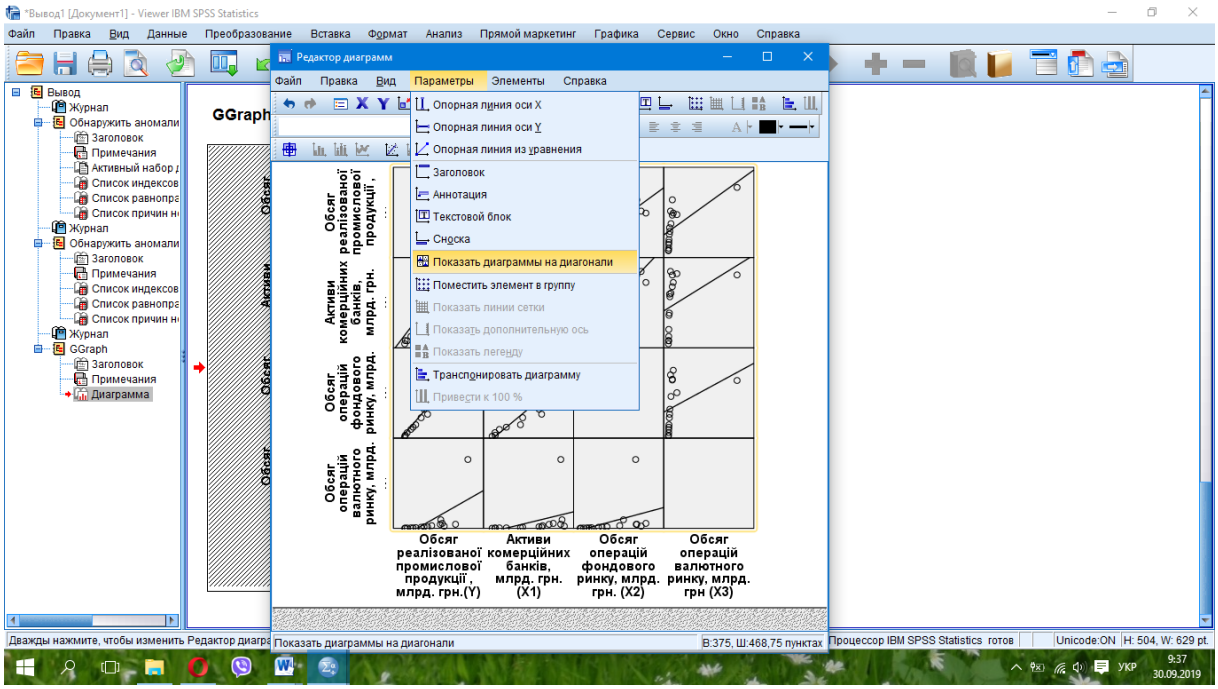


Рис. 6.6. Відображення діаграми на діагоналі діаграми розсіювання

Обсяг операцій валютного ринку має найгірший розподіл, є випадковий викид, який у подальшому вплине на якість моделі. Графіки підбору дають можливість оцінити, яку функцію необхідно будувати, на рисунку лінійна залежність.

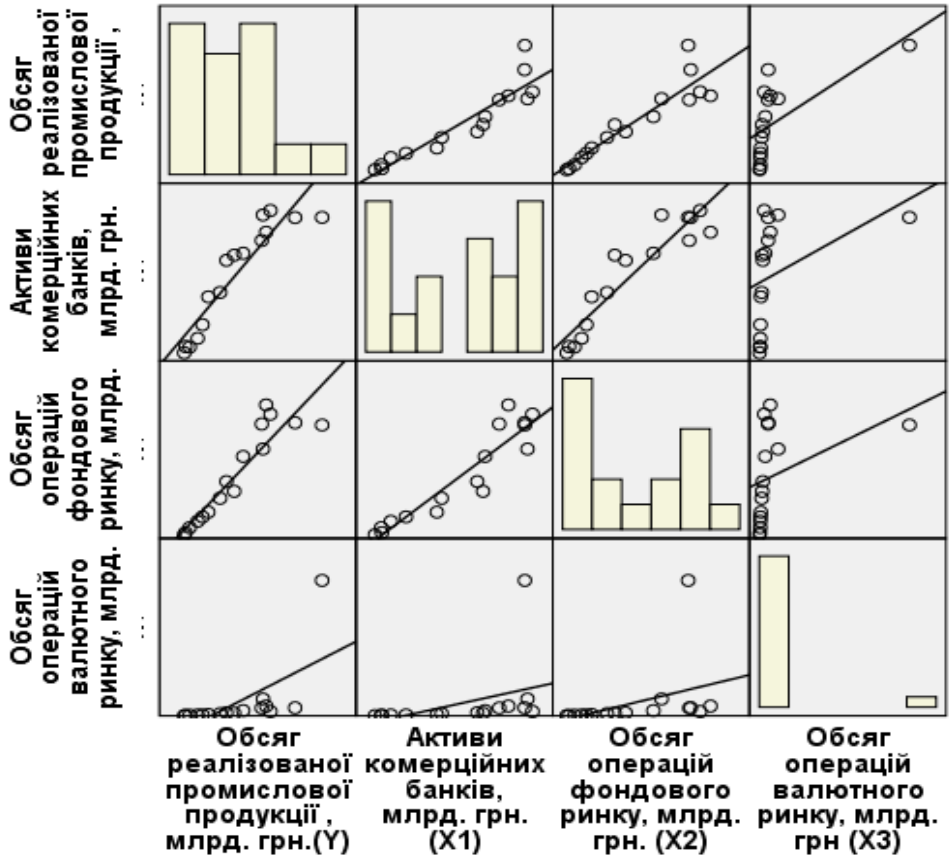


Рис. 6.7. Графічний аналіз взаємозв'язку між показниками

Графіки підбору дають можливість оцінити, яку функцію необхідно будувати, на рисунку лінійна залежність.

4. **Кореляційний етап** – визначає, який зв'язок між досліджуваними змінними його структури, форми та щільності. Детальний кореляційний аналіз за цим прикладом (табл. 6.2) описаний у розділі 5.

За його результатами вилучаємо обсяг операцій фондового ринку й аналогічно будуємо матрицю парних коефіцієнтів кореляції з трьома вхідними даними. Матриця парних коефіцієнтів кореляції вказує на відсутність між екзогенними змінними лінійного зв'язку (розрахований парний коефіцієнт кореляції не перевищує значення 0,9), тобто відсутня мультиколінеарність.

Мультиколінеарність – це лінійна залежність між незалежними змінними (факторами), вона не дає можливості визначити вплив кожного з них на результативну ознаку.

Таблиця 6.2

Матриця парних коефіцієнтів кореляції Пірсона

		Обсяг реалізованої промислової продукції, млрд грн (Y)	**Активи комерційних банків, млрд грн (X1)**	**Обсяг операцій валютного ринку, млрд грн (X3)**
Обсяг реалізованої промислової продукції, млрд грн (Y)	Кореляція Пірсона	1	,924**	,640**
	Знач. (двостороння)		,000	,008
	N	16	16	16
Активи комерційних банків, млрд грн (X1)	Кореляція Пірсона	,924**	1	,387
	Знач. (двостороння)	,000		,139
	N	16	16	16
Обсяг операцій валютного ринку, млрд грн (X3)	Кореляція Пірсона	,640**	,387	1
	Знач. (двостороння)	,008	,139	
	N	16	16	16

** Кореляція значима на рівні 0,01 (двостороння).

Причини виникнення мультиколінеарності:

1. Помилкове включення в рівняння двох і більше лінійно-залежних змінних.

2. Два і або більше фактори, які в нормальній ситуації слабо корелюють, стають у конкретних умовах вибірки сильно корельованими.

3. У модель включається змінна, що сильно корелює із залежною змінною (результативною ознакою), вплив інших факторів становиться незначним – це домінантна змінна.

Наслідки мультиколінеарності:

1. Оцінки коефіцієнтів залишаються незміщеними.

2. Стандартні похибки коефіцієнтів збільшуються.

3. Розрахунок *t-статистики* занижений.

4. Оцінки втрачають стійкість до зміни специфікації і зміни окремих спостережень (втрачається стійкість моделі). Загальна якість рівняння, а також оцінки змінних, не пов’язаних із мультиколінеарністю, залишаються незмінними.

Хоча надійних методів тестування мультиколінеарності не існує, є декілька її *ознак*:

1) незначні зміни у вихідних даних (наприклад, додавання нових даних) призводять до істотних змін оцінок коефіцієнтів регресії;

2) високе значення коефіцієнта детермінації за незначущості параметрів за *t*-статистикою Стьюдента;

3) у моделі між двома екзогенними змінними – велике значення парного коефіцієнта кореляції.

Усі ці ознаки мультиколінеарності мають один спільний недолік: жодна з них чітко не розмежовує випадки, коли мультиколінеарність істотна і коли нею можна знехтувати. На жаль, немає універсальних методів, дослідник сам обирає напрями усунення мультиколінеарності:

1) використати додаткову або первинну інформацію;

2) об'єднати інформацію;

3) відкинути змінну з високою кореляцією;

4) перетворити дані;

5) збільшити обсяг спостережень;

6) перетворити мультиколінеарні змінні, а саме: використати нелінійні форми, лінійні комбінації декількох змінних, перші різниці замість самих змінних.

При цьому простежується прямий лінійний зв'язок між ендогенною змінною і екзогенними. Усі розраховані значення вищі за критичне значення коефіцієнта кореляції:

$$r_{\text{кр}} = \sqrt{\frac{2{,}145^2}{2{,}145^2+(16-2)}} = 0{,}4973.$$

Критичне значення *t*-статистика Стьюдента обирається з таблиці для $a = 0{,}05$ і $\gamma = (n - 2) = 16 - 2 = 14$ ступенів свободи його значення дорівнює 2,145.

Зауважимо, що, крім застосування методу критичного коефіцієнта кореляції, для визначення переліку екзогенних змінних для регресійної моделі застосовують й інші методи відбору змінних, а саме:

1) примусовий – усі незалежні змінні включаються у рівняння;

2) покроковий – покрокове включення, яке проводиться у порядку зростання *p*-рівня;

3) виключення – покроковий метод, спочатку включають усі незалежні змінні в рівняння регресії, а потім вилучають ті, рівень значимості яких вище критичного. Як правило, критичним значенням є *p*-рівень до 0,1.

5. **Параметричний етап** (знаходження рівняння регресії) – визначення загального виду, структури шуканих зв'язків між *Y* та *X* у вигляді певної параметричної сукупності функції: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$.

На цьому етапі будують рівняння регресії. Функція у SPSS: **«Анализ» – «Регресия» – «Линейная».**

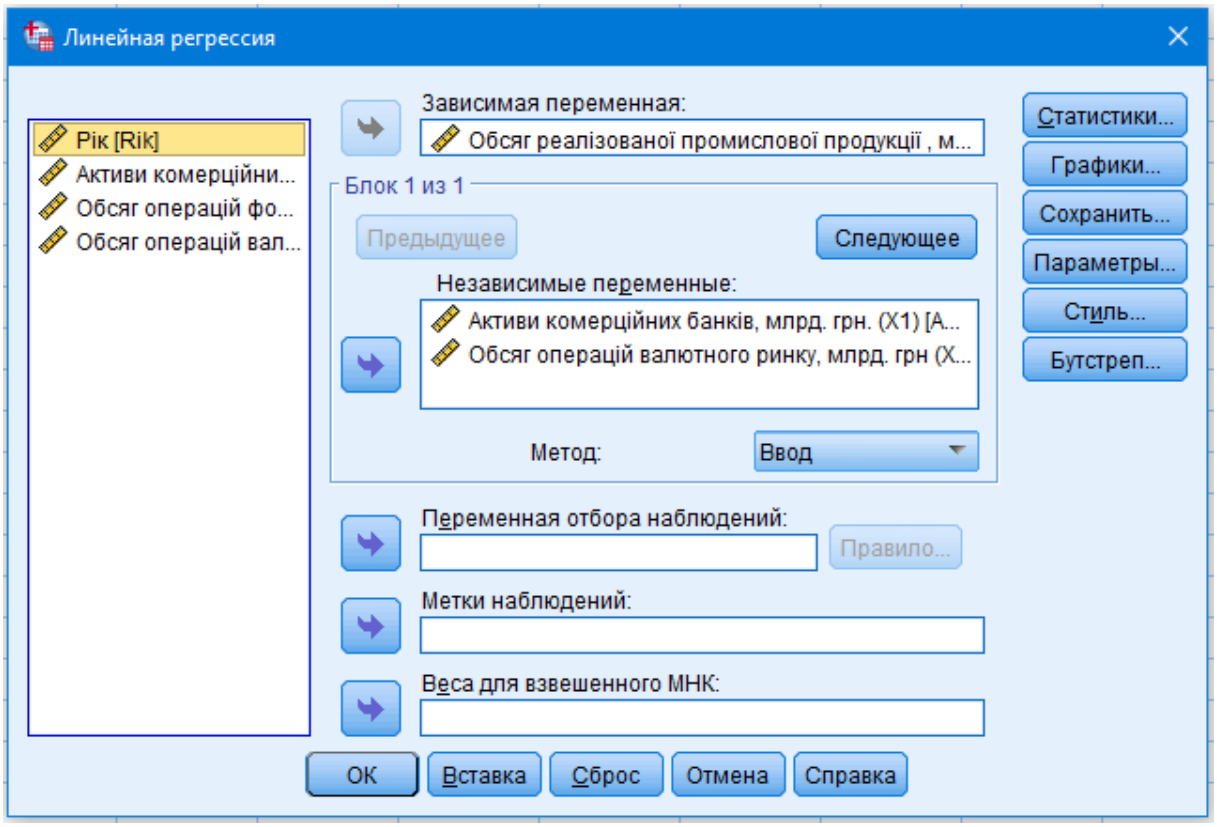


Рис. 6.8. Вибір ендогенної і екзогенних змінних

Обираємо з переліку даних ендогенну й екзогенні змінні і за допомогою стрілки переміщаємо в праві діалогові вікна. Обов'язково необхідно задати статистику моделі (вкладка «Статистики»), графіки (вкладка «Графіки») й обрати основні характеристики моделі і залишків (вкладка «Сохранить»).

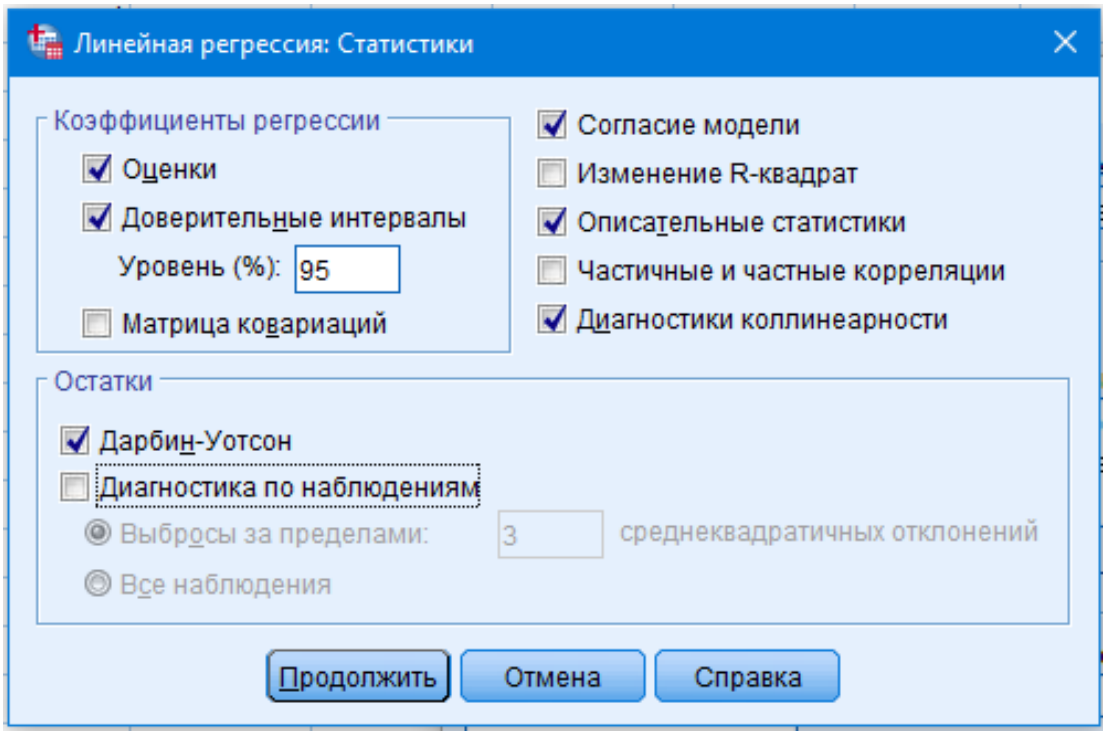


Рис. 6.9. Вивід статистики моделі

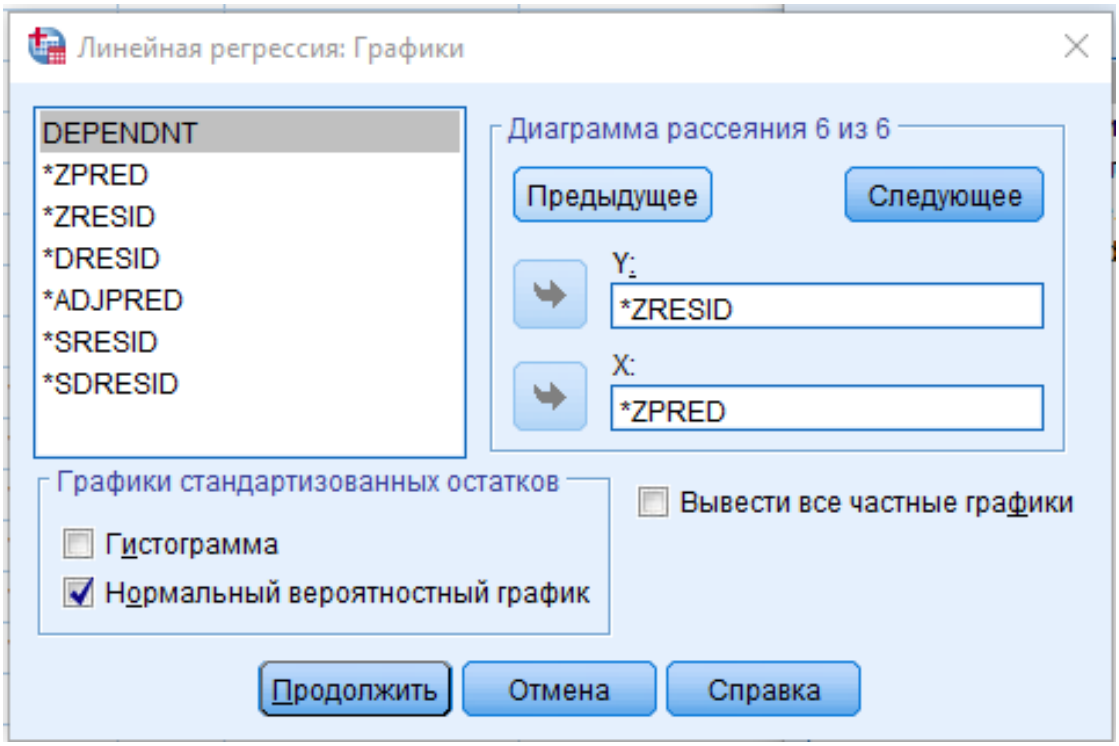


Рис. 6.10. Побудова графіка нормального розподілу залишків для ендогенної змінної

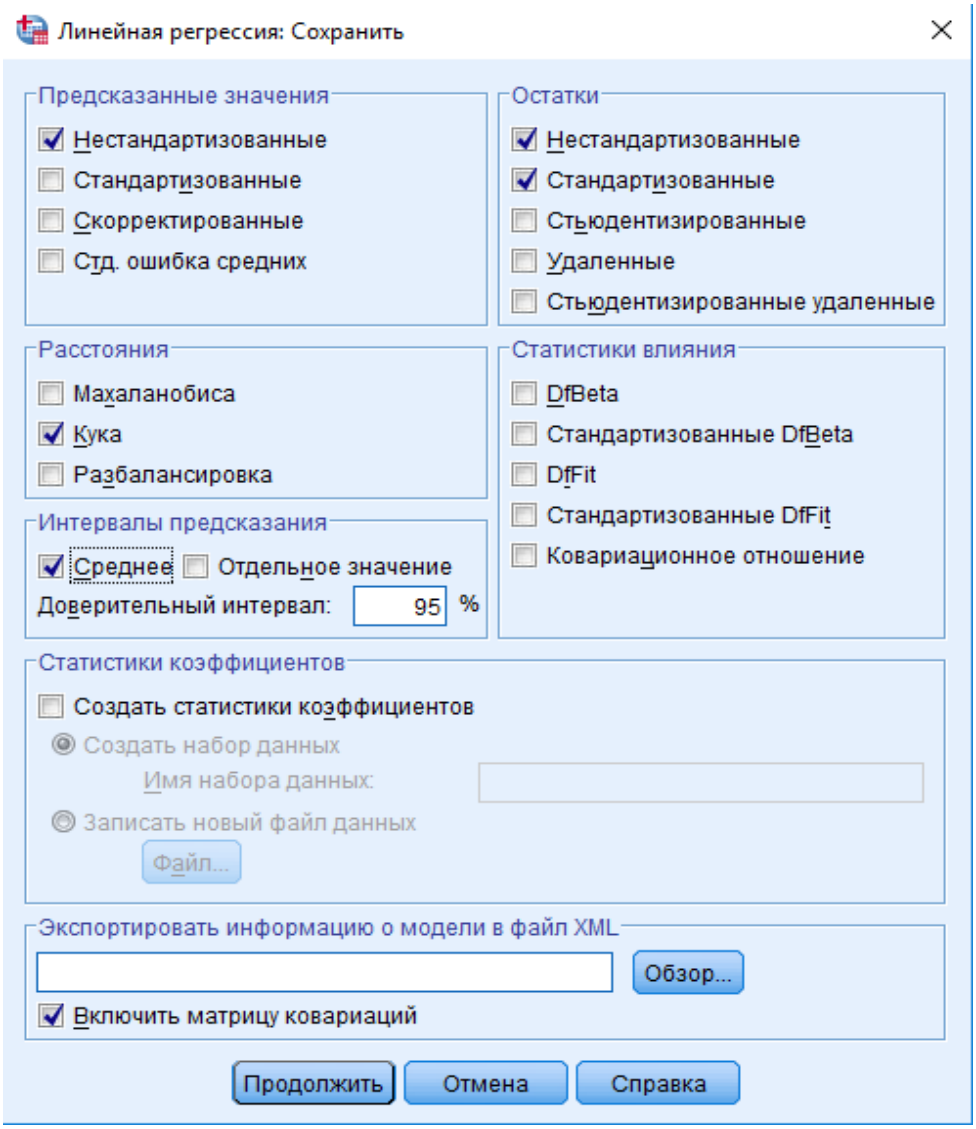


Рис. 6.11. Аналіз залишків

Натиснувши кнопки «Продолжить» і «ОК», отримаємо результат аналізу (табл. 6.3).

Регресійна модель має такий вигляд:

Y=112,842+0,999x_1+0,156x_2.

Зростання активів комерційних банків на 1 млрд грн призведе до зростання обсягу реалізованої промислової продукції на 0,999 млрд грн при сталості інших факторів. А зростання обсягів операцій валютного ринку на 1 млрд грн призведе до зростання обсягу реалізованої промислової продукції на 0,156 млрд грн, відповідно.

Стандартизовані коефіцієнти бета визначають, яка екзогенна зміна має найбільший вплив (еластичність за стандартним відхиленням) на ендогенну змінну.

Таблиця 6.3

Коефіцієнти[a]

Модель		Нестандартизовані коефіцієнти		Стандартизованні коефіцієнти	т	Значимість	95,0 % Довірчий інтервал для В		Статистика колінеарності	
		В	Стандартна похибка	Бета			Нижня межа	Верхня межа	Допуск	VIF
1.	(Константа)	112,842	70,284		1,606	0,132	–38,998	264,681		
	Активи комерційних банків, млрд грн (X1)	0,999	0,086	0,796	11,587	0,000	0,812	1,185	0,850	1,176
	Обсяг операцій валютного ринку, млрд грн (X3)	0,156	0,032	,332	4,833	0,000	0,086	0,225	0,850	1,176

а. Залежна змінна: обсяг реалізованої промислової продукції, млрд грн (Y).

Стандартизований коефіцієнт показує в 2,4 раза більший вплив на обсяг реалізованої промислової продукції активів комерційних банків (0,796), ніж вплив валютного ринку (0,332).

6. **Проблемний етап** – аналіз моделі на відсітність мультиколінеарності, гетероскедастичності та автокореляції, їх усунення у разі виявлення.

Мультиколінеарність діагностується у разі аналізу матриці парних коефіцієнтів кореляції. Крім того, в SPSS вбудований алгоритм перевірки на мультиколінеарність через дисперсійно-інфляційний фактор (параметр VIF), значення якого за факторами повинно бути меншим за граничний показник (VIF = 10, при p = 0,01), а значення толерантності (допуск) вище 0,1.

Функція в SPSS проставляється під час побудови лінійної регресії у вкладці **«Статистика» – «Диагностика коллинеарності».**

У нашому випадку мультиколінарність відсутня: VIF = 1,176 < 10, допуск 0,850 > 0,1.

Гетероскедастичність – це змінність дисперсії вільного члена, виникає у разі перехресної вибірки, рідше – у часовому ряді, тобто у системі наявні різні дисперсії помилок.

Джерело гетероскедастичності:

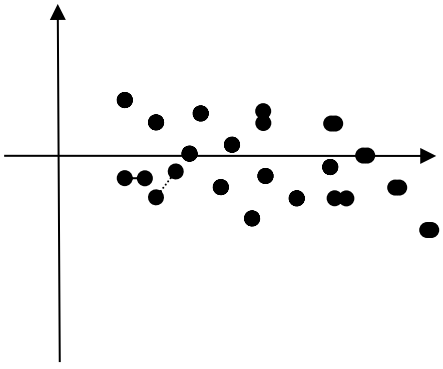
1. Дисперсія зростає із зростанням одного із факторів.
2. У часовому ряді, коли факторна залежна змінна має більший інтервал якісно неоднорідних значень або високий темп зростання.
3. Якщо якість даних варіює всередині вибірки.

Наслідки: оцінки незміщені, але збільшується дисперсія розподілення оцінок коефіцієнтів і призводить до недооцінки стандартних похибок (заниження).

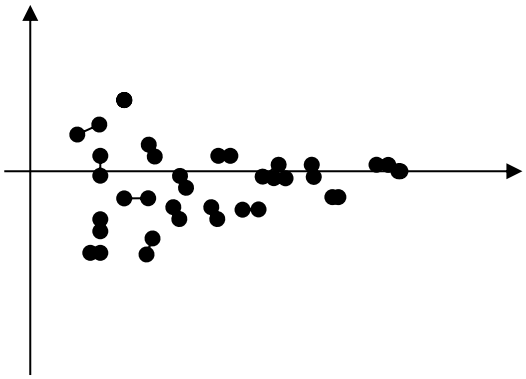
Аналіз графіків залишків є одним із методів вимірювання гетероскедастичності в моделі (рис. 6.12).

Засоби усунення: використати зважений МНК, перерозподілити змінні або розрахувати похибки з поправкою на гетероскедастичність (тест Уайта).

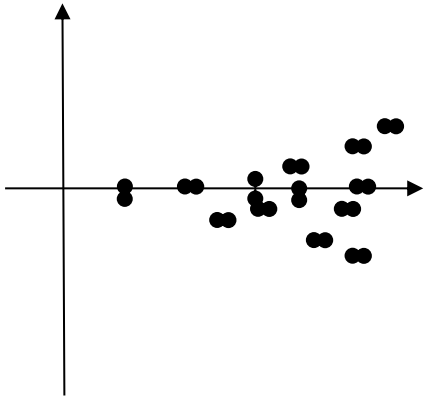
Провести діагностику залишків у SPSS досить зручно. Регресійний аналіз має вмонтовані функції, які досліджують стандартизовані залишки (перевірка на гетероскедастичність). Аналізуючи стандартизовані залишки (z-scores) можна скористатися результатами статистики залишків або діаграмою розсіювання (вкладка «Графики»).

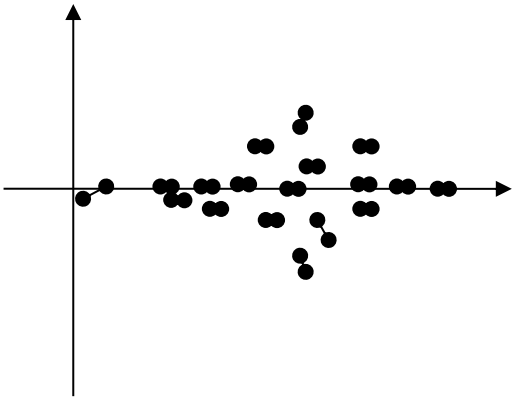


а) гомоскедастичність



б) гетероскедастичність





в) гетероскедастичність г) гетероскедастичність.

Рис. 6.12. Візуальне представлення залишків моделі

Межі z-scores:

95 % – (–1,96; +1,96)

99 % – (–2,58; +2,58)

99,9 % – (–3,29; +3,29)

|стд. залишки| > 3 – викликають підозру про ймовірність того, що це значення попало у вибірку випадково.

Якщо більше 1 % спостережень за абсолютною величиною перевищують 2,5, то модель погано представляє реальні дані.

Якщо більше 5 % спостережень за абсолютною величиною перевищують 2, то модель погано представляє реальні дані (рис. 6.13).

Стандартизоване прогнозне значення не повинно виходити за межі (–3; 3).

У побудованій моделі відсутня гетероскедастичність.

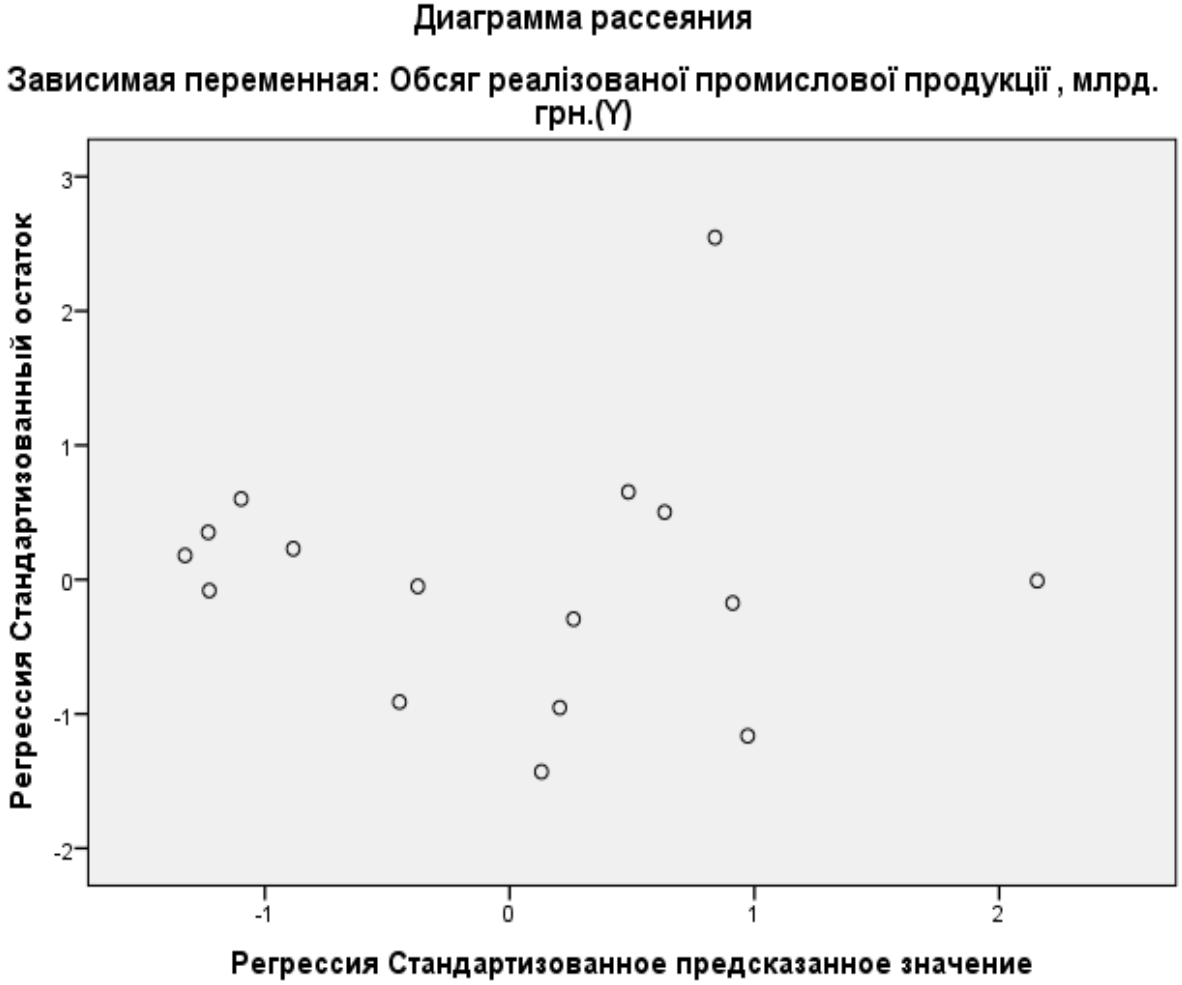


Рис. 6.13. Діаграма розсіювання стандартизованого прогнозного значення

Автокореляція – це залежність випадкового члена ряду спостереження від минулих значень.

Наслідки автокореляції:

1. Збільшуються дисперсії оцінок коефіцієнтів.
2. Занижуються стандартні похибки коефіцієнтів.

Методи вимірювання:

1. Статистика Дарбіна-Уотсона визначає автокореляцію першого порядку (однофакторні моделі) у часових рядах (не використовують для визначення: сезонної автокореляції, для аналізу: лагової моделі), вона вивчає залишки рівняння регресії:

$$DW = \frac{\sum_{2}^{T}(\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{1}^{T}\varepsilon_t^2}.$$

Розрахований критерій полягає в межах від: $0 \le DW \le 4$, при цьому якщо

– $DW < 2$ і значення близькі до нуля – присутня позитивна автокореляція;

– $DW > 2$ $DW > 2$ і значення наближаються до 4 – присутня від'ємна автокореляція;

– $DW =$ 2або наближений до 2 – автокореляція відсутня (табл. 6.4).

Таблиця 6.4

Межі та характеристика критерію Дарбіна-Уотсона

Позитивна автокореляція	**Гіпотеза**	**Негативна автокореляція**
$DW < d_L$	Присутня автокореляція	$DW > 4 - d_L$
$DW > d_U$	Відсутня автокореляція	$DW < 4 - d_U$
$d_L \le DW \le d_U$	Зона невизначеності	$4 - d_U \le DW \le 4 - d_L$

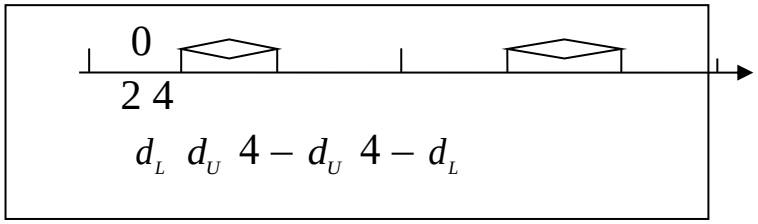


Рис. 6.14. Графічне використання тесту Дарбіна-Уотсона

Засоби усунення: використати узагальнюючий МНК.

Табличне значення d-розподілу для (додаток Ж):

$$d_l(0{,}05;16;2)=0{,}982$$

$$d_u(0{,}05;16;2)=1{,}539$$

$$1{,}539<1{,}803<(4-1{,}539)=2{,}461$$

$d_u < DW < 4 - d_u$ – автокореляція відсутня (табл. 6.5).

Таблиця 6.5

Результати моделювання[b]

Модель	R	R-квадрат	Скорегований R-квадрат	Стандартна похибка оцінки	Дарбін-Уотсон
1	,974[a]	,948	,940	142,821	1,803

a. Предиктори: (константа), Активи комерційних банків, млрд грн (X1), Обсяг операцій валютного ринку, млрд грн (X3),

b. Залежна змінна: Обсяг реалізованої промислової продукції, млрд грн (Y).

У разі потрапляння тесту Дарбіна-Уотсона в зону невизначеності або побудови моделі з лаговими змінними використовують тест Бреуша-Годфрі.

7. **Ідентифікаційний етап** – аналіз достовірності оцінок невідомих параметрів, що входять у регресійну модель.

Коефіцієнт кореляції множинної регресії r = 0,974 вказує на лінійний прямий і сильний зв'язок.

Коефіцієнт детермінації полягає у частці поясненої дисперсії в загальній дисперсії. Якщо R^2 досягає свого найбільшого можливого значення, то одночасно мінімізується сума квадратів залишків:

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS} = 1 - \frac{\sum e_i^2}{\sum \left(y_i - \bar{y}\right)^2} = 1 - \frac{SS_{ост}}{SS_{регр} + SS_{ост}} = \frac{SS_{регр}}{SS_{регр} + SS_{ост}},$$

$$r = \sqrt{R^2}.$$

Коефіцієнт детермінації R^2 лежить у межах від 0 до 1. Якщо R^2 = 0, то це означає, що залежність між Y та X відсутня, тобто X не впливає на Y. Коли R^2=1, в системі існує функціональний зв'язок між Y та X, тобто всі точки будуть лежати на підібраній прямій, що є можливим, коли виконується умова ESS = 0.

Розрахований R^2 = 0,948 це означає, що варіація результативної ознаки Y на 94,8 % залежить від варіації факторних ознак, а

5,2 % – це варіація факторів, які не включені в модель, включаючи стохастичну змінну.

Скоригований коефіцієнт детермінації (або нормований) використовують для оцінки реальної тісноти зв'язків між результативною і факторною ознаками, для порівняння моделі з різною кількістю параметрів (X, при цьому ряд спостереження повинен бути однаковий). Звертають увагу на близькість скоригованого і звичайного коефіцієнтів детермінації. Якщо ці показники мають високі значення і несуттєво вирізняються, модель вважається гарною $\overline{R}^2 < R^2$:

$$1-\left(1-R^2\right)\cdot\frac{n-1}{n-m-1},$$

де n– кількість спостережень, m– кількість факторів (x).

Стандартна похибка ($S=\sigma_\varepsilon$) – це оцінка середнього квадратичного відхилення коефіцієнта регресії від його істинного значення. Коефіцієнт буде значимим, якщо є достатня ймовірність, що його істинне значення відрізняється від нуля (табл. 6.6). Коли характеризують різні моделі з високим значенням коефіцієнта детермінації (що мають різну кількість факторів, кількість спостережень (n) однакова) обирають для прогнозу ту, в якій стандартна похибка менша:

$$S=\sigma_u=\sqrt{\frac{\sum e_t^2}{n-m-1}}=142{,}821.$$

Таблиця 6.6

Дисперсійний аналіз ANOVA[a]

Модель		Сума квадратів *SS*	ст. св.	Середній квадрат *MS*	*F*	Значимість *F*
1	Регресія ***R***	4822363,656	2	2411181,828	118,208	,000[b]
	Залишок ***E***	265170,235 $\hat{\varepsilon}^2$	13 (n-k)	20397,710 $S^2=\sigma_\varepsilon^2$		
	Усього ***T***	5087533,891	15 (n-1)			

a. Залежна змінна: обсяг реалізованої промислової продукції, млрд грн (Y).

b. Предиктори: (константа), обсяг операцій валютного ринку, млрд грн (X3), активи комерційних банків, млрд грн (X1).

MSE – середній квадрат відхилень відносно регресії ($S^2 = \sigma_\varepsilon^2$) – дає оцінку залишкової дисперсії відносно регресії, яка базується на *n* – *k* ступенів свободи. У моделі з кращим підбиранням $S^2 = \sigma_\varepsilon^2$ має менші значення. *MSR* – середня сума квадратів відхилення, що обумовлена регресією.

Сума квадратів відхилень:

$$RSS = SS_{регр} = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}_i\right)^2,$$

$$ESS = SS_{ост} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2,$$

$$TSS = SS_{итого} = SS_{регр} + SS_{ост}.$$

Регресія буде значимою, якщо сума квадратів відхилень регресії відносно середнього буде більшою порівняно з сумою квадратів відхилень залишків, тобто

$$RSS > ESS,$$

$$4\,822\,363{,}656 > 265\,170{,}235.$$

Регресійна модель значима (табл. 6.7).

Таблиця 6.7

Коефіцієнти[a]

Модель		Нестандартизовані коефіцієнти		Стандартизованні коефіцієнти	т	Значимість	95,0 % Довірчий інтервал для В		Статистика колінеарності	
		В	Стандартна похибка	Бета			Нижня межа	Верхня межа	Допуск	VIF
1	(Константа)	112,842	70,284		1,606	,132	-38,998	264,681		
	Активи комерційних банків, млрд грн (X1)	,999	,086	,796	11,587	,000	,812	1,185	0,850	1,176
	Обсяг операцій валютного ринку, млрд грн (X3)	,156	,032	,332	4,833	,000	,086	,225	0,850	1,176

Розраховане значення *t*-статистики Стьюдента дає можливість оцінити параметри моделі. Вона порівнює значення коефіцієнта з його стандартною похибкою:

$$t_i = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}} = \frac{r\sqrt{n-k}}{\sqrt{1-r^2}}.$$

Порядок перевірки значимості коефіцієнта за *t*-статистикою:

1. Обираємо рівень значимості α (1 % або 5 %).

2. Розраховуємо кількість ступенів свободи ($n-k$).

3. За таблицею розподілу Стьюдента визначаємо критичне значення *t* для *a*, *n* – *k*, де *k* = *m* + *1* (додаток А).

4. Якщо розраховане значення *t*-статистики за модулем (за абсолютною величиною) більше табличного значення, то розраховані коефіцієнти рівняння є істотними для обраного рівня значимості а α і рівняння можна використовувати для прогнозу. В іншому випадку коефіцієнти неістотні.

$t_{табл}(0{,}05;\ 16–2–1=13)=2{,}16$.

$t_0 = 11{,}587 > t_{табл} = 2{,}16$ p_1 =0,00 <0,05

$t_1 = 4{,}833 > t_{табл} = 2{,}16$ p_2 = 0,00<0,05

t-статистика розрахована значно більша за критичне табличне значення за всіма параметрами, модель адекватна. Для перевірки адекватності параметрів можна застосовувати значимість оцінок *t*-статистики Стьюдента. Порівнюючи *p*-значення з рівнем значущості $\alpha = 0{,}05$, маємо: якщо *p*-значення більше або дорівнює *a*, то коефіцієнт рівняння регресії незначущий, і навпаки. У нашому випадку *t*-статистика Стьюдента значима, модель адекватна.

Порядок перевірки значимості коефіцієнта за *F*-статистикою:

$$F = t^2 = \frac{R^2(n-k)}{1-R^2} = \frac{RSS}{ESS/(n-k)} = \frac{MSR}{MSE}.$$

1. Обираємо рівень значимості α (1 або 5 %).

2. Розраховуємо кількість ступенів свободи: $f_1 = m$ і $f_2 = n – k$.

3. За таблицею розподілу Фішера визначаємо критичне значення *F* для *a*, $f_1 = m$ і $f_2 = n – k$.

4. Якщо розраховане значення *F*-статистики за модулем (за абсолютною величиною) більше табличного значення (додаток Г), то розраховані коефіцієнти рівняння є істотними для обраного рівня значимості *a* і рівняння можна використовувати для прогнозу. В іншому випадку коефіцієнти неістотні.

$$F_{табл}(0{,}05; f_1 = 2;\ f_2 = 16 - 2 - 1 = 13) = 3{,}81.$$

Розраховане значення 118,208 (*p*-значення = 0,000), що більше за критичне значення, модель адекватна.

Стандартна похибка параметрів моделі вказує, на скільки будуть варіювати коефіцієнти від вибірки до вибірки.

Чим менше довірчі інтервали, тим модель більш репрезентативна.

Відстань Кука визначає силу впливу викиду на рівнях (N від 0 до 1), якщо немає впливових викидів, модель гарна. Розраховані значення відстані Кука коливаються від 0 (min) до 0,512 (max), що вказує на відсутність впливових викидів (табл. 6.8).

Таблиця 6.8

Статистика залишків[a]

	Мінімум	**Максимум**	**Середнє**	**Середньокв. відхилення**	**N**
Прогнозоване значення	185,1150	2159,2649	937,0963	567,00168	16
Стандартне прогнозоване значення	–1,326	2,155	0,000	1,000	16
Стандартна похибка прогнозованого значення	38,438	142,286	57,077	24,587	16
Скореговане прогнозоване значення	178,3041	2323,4541	944,4897	592,44419	16
Залишок	–204,26184	363,78345	,00000	132,95870	16
Стандартний залишок	**–1,430**	**2,547**	0,000	0,931	16
Відстань Махалонобиса	0,149	13,951	1,875	3,298	16
Відстань Кука	**0,000**	**0,512**	00,079	0,158	16
Значення центрованої балансировки	0,010	0,930	0,125	0,220	16

а. Залежна змінна: обсяг реалізованої промислової продукції, млрд грн (Y).

Більш повну інформацію про розподіл у моделі дає графік нормального розподілу стандартизованих залишків лінійної регресії. Залишки розподілені на прямій, що вказує на нормальний розподіл.

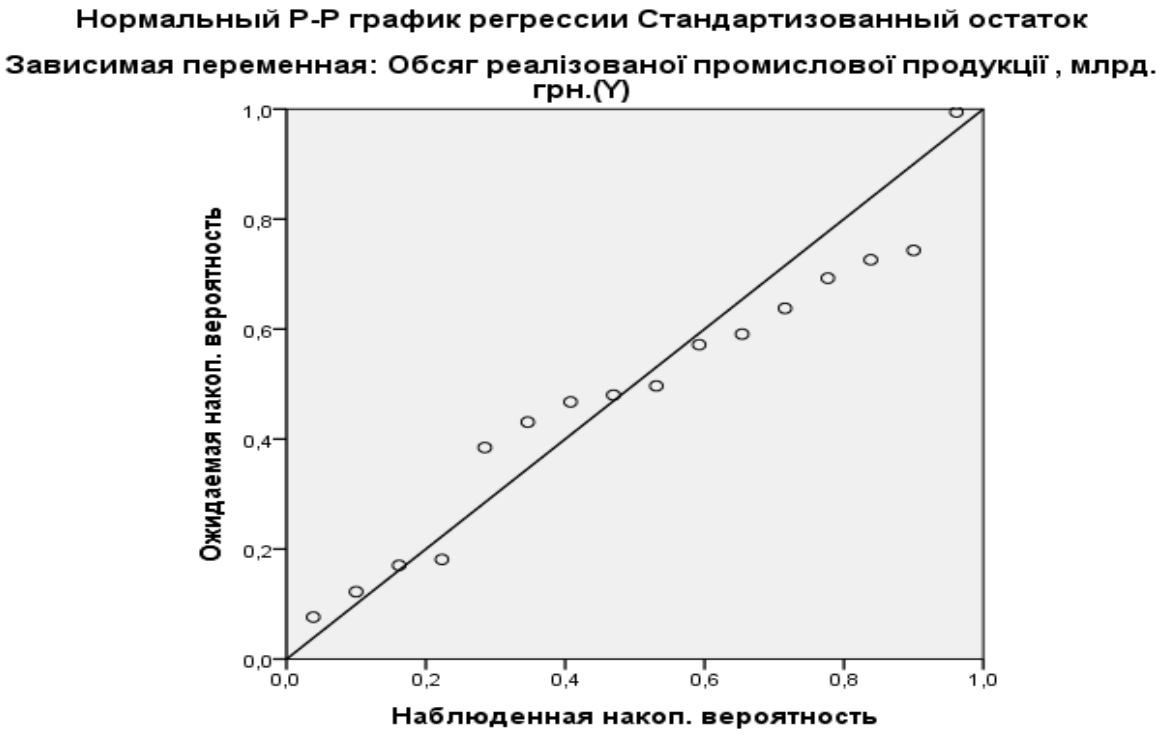


Рис. 6.15. Графік нормального розподілу стандартизованих залишків моделі

Файл Правка Вид Данные Преобразование Анализ Прямой маркетинг Графика Сервис Окно

21 : PRE_1

	Rik	PP	AKB	OVR	PRE_1	RES_1	ZRE_1	COO_1	LMCI_1	UMCI_1
1	2001	210,84	70,27	13,48	185,11499	25,72501	,18012	,00362	43,94627	326,28372
2	2002	229,63	125,43	20,44	241,28421	-11,65421	-,08160	,00062	108,26110	374,30731
3	2003	289,12	120,99	32,91	238,78981	50,33019	,35240	,01180	105,02329	372,55633
4	2004	400,76	196,95	35,01	314,97538	85,78462	,60065	,02697	192,05350	437,89725
5	2005	468,56	316,97	40,90	435,75184	32,80816	,22972	,00276	328,41656	543,08711
6	2006	551,73	561,13	55,72	681,89214	-130,16214	-,91137	,02472	596,72401	767,06027
7	2007	717,08	599,40	82,24	724,23637	-7,15637	-,05011	,00007	641,19576	807,27698
8	2008	917,04	926,09	99,31	1053,14683	-136,10683	-,95299	,03183	961,85942	1144,43424
9	2009	806,55	880,30	121,13	1010,81184	-204,26184	-1,43020	,06405	923,77103	1097,85266
10	2010	1043,11	942,09	201,41	1085,00708	-41,89708	-,29335	,00291	995,10567	1174,90848
11	2011	1305,31	1054,28	298,02	1212,07535	93,23465	,65281	,01818	1113,06202	1311,08867
12	2012	1367,93	1127,19	370,16	1296,10962	71,82038	,50287	,01285	1189,91574	1402,30349
13	2013	1322,41	1278,10	638,34	1488,53388	-166,12388	-1,16317	,09846	1366,80751	1610,26025
14	2014	1428,84	1316,85	165,80	1453,72925	-24,88925	-,17427	,00321	1315,09017	1592,36833
15	2015	1776,60	1254,39	303,79	1412,81656	363,78344	2,54714	,51156	1287,50547	1538,12765
16	2016	2158,03	1256,30	5090,3	2159,26487	-1,23487	-,00865	,44385	1851,87360	2466,65613

Рис. 6.16. Результати моделювання у вкладці «Данные»

PRE_1 (Unstandardized Predicted Value) – прогнозний $\hat{Y}$

RES_1(Unstandardized Residual) – різниця між $Y–\hat{Y}$

ZRE_1 (Standardized Residual) – стандартна похибка

COO_1(Cook’s Distance) – відстань Кука

LMCI_1 (95 % L CI for PP mean) – нижній довірчий інтервал

UMCI_1 (95 % U CI for PP mean) – верхній довірчий інтервал.

8. **Верифікація моделі** – зіставлення реальних і розрахункових даних, перевірка адекватності моделі, оцінка точності та стійкості отриманих рівнянь зв’язку, побудова прогнозів та сценаріїв розвитку.

Застосовуємо випадковий відбір спостережень у співвідношенні 70 % і 30 %: **Данные**» – «**Отобрать наблюдения**» – «**Случайная подвыборка**». Поставити мітку «**Примерно 70 % от всех наблюдений**».

Побудувати декілька разів для кожного випадку лінійну регресію: виконати побудову моделі: «**Анализ**» – «**Регресия**» – «**Линейная**». Порівняти результати, якщо показники суттєво не змінюють, модель стійка (табл. 6.9–6.11).

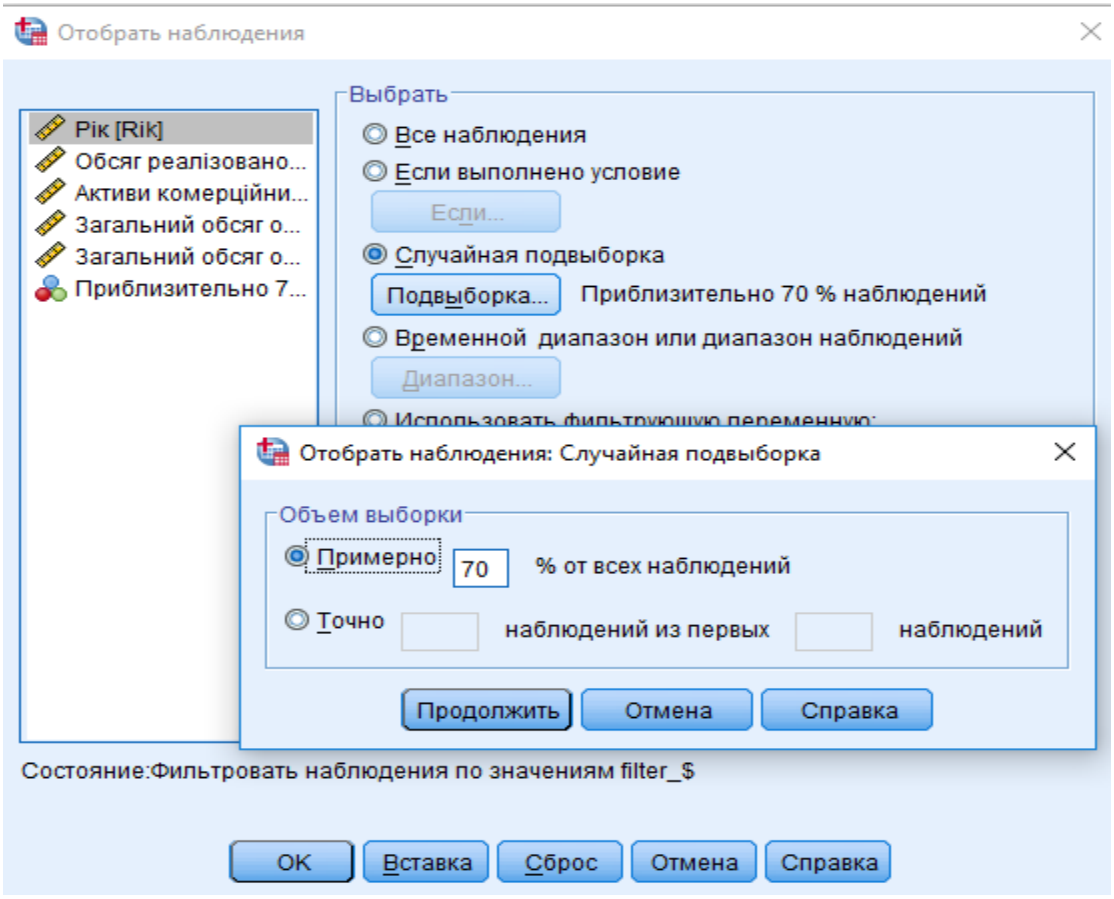


Рис. 6.17. Випадкова вибірка спостережень

Таблиця 6.9

Результати регресійної моделі[b]

Модель	R	R-квадрат	Скорегований R-квадрат	Стандартна похибка оцінки	Дарбін-Уотсон
1	0,971[a]	0,943	0,931	163,49961	2,396

a. Предиктори: (константа), обсяг операцій валютного ринку, млрд грн (X3), активи комерційних банків, млрд грн (X1);

b. Залежна змінна: обсяг реалізованої промислової продукції, млрд грн (Y)

Таблиця 6.10

ANOVA[a]

Модель		Сума квадратів	ст.св.	Середній квадрат	F	Значимість
1	Регресія	4009702,386	2	2004851,193	74,998	0,000[b]
	Залишок	240589,116	9	26732,124		
	Всього	4250291,502	11			

a. Залежна змінна: обсяг реалізованої промислової продукції, млрд грн(Y)

b. Предиктори: (константа), обсяг операцій валютного ринку, млрд грн (X3), активи комерційних банків, млрд грн (X1)

Таблиця 6.11

Коефіцієнти[a]

Модель		Нестандартизовані коефіцієнти		Стандартизовані коефіцієнти	т	Значимість	95,0 % Довірчий інтервал для B		Статистика колінеарності	
		B	Стандартна похибка	Бета			Нижня межа	Верхня межа	Допуск	VIF
1	(Константа)	113,976	93,521		1,219	,254	-97,582	325,534		
	Активи комерційних банків, млрд грн (X1)	1,018	,116	,771	8,763	,000	,755	1,281	,812	1,231
	Обсяг операцій валютного ринку, млрд грн (X3)	,150	,038	,344	3,911	,004	,063	,236	,812	1,231

a. Залежна змінна: обсяг реалізованої промислової продукції, млрд грн (Y)

Висновок: модель стійка, якість моделі не погіршилася, усі якісні характеристики вказують на адекватність регресійної моделі. Змінився лише розрахований коефіцієнт Дарбіна-Уотсона: 1,539<2,396<(4-1,539)=2,461, при цьому він знову ж таки вказує на відсутність автокореляції. У моделі відсутня мультиколінеарність.

Перелік питань для самоконтролю

1. Поясніть, чим відрізняється парна регресія від множинної.
2. Охарактеризуйте види змінних величин у регресійних моделях.
3. Поясніть, чим відрізняється коефіцієнт детермінації і кореляції.
4. Охарактеризуйте основні етапи побудови регресійної моделі.
5. Назвіть критерії адекватності регресійної моделі.
6. Поясність сутність мультиколінеарності та шляхи її усунення.
7. Поясніть, чим небезпечна автокореляція і як її виміряти.
8. Проведіть аналіз графіків залишків регресійної моделі.
9. Поясніть сутність дисперсійного аналізу ANOVA.
10. Матриця парних коефіцієнтів кореляції, який принцип її побудови і застосування.
11. Доведіть доцільність побудови графіка нормального розподілу в регресійному аналізі.
12. Назвіть показники, які характеризують якісну сторону регресійної моделі.
13. Охарактеризуйте сутність гетероскедастичності, поясніть її вплив на регресійну модель.
14. Поясніть сутність дисперсійно-інфляційного фактора, для чого його застосовують у регресійних моделях.
15. Поясніть, у чому сутність автокореляції і які її наслідки для моделі.

Тести

1. Однією з умов побудови багатофакторної моделі є достатня кількість одиниць у сукупності:

а) як мінімум, у 8 разів більше, ніж кількість факторів;

б) 10;

в) 5;

г) правильна відповідь відсутня.

2. Якщо між двома або більше факторами моделі існує мультиколінеарність, то варто:

а) ігнорувати цей факт;

б) розрухувати критерій Дарбіна-Уотсона;

в) вилучити один із факторів;

г) правильна відповідь відсутня.

3. Змінність дисперсії вільного члена – це:

а) мультиколінеарність;

б) автокореляція;

в) гетероскедастичність;

г) правильна відповідь відсутня.

4. Лінійна залежність члена ряду спостереження від минулих значень – це:

а) мультиколінеарність;

б) автокореляція;

в) гетероскедастичність;

г) гомоскедастичність.

5. Зіставлення реальних і розрахункових даних, оцінка точності та стійкості, побудова прогнозів та сценаріїв розвитку визначається на етапі регресійного аналізу:

а) проблемному;

б) ідентифікаційному;

в) верифікаційному;

г) формалізаційному.

6. Критерій Фішера визначається для визначення:

а) мультиколінеарності;

б) автокореляції;

в) адекватності моделі;

г) гетероскедастичності.

7. Якщо розрахований критерій Дарбіна-Уотсона потрапляє в інтервал [4- d_u; 4-d_l], то автокореляція:

а) відсутня;

б) зона невизначеності;

в) присутня;

г) можна визначити наближено.

8. Формулювання та формалізація бази моделі щодо достатньої кількості одиниць у сукупності визначається на етапі регресійного моделювання:

а) ідентифікаційному;

б) інформаційному;

в) параметричному;

г) постанов чому.

9. Якщо розрахований критерій Дарбіна-Уотсона потрапляє в інтервал [d_u; 4- d_u], то автокореляція:

а) відсутня;

б) зона невизначеності;

в) присутня;

г) можна визначити наближено.

10. Критерій Дарбіна-Уотсона міститься у межах:

а) [–1; 1];

б) [0; 1];

в) [0; 4];

г) [0; 2].

11. Коефіцієнт детермінації міститься у межах:

а) [–1; 1];

б) [0; 1];

в) [0; 4];

г) [0; 2].

12. На якому етапі моделювання знаходять рівняння регресії:

а) параметричному;

б) кореляційному;

в) верифікаційному;

г) ідентифікаційному.

13. Лінійна залежність між незалежними змінними – це:

а) мультиколінеарність;

б) автокореляція;

в) гетероскедастичність;

г) гомоскедастичність.

14. Наявність гетероскедастичності в моделі можна знайти, проаналізувавши:

а) графік підбору;

б) графіки залишків моделі;

в) графік нормального розподілу;

г) правильна відповідь відсутня.

15. Якщо розрахований критерій Дарбіна-Уотсона потрапляє в інтервал $[d_l; d_u]$, то автокореляція:

а) відсутня;

б) зона невизначеності;

в) присутня;

г) можна визначити наближено.

16. На скільки відсотків обрані фактори характеризують модель, якщо коефіцієнт детермінації дорівнює 0,68:

а) 32 %;

б) 68 %;

в) 0,68 %;

г) 0,32 %.

17. Коефіцієнт детермінації 0,95 – це означає:

а) факторна змінна на 95 % залежить від результативної і на 5 % від інших величин, не врахованих у моделі;

б) результативна змінна на 95 % залежить від факторних і на 5 % від інших величин, не врахованих у моделі;

в) тісний прямий лінійний зв'язок;

г) тісний обернений лінійний зв'язок.

18. Ступені вільності критерію Фішера для перевірки рівня значущості параметрів моделі, яка досліджується за даними 26 спостережень і 3-х факторів, дорівнюють:

а) 26 і 3;

б) 22 і 3;

в) 29 і 3;

г) 22 і 4.

19. Якщо розрахований критерій Дарбіна-Уотсона потрапляє в інтервал [0; dl], то автокореляція:

а) відсутня;

б) зона невизначеності;

в) присутня;

г) можна визначити наближено.

20. Мультиколінеарність – це:

а) змінність дисперсії вільного члена;

б) лінійна залежність між незалежними змінними;

в) нелінійна залежність між незалежними змінними;

г) лінійна залежність між залежними змінними.

21. На якому етапі регресійного аналізу визначається структура, форма та щільність зв'язків:

а) параметричному;

б) кореляційному;

в) верифікаційному;
г) проблемному.

22. Обчислення достовірності оцінок параметрів, що входять у модель, визначається на етапі регресійного аналізу:
а) ідентифікаційному;
б) кореляційному;
в) верифікаційному;
г) параметричному.

23. Якщо розрахований критерій Дарбіна-Уотсона потрапляє в інтервал $[4–d_u;4]$, то автокореляція:
а) відсутня;
б) зона невизначеності;
в) присутня;
г) можна визначити наближено.

24. Прийняття гіпотези взаємозв'язку, вибір загального виду моделі, специфікація зв'язків у рівняннях:
а) формалізаційному;
б) постановчому;
в) параметричному;
г) кореляційному.

25. Ознака, яка характеризує наслідок дії фактора або факторів, залежна змінна – це змінні:
а) екзогенні;
б) ендогенні;
в) фіктивна;
г) латентні.

26. Ознака, яка характеризує причини та умови, які необхідні для виникнення певного наслідку, що не перебуває під його впливом – це змінні:
а) екзогенні;
б) ендогенні;

в) латентні;
г) фіктивні.

27. Економічні величини, які не входять у рівняння регресійних моделей, але впливають на залежні змінні – це змінні:
а) фіктивні;
б) наперед визначені;
в) лагові;
г) латентні.

28. Категорійні змінні, які мають якісні характеристики, ці змінні бінарного типу – це змінні:
а) наперед визначені;
б) лагові;
в) латентні;
г) фіктивні.

29. Змінну, значення якої відстає на один або декілька періодів, – це змінні:
а) лагові;
б) фіктивні;
в) латентні;
г) екзогенні.

Економічна інтерпретація регресійного аналізу

Приклад. 1. Здійснити статистичну оцінку динаміки розвитку будівельної галузі України, провести регресійний аналіз та здійснити якісний прогноз.

Аналіз динаміки введення в експлуатацію загальної площі будівель за період від 2000 до 2016 рр. (рис. 1) вказує на існування двох періодів, упродовж яких галузь розвивалась по-різному.

Перший період – це період з 2000 до 2008 рр., який можна охарактеризувати стабільним зростанням галузі, на що вказує динаміка її основних показників (рис. 1 і рис. 2).

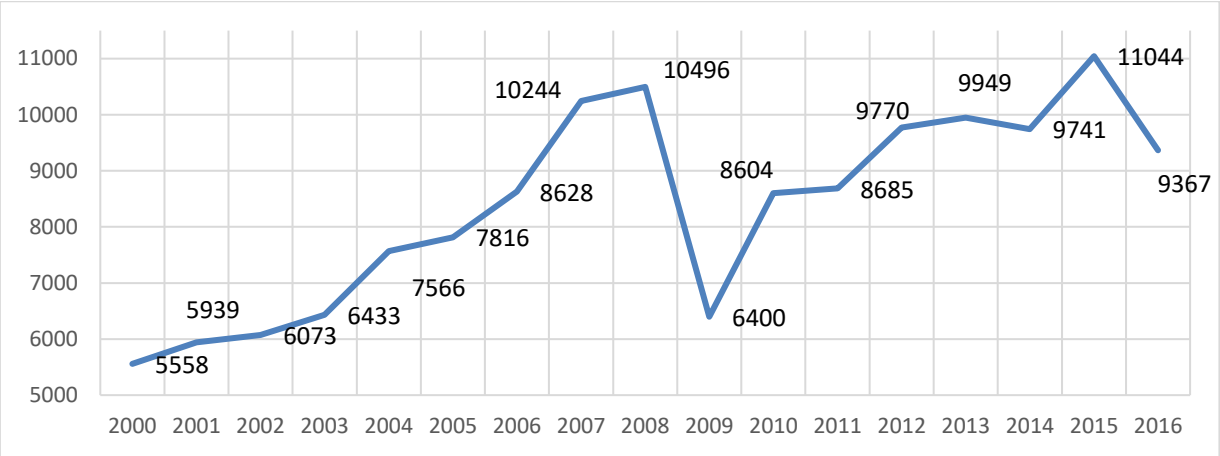


Рис. 1. Динаміка введення в експлуатацію загальної площі, тис. м^2

Джерело: побудовано авторами за даними[7].

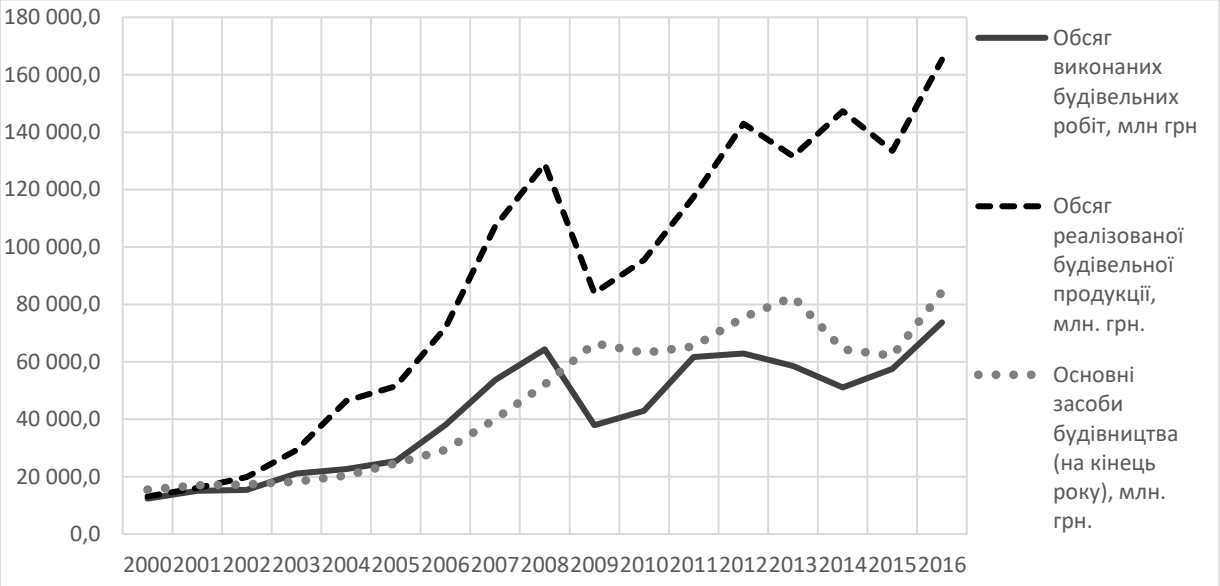


Рис. 2. Динаміка основних показників будівельної галузі

Джерело: побудовано авторами за даними[8].

Незважаючи на те, що темпи їхнього зростання не були сталими, спостерігалось падіння обсягу реалізованої будівельної продукції 2005 року й обсягу виконаних будівельних робіт 2002, 2004 років (рис. 3), проте їхні значення перевищували одиницю.

[7]Офіційний сайт Державної служби статистики України. URL : http://www.ukrstat.gov.ua.

[8]Там само.

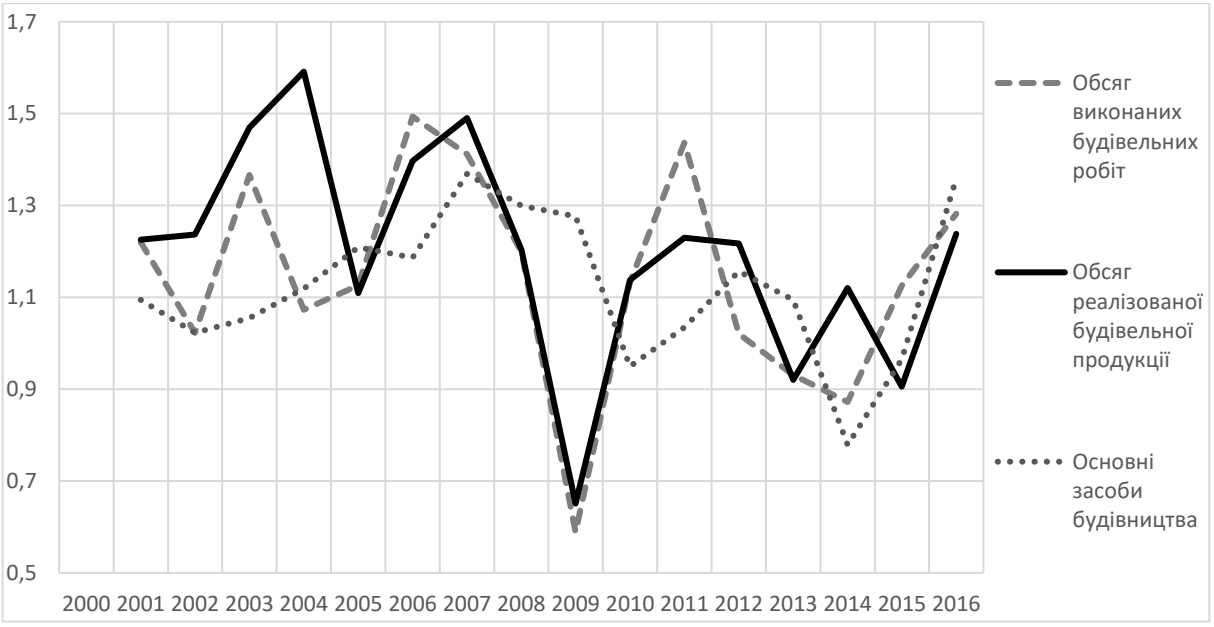


Рис. 3. Динаміка темпів зростання основних показників будівельної галузі

Джерело: побудовано авторами за даними[9].

Так, для загальної площі введеної в експлуатацію цей показник у середньому становив 1,08 ± 0,06, а середній річний приріст – 652,67 ± 142,82 тис. кв метрів; для основних засобів будівництва – 1,17± 0,11, середній річний приріст – 4104,78 ± 1787,11 млн грн; для кількості будівельних підприємств – 1,09 ± 0,04 і 2731,38 ± 432,69, для обсягу виконаних будівельних робіт – 1,24 ± 0,16 і 6223,31 ± 2257,25 млн грн; для обсягу реалізованої будівельної продукції – 1,34 ± 0,15 і 14387,78 ± 4170,07 млн грн відповідно (табл. 1). Після 2009 року, в якому відбулося різке падіння усіх показників, крім основних засобів виробництва, розпочався другий період, який можна охарактеризувати як період зростання і відновлення основних показників будівельної галузі, проте він був нестабільний. Хоча середні темпи зростання (крім кількості будівельних підприємств) перевищували одиницю, проте в певні роки їх значення було менше одиниці (табл. 1). У цей період також зменшилися середні прирости за рік порівняно з попереднім періодом. Кількість будівельних підприємств взагалі проявила тенденцію до їх зменшення. Проте динаміка розвитку

[9]Там само.

будівельної галузі упродовж 2010–2016 років не буде вигляди оптимістично, якщо врахувати зростання курсу долара США і звести обсяг виконаних будівельних робіт, обсяг реалізованої продукції і основні засоби будівництва до цієї грошової одиниці.

Таблиця 1

Середні прирости та середні темпи зростання основних показників будівельної галузі в період з 2000 до 2016 роки

Основні показники будівельної галузі	2000–2008			2009–2016		
	середній приріст за рік	середні темпи зростання		середній приріст за рік	середні темпи зростання (2010–2016)	
		у грн	у дол.		у грн	у дол.
Обсяг виконаних будівельних робіт, млн грн	6223,31±2257,25	1,24±0,16	1,24 ±0,16	3426,49±3208,25	1,11±0,18	0,96 ±0,27
Обсяг реалізованої будівельної продукції, млн грн	14387,78±4170,07	1,34±0,15	1,34±0,16	9965,93±4771,20	1,11±0,13	0,97±0,25
Основні засоби будівництва (наприкінці року), млн грн	4104,78±1787,11	1,17±0,11	1,17 ±0,11	1408,81±3374,07	1,05±0,17	0,92 ±0,26
Кількість будівельних підприємств, одиниць	2731,38±432,69	1,09 ±0,04		-1233,21±1654,83	0,99± 0,16	
Введення в експлуатацію загальної площі, тис. м^2	652,67±142,82	1,08±0,06		432,33±352,06	1,07±0,14	

Джерело: розраховано за даними[10].

Рис. 4. Динаміка темпів зростання основних показників будівельної галузі, виражених у доларах США

Джерело: побудовано авторами за даними[11].

10 Там само.

11 Там само.

Таблиця 2

Парні коефіцієнти кореляції

Показники	Курс гривні щодо долара США	Обсяг виконаних будівельних робіт	Обсяг реалізованої будівельної продукції	Основні засоби будівництва	Кількість будівельних підприємств	Індекс реальної заробітної плати	Середня заробітна плата	Сукупні ресурси домогосподарств	Загальна площа
Курс гривні щодо долара США	1,00								
Обсяг виконаних будівельних робіт	0,57	1,00							
Обсяг реалізованої будівельної продукції	0,62	0,97	1,00						
Основні засоби виробництва	0,59	0,88	0,91	1,00					
Кількість будівельних підприємств	0,13	0,69	0,63	0,64	1,00				
Індекс реальної заробітної плати	–0,53	–0,34	–0,43	–0,47	–0,13	1,00			
Середня заробітна плата	0,85	0,86	0,92	0,91	0,43	–0,54	1,00		
Сукупні ресурси домогосподарств	0,79	0,89	0,94	0,95	0,50	–0,54	0,99	1,00	
Загальна площа	0,45	0,87	0,88	0,67	0,53	–0,30	0,72	0,74	1,00

Джерело: розраховано авторами за даними[12].

[12]Там само.

Як видно з таблиці 1, середні темпи зростання цих показників у цьому періоді менше одиниці, починаючи з 2011–2012 років відбувалось різке їх зменшення (рис. 4) і, майже досягнувши кризового значення 2009 року, в 2016 році цей процес зупинився.

Для подальшого аналізу застосовуємо економетричні методи. Матриця парних коефіцієнтів кореляції показує, що введення в експлуатацію загальної площі корелюється з обсягом виконаних будівельних робіт (коефіцієнт кореляції 0,87), обсягом реалізованих будівельних робіт (0,88), основними засобами будівництва (0,67), середньою заробітною платою (0,72), сукупними ресурсами домогосподарств (0,74) (табл. 2).

Введення до розгляду офіційного середньорічного курсу гривні, долара США, індексу реальної заробітної плати (до попереднього року, %), середньої заробітної плати в розрахунку на одного штатного працівника, сукупних ресурсів у середньому за місяць в розрахунку на одне домогосподарство нами зроблено з метою дослідження можливого впливу на будівельну галузь показників, які формують попит. Оскільки заробітна плата в Україні становить постійну частку в загальних доходах населення, то в модель, з метою усунення мультиколінеарності, включено лише середню заробітну плату.

Побудована модель (модель 1) має вигляд:

$$Y = 6483{,}53 - 2{,}18 \cdot X_1 + 0{,}04 \cdot X_2 - 0{,}10 \cdot X_3 + 0{,}05 \cdot X_4 + 1{,}93 \cdot X_5,$$

де Y – введення в експлуатацію загальної площі, тис. м^2.

Коефіцієнт детермінації ($R^2 = 0{,}89$) та критерій Фішера ($F = 18{,}44$) вказують на те, що модель є адекватною і має гарні пояснювальні властивості. Проте при рівні значущості 0,1 ряд коефіцієнтів є незначущими (табл. 3, рис. 5).

Це можна пояснити наявністю в моделі мультиколінеарністю (впливом курсу гривні на інші фактори). Щоб усунути цей недолік, виразимо фактори, що вимірюються в грошових одиницях, у доларах США.

У результаті отримуємо модель (модель 2):

$$Y = 6573{,}58 + 0{,}21 \cdot X_2 - 0{,}95 \cdot X_3 + 0{,}02 \cdot X_4 + 21{,}29 \cdot X_5.$$

Таблиця 3

Статистичні характеристики моделі 1

Фактори	Коефіцієнти	t-статистика	Р-значення	Нижня межа 90,0 %	Верхня межа 90,0 %
Y – перетин	6483,53	6,11	0,00	4577,00	8390,06
X_1 – офіційний середньорічний курс гривні (за 100 доларів США)	–2,18	–1,17	0,27	–5,52	1,16
X_2 – обсяг реалізованої будівельної продукції, млн грн	0,04	1,57	0,15	–0,01	0,08
X_3 – основні засоби будівництва, млн грн	–0,10	–1,99	0,07	–0,19	–0,01
X_4 – кількість будівельних підприємств, одиниць	0,05	0,70	0,50	–0,07	0,17
X_5 – середня заробітна плата, грн	1,93	1,00	0,34	–1,55	5,40

Джерело: розраховано авторами за даними[13].

Коефіцієнт детермінації (R^2 = 0,91) та критерій Фішера (F = 29,81) вказують на те, що модель є адекватною і має гарні пояснювальні властивості. Як видно з таблиці 4, в цілому статистичні характеристики моделі значно покращилися (рис. 5). Зауважимо, що в першій і другій моделях коефіцієнт при змінній X_4 є незначущим. Це вказує на те, що кількість будівельних підприємств не має статистичного впливу на введення в експлуатацію загальної площі і в подальшому цей фактор з моделі можна вилучити.

Аналіз будь-якого ринка має включати дослідження взаємозв’язку ціни продукції та попиту і пропозиції. Офіційна статистика не наводить даних стосовно ціни квадратного метра будівель упродовж досліджуваного періоду. Для оцінки цього показника поділимо обсяг реалізованої будівельної продукції, виражену в грошових одиницях, на величину введеної в експлуатацію загальної площі. Цю величину назвемо оцінкою вартості квадратного метра будівель і будемо її виражати в гривнях за квадратний метр (або доларів США за квадратний метр).

[13] Там само.

Таблиця 4

Статистичні характеристики моделі 2

Фактори	Коефіцієнти	t-статистика	P-значення	Нижня межа 90,0 %	Верхня межа 90,0 %
Y – перетин	6573,58	7,44	0,00	4998,81	8148,35
X_2 – обсяг реалізованої будівельної продукції, млн дол. США	0,21	4,02	0,00	0,12	0,30
X_3 – основні засоби будівництва, млн дол. США	–0,95	–6,48	0,00	–1,21	-0,69
X_4 – кількість будівельних підприємств, одиниць	0,02	0,43	0,68	-0,05	0,08
X_5 – середня заробітна плата, дол. США	21,29	6,82	0,00	15,73	26,86

Джерело: розраховано авторами за даними[14].

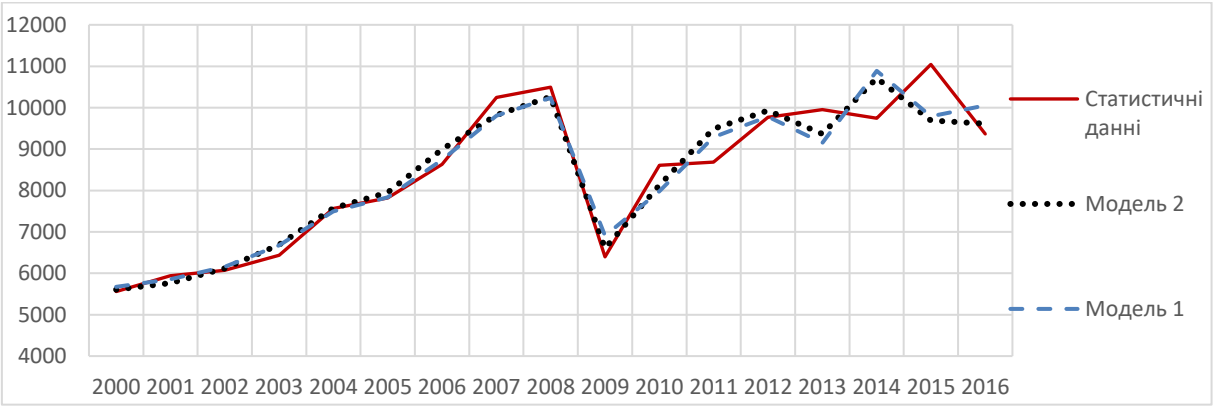


Рис. 5. Графіки моделі 1 і моделі 2 (введення в експлуатацію загальної площі, тис. м^2)

Джерело: побудовано авторами за даними[15].

[14]Там само.

[15]Там само.

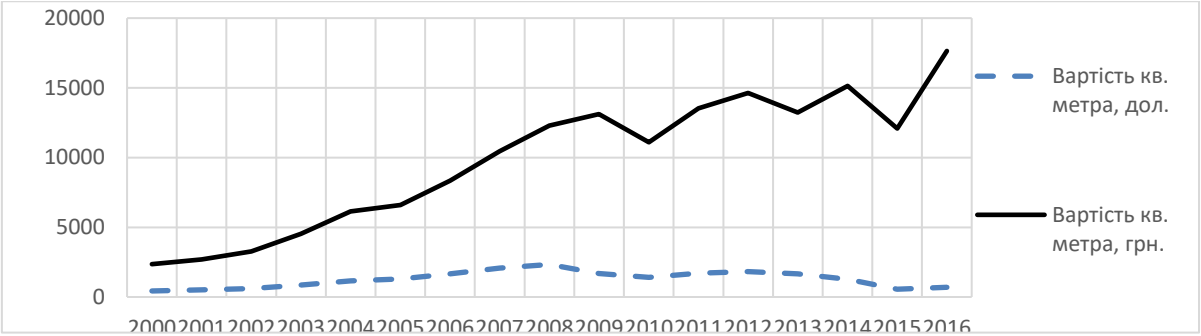


Рис. 6. Динаміка оцінки вартості квадратного метра будівель

Джерело: побудовано авторами за даними[16].

Як видно з графіка (рис. 6), динаміка оцінки вартості квадратного метра будівель має чітко виражені дві тенденції. Першу в період з 2000 до 2009 рр. У цей період відбувалось поступове, майже з однаковими темпами, зростання досліджуваного показника, вираженого як у національній валюті, так і в доларах США (в середньому 1,22 і 1,24, відповідно). У другому періоді спостерігається як нестабільність динаміки, так і зменшення темпів зростання (в середньому 1,06 і 0,90 відповідно). Ввівши цей фактор, отримуємо таку модель (модель 4):

$$Y = 7355{,}38 - 2{,}35 \cdot X_1 + 0{,}39 \cdot X_2 - 0{,}71 \cdot X_3 + 17{,}83 \cdot X_4 .$$

Ця модель має гарні статистичні оцінки (табл. 5), коефіцієнт детермінації дорівнює 0,93, критерій Фішера – 40,36, усі коефіцієнти є значущими.

На рис. 7 зображено статистичні дані і графіки, які ілюструють модель 3 і модель 4 (модель 3 – це модель, яка аналогічна моделі 4, в якій фактори представлені в національній валюті, її аналітичний вираз ми не наводимо, оскільки, як було вже вказано вище, в ній присутня мультиколінеарність, яка погіршує статистичні оцінки коефіцієнтів).

[16] Там само.

Таблиця 5

Статистичні характеристики моделі 4

Фактори	Коефіцієнти	t-статистика	*P*-значення	Нижня межа 90,0 %	Верхня межа 90,0 %
Y – перетин	7355,38	18,95	0,00	6663,72	8047,05
X_1 – вартість кв метра, дол.	–2,34	–2,03	0,07	–4,41	–0,28
X_2 – обсяг реалізованої будівельної продукції, млн дол.	0,39	4,04	0,00	0,22	0,56
X_3 – основні засоби будівництва, млн дол.	–0,71	–4,14	0,00	–1,01	–0,40
X_4 – середня заробітна плата, дол	17,83	5,51	0,00	12,06	23,60

Джерело: розраховано авторами за даними[17].

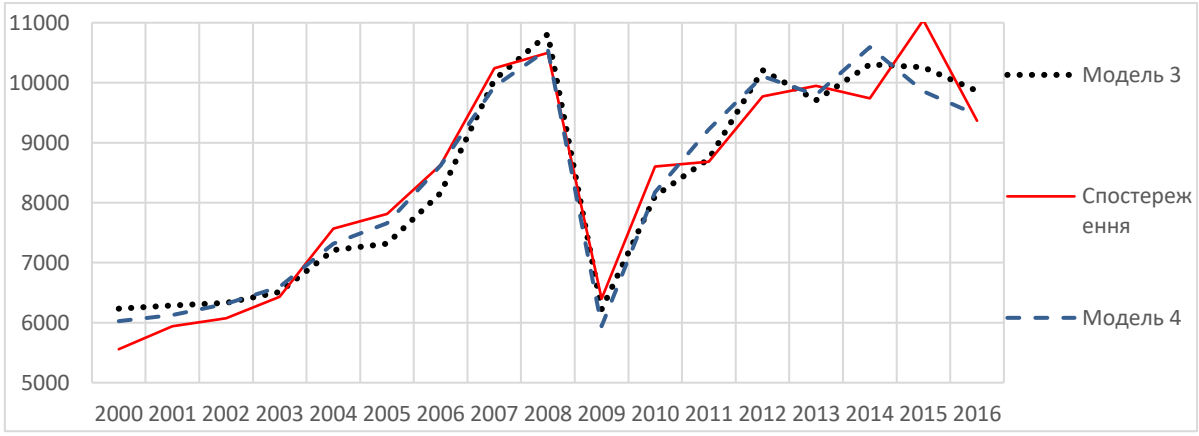


Рис. 7. Графіки моделі 3 і моделі 4 (введення в експлуатацію загальної площі, тис. м^2)

Джерело: побудовано авторами за даними[18].

Аналіз графіків показує, що модель 4 краще описує докризовий період, а модель 3 – післякризовий період. З цього можна зробити висновок, що учасники ринку в докризовий період більше орієнтувались на ціни, виражені в доларовому еквіваленті. Після 2009 року, коли відбулося знецінення національної валюти,

[17] Там само.

[18] Там само.

реальні доходи громадян почали зменшуватись, учасники ринку стали орієнтуватись на ціни, які виражені в гривнях.

Проведений аналіз дає можливість зробити такі висновки. На стан будівництва в Україні значний вплив мають: середньорічний курс гривні щодо іноземних валют, обсяг реалізованої будівельної продукції, основні засоби будівництва, середня заробітна плата в розрахунку на одного штатного працівника, вартість квадратного метра будівель. Кількість будівельних підприємств суттєвого впливу не має. На цьому етапі збільшення вартості квадратного метра житла на один долар США призводить до зменшення виробництва в будівельній галузі в середньому на 2,34 ± 2,06 квадратних метра. Збільшення середньої заробітної плати на один долар США призводить до збільшення введення в експлуатацію загальної площі на 17,83 ± 5,76 кв. метра (табл. 5). Збільшення офіційного курсу гривні до іноземних валют на 1 долар США призводить до зменшення в середньому на 2,18 ± 3,34 кв. метра (табл. 3). Кризові явища в економіці, а саме знецінення національної валюти, зменшення реальних доходів громадян, закриття банків, призвели до зменшення інвестицій у будівельну галузь, падіння попиту і, як наслідок, падіння обсягів будівництва. Так, у праці [3] вказується, що основною причиною, з якої будівництво тимчасово призупинено або законсервовано, є відсутність фінансування. Значний знос виробничих фондів будівельних підприємств впливає на зростання собівартості будівельної продукції і, як наслідок, призводить до збільшення вартості і скорочення платоспроможного попиту на нього. Оновлення виробничих фондів потребує додаткових коштів, які не йдуть безпосередньо на будівництво. Саме цим пояснюється від'ємне значення коефіцієнта за основними засобами будівництва в побудованих моделях.

Однією з цілей побудови математичних моделей економічних процесів є здійснення на їх основі прогнозів. Особливість побудованих і досліджених вище моделей полягає у тому, що на їхній основі важко здійснювати прогнози, оскільки для цього необхідно спрогнозувати спочатку ендогенні змінні, якщо їх значення наперед невідоме. У результаті цього виникає необхідність будувати «прогнози на прогнозах».

Одним із припущень щодо пошуку прогнозних значень є збереження тенденцій, що можна пояснити, зокрема, інертністю економічних процесів. Однією з моделей прогнозування, що використовуються на практиці під час збереження тенденцій, є модель тренда, в якій залежною змінною виступає досліджуваний показник, а незалежною – час або номер спостереження цього показника. Таким чином, трендова модель – це математичний опис тимчасової тенденції. Проведене дослідження вказує на існування двох тенденцій у будівельній галузі, тому логічно будувати трендову модель на статистичних даних другого періоду (2010–2016 рр.). Проте мала кількість спостережень призведе до низької статистичної якості моделі і, як наслідок, прогнози будуть неефективні. Щоб усунути цей недолік, нами побудована модель на даних за весь досліджуваний період, але при цьому введено дві фіктивні змінні. Змінна D_1 набуває значення 0 у докризовий період і значення 1 в інший. Змінна D_2 набуває значення 1 в кризові роки і значення 0 в усіх інших. Побудована на основі цього модель має вигляд:

$$Y = 4375{,}89 + (652{,}67 - 220{,}33 \cdot D_1) \cdot t - 580{,}22\, D_1 - 1748{,}67 \cdot D_2,$$

де Y – введення в експлуатацію загальної площі, тис. м^2;

t – номер спостереження.

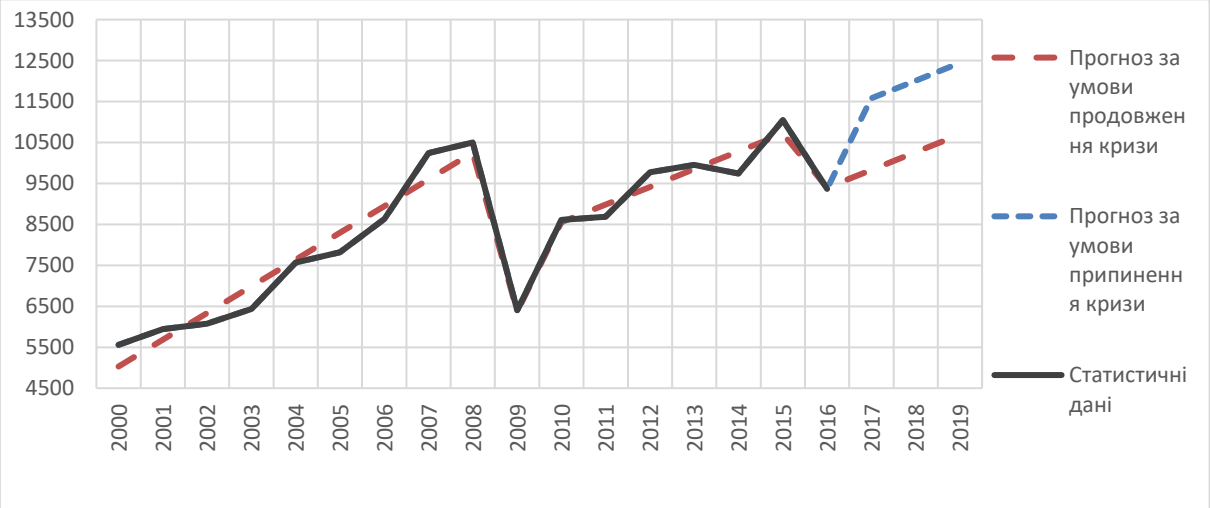


Рис. 8. Трендова модель та прогнози

Джерело: побудовано авторами за даними[19].

[19] Там само.

Критерій Фішера становить 66,95, коефіцієнт детермінації дорівнює 0,96, що вказує на адекватність побудованої моделі та якісні її статистичні властивості. На основі цієї моделі побудовано прогноз розвитку будівництва в Україні на період 2018–2019 рр. (рис. 8, табл. 6).

Таблиця 6

Прогнозні значення введення в експлуатацію загальної площі, тис. м^2 (розраховано для рівня значущості 0,1)

Показник	2018 рік	2019 рік
Збереження кризи	10261,33 ± 857,79	10693,67 ± 911,31
Припинення кризи	12010 ± 755,5	12442,33 ± 815,35

Джерело: розраховано авторами за даними[20].

Прогноз показує, що під час збереження тенденцій варто очікувати подальше поступове зростання виробництва в будівельній галузі.

Висновки. Незважаючи на економічну нестабільність в Україні, будівельний ринок продовжує нарощувати свої обороти. При цьому будівельній галузі притаманні ряд проблем, які необхідно вирішувати.

Однією з проблем аналізу галузі є відсутність якісної, комплексної, достовірної інформаційної про всіх учасників ринкових відносин. Створення єдиної інформації системи може забезпечити збір, систематизацію і накопичення якісної, достовірної інформації.

Статистично доведено, що зменшення кількості підприємств не має негативного пливу на будівельну галузь у цілому. Причини такого явища потребують подальшого дослідження.

Найбільш проблемною і вагомою у будівельний галузі на сьогодні є її фінансова складова. Низька платоспроможність покупців, відсутність стабільних доходів, зменшення купівельної спроможності, закриття банків, знецінення національної валюти негативно впливає на попит, фінансування і, відповідно, на виробництво в

[20]Там само.

досліджуваній галузі. В умовах дефіциту грошових потоків підписання довгострокових договорів не є можливим, що також має негативний вплив.

Приклад 2. Побудувати регресійну модель залежності промислового сектору України від фінансового за даними таблиці 7.

Таблиця 7

Основні показники діяльності економіки України за 2001–2016 рр.

Роки	Y	X_1	X_2	X_3	X_4	X_5
2001	210,842	70,27	68,50	13,48	5,37	304,90
2002	229,634	125,43	108,60	20,44	5,33	370,33
2003	289,117	120,99	203,00	32,91	5,33	454,56
2004	400,757	196,95	321,30	35,01	5,32	589,63
2005	468,562	316,97	403,80	40,90	5,12	806,18
2006	551,729	561,13	492,80	55,72	5,05	1041,44
2007	717,076	599,40	754,30	82,24	5,05	1351,00
2008	917,035	926,09	883,40	99,31	5,27	1806,00
2009	806,55	880,30	1067,30	121,13	7,79	1906,00
2010	1043,11	942,09	1537,80	201,41	7,94	2239,00
2011	1305,31	1054,28	2147,50	298,02	7,97	2633,00
2012	1367,93	1127,19	2506,46	370,16	7,99	3025,00
2013	1322,41	1278,10	1676,97	638,34	7,99	3234,00
2014	1428,839	1316,85	2331,94	165,80	11,89	3476,00
2015	1776,603	1254,39	2171,59	303,79	21,84	4195,00
2016	2158,03	1256,30	2127,55	5090,33	25,55	5070,00
Середнє	937,10	751,67	1175,18	473,06	8,80	2031,38

Джерело: складено за даними[21].

[21]Офіційний сайт Державної служби статистики України. URL : http://www.ukrstat.gov.ua.
Офіційний сайт Національного банку України. URL : http://nbuportal.bank.gov.ua.
Офіційний сайт Національної комісії, що здійснює державне регулювання у сфері ринків фінансових послуг. URL : http://nfp.gov.ua/content/rzviti-nackomfinposlug.html.
Офіційний сайт Світового банку. URL : http://www.worldbank.org/ru/country/russia.

За допомогою виробничого методу розрахунку ВВП проаналізуємо економічне середовище України. 2017 року майже 53 % у складі ВВП займають чотири позиції. Лідером є податки на продукти (15,86 %), звідси випливає, що споживання переважає над виробничими функціями у країні. Значний податковий тягар несе кінцевий споживач, купівельна спроможність якого на сьогодні суттєво зменшилася. Більшу частку ВВП становить оптова та роздрібна торгівля (14,13 %). На жаль, сьогодні переробна промисловість України займає лише третю позицію і це лише 12,39 % ВВП. Україна стверджується як сільськогосподарська країна (сільське, лісове та рибне господарство лише на 2,16 % менші за переробну промисловість) відсоток ВВП від цього виду діяльності становить 10,23 %.

Якщо порівнювати ситуацію з 2008 роком (до фінансової кризи), то переробна промисловість займала лідируючу позицію (17,38 % ВВП), торгівля була на другому місці (13,85 %), третю позицію займали податки на продукти (13,34 %), замикала четвірку лідерів діяльність транспорту і зв'язку (9,18 %). 2008 року сільське господарство займало 6,87 % у ВВП, а це 65,148 млрд грн, що майже в п'ять разів менше від рівня 2017 року (305,194 млрд грн).

Аналізуючи промисловий сектор економіки, необхідно застосовувати два підходи, а саме, з одного боку, визначати фактори впливу на нього фінансових важелів економіки, які впливають на ведення фінансово-господарської діяльності конкретного підприємця. З іншого – дослідити фактори впливу на результати виробництва зі сторони кінцевого споживача. Перша множинна регресійна модель побудована за статистичними даними 2001–2016 рр., де ендогенну змінну визначає результат діяльності виробничого сектору економіки, а саме обсяг реалізованої промислової продукції, млрд грн ($У$).

Екзогенними змінними є елементи фінансового ринку, а саме: активи комерційних банків, млрд грн (X_1), загальний обсяг операцій фондового ринку, млрд грн (X_2), загальний обсяг операцій валютного ринку, млрд грн (X_3) і фактор часу (t), який має безпосередній вплив як на фінансовий, так і промисловий сектори економіки, адже від залучення коштів у діяльність суб'єктів підп-

риємницької діяльності до отримання результатів від неї проходить певний проміжок часу (табл. 7). Побудована матриця парних коефіцієнтів кореляції вказує на пряму лінійну залежність між факторними і результативними ознаками (табл. 8).

Упродовж шістнадцяти років середньорічний приріст обсягу промислової продукції становив 119,24 млрд грн. Цей приріст відносний, необхідно враховувати різкий стрибок зростання показника за 2015–2016 рр., який спровокований інфляційною складовою обсягом реалізованої промислової продукції (рис. 9).

Таблиця 8

Матриця парних коефіцієнтів кореляції між обсягом реалізованої промислової продукції та рядом факторів

Показники	***Y***	X_1	X_2	X_3	***t***
Y	1	0,9242	0,9218	0,6400	0,9748
X_1	0,9242	1	0,9327	0,3871	0,9727
X_2	0,9218	0,9327	1	0,3809	0,9418
X_3	0,6400	0,3871	0,3809	1	0,5106
t	0,9748	0,9727	0,9418	0,5106	1

Джерело: розраховано автором за даними таблиці 7.

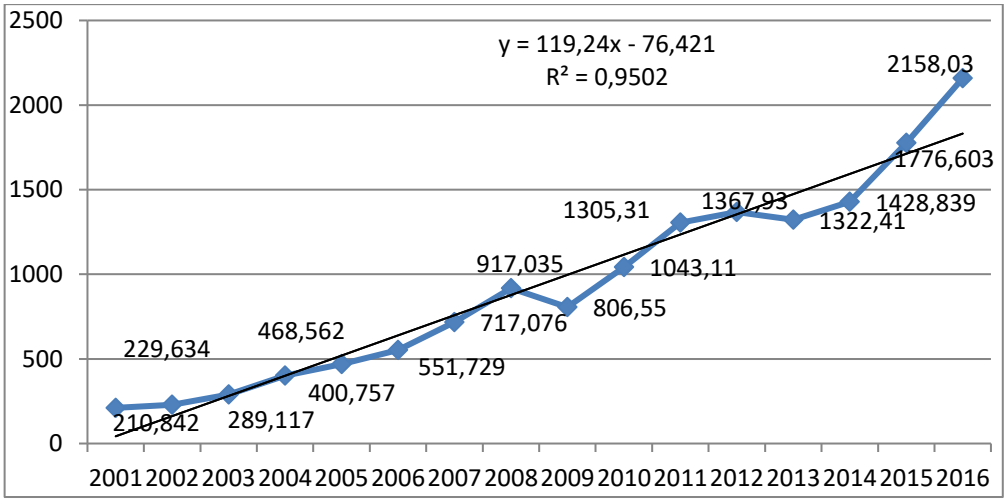


Рис. 9. Динаміка обсягу реалізованої промислової продукції за 2001–2016 рр.

Джерело: розраховано автором.

Динаміка обсягу промислової продукції вказує на зміну основної тенденції і 2009 року, тому в модель введено фіктивну змінну (*D*), яка фіксує вплив фінансової кризи. Для побудови моделі застосуємо рідж-регресію (гребеневу регресію). Вона дає можливість розмежувати оцінки коефіцієнтів моделі за наявності мультиколеарності. Побудована модель має високий коефіцієнт множинної кореляції (R = 0,9929), що вказує на тісний лінійний зв'язок між факторами. Обсяг виробництва на 98,59 % залежить від факторів моделі і лише на 1,41 % від інших факторів, включаючи стохастичну змінну (R^2 = 0,9859). Розрахований критерій Фішера 140,04 значимий (p = 6,38E-09). У моделі відсутня автокореляція, статистика Дарбіна-Уотсона дорівнює 2,04 (du = 2.1). Найбільш результативним є значення гребня 0,02, при якому значення VIF менші 10, а R^2 = 0,977. Рівняння регресії матиме такий вигляд:

$$Y = 30,7068 + 0,1779 X_1 + 0,1537 X_2 + 0,1117 X_3 - 133,1880 D + 59,6236 t \,. \quad (1)$$

Побудована модель вказує: якщо активи комерційних банків збільшаться на 1 млрд грн, то обсяг реалізованої промислової продукції збільшиться в середньому на 0,1779 млрд грн за умови сталості інших факторів моделі; якщо загальний обсяг операцій фондового ринку збільшиться на 1 млрд грн, то обсяг реалізованої промислової продукції збільшиться в середньому на 0,1537 млрд грн відповідно; якщо загальний обсяг операцій валютного ринку збільшиться на 1 млрд грн, то обсяг реалізованої промислової продукції зросте в середньому на 0,1117 млрд грн. Крім того, втрати в обсягу реалізованої продукції промисловості в кризовий період для України становили у середньому 133,188 млрд грн, а щорічна, середня динаміка приросту виробництва за шістнадцять років становить 59,6236 млрд грн. Цей приріст значно менший за результат лінійного тренду на рис. 1, при цьому достовірніший.

Порівняємо вплив пояснюючих змінних на обсяг промислової продукції за допомогою стандартизованих коефіцієнтів (бета-коефіцієнтів) (табл. 9). Відповідно до результатів регресійної моделі найбільший вплив на обсяг промислової продукції має час, адже з часом вартість капіталу зменшується. Тому для різних типів

грошових потоків повинна застосовуватися відповідна ставка дисконтування або капіталізації.

Таблиця 9

Стандартизовані регресійні коефіцієнти першої моделі

Рідж-параметр	X_1	X_2	X_3	D	t
0,02	0,1418	0,2332	0,2383	–0,0572	0,4874
Коефіцієнт еластичності	0,1427	0,1927	0,0564	–	–

Джерело: розраховано автором.

Другу позицію займає валютний ринок. Штучне стримування реального курсу гривні призводить до падіння конкурентоспроможності промислової продукції на внутрішньому і зовнішньому ринках і, як наслідок, до скорочення її реального випуску. Це пояснюється тим, що короткостроковий ефект реального зміцнення гривні однозначно негативно відбивається на динаміці виробництва, але довгостроковий ефект цього укріплення, навпаки, позитивний. З іншого боку, інфляційне збільшення обсягів реалізованої промислової продукції має короткострокову перспективу, адже при цьому суттєво зменшується купівельна спроможність населення. Підприємства, які займаються зовнішньоекономічною діяльністю й отримують виручку у доларах, мають суттєві переваги перед іншими.

Вплив фондового ринку на промисловий сектор попри його недосконалість (низький рівень капіталізації, ліквідності та прозорості фондового ринку; недосконала ринкова інфраструктура; незахищеність прав інвесторів) позитивний і займає третю позицію. Випуск акцій підприємств і залучення за їх допомогою інвесторів дає можливість отримувати оборотні кошти для ведення фінансово-господарської діяльності підприємства. При цьому необхідно враховувати, що фондовий ринок не користується попитом у населення через відсутність в останнього достатніх коштів для інвестицій.

Активи комерційних банків не виконують стимулюючої функції для виробництва. В Україні надто жорсткі умови кредитування промислових підприємств. Підприємства, які отримали кредит, або несуть непомірний тягар боргу і потрапляють у стан фактичного банкрутства, або спочатку не мають наміру повернути кредит і навмисне завдають шкоди банку.

Вплив фінансової кризи займає останню позицію. Він негативно впливає на економіку взагалі, знижує обсяги реалізованої продукції. Сучасна фінансова криза обумовлена специфікою прояву бімодальної основи фінансового капіталу. Двоїстий характер фінансового капіталу в сучасності виражається у все більш чіткому поділі цього капіталу на реальний капітал, акумульований у грошових фондах промислових підприємств, і трансфертний капітал, представлений диверсифікованим спектром фондових і кредитних інструментів, що обертаються на фінансових ринках. Темпи зростання фінансового ринку випереджають темпи зростання виробництва, що спричиняє зниження платоспроможності промислових підприємств. Вимірюємо вплив кожного фактора на ендогенну змінну за допомогою коефіцієнта еластичності $E_{y/x_i} = \beta_i \frac{\bar{x}_i}{\bar{y}}$ (де β – параметри рівняння регресії). Зростання обсягів операцій фондового ринку на 1 % збільшить обсяг реалізованої промислової продукції на 19,27 %, збільшення активів комерційних банків збільшить обсяг на 14,27 %, відповідно, аналогічна зміна обсягів операцій валютного ринку збільшить обсяги реалізації на 5,64 % (табл. 3).

Аналізуючи промисловий сектор економіки, необхідно розглядати його не тільки зі сторони виробництва, а й споживання готової продукції населення. Звідси доцільно побудувати іншу регресійну модель, в основі якої екзогенними змінними будуть X_4 – офіційний, середній за період курс долара до гривні (який дасть змогу виміряти інфляційний вплив на обсяг реалізованої промислової продукції) і X_5 – середньомісячна заробітна плата одного працівника (яка визначає платоспроможність населення) (табл. 1). При цьому використаємо два періоди: до і післякризовий і врахуємо зміну кута основної тенденції 2009 року. Ця методика і її порівняльний аналіз детально вже розглядалися у прогнозуванні будівельної галузі. Модель матиме такий вигляд:

$$Y = \beta_0 + \beta_1 x_4 + \beta_2 x_5 + (\alpha_0 + \alpha_1 D_1)t + \beta_4 D_1 + \beta_5 D_2 + \varepsilon, \quad (2)$$

де Y – обсяг реалізованої промислової продукції, млрд грн; β_0 – параметр множинної регресії в точці перетину результативної ознаки

з віссю y; β_1– параметр множинної регресії при X_4; β_2 – параметр множинної регресії при X_5; $(a_0 + a_1) = \beta_3$ – при цьому a_0 – загальний тренд за шістнадцять років; a_1 – проміжний коефіцієнт, що враховує зміну напряму тенденції; t – тренд; $\beta 4$ і β_5 – параметри множинної регресії за фіктивними змінами; D_1 – розмежування часового ряду на до і післякризовий період; D_2 – вплив світової економічної кризи 2009 року; ε – похибка моделі. Побудована модель має тісний лінійний зв'язок між факторами (R = 0,9983). Обсяг виробництва на 99,67 % залежить від факторів моделі і лише на 0,33 % від інших величин, включаючи стохастичну змінну (R^2 = 0,9967). Розрахований критерій Фішера 446,39 значимий (p = 1,29E-10). Найбільш результативним є значення гребня 0,015. Пояснимо результати моделювання. Загальне рівняння:

$$Y = 32{,}794 + 11{,}22x_4 + (42{,}858 + 123{,}655D_1)t - 11{,}323D_1 - 150{,}202D_2 \quad (3)$$

Світова фінансова криза негативно вплинула на купівельну спроможність населення, в результаті чого обсяги реалізації промислової продукції скоротилися на 150,202 млрд грн.

До економічної кризи (2001–2008 рр.):

$$Y = 32{,}794 + 11{,}26x_4 + 0{,}227x_5 + 42{,}868t - 150{,}202D_2 \quad (4)$$

Абсолютний середньорічний приріст реалізованої промислової продукції упродовж восьми років становив 42,868 млрд грн.

Після економічної кризи (2009–2016 рр.):

$$Y = 21{,}471 + 11{,}226x_4 + 0{,}227x_5 + 166{,}523t - 150{,}202D_2 \quad (5)$$

Абсолютні темпи зростання у післякризовому періоді в 3,9 раза вищі від докризового. Так, упродовж восьми років щорічно в середньому реалізовувалося промислової продукції на 166,523 млдр грн. Зростання середнього курсу долара до гривні на одиницю впливу призведе до зростання обсягів реалізації промислової продукції на 11,226 млрд грн – це і є інфляційна складова, яка, з одного боку, збільшує обсяги доходів суб'єктів господарювання, а з іншого – зменшує купівельну спроможність населення. Збільшення середньомісячної зарплати 1-го працівника на 1 грн

збільшить обсяги продажу на 0,227 млрд грн. За результатами моделювання, найсильніший вплив на обсяг реалізованої продукції має рівень середньої зарплати конкретного працівника (табл. 10).

Таблиця 10

Стандартизовані регресійні коефіцієнти другої моделі

Рідж-параметр	X_4	X_5	D_2
0,015	0,1184	0,5697	–0,0645
Коефіцієнт еластичності	0,1054	0,4921	–

Джерело: розраховано автором.

Його купівельна спроможність визначає обсяги чистого доходу від реалізації промислової продукції суб'єкта господарювання. Вплив курсу долара майже в п'ять разів менший. Третю позицію займає фінансова криза і її вплив найменший, але негативний.

Висновок. Проведений порівняльний аналіз вмісту ВВП за виробничим методом показав, що міжнародна фінансова криза, політична нестабільність у країні, воєнні дії на Сході суттєво погіршили економічне становище в Україні. 2017 року майже на 5 % зменшився відсоток переробної промисловості порівняно з 2008 роком, що перемістило цю категорію з позиції лідера на третє місце. Торгівля зберегла своє друге місце, збільшивши вміст у ВВП на 0,28 %. Коли податки на продукти займають першість у складі ВВП, можна стверджувати про бездіяльність уряду, в країні якого процвітає економічна, політична і фінансова криза. Необхідно зазначити, що населення несе основне податкове навантаження, при цьому більшість з них більшу частку своїх доходів витрачає на продукти харчування і комунальні платежі.

Результати регресійної моделі показали, що серед фінансових важелів найбільший вплив на результати промислової діяльності має валютний ринок, на другому місці фондовий ринок, третю позицію займає банківський сектор економіки. При цьому обсяг реалізованої промислової продукції найбільш чутливим є до змін на фондовому ринку, банківський сектор за впливом на другому місці, третю позицію займає валютний ринок.

Розвиток промислового сектору напряму пов’язаний із добробутом населення країни. Результати другої моделі вказують на пряму сильну лінійну залежність між офіційним середнім курсом долара до гривні і середньомісячною зарплатою працівника в Україні (r = 0,8523). При цьому обсяг реалізованої промислової продукції майже в п’ять разів є більш чутливим до змін середньомісячної зарплати працівника (коефіцієнт еластичності 0,4921), ніж до змін курсу долара (коефіцієнт еластичності 0,1054).

РОЗДІЛ 7
КЛАСТЕРНИЙ АНАЛІЗ

7.1. Сутність та види кластерного аналізу

Кластерний аналіз дозволяє виявити групи (кластери) об'єктів за заданими змінними.

Завдання кластерного аналізу полягає у формуванні груп:

• однорідних всередині, об'єкти подібні між собою (умова внутрішньої гомогенності);

• відмінних від об'єктів в інших групах (умова зовнішньої гетерогенності).

У процесі виконання кластерного аналізу виявляються статистичні зв'язки між аналізованими змінними, які вказують на подібність змінних, і потім об'єднання виявлених чинників у групи на основі рівня відмінностей між ними. Кількість кластерів залежить від параметрів, що задаються, подібності змінних, що об'єднуються в один кластер.

Зазвичай параметри кластеризації мають різну величину виміру, що не дозволяє їх об'єднувати. Усунути цей недолік дозволяє стандартизація параметрів. Спосіб стандартизації обирається залежно від соціально-економічного змісту та статистичної природи показників. Варіанти процедур стандартизації за допомогою відношень змінюються і коригуються залежно від мети дослідження та змісту первинних даних. Основні способи стандартизації наведено у таблиці 7.1.

Таблиця 7.1

Способи стандартизації параметрів

Спосіб стандартизації	**Стимулятори (max)**	**Дестимулятори (min)**
Спосіб співвідношень: $\alpha = x_{\max};\ x_{\min};\ x_{st};\ \bar{x}$	$z_{ij} = \frac{x_{ij}}{\alpha}$	$z_{ij} = \frac{\alpha}{x_{ij}}$
Класичний спосіб стандартизації	$\frac{x_{ij} - \bar{x}_i}{\sigma_i}$	$\frac{\bar{x}_i - x_{ij}}{\sigma_i}$
Стандартизація варіаційним розмахом	$z_{ij} = \frac{x_{ij} - x_{\min}}{x_{\max} - x_{\min}}$	$z_{ij} = \frac{x_{\max} - x_{ij}}{x_{\max} - x_{\min}}$

Якщо всі показники є стимуляторами, то стандартизацію доцільно здійснювати відношенням: $z_{ij} = \frac{x_{ij}}{x_{\max}}$. Якщо базою порівняння є середній рівень показника, тобто $\alpha = \bar{x}$. У порівняльному аналізі соціально-економічних явищ поширені оцінки, розраховані на основі відхилень $(x_{ij} - \alpha)$, стандартизованих варіаційним розмахом $(x_{\max} - x_{\min})$. При цьому для стимуляторів $\alpha = x_{\min}$, для дестимуляторів $\alpha = x_{\max}$:

$$z_{ij} = \frac{x_{ij} - x_{\min}}{x_{\max} - x_{\min}}, \; z_{ij} = \frac{x_{\max} - x_{ij}}{x_{\max} - x_{\min}}.$$

Стандартизоване у такий спосіб значення z_{ij} показує відносну позицію *j*-ої одиниці сукупності в діапазоні варіації за *i*-м показником. За високими значеннями показника z_{ij} наближається до 1, за низькими – до 0. Якщо усі величини мають однакову величину виміру, їх стандартизувати непотрібно.

7.2. Етапи кластерного аналізу в SPSS

Програма SPSS реалізує три методи кластерного аналізу:

1. Двоетапний кластерний аналіз. Дозволяє виявити групи об'єктів (кластерів) за заданими параметрами, якщо ці групи дійсно існують. При цьому програма автоматично визначає кількість існуючих кластерів. Якщо однозначно визначити це неможливо, усі об'єкти розміщують в один.

2. Кластеризація К-середніми. Розбиває за заданими параметрами всю множину об'єктів на задану дослідником кількість кластерів так, щоб середнє значення для кластерів за кожним параметром максимально відрізнялися.

3. Ієрархічна кластеризація. Найбільш гнучкий метод, який дозволяє детально дослідити структуру відмінностей між об'єктами й обрати найбільш оптимальну кількість кластерів.

Етапи кластерного аналізу:

1. Вибір параметрів-критеріїв для кластеризації.

2. Вибір способу виміру відстані між об'єктами (кластерами).

Початковою умовою є віднесення кожного об'єкта до одного відповідного кластеру. Зазвичай застосовують квадрат евклідової відстані, відповідно до якої відстань між об'єктами дорівнює сумі квадратів різниці між значеннями однойменних параметрів об'єктів. У процесі аналізу сума квадратів різниць розраховується для всіх змінних. Отримані відстані використовуються програмою для формування кластерів. Крім евклідової, існують й інші види відстаней, а саме кореляція Пірсона, косинус, квадрат евклідової відстані, Чебишева, Манхетенська, Мінковського, користувацька відстані.

Необхідно зазначити, що всі параметри стандартизують, і їхня вага стає однаковою. У програмному продукті застосовуються такі методи стандартизації: z-оцінки, діапазон від –1 до +1, діапазон від 0 до 1, максимальна величина 1, середнє 1, стандартне відхилення.

3. *Формування кластерів.*

Існує два основні методи формування кластерів:

– метод злиття. Початкові кластери збільшуються шляхом об'єднання доти доки, не буде сформований єдиний кластер, який містить усі дані;

– метод дроблення будується на зворотній операції: спочатку усі дані об'єднуються в один кластер, який потім ділиться на частини доти, доки не буде досягнутий бажаний результат.

Щодо замовчування програмою застосовується метод злиття, який має такі сім способів об'єднання об'єктів, а саме міжгрупового і внутрішньогрупового зв'язків, найближчого і найдальшого сусідів, центроїдної і медіанної кластеризацій, Уорда.

Наприклад, міжгруповий зв'язок об'єднує середні всередині груп. Спочатку розраховується найменше середнє значення відстані між всіма парами груп, а потім об'єднує дві групи, що найбільш близькі між собою. На першому кроці, коли усі кластери є одиничними об'єктами, ця операція зводиться до звичайного попарного порівняння відстані між об'єктами. Термін «середнє

значення» має сенс на другому етапі, коли сформовані кластери, які містять більше одного об'єкта.

4. Інтерпретація результатів.

Залежить від цілей дослідника.

Приклад. Провести кластерний аналіз трудового потенціалу за регіонами України. Для аналізу застосувати такі змінні: зайняте населення, тис. осіб; попит на робочу силу на кінець періоду, тис. осіб; постійне населення, середня чисельність, тис. осіб; середньомісячна заробітна плата, грн; рівень зареєстрованого безробіття на кінець року, у % до населення працездатного віку; наявний дохід у розрахунку на одну особу, грн; індекси споживчих цін, %; доходи населення, млн грн (додаток З).

З ряду спостереження вилучено м. Київ у зв'язку із суттєвим коливанням усіх його значень від загальної тенденції.

Функція в SPSS: «**Анализ**» – «**Классификация**» – «**Иерархическая кластеризация**» (рис. 7.1).

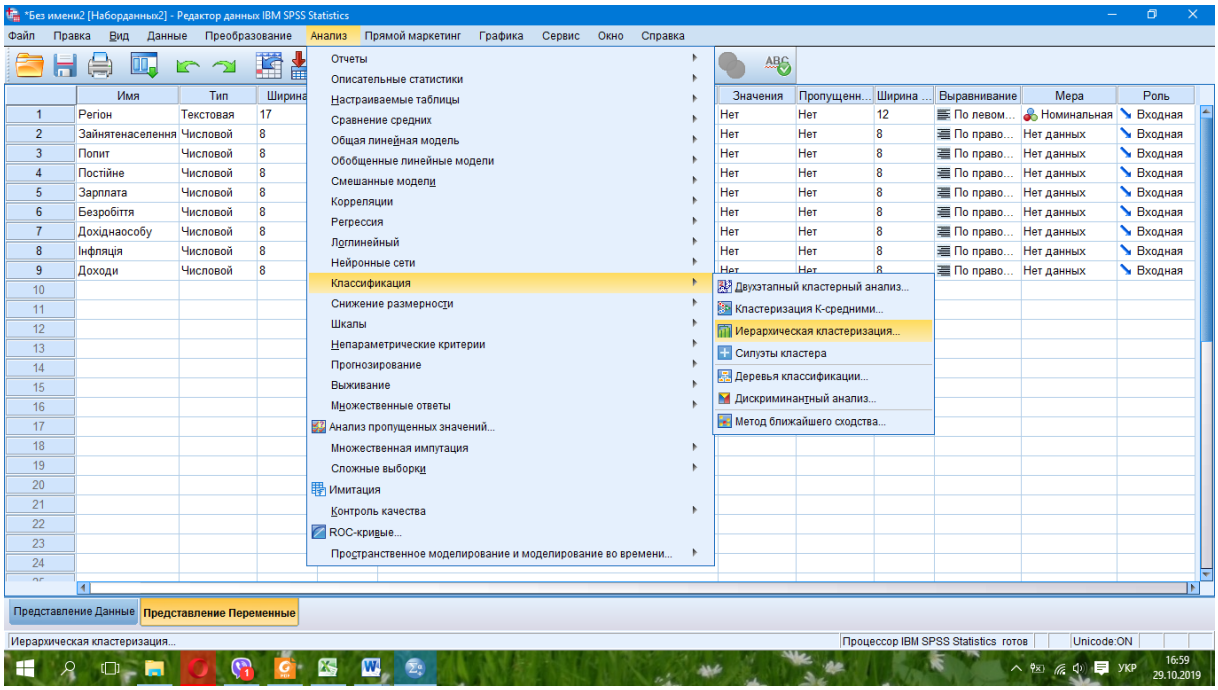


Рис. 7.1. Метод ієрархічної кластеризації

З лівого діалогового вікна обираємо параметри і за допомогою стрілки переміщуємо праворуч (рис. 7.2). Задати спосіб ідентифікації змінних, для нашого спостереження це регіони України, переміщуємо його в нижнє вікно праворуч.

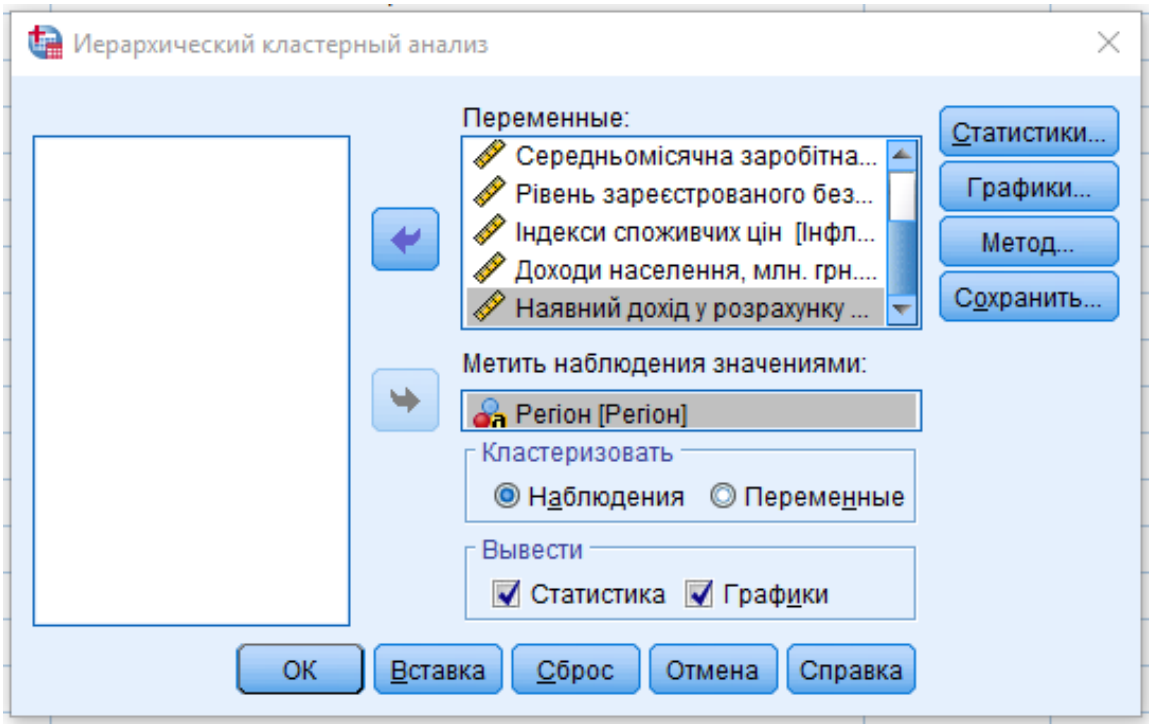


Рис. 7.2. Внесення параметрів для проведення ієрархічної кластеризації

Обираючи вкладку статистики, можна встановити мітки, які забезпечать візуалізацію процесів кластеризації, а саме:

– мітка «Порядок агломерації» забезпечить візуалізацію покрокового кластерного аналізу (рис. 7.3).

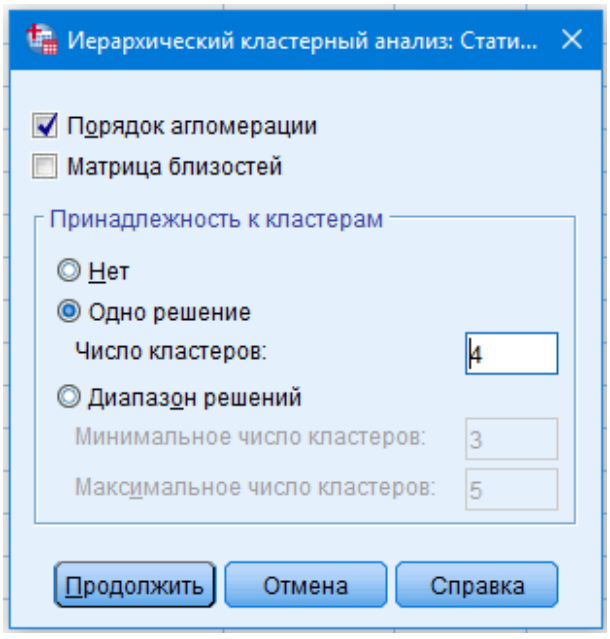


Рис. 7.3. Статистичні характеристики ієрархічної кластеризації

У таблиці «Кроки агломерації» друга колонка Кластер об'єднаний і складається із стовпців Кластер 1 і Кластер 2, які відповідають номерам кластерів, що об'єднані на цьому кроці. Після об'єднання кластера присвоюється номер, який відповідає номеру в колонці Кластер 1. Наступна колонка «Коефіцієнти містять значення відстані між кластерами, які об'єднуються на даному кроці. Колонка «Етап» першого появлення кластера показує, на якому кроці до цього з'явилися перший і другий об'єднані кластери. Остання колонка «Наступний етап» показує, на якому кроці знову з'явиться кластер, що утворений на цьому кроці:

– мітка «Матрица близостей» дасть змогу візуалізувати інформацію про відстань між об'єктами і кластерами.

– наступним є вибір приналежності до кластерів: мітка «Нет» – забезпечить вивід усіх кластерів у результаті; «Одно решение» – дає можливість задати точну кількість кластерів у рішенні; «Диапазон решений» – забезпечує вивід декількох рішень з різною кількістю кластерів, задається діапазон від і до.

Відкривши вкладку «Графики», необхідно проставити відмітки «Дендрограмма», яка графічно відобразить відносну величину різності між параметрами або кластерами на кожному кроці процесу (рис. 7.4). «Сосульчатая» діаграма в програмі має ряд недоліків, тому ставиться відмітка «Нет».

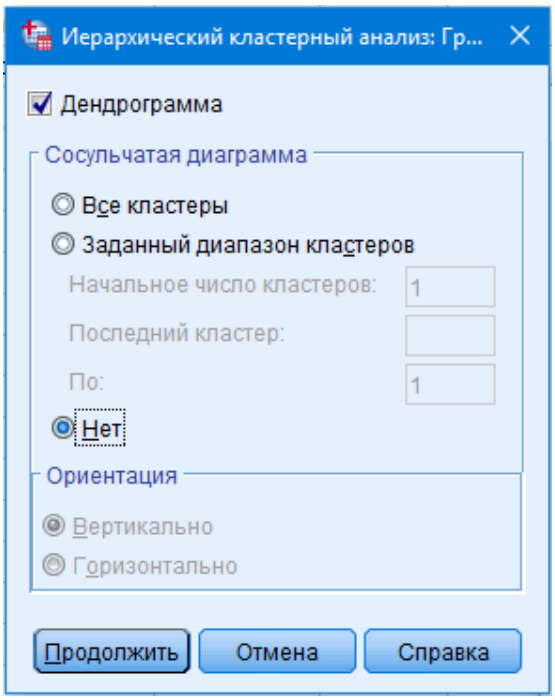


Рис. 7.4. Графічна візуалізація ієрархічної кластеризації

Вкладка «Метод» дає можливість обрати **методи зв'язку** – методи формування кластерів, при яких об'єкти об'єднуються в групу на основі розрахованої між ними відстані, тобто спосіб об'єднання параметрів (рис. 7.4):

«Межгрупповые связи» – сутність методу в об'єднані на кожному кроці підлягають кластери або об'єкти, відстань між якими мінімальна.

«Ближайший сосед» – в основі полягає вибір змінних, відстань між якими мінімальна.

«Дальний сосед» – в основі полягає вибір змінних, відстань між якими максимальна.

«Центроидная кластеризация» – оцінюється відстань між центроїдами (середніми) груп змінних.

«Метод Уорда» – як міра, використовується квадрат евклідової відстані, який повинен бути мінімальним тощо.

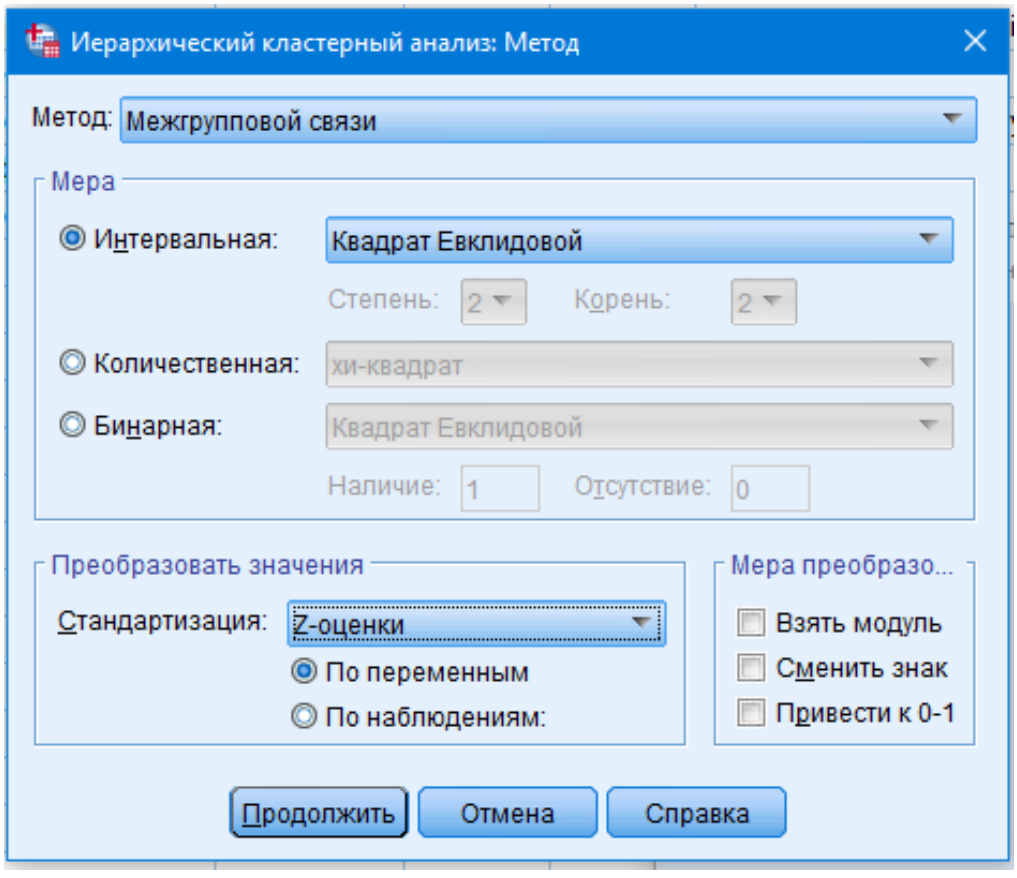


Рис. 7.5. Встановлення методів ієрархічної кластеризації

У списку «Интервальная» можна обрати спосіб розрахунку відстані між об’єктами, а саме:

– «Квадрат расстояния Евклида» – відстань між об’єктами розраховується як різниця квадратів відповідних змінних цих об’єктів, що беруть участь в аналізі;

– «Косинус» – метод виміру відстані базується на косинусах векторів змінних;

– «Корреляция Пирсона» – метод виміру відстані за допомогою кореляції векторів змінних;

– «Чебышева» – розрахунок відстані як максимуму абсолютного значення різниці між елементами;

– «Настроенная» – дозволяє задати користувацький вимір відстані тощо.

Спосіб стандартизації обирається зі списку «Стандартизація», зауважимо, що проведення стандартизації різними методами може дати різні результати досліднику. Загальноприйнятим є метод стандартизації z-значення.

Для збереження результатів кластерного аналізу потрібно відкрити вкладку «Сохранить» і проставити мітки, яку кількість кластерів потрібно зберегти. Дослідник може отримати одне значення, а може декілька у встановленому ним діапазоні (рис. 7.6).

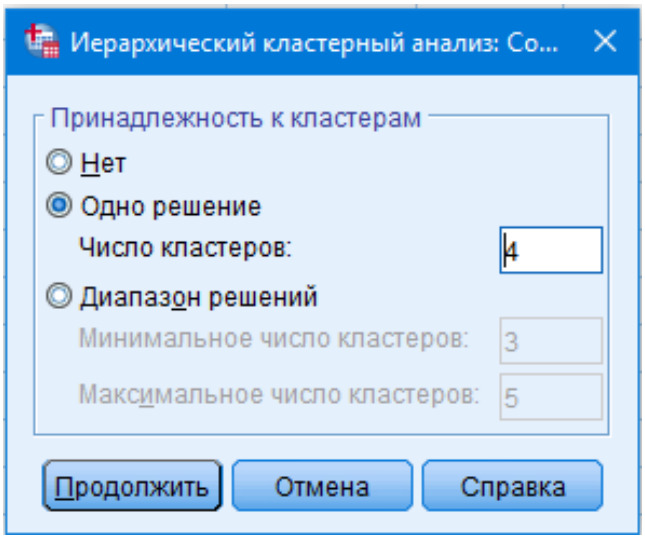


Рис. 7.6. Збереження ієрархічної кластеризації

Натиснувши кнопку «Продолжить», отримаємо такий результат (табл. 7.2).

Таблиці 7.2

Порядок агломерации (кластеров)

Етап	Об'єднаний кластер		Коефіцієнти	Етап першої появи кластера		Наступний етап
	Кластер 1	Кластер 2		Кластер 1	Кластер 2	
1	**18**	**20**	,566	0	0	**5**
2	5	16	,584	0	0	4
3	17	22	,624	0	0	7
4	5	21	,697	2	0	7
5	**18**	**24**	,870	1	0	8
6	1	13	1,354	0	0	9
7	5	17	1,485	4	3	9
8	2	18	1,514	0	5	10
9	1	5	1,831	6	7	12
10	2	23	2,286	8	0	11
11	2	8	2,838	10	0	13
12	1	10	2,895	9	0	13
13	1	2	3,914	12	11	18
14	12	14	4,186	0	0	16
15	7	15	5,199	0	0	17
16	12	19	6,017	14	0	20
17	7	9	7,875	15	0	19
18	1	11	10,892	13	0	19
19	1	7	12,709	18	17	21
20	3	12	13,013	0	16	22
21	1	6	18,555	19	0	22
22	1	3	27,703	21	20	23
23	1	4	39,410	22	0	0

Як видно з таблиці, на першому етапі в кластер об'єднується пара об'єктів, відстань між якими є найменшою. На другому – знову розраховується відстань між об'єктами і об'єднують у кластер пару найбільш близьких об'єктів, у результаті отримується або один кластер із трьох об'єктів, або два кластери із двох об'єктів. Процес злиття продовжується доти, доки усі об'єкти не потраплять в один кластер. На етапі 1 об'єднані об'єкти 18 і 20. Відстань між об'єктами дорівнює 0,566. Не один із двох об'єктів не належить до будь-якого кластера, про що свідчать нулі в стовпчику кластер 1 і 2 (етап першої появи кластера). Наступним етапом для цього кластера є етап 5, на якому до кластера приєднується об'єкт 24 тощо.

Належність до кластерів подано в таблиці 7.3.

Таблиця 7.3

Належність до кластерів

Спостереження	Кластери 4
1:Вінницька	1
2:Волинська	1
3:Дніпропетровська	2
4:Донецька	3
5:Житомирська	1
6:Закарпатська	4
7:Запорізька	1
8:Івано-Франківська	1
9:Київська	1
10:Кіровоградська	1
11:Луганська	1
12:Львівська	2
13:Миколаївська	1
14:Одеська	2
15:Полтавська	1
16:Рівненська	1
17:Сумська	1
18:Тернопільська	1
19:Харківська	2
20:Херсонська	1
21:Хмельницька	1
22:Черкаська	1
23:Чернівецька	1
24:Чернігівська	1

Для зручності трактування результату необхідно зобразити розподіл за кластерами за допомогою діаграми (рис. 7.7). Для цього потрібно активізувати таблицю, натиснувши двічі ліву клавішу мишки, виділити стовпчик і натиснути праву кнопку мишки. З контекстного меню обрати «Создать диаграмму» – «Столбчатую».

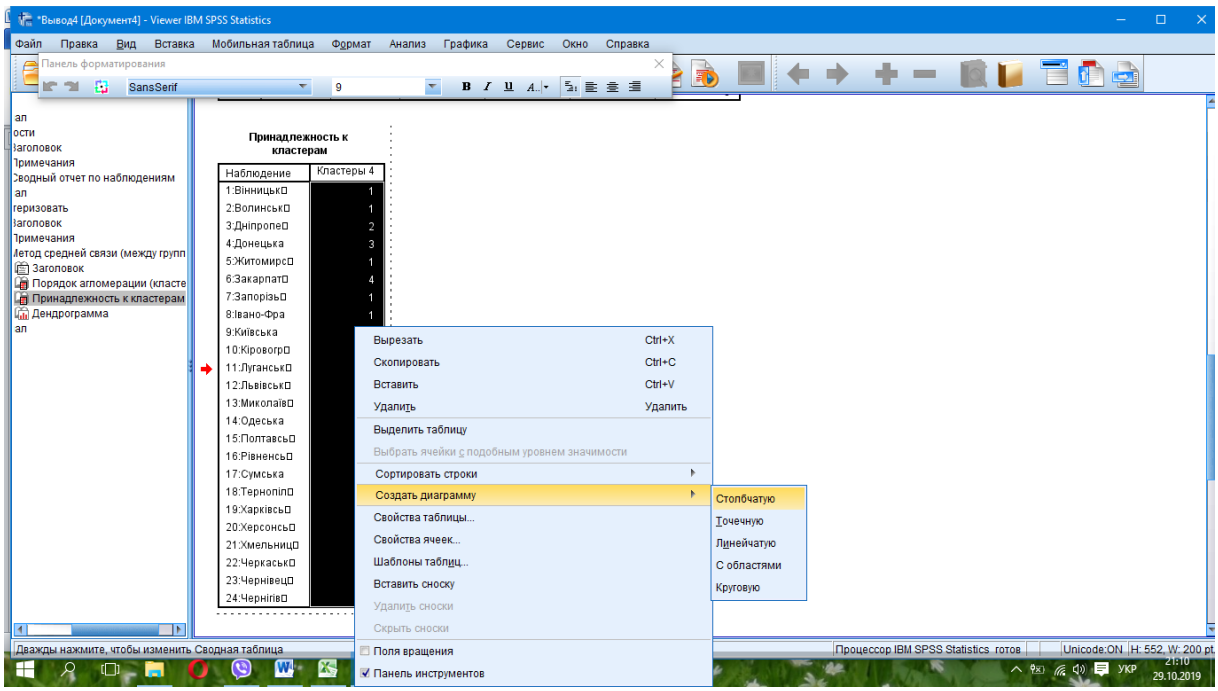


Рис. 7.7. Побудова діаграми

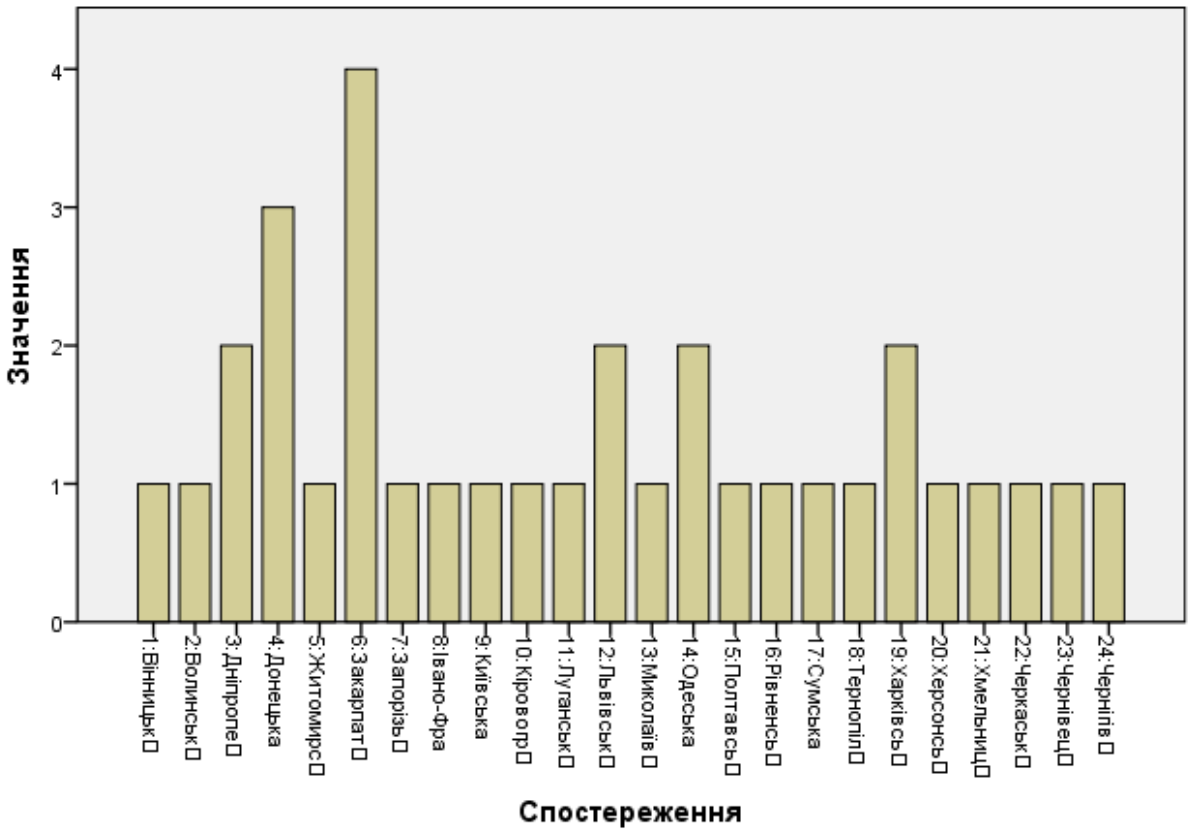


Рис. 7.8. Діаграма ієрархічної кластеризації

Дендрограма показує процес кластеризації за допомогою деревовидної структури (рис. 7.9).

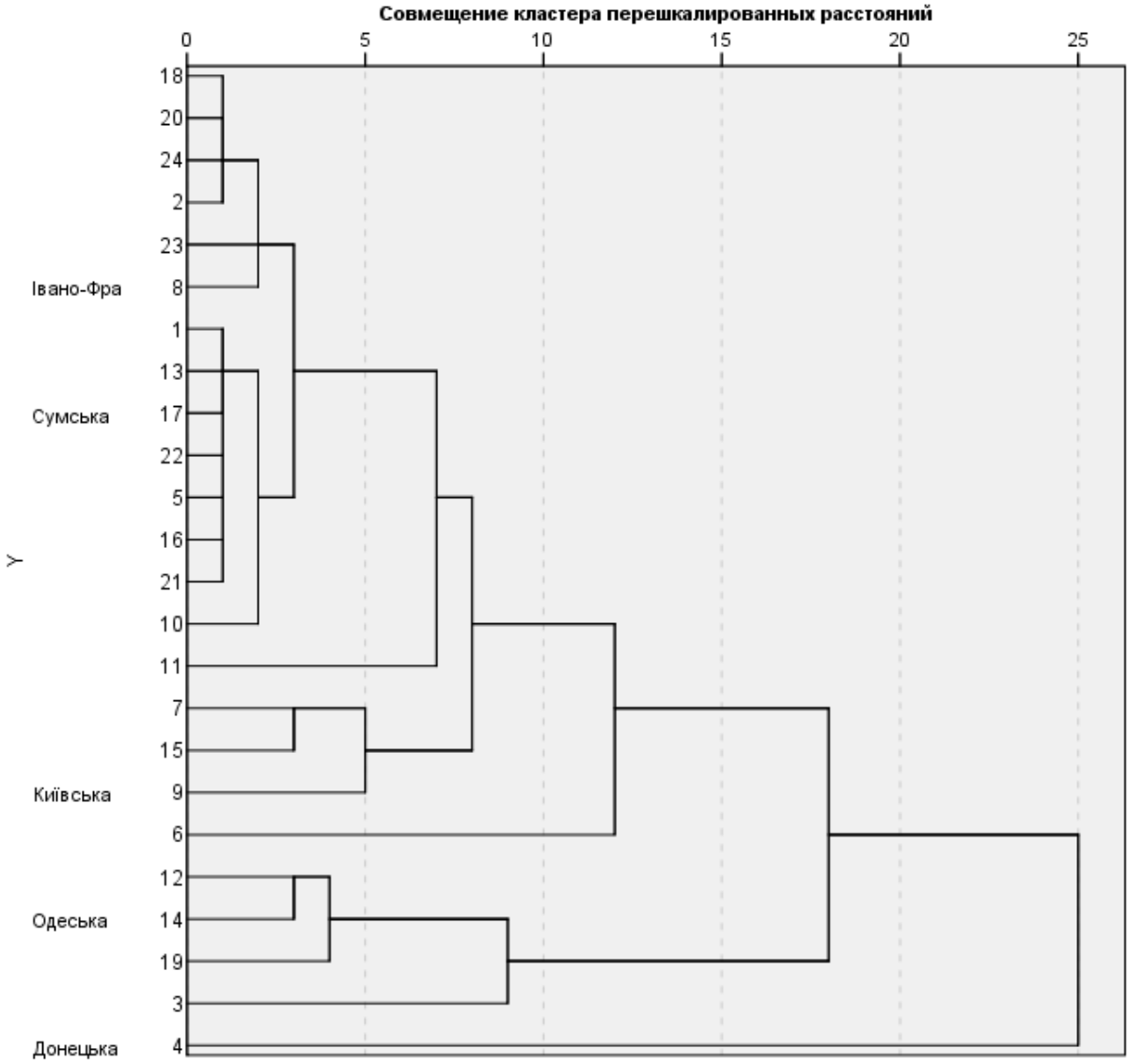


Рис. 7.9. Дендрограма ієрархічної кластеризації

Дає можливість досліднику проаналізувати всі етапи кластеризації, а саме визначити їх кількість і відстані. Шкала від 0 до 25 умовна. На дендрограмі будь-яке рішення характеризується вертикальною лінією, кількість точок перетину якої з деревом відповідає кількості кластерів на поточному етапі. Отже, якщо розмітити лінію на 15, то це 18 і 19 кроки, в яких отримується 3 кластери. Дендрограма дозволяє розрахувати оптимальну кількість кластерів, яку необхідно було будувати. Для цього від кількості параметрів віднімаємо кількість кроків, тобто 24 – 20 = 4 кластери.

Висновок. Виробничий потенціал України значно відрізняється за регіонами, кластерний аналіз показав, що в перший кластер потрапила лише Закарпатська область, яка має низький рівень

безробіття і високий рівень середньої зарплати. У другий Донецька область як одна з потужних промислових областей, при цьому вона має максимальну кількість постійного населення й індекс споживчих цін. У третій кластер потрапило чотири промислових регіони, а саме Дніпропетровський (має максимальну кількість зайнятого населення), Львівський (притаманний мінімальний індекс споживчих цін), Одеський (майже мінімальний рівень безробіття) і Харківський (друге місце щодо зайнятості населення). Усі інші регіони потрапили в четвертий кластер.

2. Розглянемо кластеризацію К-середнім. Функція в SPSS: «**Анализ**» – «**Классификация**» – «**Кластеризация К-средними**» (рис. 7.10).

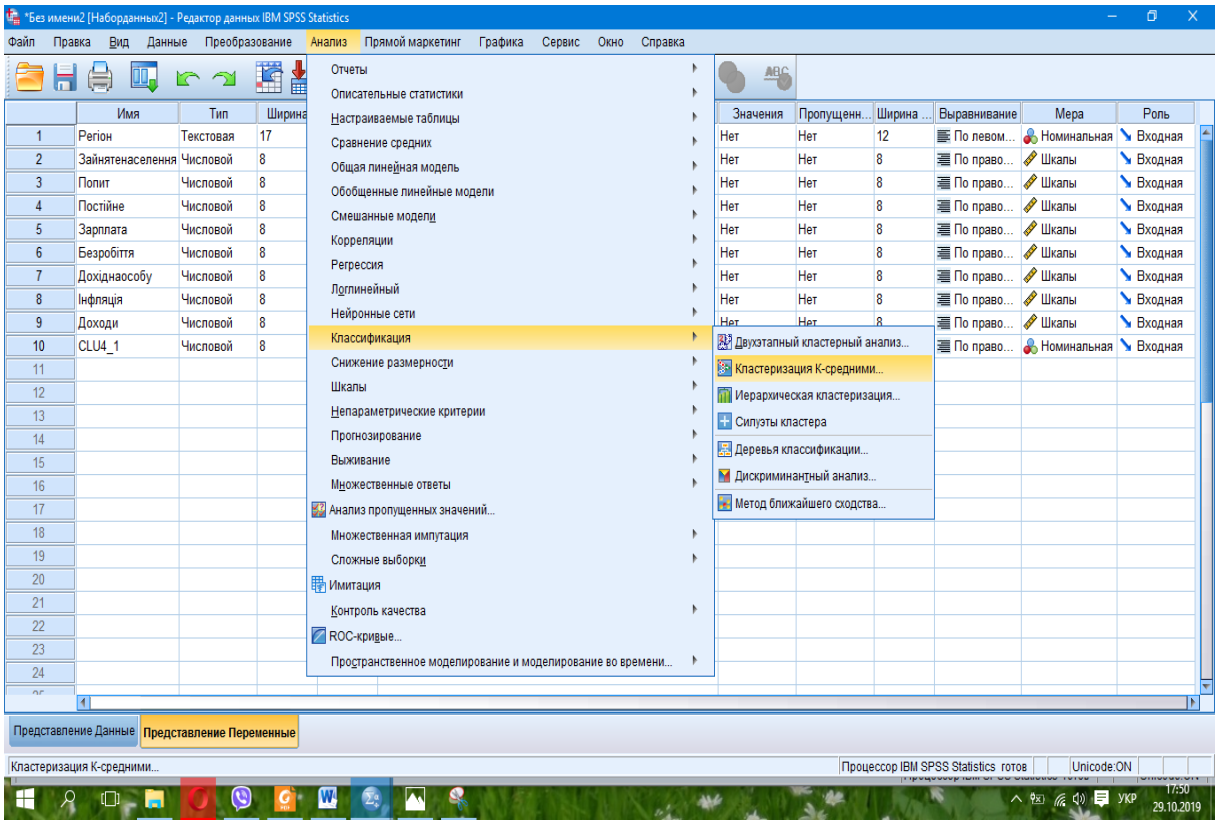


Рис. 7.10. Метод кластеризації К-середнім

Обравши функцію зі списку, отримаємо діалогове вікно, яке надасть змогу обрати параметри і методи дослідження (рис. 7.11).

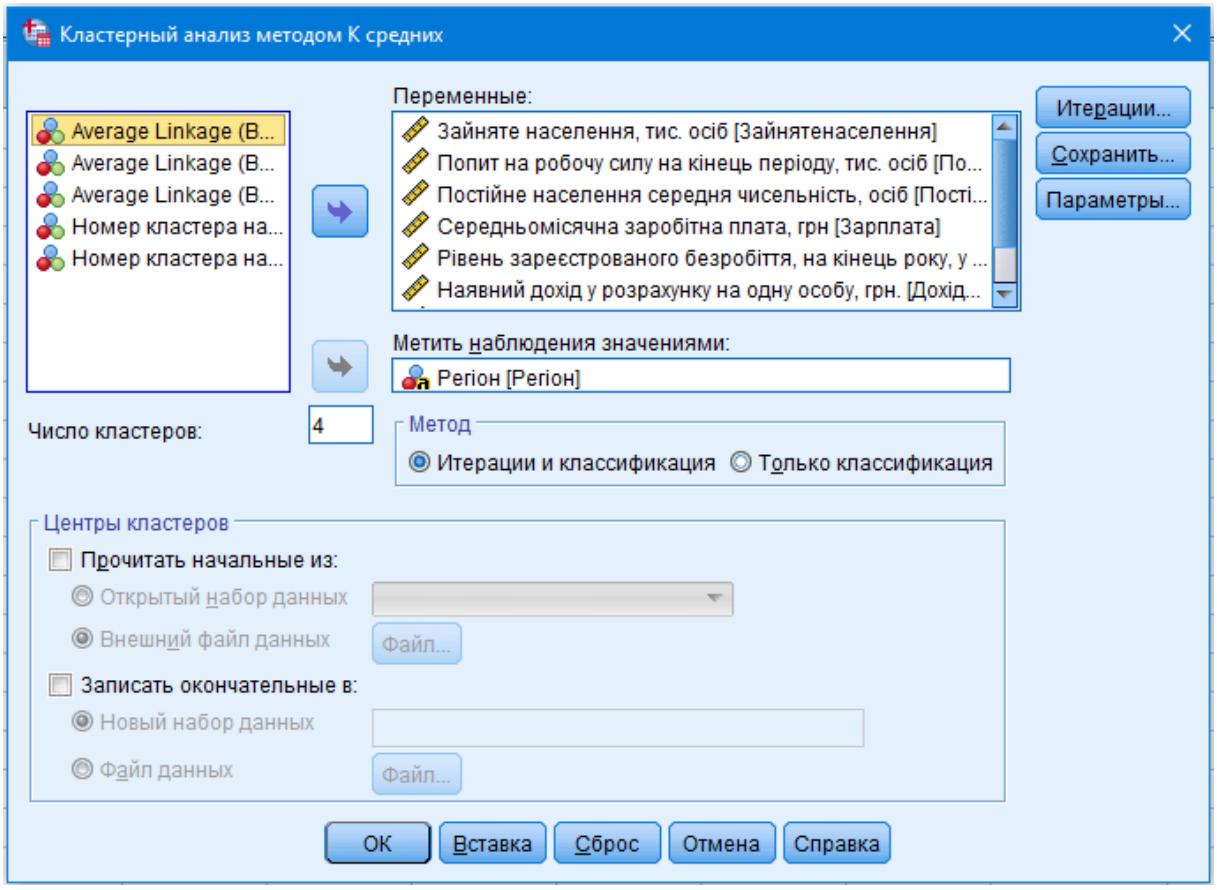


Рис. 7.11. Діалогове вікно вибору параметрів для кластеризації К-середнім

Обов’язково необхідно поставити мітку «Принадлежание к кластеру», яка міститься у вкладці «Сохранить» (рис. 7.12).

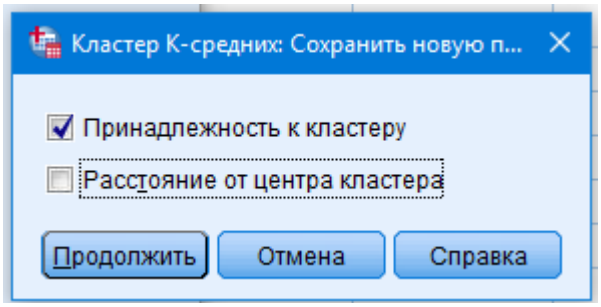


Рис. 7.12. Збереження результатів кластеризації К-середнім

Натиснувши вкладку «Параметры», поставити мітку про відображення статистичних результатів центрованих клатерів початкових і кінцевих (рис. 7.13).

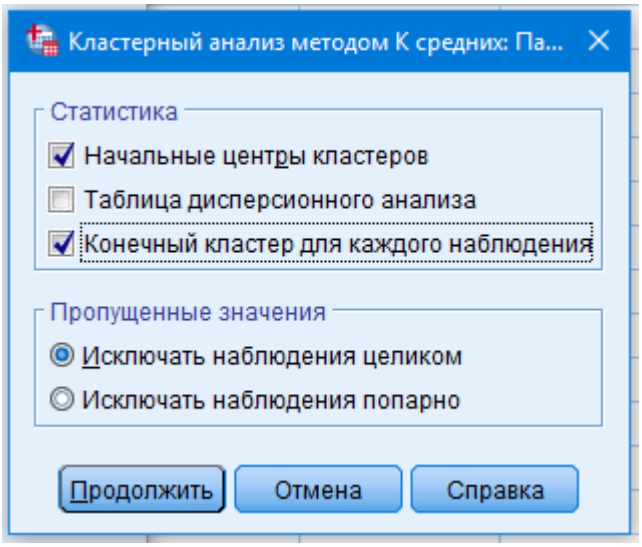


Рис. 7.13. Деталізація параметрів кластеризації К-середнім

Результат отримаємо після натискання кнопки «Продолжить».

Приклад. За даними наведеного вище прикладу проведемо кластеризацію за допомогою методу К-середнього.

Початкові центри кластерів вказують, що третій кластер має максимальні значення за більшістю показників, крім рівня зареєстрованого безробіття й індексу споживчих цін, які наближені до мінімального значення (табл. 7.4).

Таблиця 7.4

Початкові центри кластерів

Показники	Кластер			
	1	2	3	4
Зайняте населення, тис. осіб	580,60	298,20	1 402,30	1 258,90
Попит на робочу силу на кінець періоду, тис. осіб	3,70	0,50	6,40	3,30
Постійне населення, середня чисельність, тис. осіб	1 399,30	2 155,22	3 215,50	2 669,17
Середньомісячна заробітна плата, грн	8 375,00	7 365,00	8 862,00	7 657,00
Рівень зареєстрованого безробіття, на кінець року, у % до населення працездатного віку	3,00	2,10	1,70	1,60
Наявний дохід у розрахунку на одну особу, грн	60217,50	20618,60	72883,40	60117,70
Індекси споживчих цін	109,30	109,30	109,20	111,20
Доходи населення, млн грн	112856,00	60086,00	305510,00	216333,00

Другий кластер містить багато мінімальних значень, а саме кількість зайнятого населення, попит на робочу силу, середньомісячну зарплату, наявний дохід у розрахунку на одну особу і доходи населення.

Метод кластеризації К-середнім дає можливість змінити розраховані центри кластерів, що дозволяє отримати більш точний результат (табл. 7.5), тому більш детально доцільно описувати кінцеві центри кластерів (табл. 7.6).

Таблиця 7.5

Хронологія ітерацій [a]

Ітерація	Зміна центрів кластерів			
	1	2	3	4
1	11380,951	23511,484	,000	26379,105
2	1069,611	3525,512	,000	8999,627
3	2044,327	2685,590	,000	6326,949
4	,000	,000	,000	,000

а. Подібність досягнута завдяки невеликим змінам центрів кластерів або без них. Максимальна зміна абсолютної координації для будь-якого центру: ,000. Поточна ітерація – 4. Мінімальна відстань між начальними центрами: 65 988,007.

Таблиця 7.6

Кінцеві центри кластерів

Показники	Кластер			
	1	2	3	4
Зайняте населення, тис. осіб	603,03	439,85	1 402,30	925,15
Попит на робочу силу на кінець періоду, тис. осіб	1,93	1,32	6,40	3,22
Постійне населення, середня чисельність, тис. осіб	1 399,84	1 169,03	3 215,50	2 531,10
Середньомісячна заробітна плата, грн	7 768,25	7 366,62	8 862,00	8 529,67
Рівень зареєстрованого безробіття, на кінець року, у % до населення працездатного віку	2,33	2,30	1,70	1,55
Наявний дохід у розрахунку на одну особу, грн	54 016,20	46 737,45	72 883,40	56 693,82
Індекси споживчих цін	109,15	109,62	109,20	110,35
Доходи населення, млн грн	99 995,50	69 660,77	305 510,00	176 593,83

Таблиця кінцевих центрів кластерів у цьому випадку суттєво не змінилася. Лідером є кластер три і в нього потрапив лише один регіон (Дніпропетровський) (табл. 7.7, рис. 7.14).

Таблиця 7.7

Належність до кластерів

Номер спостереження	Регіон	Кластер	Відстань
1	Вінницька	1	12958,435
2	Волинська	2	5927,666
3	Дніпропетровська	3	,000
4	Донецька	4	25006,493
5	Житомирська	2	15054,759
6	Закарпатська	2	6725,481
7	Запорізька	4	29749,357
8	Івано-Франківська	1	14212,141
9	Київська	4	31634,745
10	Кіровоградська	2	6693,619
11	Луганська	2	27836,368
12	Львівська	4	9459,367
13	Миколаївська	2	14774,424
14	Одеська	4	12554,399
15	Полтавська	1	14290,460
16	Рівненська	2	3312,028
17	Сумська	2	14100,054
18	Тернопільська	2	8614,133
19	Харківська	4	39897,573
20	Херсонська	2	3821,417
21	Хмельницька	1	12840,760
22	Черкаська	2	12883,308
23	Чернівецька	2	18785,186
24	Чернігівська	2	4303,059

З максимальних значень за параметрами його можна охарактеризувати як потужний промисловий центр України із відносно високими доходами населення – «найкраще місце працевлаштування» (рис. 7.13).

Візуалізацію результатів розподілу можна отримати, побудувавши додатково за таблицею 7.6 діаграму (рис.7.15).

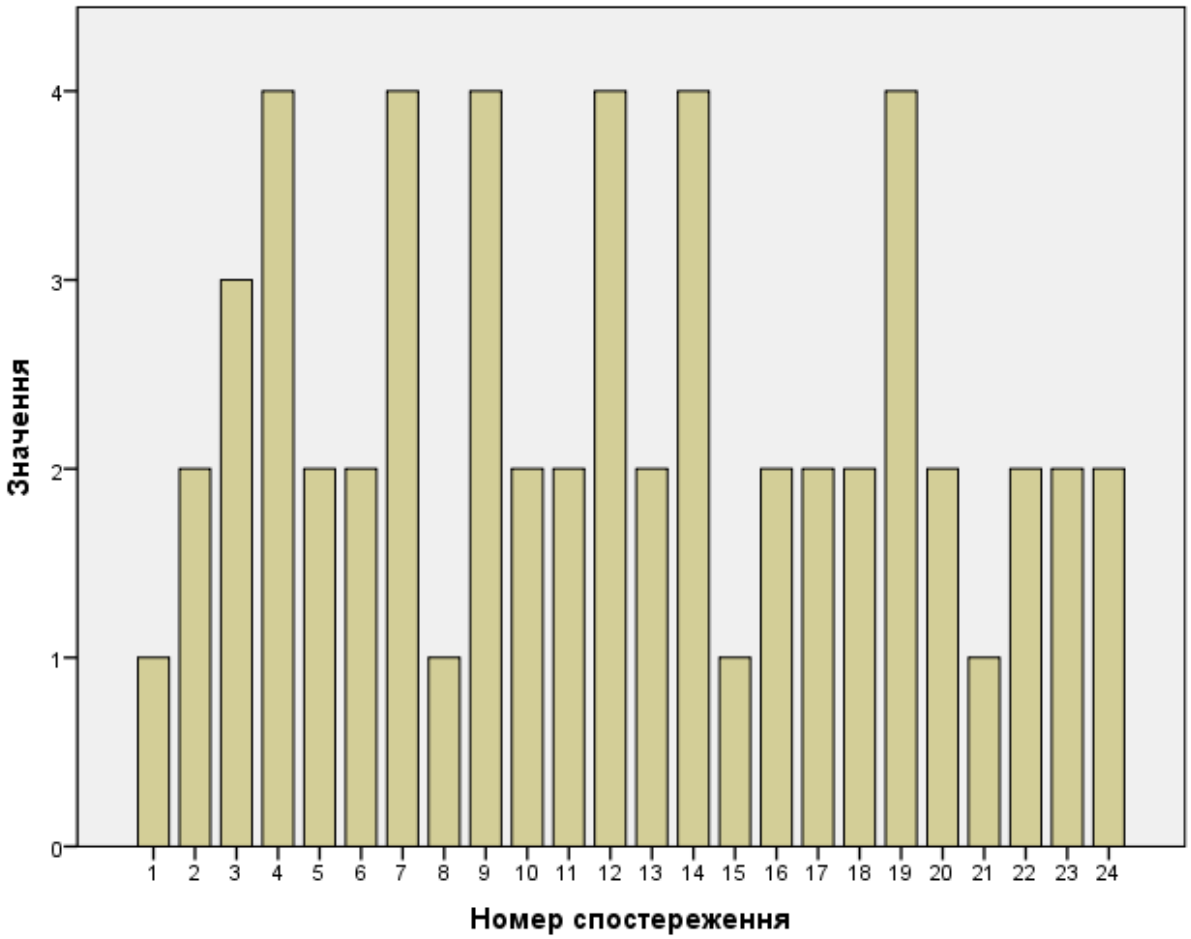


Рис. 7.14. Діаграма кластера за кластеризацією К-середнім

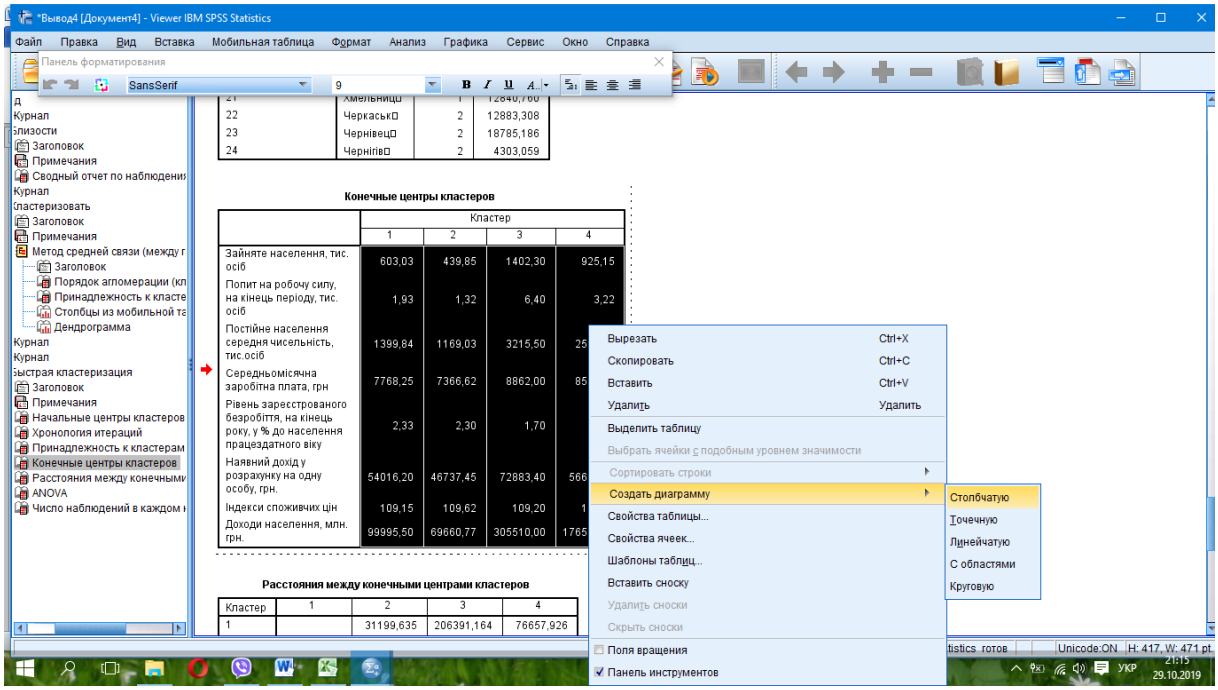


Рис. 7.15. Етапи побудови діаграми кінцевих центрів кластерів

Другий кластер містить майже всі мінімальні значення центрів кластерів, його можна назвати «найгірше місце працевлаштування» (рис. 7.16).

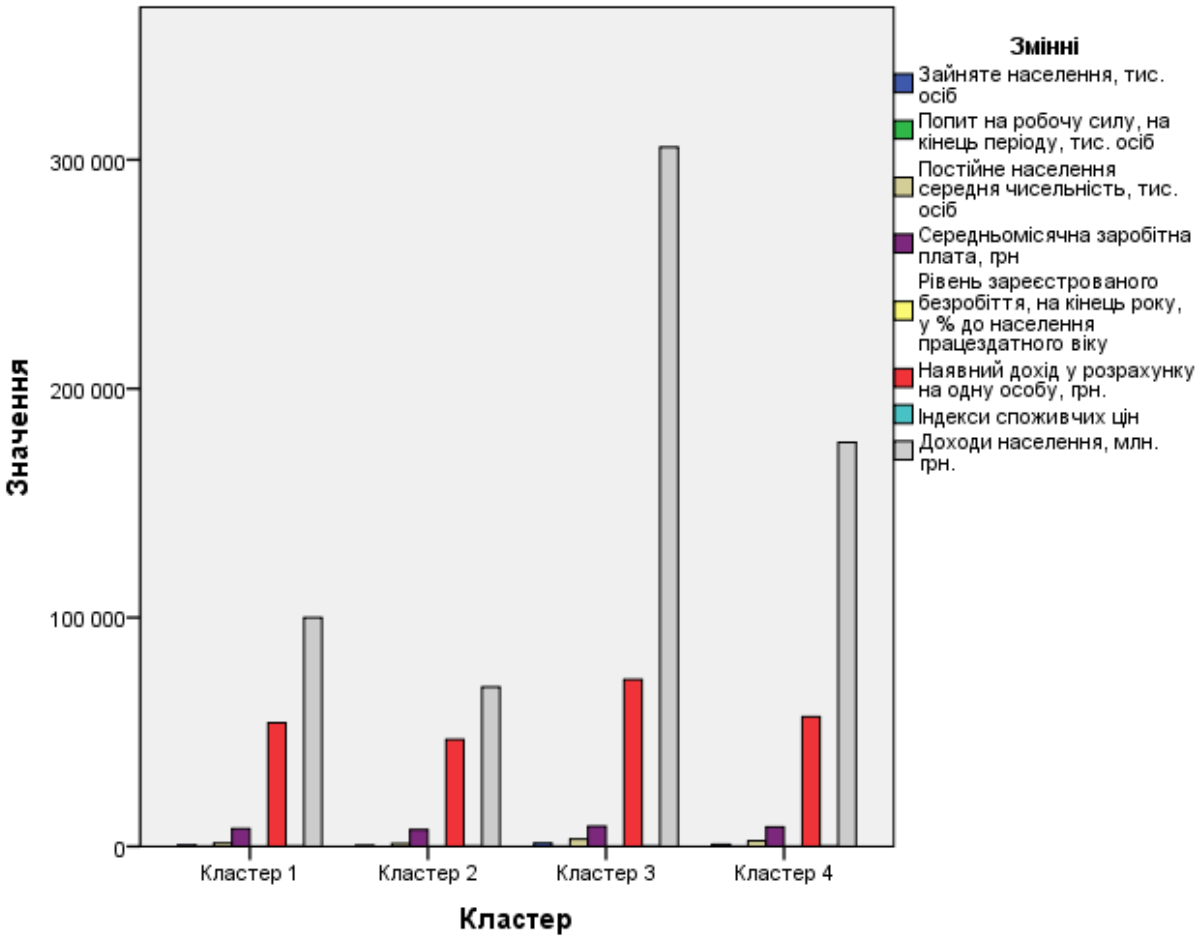


Рис. 7.16. Діаграма кінцевих центрів кластерів

Аутсайдерами є більше половини регіонів (13) України, які потрапили в другий кластер, а саме Волинська, Житомирська, Закарпатська, Кіровоградська, Луганська, Миколаївська, Рівненська, Сумська, Тернопільська, Херсонська, Черкаська, Чернігівська, Чернівецька області (табл. 7.7, рис. 7.14). Це сільськогосподарські регіони, які мають низький рівень розвитку промислового сектору, крім Луганської області, яка потрапила в цей кластер через воєнні дії в Україні, що негативно вплинуло на його соціально-економічні показники.

Наступні два кластери суттєво не вирізняються за кількістю регіонів (табл. 7.8).

Так, у перший кластер увійшли чотири регіони, а саме Вінницький, Івано-Франківський, Полтавський і Хмельницький – це

області з максимальним рівнем зареєстрованого безробіття. У четвертому кластері характерним є параметр індексу споживчих цін, тому можна надати кластеру назву «найдорожчі ціни», має шість областей, а саме Донецьку, Запорізьку, Київську, Львівську, Одеську, Харківську.

Таблиця 7.8

Кількість спостережень у кожному кластері

Кластер	1 «найбільше безробіття»	4
	2 «найгірше місце працевлаштування»	13
	3 «найкраще місце працевлаштування»	1
	4 «найдорожчі ціни»	6
Валідні		24
Пропущені		0

Програма також визначає відстані між кінцевими центрами кластерів (табл. 7.8) і проводить дисперсійний аналіз ANOVA (табл. 7.9)

Таблиця 7.9

Відстань між кінцевими центрами кластерів

Кластер	1	2	3	4
1		31199,635	206391,164	76657,926
2	31199,635		237309,546	107411,605
3	206391,164	237309,546		129931,856
4	76657,926	107411,605	129931,856	

При цьому необхідно зауважити, що F-критерій потрібно застосовувати лише з описовою ціллю у зв’язку з тим, що кластери вибрані так, щоб різниця між спостереження в різних кластерах була максимальною. Розраховані рівні значимості не скореговані для цього, і тому їх не можна застосовувати для перевірки гіпотези про рівність середніх кластерів.

Таблиця 7.10

ANOVA

Показники	Кластер		Похибка		F	Значимість
	Середній квадрат	ст. св.	Середній квадрат	ст. св.		
Зайняте населення, тис. осіб	530681,072	3	15183,617	20	34,951	0,000

Попит на робочу силу, на кінець періоду, тис. осіб	11,313	3	1,626	20	6,958	0,002
Постійне населення середня чисельність, тис. осіб	3453248,962	3	261911,342	20	13,185	0,000
Середньомісячна заробітна плата, грн	2254584,155	3	265820,558	20	8,482	0,001
Рівень зареєстрованого безробіття, на кінець року у % до населення працездатного віку	0,899	3	0,383	20	2,345	0,103
Наявний дохід у розрахунку на одну особу, грн	313090976,944	3	94737684,961	20	3,305	0,041
Індекси споживчих цін	1,357	3	0,785	20	1,727	0,194
Доходи населення, млн грн	28782706060,828	3	268217224,107	20	107,311	0,000

Висновок. У випадку, коли потрібно швидко знайти роботу з високим рівнем доходу, потрібно їхати в Дніпропетровський регіон. Складно буде шукати роботу у Вінницькій, Івано-Франківській, Полтавській і Хмельницькій областях. Найдорожчий рівень життя у Донецькому, Запорізькому, Київському, Львівському, Одеському, Харківському регіонах. Кластерний аналіз показав, що більшість регіонів України (13) потрапили в кластер «найгірше місце працевлаштування».

Перелік питань для самоконтролю

1. Поясніть сутність кластерного аналізу.
2. Назвіть завдання, які стоять перед дослідником кластерного аналізу.
3. Для чого застосовується стандартизація показників і які методи стандартизації ви знаєте?
4. Назвіть методи кластерного аналізу, які передбачені в програмному продукті SPSS.
5. Поясніть сутність двоетапного кластерного аналізу.
6. Назвіть особливість кластеризації за допомогою К-середнім.
7. Охарактеризуйте особливості застосування ієрархічної кластеризації.

8. Назвіть основні етапи кластерного аналізу і поясніть їхню сутність.

9. Поясніть сутність поняття «Порядок агломерації».

10. Для чого розраховується в кластерному аналізі «Матриця близькості»?

11. Для чого використовують «Дендрограму» в кластерному аналізі?

12. Назвіть основні способи об'єднання параметрів.

13. Назвіть основні способи розрахунку відстані між об'єктами у кластерному аналізі.

Тести

1. Назвіть базові методи формування кластерів:

а) злиття і дроблення;

б) інтервального розподілу;

в) міжгрупових зв'язків;

г) правильної відповіді немає.

2. Етапи кластерного аналізу:

а) вибір параметрів і способу виміру відстані, формування кластерів та інтерпретація результатів;

б) вибір параметрів, формування кластерів та інтерпретація результатів;

в) вибір способу виміру відстані, формування кластерів та інтерпретація результатів;

г) правильної відповіді немає.

3. Назвіть методи кластерного аналізу, які функціонують у SPSS:

а) двоетапний кластерний аналіз, кластеризація К-середніми та ієрархічна кластеризація;

б) одноетапний кластерний аналіз, кластеризація К-середніми та ієрархічна кластеризація;

в) двоетапний кластерний аналіз, кластеризація R-середніми та ієрархічна кластеризація;

г) правильної відповіді немає.

4. Надає змогу візуалізувати інформацію про відстань між об’єктами і кластерами:

а) матриця близькості;

б) діаграма кластерів;

в) z-оцінки;

г) правильної відповіді немає.

5. Графічно відображає відносну величину різності між параметрами або кластерами на кожному кроці процесу:

а) дендрограма;

б) матриця близькості;

в) діаграма;

г) правильної відповіді немає.

6. Метод, в якому об’єднуються на кожному кроці кластери з мінімальною відстанню:

а) міжгрупових зв’язків;

б) центроідної кластеризації;

в) метод Уорда;

г) правильної відповіді немає.

7. В основі методу полягає вибір змінних, відстань між якими максимальна:

а) міжгрупових зв’язків;

б) центроідної кластеризації;

в) метод Уорда;

г) правильної відповіді немає.

8. В основі методу полягає вибір змінних, відстань між якими максимальна:

а) міжгрупових зв’язків;

б) центроідної кластеризації;

в) дальнього сусіда;

г) правильної відповіді немає.

9. В основі методу полягає вибір змінних, відстань між якими мінімальна:

а) близького сусіда;

б) центроідної кластеризації;

в) метод Уорда;

г) правильної відповіді немає.

10. Метод, в якому оцінюється відстань між середніми значеннями груп змінних:

а) близького сусіда;

б) центроідної кластеризації;

в) метод Уорда;

г) далекого сусіда.

11. Метод, в якому як міра використовується квадрат евклідової відстані, який повинен бути мінімальним:

а) близького сусіда;

б) центроідної кластеризації;

в) метод Уорда;

г) далекого сусіда.

12. Метод, в якому відстань між об’єктами розраховується як різниця квадратів відповідних змінних цих об’єктів:

а) квадрат відстані Евкліда;

б) квадрат відстані Чебишева;

в) кореляція Пірсона;

г) метод Уорда.

13. Метод виміру відстані за допомогою кореляції векторів змінних:

а) метод Чебишева;

б) квадрат відстані Евкліда;

в) метод косинусу;

г) правильної відповіді немає.

14. Розрахунок відстані як максимуму абсолютного значення різниці між елементами:

а) метод Чебишева;

б) квадрат відстані Евкліда;

в) метод косинусу;

г) правильної відповіді немає.

Економічна інтерпретація кластерного аналізу

Приклад 1. Дослідити рівень розвитку трудового потенціалу регіонів України та порівняти їх за допомогою кластерного аналізу, визначити найбільш привабливі регіони для працевлаштування (табл. 1), а також оцінити зайнятість населення за видами економічної діяльності (табл. 4).

Таблиця 1

Статистичні дані трудового потенціалу за 2018 рік

Показники	Зайняте населення, тис. осіб	Попит на робочу силу на кінець періоду, тис. осіб	Постійне населення, середня чисельність, тис. осіб	Середньомісячна заробітна плата, грн	Рівень зареєстрованого безробіття на кінець року у % до населення працездатного віку	Наявний дохід у розрахунку на одну особу, грн	Індекси споживчих цін	Доходи населення, млн грн
Вінницька	652,7	0,9	1561,016	7801	2,7	54992	109	112916
Волинська	371,1	2,6	1034,165	7324	2	46475,1	109,9	63741
Дніпропетровська	1402,3	6,4	3215,499	8862	1,7	72883,4	109,2	305510
Донецька	741	0,8	4170,296	9686	1,4	31888	112,3	174157
Житомирська	516,7	2,2	1226,485	7372	2,6	52135,9	109,1	83714
Закарпатська	502,4	1,2	1254,645	8070	0,9	40471,6	112,2	67323
Запорізька	732,2	0,8	1713,714	8726	2,4	67982,5	109,2	149083
Івано-Франківська	656,8	1,4	1372,648	7551	1,5	48367,7	109,1	86956
Київська	755,7	4,9	1755,333	9097	1,6	63498,4	110	145715
Кіровоградська	380,5	1,8	944,484	7191	3,4	51018	109	64523
Луганська	298,2	0,5	2155,221	7365	2,1	20618,6	109,3	60086
Львівська	1061,2	6,2	2507,445	8001	1,3	55510,7	110,1	185963
Миколаївська	496,2	1,4	1135,495	8160	2,8	55543,9	109,4	81497

Одеська	1001,9	3,3	2370,631	8011	1	61165,6	109,3	188312
Полтавська	580,6	3,7	1399,296	8375	3	60217,5	109,3	112856
Рівненська	473,6	1,3	1157,914	7469	2,5	47729,1	109,3	72819
Сумська	485,1	1,4	1085,659	7324	2,9	55934,4	109,7	80348
Тернопільська	410,8	1,2	1045,845	6969	1,9	43512,5	109,7	61684
Харківська	1258,9	3,3	2669,167	7657	1,6	60117,7	111,2	216333
Херсонська	448,2	0,6	1040,878	7058	1,9	50109,4	109,5	67894
Хмельницька	522	1,7	1266,394	7346	2,1	52487,6	109,2	87254
Черкаська	522,6	0,5	1209,728	7478	2,9	50292,6	109,7	82043
Чернівецька	382,9	1,3	902,473	6991	1,6	42850,4	108,7	51288
Чернігівська	429,7	1,2	1004,37	6995	2,4	50895,4	109,6	68630

Джерело: за даними[22].

Втрата трудового потенціалу в результаті безвізового режиму сьогодні одна із основних проблем в Україні. Він виступає однією з основних умов забезпечення соціально-економічного розвитку регіону та країни в цілому. Високоефективне використання трудового потенціалу, його розвиток та рівномірний розподіл за регіонами дозволить підвищити економічний рівень розвитку країни в цілому, що, в свою чергу, сприятиме зменшенню рівня безробіття та відпливу трудових ресурсів з країни, дозволить сформувати належний рівень оплати праці, покращити умови праці та підвищити рівень життя населення.

Застосування кластерного аналізу дозволяє об'єднати визначений перелік параметрів у групи за спорідненими ознаками. На першому етапі кластерного аналізу обрані наявні статистичні дані, які характеризують трудовий потенціал регіонів України. Для аналізу застосовані такі параметри за 2018 рік: зайняте населення, тис. осіб; попит на робочу силу, на кінець періоду, тис. осіб; постійне населення середня чисельність, тис. осіб; середньомісячна заробітна плата, грн; рівень зареєстрованого безробіття на кінець року у відсотках до населення, працездатного віку; наявний дохід у розрахунку на одну особу, грн; індекси споживчих цін, %; доходи населення, млн грн. Із ряду спостереження вилучено м. Київ у зв'язку із суттєвим коливанням усіх його значень від загальної тенденції (табл. 1).

[22]Офіційний сайт Державної служби статистики України. URL : http://www.ukrstat.gov.ua/operativ/gdn/dvn/arh_dvn2001.html

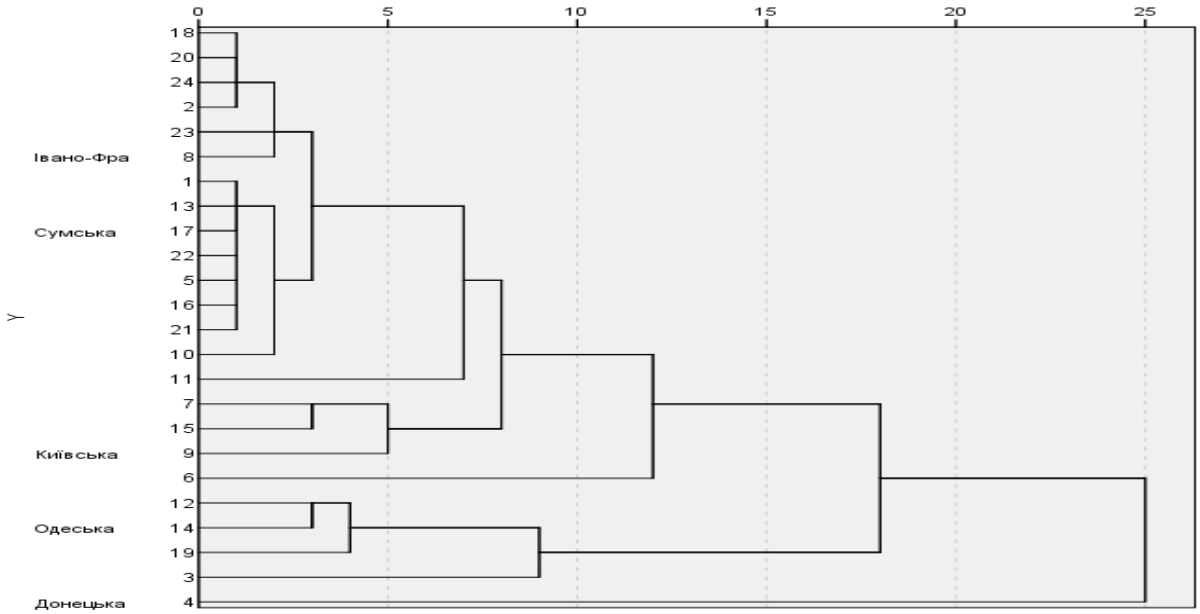


Рис. 1. Дендрограма ієрархічної кластеризації

Джерело: розраховано авторами.

На другому етапі побудована дендрограма ієрархічної кластеризації в програмному продукті IBM SPSS STATISTICS, яка дозволила визначити оптимальну кількість кластерів (рис. 1).

За допомогою деревовидної структури розраховано кількість здійснених кроків (20). Знайдено різницю між кількістю параметрів (кількість регіонів) і кількістю кроків та отримано кількість кластерів, яку доцільно будувати 4 (24 – 20 = 4 кластери).

На третьому етапі проведено кластеризацію за допомогою методу К-середнього, який дозволяє отримати повну характеристику кластерів [4].

Аналізуючи отримані результати, зазначимо, що третій кластер має максимальні значення за більшістю показників, крім рівня зареєстрованого безробіття та індексу споживчих цін, які наближені до мінімального значення (табл. 2, рис. 2).

Він є лідером серед кластерів і в нього потрапив лише один Дніпропетровський регіон (рис. 3). За максимальним значенням щодо параметрів його можна охарактеризувати як потужний промисловий центр України з відносно високими доходами населення – «найкраще місце працевлаштування» (рис. 3).

Таблиця 2

Центри кластерів трудового потенціалу

Показник	Кластер			
	1	2	3	4
Зайняте населення, тис. осіб	603,03	439,85	1 402,30	925,15
Попит на робочу силу, на кінець періоду, тис. осіб	1,93	1,32	6,40	3,22
Постійне населення, середня чисельність, тис. осіб	1 399,84	1 169,03	3 215,50	2 531,10
Середньомісячна заробітна плата, грн	7 768,25	7 366,62	8 862,00	8 529,67
Рівень зареєстрованого безробіття, на кінець року, у % до населення працездатного віку	2,33	2,30	1,70	1,55
Наявний дохід у розрахунку на одну особу, грн	54 016,20	46 737,45	72 883,40	56 693,82
Індекси споживчих цін	109,15	109,62	109,20	110,35
Доходи населення, млн грн	99 995,50	69 660,77	305 510,00	176 593,83

Джерело: розраховано авторами.

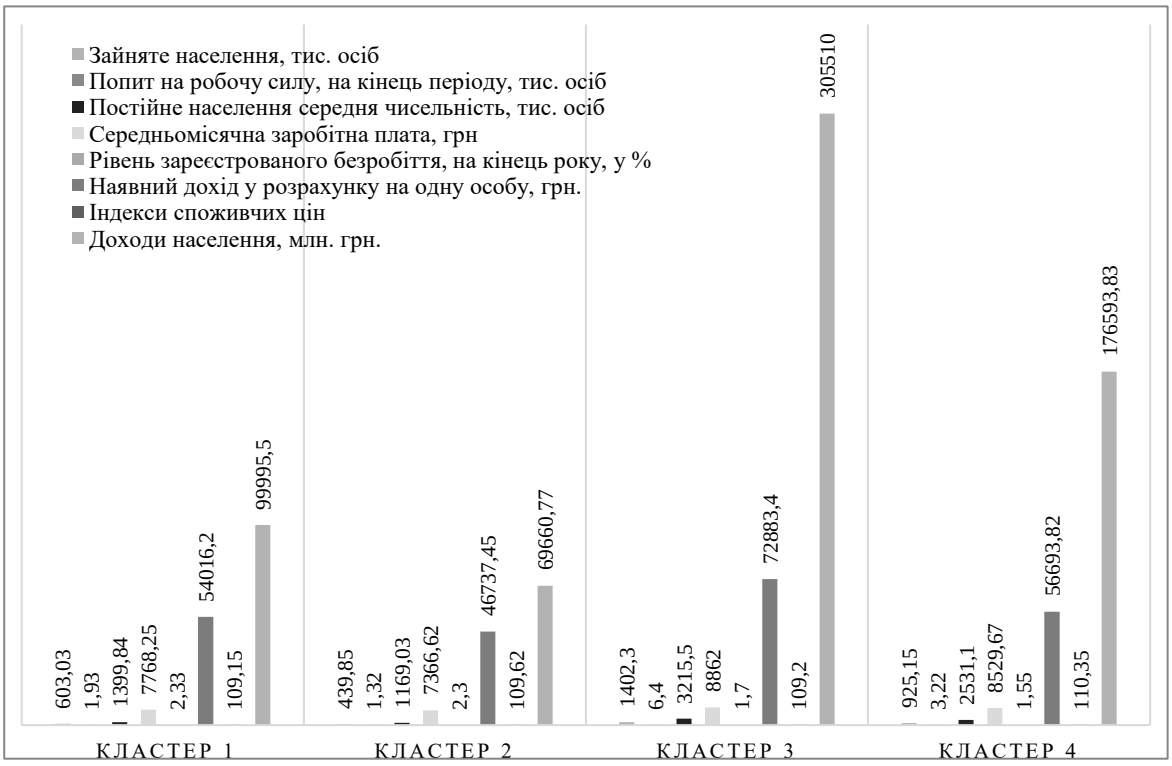


Рис. 2. Діаграма центрів кластерів

Джерело: розраховано авторами.

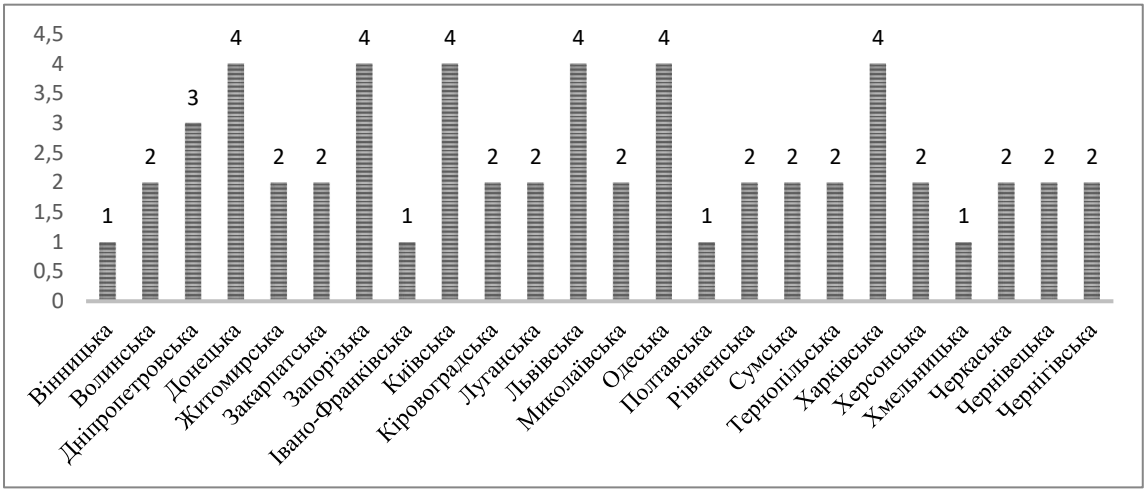


Рис. 3. Діаграма кластерів трудового потенціалу за К-середнім

Джерело: розраховано авторами.

Другий кластер містить чимало мінімальних значень, а саме кількість зайнятого населення, попит на роботу силу, середня чисельність постійного населення, середньомісячна зарплата, наявний дохід у розрахунку на одну особу і доходи населення. Його можна назвати «найгірше місце працевлаштування» (рис. 3).

Аутсайдерами є більша половина регіонів (13) України, які потрапили в другий кластер, а саме: Волинська, Житомирська, Закарпатська, Кіровоградська, Луганська, Миколаївська, Рівненська, Сумська, Тернопільська, Херсонська, Черкаська, Чернігівська та Чернівецька області (табл. 3, рис. 3).

Це сільськогосподарські регіони, які мають низький рівень розвитку промислового сектору, крім Луганської області, яка потрапила в цей кластер через воєнні дії в Україні, що негативно вплинуло на його соціально-економічні показники.

Наступні два кластери суттєво не вирізняються за кількістю регіонів (табл. 3). Так, у перший кластер увійшли чотири регіони, а саме: Вінницький, Івано-Франківський, Полтавський і Хмельницький, це області з максимальним рівнем зареєстрованого безробіття. Для четвертого кластера характерним є параметр індексу споживчих цін, тому можна надати кластеру назву «найдорожчі ціни», він має шість областей, а саме: Донецьку, Запорізьку, Київську, Львівську, Одеську та Харківську.

Таблиця 3

Кількість спостережень у кожному кластері

Кластер	К-ть	Регіон
1 «найбільше безробіття»	4	Вінницький, Івано-Франківський, Полтавський і Хмельницький
2 «найгірше місце працевлаштування»	13	Усі інші
3 «найкраще місце працевлаштування»	1	Дніпропетровський
4 «найдорожчі ціни»	6	Донецький, Запорізький, Київський, Львівський, Одеський та Харківський
Валідні	24	

Джерело: розраховано авторами.

На четвертому етапі проаналізовано зайнятість населення за видами економічної діяльності (рис. 4) для визначення регіонів із розвиненим галузевим аспектом. Для цього спочатку проаналізовано структуру зайнятого населення за видами економічної діяльності 2018 року. Найбільшу питому вагу в структурі зайнятості населення мають такі види діяльності: оптова та роздрібна торгівля, ремонт автотранспортних засобів і мотоциклів (22 %); сільське, лісове та рибне господарство (18 %); промисловість (15 %); освіта (9 %); однакову питому вагу мають транспорт, складське господарство, поштова та кур'єрська діяльність; державне управління й оборона, обов'язкове соціальне страхування; охорона здоров'я та надання соціальної допомоги (6 %).

Перераховані показники в загальній сумі становлять 82 % зайнятого населення, тому використано лише їх для кластерного аналізу К-середнім (табл. 4).

Центри кластерів вказують, що третій кластер включає в себе майже всі максимальні значення за кількістю зайнятого населення за видами економічної діяльності, крім державного управління та оборони (табл. 5).

Його можна охарактеризувати як потужний промисловий, торговельний центр України, осередок освіти та охорони здоров'я, який має найбільшу кількість працевлаштованого населення за

всіма видами економічної діяльності, це – регіони «лідери соціально-економічного розвитку». Їх лише два: Дніпропетровська і Харківська області (рис. 5).

Рис. 4. Структура зайнятого населення за видами діяльності 2018 року

Джерело: розраховано авторами.

Другий кластер містить практично всі показники зайнятого населення за видами економічної діяльності з мінімальними значеннями, а саме: освіта, охорона здоров’я та надання соціальної допомоги, сільське господарство, оптова та роздрібна торгівля, транспорт та промисловість. Його можна назвати «аутсайдери соціально-економічного розвитку». Аутсайдерами є більша половина регіонів (16) України, яка потрапила в другий кластер, а саме: Полтавська, Миколаївська, Кіровоградська, Луганська, Житомирська, Закарпатська, Волинська, Рівненська, Сумська, Тернопільська, Івано-Франківська, Херсонська, Хмельницька, Черкаська, Чернівецька та Чернігівська області (рис. 6). Сюди входять регіони, які мають низький рівень розвитку промислового і сільськогосподарського секторів, крім Луганської області, яка потрапила в цей кластер через воєнні дії в Україні, що негативно вплинуло на її соціально-економічні показники.

Таблиця 4

Зайняте населення за видами економічної діяльності та регіонами 2018 року

Показники	Оптова та роздрібна торгівля; ремонт автотранспортних засобів і мотоциклів	Сільське господарство, лісове господарство та рибне господарство	Освіта	Транспорт, складське господарство, поштова та кур'єрська діяльність	Державне управління й оборона; обов'язкове соціальне страхування	Охорона здоров'я та надання соціальної допомоги	Промисловість
Вінницька	122,6	215	59,1	41,3	37,2	41,7	68,4
Волинська	86,5	74,3	44,6	20,4	24,6	26,4	49,9
Дніпропетровська	360	107,4	111,8	87,3	65,4	80,9	336,5
Донецька	171,7	64,1	50,9	57,8	45,6	42,9	191,9
Житомирська	128,2	77	45,1	39,2	39,6	32,2	75,7
Закарпатська	86,3	130,8	47,7	23,8	22,6	28,1	58,6
Запорізька	158,8	120,8	53,2	35,8	36,7	45,5	163,9
Івано-Франківська	108,3	168	53	23,4	21,6	36,2	67,8
Київська	168,3	48,5	62,3	68,2	55	57,1	130,9
Кіровоградська	61,1	109	35,7	24,3	25,1	26,2	49,2
Луганська	79	39,6	18	17,1	25,5	15	59,3
Львівська	203,4	194,8	98,3	62,4	56,3	71,8	155,4
Миколаївська	93,4	142,5	40,3	29,5	36,2	25,8	64,9
Одеська	239,3	163,2	88,8	107,3	59,2	57,6	85,3
Полтавська	124,7	125,6	47	35,8	33,1	39,6	97,8
Рівненська	124,6	85,9	48	25,9	21,3	31,2	62,7
Сумська	98,5	116,8	41,1	26,7	26	30,1	70,2
Тернопільська	78,1	128,4	44,4	19,8	18,5	27,4	36,5
Харківська	307,7	172,4	105,7	71,2	57,6	68,8	232,1
Херсонська	99	136,4	38,5	22,7	28,3	24,6	39,8
Хмельницька	116,5	146,7	48,4	25,3	31,9	32,2	64,1
Черкаська	95,8	146,3	44,7	28,4	23,3	35,5	73,8
Чернівецька	70,9	111,7	35,3	16	17,3	22,8	39,6
Чернігівська	97,8	107,9	34,7	19,3	32,5	32,1	53,1

Джерело: за даними[23].

[23]Офіційний сайт Державної служби статистики України. URL : http://www.ukrstat.gov.ua/operativ/gdn/dvn/arh_dvn2001.html

Таблиця 5

Центри кластерів зайнятого населення за видами економічної діяльності

Показник	Кластер			
	1	2	3	4
Освіта	188,43	96,79	**333,85**	166,27
Державне управління й оборона; обов'язкове соціальне страхування	**191,00**	115,43	139,90	77,80
Охорона здоров'я та надання соціальної допомоги	82,07	41,66	**108,75**	55,47
Сільське господарство, лісове господарство та рибне господарство	70,33	24,85	**79,25**	53,93
Оптова та роздрібна торгівля; ремонт автотранспортних засобів і мотоциклів	50,90	26,71	**61,50**	45,77
Транспорт, складське господарство, поштова та кур'єрська діяльність	57,03	29,09	**74,85**	48,50
Промисловість	103,03	60,19	**284,30**	162,23

Джерело: розраховано авторами.

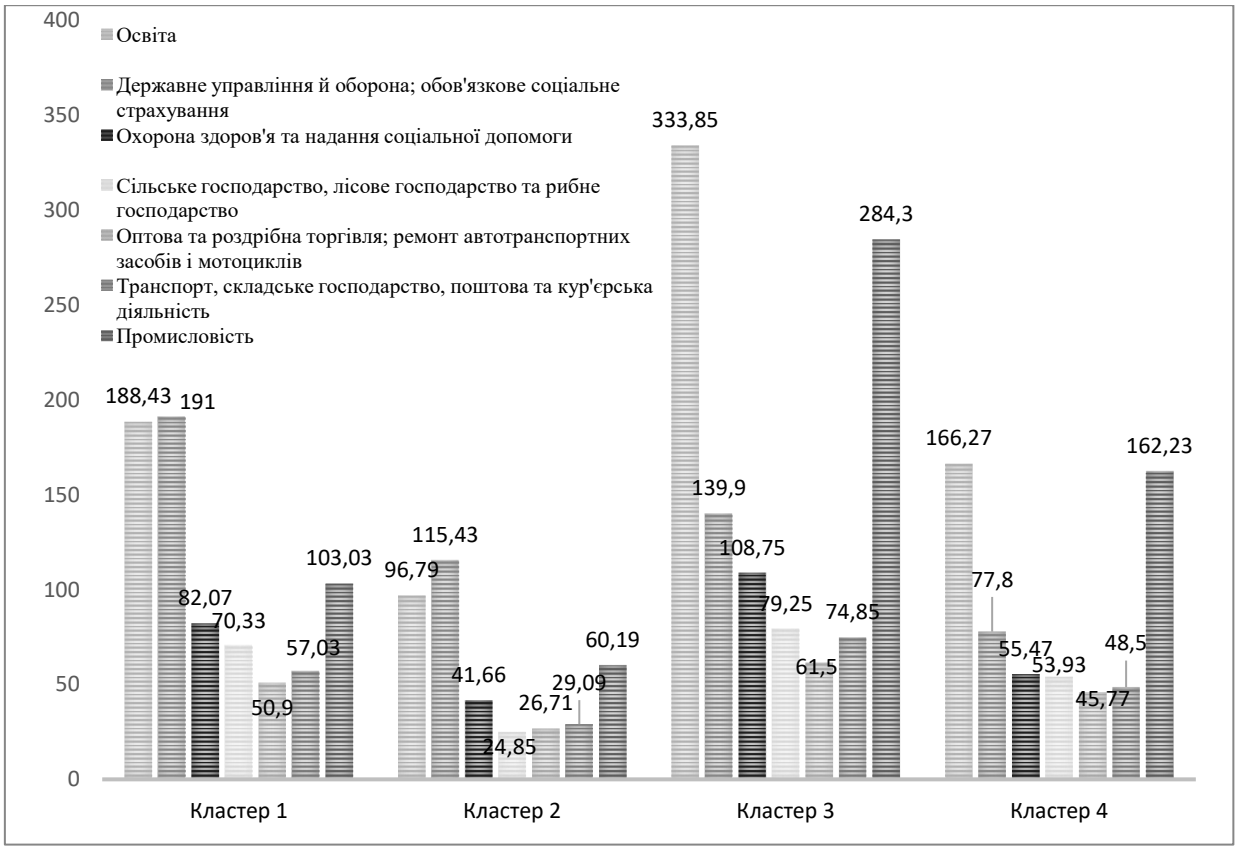


Рис. 5. Діаграма центрів кластерів зайнятого населення за видами діяльності

Джерело: розраховано авторами.

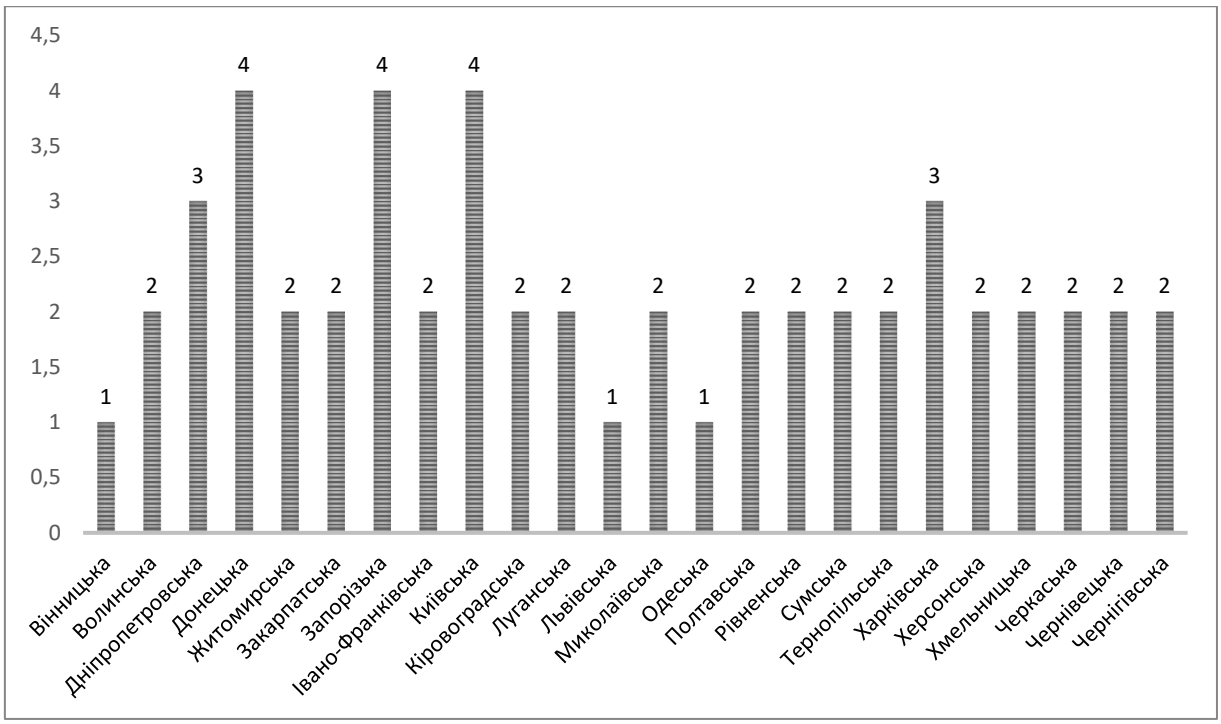


Рис. 6. Діаграма кластера за К-середнім

Джерело: розраховано авторами.

Варто зазначити, що до 2 кластера також ввійшли більшість областей західної частини України, це можна пояснити тим, що на сьогодні більшість жителів цього регіону відправляються на заробітки за кордон до країн-сусідів, зокрема до Польщі, Румунії та інших, що в загальному призводить до зниження рівня зайнятості населення.

Наступні два кластери не вирізняються за кількістю регіонів (табл. 6).

Таблиця 6

Кількість спостережень у кожному кластері

Кластер	К-ть	Регіон
1 «центр соціально-економічного розвитку»	3	Вінницька, Львівська, Одеська
2 «аутсайдери соціально-економічного розвитку»	16	Усі інші
3 «лідери соціально-економічного розвитку»	2	Дніпропетровська, Харківська
4 «периферія соціально-економічного розвитку»	3	Донецька, Запорізька, Київська
Валідні	24	

Джерело: розраховано авторами.

Перший кластер має наближені до максимального значення показники розподілу, а за державним управлінням й обороною взагалі максимальне. У перший кластер увійшли Вінницька, Львівська і Одеська області. Вони мають потужний галузевий аспект.

Четвертий кластер має характеристики, які наближаються до мінімальних (крім промислового виду діяльності), це Донецька, Запорізька і Київська області.

Висновок: результати дослідження показали, що майже четверта частина всього населення займається торгівлею та ремонтом автотранспорту (22 %), на другому місці – зайнятість у сільському господарстві (18 %), третю позицію займає промисловий сектор економіки (15 %). Фактично можна зазначити, що актуальною є підтримка держави для розвитку промислового і сільськогосподарського секторів економіки, які в майбутньому забезпечать гідне життя населенню країни. Першість торгівлі характерна для малорозвинених країн. У подальшому доречно застосувати цю методику для аналізу обсягів виробництва за видами економічної діяльності. Проведення порівняльного кластерного аналізу за різними періодами дозволить оцінити в динаміці вплив фінансової (2008 р.) і політичної (2014 р.) криз в Україні тощо.

Результати кластерного аналізу трудового потенціалу населення показали, що лідерами соціально-економічного розвитку є Дніпропетровська та Харківська області, при цьому в першій найпростіше можна знайти достойну роботу, а останній притаманне дороге проживання. Кластер «центр соціально-економічного розвитку» включає три регіони: Вінницький, Львівський і Одеський, при цьому в першому присутній найбільший рівень безробіття, а в двох останніх – найбільший індекс споживчих цін. У третій кластер «периферія соціально-економічного розвитку» потрапили Донецька, Запорізька і Київська області, для яких характерною ознакою також є високий рівень інфляції. Фактично це промислові області, в яких не використані власні можливості. Донецька область має цей результат через події на Сході країни. Для Київської області такі низькі показники отримані через те, що більшість населення працює в м. Київ, через територіальну близькість, де

присутні широкі можливості щодо працевлаштування та рівня доходів. Аутсайдерами за розподілом трудових ресурсів є більша половина регіонів України (16), при цьому в трьох з них: Івано-Франківській, Полтавській і Хмельницькій – найбільший рівень безробіття. Якщо ситуація в країні найближчим часом не зміниться, прискориться відтік працездатного населення за кордон. Введення безвізового режиму спростило бюрократичні перепони щодо працевлаштування в країнах ближнього зарубіжжя, де за аналогічну працю плата на порядок вища ніж в Україні. Отримані результати мають прикладний аспект і можуть виступати основою для розроблення стратегії соціально-економічного розвитку як окремого регіону, так і країни в цілому.

Приклад 2. Оцінити доходи населення України за допомогою кластерного аналізу та вплив політичної кризи на них.

Політична та економічна нестабільність в Україні збільшує варіаційний розмах між багатими і бідними верствами населення. Відсутність середнього класу призводить до соціальних вибухів, які періодично руйнують економічне середовище країни, це революція 2004 року, воєнні дії на Сході 2014 року, які призвели до девальвації гривні. І як наслідок, знецінюються заощадження і доходи населення, зменшується їхня купівельна спроможність. З погіршенням формування достатнього рівня доходу в населення України виникають думки про міграцію, що спричиняє втрату трудового потенціалу і негативно впливає на економіку країни загалом.

Спочатку проаналізуємо динаміку середньорічної заробітної плати і доходів населення за 2002–2018 рр. Доходи населення доцільно розглядати без інфляційної складової, щоб оцінити реальну їх динаміку. При цьому порівняємо номінальні і реальні дані, знайдемо їх за допомогою індексу споживчих цін. Для розрахунку реальних доходів доцільно розрахувати кумулятивний коефіцієнт у цінах базового 2001 року [8]:

$$R_i = \prod_{j \le i} \left(1 + r_j / 100\right), \qquad (1)$$

де r_j – щорічні темпи інфляції у %, П – визначає добуток виразу.

Результати кумулятивного індексу споживчих цін і реальних показників доходів подано в таблиці 7.

Таблиця 7

Вихідні дані для статистичного аналізу

Показники	Середньомісячна заробітна плата, грн	Доходи населення України, млн грн	Індекси споживчих цін*	Кумулятивний індекс	Реальна середньомісячна заробітна плата, грн	Реальні доходи населення України, млн грн
2002	376	185073	99,4	0,99	378,27	1861,90
2003	462	215672	108,2	1,08	429,56	1993,27
2004	590	274241	112,3	1,21	488,49	2442,04
2005	806	381404	110,3	1,33	605,01	3457,88
2006	1041	472061	111,6	1,49	700,19	4229,94
2007	1351	623289	116,6	1,73	779,33	5345,53
2008	1806	845641	122,3	2,12	851,84	6914,48
2009	1906	894286	112,3	2,38	800,54	7963,37
2010	2239	1101175	109,1	2,60	861,97	10093,26
2011	2633	1266753	104,6	2,72	969,07	12110,45
2012	3025	1457864	99,8	2,71	1115,58	14607,86
2013	3265	1548733	100,5	2,73	1198,10	15410,28
2014	3480	1516768	124,91	3,40	1022,33	12142,89
2015	4195	1772016	143,32	4,88	859,88	12364,05
2016	5183	2051331	112,42	5,48	945,02	18247,03
2017	7104	2652082	113,72	6,24	1139,01	23321,16
2018	8865	3219518	109,82	6,85	1294,26	29316,32

*(грудень до грудня попереднього року).

Джерело: розраховано автором за даними[24].

Аналізуючи динаміку номінальної середньомісячної заробітної плати за сімнадцять років, необхідно зазначити, що вона зростала на 435,23 грн щорічно, при цьому реальний показник вказує середній абсолютний приріст у дев'ять разів менше, а саме щороку приріст становив лише 47,91 грн. Коефіцієнти апроксимації наближаються до одиниці, що вказує на можливість застосування лінійної функції для аналізу (рис. 1). 2013 року реальний рівень зарплати мав більшу купівельну спроможність, ніж 2018 року, коли їх розмір був на 100 грн більший.

[24]Офіційний сайт Державної служби статистики України. URL : http://www.ukrstat.gov.ua/operativ/gdn/dvn/arh_dvn2001.html

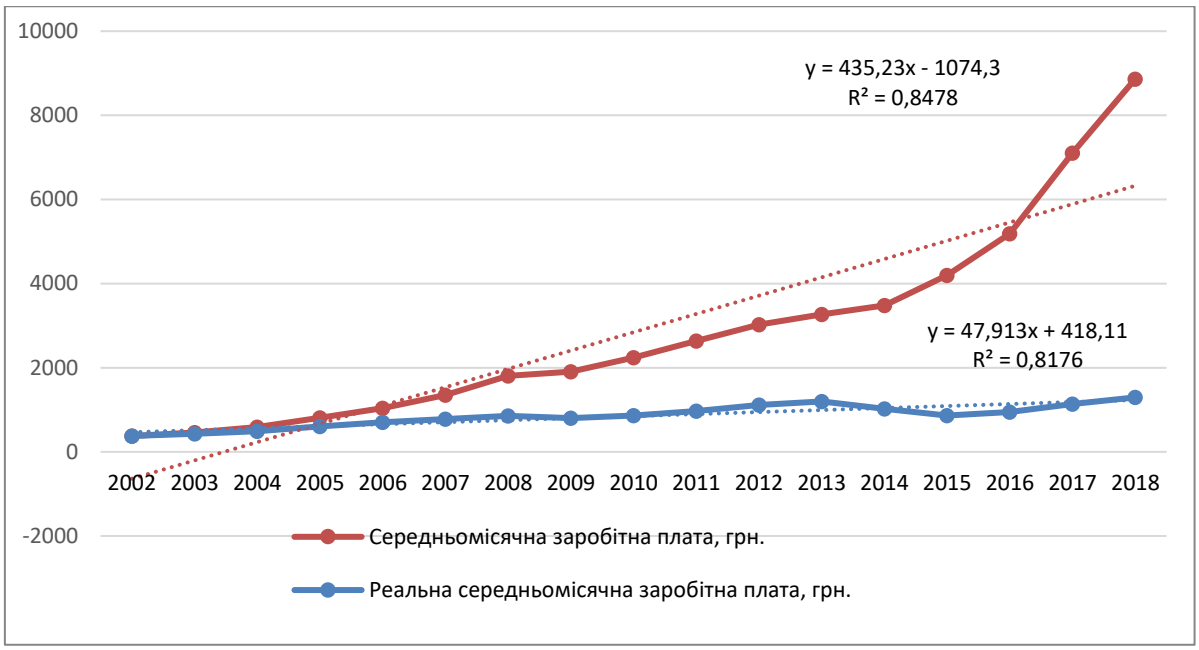


Рис. 7. Динаміка середньомісячної заробітної плати на 2002–2018 рр., грн

Джерело: розраховано автором за даними таблиці 7.

Аналізуючи номінальний дохід населення України за цей період, необхідно зазначити суттєве зростання динаміки з 2014 року, яке відбулося за рахунок девальвації гривні в країні. Номінальні доходи за 2002–2018 рр. щорічно зростали на 165,437 млрд грн (коефіцієнт апроксимації 0,92), при цьому реальний приріст становив лише 16,218 млрд грн (коефіцієнт апроксимації 0,56) (рис. 8).

Для кластерного аналізу взяті показники: зайнятості населення (зайняте населення, тис. осіб; рівень зареєстрованого безробіття на кінець року у відсотках до населення працездатного віку), доходів (середньомісячна заробітна плата, грн; наявний дохід у розрахунку на одну особу, грн; доходи населення, млн грн; витрати на персонал підприємств, млн грн), кількість населення (постійне населення, середня чисельність, тис. осіб), індекси споживчих цін і обсяг реалізованої продукції (товарів, послуг) суб'єктів господарювання, млн грн, як показник витрачених наявних коштів населення. Для розрахунку обрано метод кластеризації К-середнім і побудовано чотири кластери.

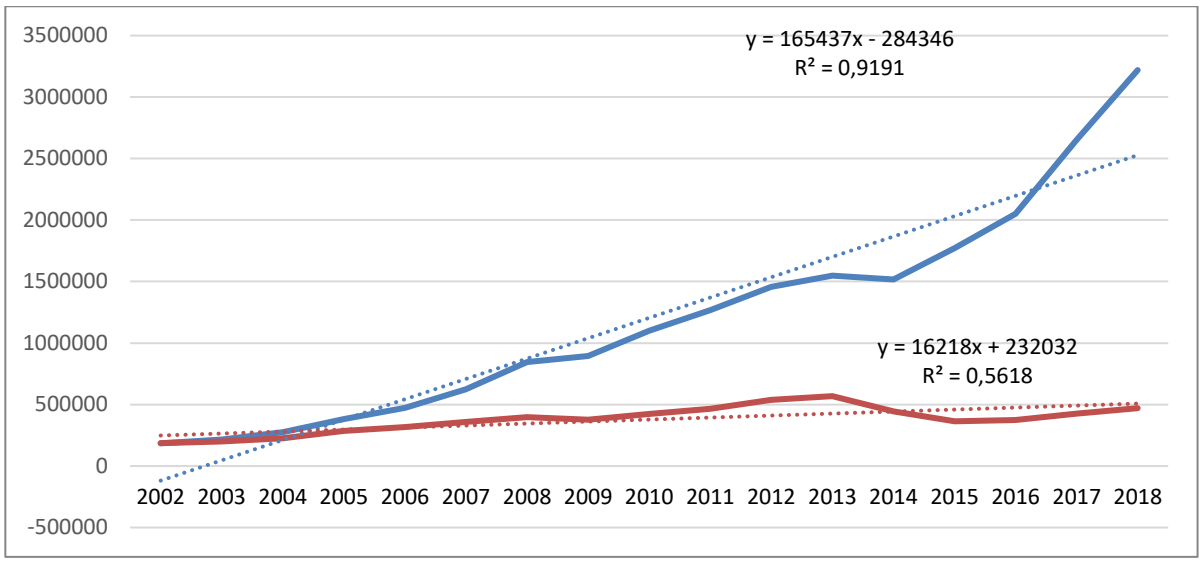


Рис. 8. Динаміка доходів населення України за 2002–2018 рр.
Джерело: розраховано автором за даними таблиці 7.

У таблиці 8 наведений результат розподілу регіонів за доходами населення і відстань між ними.

Таблиця 8

Належність до кластера

Спосте-реження	Регіон	2013 р.		2018 р.	
		Кластер	Відстань	Кластер	Відстань
1	Вінницька	1	19425,460	2	67437,357
2	Волинська	1	23528,716	1	82822,997
3	Дніпропетровська	3	,000	4	,000
4	Донецька	4	,000	3	31083,832
5	Житомирська	1	15066,495	1	25586,737
6	Закарпатська	1	19250,555	1	28133,907
7	Запорізька	2	28908,469	2	91371,399
8	Івано-Франківська	1	14495,474	1	18256,722
9	Київська	2	90591,709	3	84874,211
10	Кіровоградська	1	15289,937	1	8423,946
11	Луганська	2	37411,591	1	67434,900
12	Львівська	2	12559,385	3	54375,101
13	Миколаївська	1	18665,713	1	61148,975
14	Одеська	2	10311,722	3	32380,756
15	Полтавська	2	38015,920	2	53190,022
16	Рівненська	1	17858,907	1	23955,796
17	Сумська	1	11994,838	1	18139,121

18	Тернопільська	1	19954,838	1	24636,175
19	Харківська	2	30551,889	3	36475,740
20	Херсонська	1	183168,884	1	21343,717
21	Хмельницька	1	13134,061	1	25161,056
22	Черкаська	1	24990,942	2	77239,814
23	Чернівецька	1	32286,774	1	65350,717
24	Чернігівська	1	12521,805	1	10930,775

Джерело: розраховано автором за даними таблиці 7

Візуально сприймати результати простіше, тому побудовано діаграму належності до кластеру регіонів України (рис. 9).

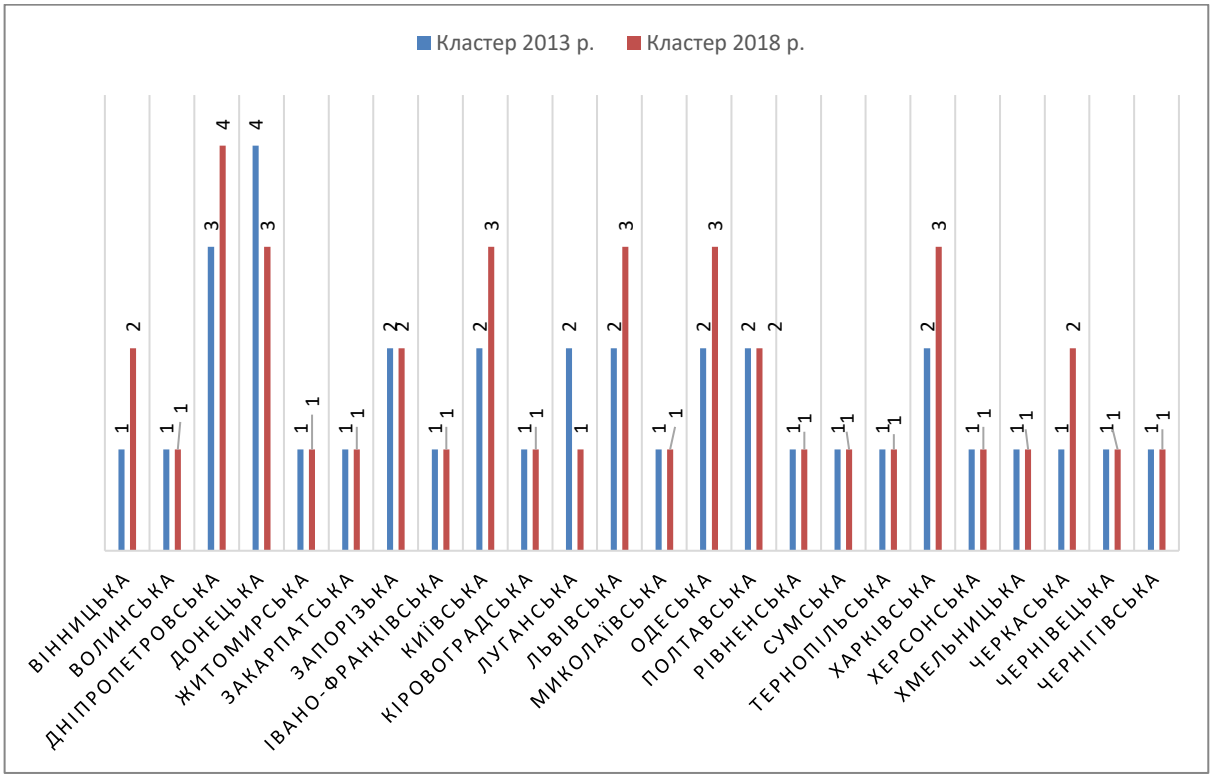


Рис. 9. Діаграма належності до кластеру 2013 і 2018 рр.

Джерело: розраховано автором за даними таблиці 7.

Аналізуючи кластери, необхідно зазначити, що першість займає четвертий кластер, це – лідер, його можна назвати «краще місце працевлаштування». У нього потрапив лише один регіон – Донецька область, яка має найбільшу кількість зайнятого населення, яке отримує максимальну середньомісячну зарплату. При цьому індекс споживчих цін найбільший серед чотирьох кластерів. У регіоні мінімальний рівень безробіття і максимальні витрати

підприємств на персонал. За рахунок максимальної кількості постійного населення наявний дохід у розрахунку на одну особу займає другу позицію з чотирьох (табл. 9).

Таблиця 9

Кінцеві центри кластерів 2013 року

Параметри	Кластер			
	1	2	3	4
Зайняте населення, тис. осіб	516,71	941,17	1508,00	19 36,40
Постійне населення, середня чисельність, тис. осіб	1 176,59	2 115,30	3289,10	4 331,00
Рівень зареєстрованого безробіття на кінець року у % до населення працездатного віку	2,31	1,73	1,60	1,20
Середньомісячна заробітна плата, грн	2 627,07	3 075,57	3 336,00	37 55,00
Наявний дохід у розрахунку на одну особу, грн	34 487,72	25 935,49	30 300,60	31 048,50
Індекси споживчих цін	99,37	99,26	99,70	100,40
Доходи населення, млн грн	32 596,80	69 344,86	12 4594,00	16 6366,00
Обсяг реалізованої продукції (товарів, послуг) суб'єктів господарювання, млн грн	39 163,43	141 458,83	444 4549,70	57 5698,90
Витрати на персонал підприємств, млн грн	3 777,51	16 150,96	39 231,40	54 468,80

Джерело: розраховано автором за даними таблиці 7.

Третій кластер займає друге місце. Для нього характерні такі ознаки: максимальний обсяг реалізації продукції суб'єктів господарювання за відносно високим рівнем інфляційних процесів. Індекс споживчих цін займає другу позицію після максимального. За всіма іншими показниками він на рівень нижчий від четвертого кластеру, але має гарні характеристики, а саме: високий рівень зайнятого населення, заробітної плати, доходів населення, низький рівень безробіття тощо. За всіма характеристиками його можна назвати «середній рівень доходів». До нього потрапив один регіон, а саме Дніпропетровський, який є потужним промисловим центром України.

Третю позицію займає другий кластер з назвою «найнижчі ціни», який включає Запорізьку, Київську, Луганську, Львівську, Одеську, Полтавську та Харківську області (табл. 4). Для нього характерні такі ознаки: мінімальний дохід у розрахунку на одну

особу та індекс споживчих цін, відносно високий рівень безробіття (друга позиція після максимального). Усі інші показники наближаються до мінімальних значень.

У перший кластер із назвою «найбільше безробіття» ввійшла більша частина областей України, а саме: Вінницька, Волинська, Житомирська, Закарпатська, Івано-Франківська, Кіровоградська, Миколаївська, Рівненська, Сумська, Тернопільська, Херсонська, Хмельницька, Черкаська, Чернівецька та Чернігівська (табл. 10). Варто зазначити, що цей кластер характеризується мінімальними значеннями майже за всіма критеріями, звертає на себе увагу максимальний рівень безробіття в цих регіонах, при цьому наявний дохід у розрахунку на одну особу має максимальне значення (табл. 9). Більшість у кластері займають регіони Західної і Північної України, які внаслідок географічного положення сприяють до працевлаштування в ближньому зарубіжжі.

Таблиця 10

Кількість спостережень у кожному кластері 2013 р.

Кластер	К-сть	Регіони
4 «краще місце працевлаштування»	1	Донецька
3 «середній рівень життя»	1	Дніпропетровська
2 «найнижчі ціни»	7	Запорізька, Київська, Луганська, Львівська, Одеська, Полтавська та Харківська
1 «найбільше безробіття»	15	Вінницька, Волинська, Житомирська, Закарпатська, Івано-Франківська, Кіровоградська, Миколаївська, Рівненська, Сумська, Тернопільська, Херсонська, Хмельницька, Черкаська, Чернівецька та Чернігівська

Джерело: розраховано автором за даними таблиці 7.

Після політичної кризи, яка спричинила до перерозподіл території України, зруйнувала стабільність економіки, призвела до девальвації гривні і стрімкого зростання інфляційних процесів, ситуація змінилася. Так, другий кластер з «найнижчих цін» перейшов у категорію «мінімальне безробіття». Донецький регіон, втративши позиції

лідера, опинився у другому кластері «середній рівень життя». Лідером 2018 року став другий за потужністю промисловий регіон – Дніпропетровський. Друге місце займають п’ять регіонів: Донецький, Київський, Львівський, Одеський та Харківський, останні чотири 2013 року були в кластері «найнижчі ціни» (табл. 11, 12).

Третю позицію займає кластер «мінімальне безробіття» з чотирма регіонами, серед яких Запорізька, Полтавська, Черкаська та Вінницька області.

Аутсайдерами є більшість регіонів України 2013 року, їх було п’ятнадцять 2018 року. Вінницька і Черкаська області покращили свої результати, при цьому Луганська область у результаті воєнних дій втратила свої позиції і поповнила цей кластер (табл. 12).

Таблиця 11

Кінцеві центри кластерів 2018 року

Параметри	**Кластер**			
	4	**3**	**2**	**1**
Зайняте населення, тис. осіб	1 402,30	963,74	622,03	455,30
Постійне населення, середня чисельність, тис. осіб	3 215,50	2 694,57	1 470,94	1 187,62
Рівень зареєстрованого безробіття на кінець року у % до населення працездатного віку	2,00	2,00	1,67	2,27
Середньомісячна заробітна плата, грн	8 862,00	8 490,40	8 095,00	7 370,36
Наявний дохід у розрахунку на одну особу, грн	72 883,40	54 436,08	58 371,15	47 010,69
Індекси споживчих цін	109,20	110,58	109,30	109,55
Доходи населення, млн грн	305 510,00	18 2096,00	114 224,50	71 268,36
Обсяг реалізованої продукції (товарів, послуг) суб’єктів господарювання, млн грн	111 5583,00	459 433,14	267 344,60	108 082,67
Витрати на персонал підприємств, млн грн	76 617,00	34 947,72	21 822,90	9 075,15

Джерело: розраховано автором за даними таблиці 7.

Таблиця 12

Кількість спостережень у кожному кластері

Кластер	К-сть	Регіони
4 «краще місце працевлаштування»	1	Дніпропетровська
3 «середній рівень життя»	5	Донецька, Київська, Львівська, Одеська та Харківська
2 «мінімальне безробіття»	4	Вінницька, Запорізька, Полтавська та Черкаська
1 «найбільше безробіття»	14	Волинська, Житомирська, Закарпатська, Івано-Франківська, Кіровоградська, Луганська, Миколаївська, Рівненська, Сумська, Тернопільська, Херсонська, Хмельницька, Чернівецька та Чернігівська

Джерело: розраховано автором за даними таблиці 7.

Висновки. Аналізуючи будь-які доходи, необхідно враховувати їх дійсну купівельну спроможність. Використаний кумулятивний індекс споживчих цін дозволяє виміряти вплив інфляційних процесів у цінах базового року, за базовий взято 2001 рік. Аналіз показав, що приріст показників завищений майже в десять разів за середнім розміром зарплати і доходами населення України.

Кластерний аналіз визначив, що більшість регіонів України мають низький рівень доходів і високий рівень безробіття. Порівняльний аналіз до і в період воєнних дій на Сході країни вказує, що ситуація в Україні суттєво не змінилася. У країні є лише один лідер, п’ять регіонів можуть забезпечити середній рівень життя, при цьому їхні доходи знецінюються за рахунок максимального рівня індексу споживчих цін. Для чотирьох областей характерний низький рівень безробіття, невисокий середній рівень зарплати, при цьому дохід у розрахунку на одну особу наближається до максимального значення. За цих умов бездіяльність державного управління, відкритість кордонів (за рахунок безвізового режиму), значно вищий рівень доходів у ближніх західних сусідів призведуть до суттєвої втрати працездатного населення, яке, оцінивши переваги іншого життя, не будуть повертатися в Україну. У подальшому доцільно проаналізувати міграційні процеси в Україні, визначити структурні зрушення і спрогнозувати наслідки бездіяльності з боку держави.

Список використаних джерел

1. Аналітика та прогнозування соціально-економічних процесів і податкових надходжень : монографія / Паянок Т. М., Лаговський В. В., Краєвський В. М. та ін. – К. : ЦП «Компринт», 2019. – 426 с.

2. Бююль А. SPSS: Искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей / А. Бююль, П.Цёфель ; пер. с нем. – СПб. : ДиаСофтЮП, 2005. – 608 с. Download (PDF / 40.47 Mb)

3. Крыштановский А. О. Анализ социологических данных с помощью пакета SPSS : учеб. пособие для вузов / А. О. Крыштановский ; ГУ-ВШЭ. – М. : Изд. дом ГУ-ВШЭ, 2006. – 281 с.

4. Наследов А. IBM SPSS Statistics 20 и AMOS : профессиональный статистический анализ данных / А. Наследов. – СПб. : Питер, 2013. – 416 с.

5. Сотніков Ю. М. Маркетингові дослідження з використанням пакета SPSS : навчальний посібник / Ю. М. Сотніков. – Одеса : Атлант, 2016. – 145 с.

6. Таганов Д. Н. Статистический анализ в маркетинговых исследованиях / Д. Н.Таганов. – СПб. : Питер, 2005. – 192 с.

7. Andry Field. Discovering Statistics using SPSS [Електронний ресурс]. – Режим доступу : https://in.sagepub.com/.../-discovering-statistics-using...spss-statastics/

8. Bentler P. M. EQS 6 Structural Equations Program Manual. Encino, CA: Multivariate Software, Inc., 2006. – 418 p.

9. Byrne B. M. Structural Equation Modeling With AMOS: Basic Concepts, Applications, and Programming. 2nd ed. Multivariate applications series. New York : Taylor & Francis Group, 2010. – 396 p.

10. Harrington D. Confirmatory factof analysis. – New York : Oxford University Press, Inc., 2009. – 122 p.

11. Joseph F. Healey. Statistics: A tool for Social Research [Електронний ресурс]. – Режим доступу : https://www.amazon.-com/Statistics-Research-Joseph-F-Healey

12. Kline R. B. Principles and practice of structural equation modeling. 3nd ed. – New York : The Guilford Press, 2011. – 432 p.

13. [Електронний ресурс]. – Режим доступу : http://www.spss.ru/
14. [Електронний ресурс]. – Режим доступу : http://www.spss.com.ua/

ГЛОСАРІЙ

Автокореляція – це залежність випадкового члена ряду спостереження від минулих значень.

Асиметрія – відносне відхилення, яке характеризує напрям і міру скошеності всередині розподілу, тобто в який бік відносно середнього зміщена більшість значень розподілу.

Вагами спостережень називають частоти і частості.

Варіантами називають різні значення властивості випадкової величини X, позначають їх через x.

Варіаційний ряд – це сукупність значень властивості, записаних у порядку зростання.

Варіаційний ряд називається дискретним, якщо будь-які його варіанти відрізняються на сталу величину, і неперервним (інтервальним), якщо варіанти можуть відрізнятись одна від одної на як завгодно малу величину.

Варіація – це мінливість (коливання) індивідуальних значень ознаки сукупності, тобто вимірювання ступеня коливання ознаки.

Верифікація моделі – співставлення реальних і розрахункових даних, перевірка адекватності моделі, оцінка точності та стійкості отриманих рівнянь зв'язку, побудова прогнозів та сценаріїв розвитку.

Вибірка (*n*) – частина генеральної сукупності, випадково відібрана для дослідження з метою отримання висновків про властивості генеральної сукупності.

Відбір методом «снежного кома» – опитування проводиться шляхом передавання анкети в межах групи (наркомани, хворі на туберкульоз та ін.).

Відносними частотами або частостями називаються відношення частот до загальної кількості спостережень.

Відстань Кука визначає силу впливу викиду на рівнях (N від 0 до 1), якщо немає впливових викидів, модель гарна. Розраховані значення відстані Кука коливаються від 0 (min) до 0,512 (max), що вказує на відсутність впливових викидів.

Генеральна сукупність (*N*) – уся сукупність об'єктів, яка цікавить дослідника.

Гетероскедастичність – це змінність дисперсії вільного члена, виникає щодо перехресної вибірки, рідше у часовому ряді, тобто у системі наявні різні дисперсії помилок.

Гістограмою частот називають ступінчату фігуру, яка складається з прямокутників, основами яких є часткові інтервали варіант довжиною $h = x_k - x_{k-1}$, а висоти дорівнюють (n_k / h) частотам або частостям $n_r(W_i)$ інтервалів. Гістограма – це стовпчикова діаграма частот, а не даних.

Дані – факти і числа, за допомогою яких можуть прийматися рішення.

Довірчий інтервал – це розрахований на основі вибірки інтервал значень ознак, який з відомою ймовірністю містить оціночний параметр генеральної сукупності.

Екзогенні змінні (факторні змінні) **–** причини та умови, які необхідні для виникнення певного наслідку (х), екзогенні зміни визначають ендогенні, але самі не перебувають під їх впливом.

Ексцес – відображає ступінь зосередженості елементів сукупності навколо центра розподілу.

Емпірична вибірка (проста, квотована) – передбачає, що в коло респондентів для відбору інформації включається «перший зустрічний».

Емпіричною функцією розподілу (функцією розподілу вибірки) називають функцію F^*(х), яка визначає для кожного значення X відносну частоту того, що властивість (випадкова величина Х) прийме значення, менше заданого х.

Ендогенна змінна (результативна змінна) **–** ознака, яка характеризує наслідок дії фактора або факторів, залежна змінна (Y).

Збурення або латентні змінні – це економічні величини, які не входять у рівняння регресійної моделі, але впливають на спільно залежні змінні. Вони формуються завдяки випадковим впливам, помилкам і припущенням.

Змінні – будь-яка характеристика об’єкта.

Кластерний аналіз – алгоритм розбиття заданої вибірки об’єктів (ситуацій) на підмножини, які називаються кластерами

так, щоб кожен кластер складався з подібних об'єктів, а об'єкти різних кластерів істотно відрізнялися.

Кластерна вибірка передбачає вибірку, в якій сукупність на першому етапі розподіляється на підгрупи, які не перетинаються між собою (кластери), а потім із цих підгруп формується випадкова вибірка.

Коефіцієнт детермінації (R^2) визначає міру впливу факторної ознаки на результативну, полягає в межах від 0 до 1. Якщо R^2 $R^2 = 0$, то це означає, що залежність між Y та X відсутня, тобто X не впливає на Y.

Коефіцієнт осциляції – характеризує відносну варіацію крайніх значень ознаки навколо середньої.

Коефіцієнти варіації: лінійний – характеризує співвідношення середнього лінійного відхилення ознаки та її середньої величини; **квадратичний** – характеризує співвідношення середнього квадратичного відхилення ознаки та її середньої величини.

Кореляційний аналіз – статистичний метод, який дозволяє визначити, чи є залежність між випадковими змінними, який вона має напрямок і наскільки вона сильна.

Кореляція (r) – це статистична залежність між випадковими величинами, що носить ймовірнісний характер, полягає в межах від -–1 до +1.

Критерієм згоди називають статистичний критерій перевірки гіпотези про можливий закон невідомого розподілу.

Кумулятивною кривою називається крива накопичених частот (частостей).

Медіана – значення ознаки, яке міститься посередині впорядкованого щодо зростання варіаційного ряду.

Мода – значення ознаки, що найбільш часто зустрічається у ряді розподілу; найпоширеніше значення ознаки, тобто варіанта, яка в ряду розподілу має найбільшу частоту.

Мультиколінеарність – це лінійна залежність між незалежними змінними (факторами), вона не дає можливості визначити вплив кожного з них на результативну ознаку.

Набір даних – дані, що відібрані для конкретного аналізу.

Накопиченою частістю $W_i^{нак}$ називають відношення накопиченої частоти $n_i^{нак}$ до загальної кількості спостережень n.

Наперед визначена змінна (лагова) – змінна, значення якої відстає на один або декілька періодів. Якщо $x_{t,k}$ – значення звичайної змінної x_k, зафіксовані в цей момент часу t, то $x_{t-1,k}$ – її лагові значення, зміщенні на один період.

Негативна лінійна залежність (негативна кореляція або зворотній зв'язок) – під час зростання значень однієї змінної зменшуються значення іншої, розраховане значення перебуває в межах $-1 < r < 0$.

Нормальний розподіл називають стандартним (або нормованим), якщо математичне сподівання $M(X) = 0$, дисперсія $D(X) = 1$. Нормальний розподіл **називають загальним**, якщо математичне сподівання M(X) і середнє квадратичне відхилення довільні.

Нульова кореляція – це відсутність зв'язку між змінними. При цьому нульова загальна кореляція може свідчити про відсутність лише лінійної залежності, а не про відсутність будь-якого статистичного зв'язку.

Основна (або нульова) – називається гіпотеза, яка позначається H_0, логічне заперечення основної гіпотези називається **альтернативною** та позначається H_1.

Параметри – це числові характеристики генеральної сукупності.

Позитивна лінійна залежність ($0 < r < 1$) – зростання рівня значень однієї змінної супроводжується підвищенням значення іншої змінної.

Полігоном відносних частот називають ламану, відрізки якої з'єднують точки (x_1, W_1), (x_2, W_2) …, (x_m, W_m).

Полігоном частот називають ламану, відрізки якої з'єднують точки (x_1, n_1), (x_2, n_2), ..., (x_m, n_m).

Помилка другого роду полягає у відхиленні альтернативної гіпотези, що дійсно правильна. Ймовірність помилки другого роду позначається β. Відповідно до означення $\beta = P(H_0/H_1)$. Величина $1-\beta$ називається потужністю критерію.

Помилка першого роду полягає в тому, що відхиляється нульова гіпотеза, яка насправді є правильною. Ймовірність цієї помилки позначається α та називається рівнем значущості критерію. За означенням $\alpha = P(H_1|H_0)$. Цю ймовірність задають, як правило, перед проведенням тесту, числове значення є стандартним α=0,01;0,05;0,001; 0,005.

Проста випадкова вибірка передбачає, що всі елементи генеральної сукупності мають рівні шанси потрапити до статистичної вибірки.

Ранжирування – це впорядкування статистичного матеріалу, розміщення варіант у порядку зростання (спадання).

Регресія – це вимірювання одностороннього стохастичного зв'язку між факторними і результативною ознаками.

Рівень значимості гіпотези називають ймовірність здійснити похибку першого роду, тобто відхилити правильну нульову гіпотезу (a = 10 % або 5 % або 1 %).

Розмах варіації (R) – це різниця між максимальним і мінімальним значенням ознаки.

Середнє значення – рівне сумі всіх значень розподілу, поділене на їх кількість.

Середній квадрат відхилень відносно регресії (*MSE*, $S^2 = \sigma_\varepsilon^2$) – дає оцінку залишкової дисперсії відносно регресії, яка базується на $n - k$ ступенів свободи.

Систематична вибірка отримується шляхом нумерації кожного члена генеральної сукупності і потім вибором k-ого номеру.

Скоригований коефіцієнт детермінації (або нормований) використовують для оцінки реальної тісноти зв'язків між результативною і факторною ознаками, для порівняння моделі з різною кількістю параметрів (X, при цьому ряд спостереження повинен бути однаковий).

Спільнозалежні змінні – це звичайні ендогенні змінні, які пояснюються регресійною моделлю в момент часу t. Між ними існують багатосторонні зв'язки і визначаються не одним рівнянням, а одночасними рівняннями моделі.

Стандартна похибка – показує, наскільки вибіркове середнє відрізняється від середньої генеральної сукупності, це показник репрезентативності.

Стандартна похибка ($S = \sigma_\varepsilon$) – це оцінка середнього квадратичного відхилення коефіцієнта регресії від його істинного значення.

Стандартна похибка параметрів моделі вказує, наскільки будуть варіювати коефіцієнти від вибірки до вибірки.

Статистика – числові характеристики вибірки.

Статистична гіпотеза (гіпотеза) – це певне твердження (припущення) відносно генеральної сукупності, що перевіряється на основі вибірки.

Статистичним критерієм перевірки гіпотези H_0 – називають правило, за яким ухвалюється рішення про прийняття або відхилення гіпотези H_0. Рішення ухвалюють на основі вибірки $X_1, X_2, \dots X_n$, за якою формують спеціальну функцію вибірки $T_n = T(X_1, X_2, \dots X_n)$, цю функцію називають **статистикою критерію**.

Статистичним розподілом вибірки називається перелік варіант, що спостерігаються, розміщених у порядку зростання, та відповідних їм частот, відносних частот.

Стратифікована вибірка передбачає, що статистична вибірка проводиться виключним чином окремо в кожній групі генеральної сукупності із збереженням пропорції співвідношення розмірів цих груп.

Чітка негативна кореляція дорівнює –1, коли прослідковується 100 % спільна варіація двох змінних, при цьому випадкові змінні повністю накладаються на пряму (яка прямує з нижнього правого кута – у верхній лівий), відсутня зона розсіювання.

Строга позитивна кореляція дорівнює +1, коли прослідковується 100 % спільна варіація двох змінних, при цьому напрямок функції з лівого нижнього кута – в правий верхній.

Точність оцінювання – ступінь допустимого відхилення величини, що оцінюється від істинної.

Фактичне значення статистики – це оцінка параметра генеральної сукупності.

Фіктивні змінні (дихотомічні) – це категорійні змінні, які мають якісні характеристики, ці змінні бінарного типу, тобто кожна змінна може приймати всього два значення – одиницю і нуль.

Частотами (позначаються n_i) – називаються числа, які показують, скільки разів зустрічаються варіанти з цього інтервалу.

ДОДАТКИ

Додаток А

Критичні точки розподілу Стьюдента (*t*-розподілу)

k \ α	0,1	0,05	0,02	0,01	0,001
1	6,3138	12,7062	31,8205	63,6567	636,6192
2	2,9200	4,3027	6,9646	9,9248	31,5991
3	2,3534	3,1824	4,5407	5,8409	12,924
4	2,1318	2,7764	3,7469	4,6041	8,6103
5	2,0150	2,5706	3,3649	4,0321	6,8688
6	1,9432	2,4469	3,1427	3,7074	5,9588
7	1,8946	2,3646	2,9980	3,4995	5,4079
8	1,8595	2,3060	2,8965	3,3554	5,0413
9	1,8331	2,2622	2,8214	3,2498	4,7809
10	1,8125	2,2281	2,7638	3,1693	4,5869
11	1,7959	2,2010	2,7181	3,1058	4,4370
12	1,7823	2,1788	2,6810	3,0545	4,3178
13	1,7709	2,1604	2,6503	3,0123	4,2208
14	1,7613	2,1448	2,6245	2,9768	4,1405
15	1,7531	2,1314	2,6025	2,9467	4,0728
16	1,7459	2,1199	2,5835	2,9208	4,0150
17	1,7396	2,1098	2,5669	2,8982	3,9651
18	1,7341	2,1009	2,5524	2,8784	3,9216
19	1,7291	2,0930	2,5395	2,8609	3,8834
20	1,7247	2,0860	2,5280	2,8453	3,8495
21	1,7207	2,0796	2,5176	2,8314	3,8193
22	1,7171	2,0739	2,5083	2,8188	3,7921
23	1,7139	2,0687	2,4999	2,8073	3,7676
24	1,7109	2,0639	2,4922	2,7969	3,7454
25	1,7081	2,0595	2,4851	2,7874	3,7251
26	1,7056	2,0555	2,4786	2,7787	3,7066
27	1,7033	2,0518	2,4727	2,7707	3,6896
28	1,7011	2,0484	2,4671	2,7633	3,6739

29	1,6991	2,0452	2,4620	2,7564	3,6594
30	1,6973	2,0423	2,4573	2,7500	3,6460
35	1,6896	2,0301	2,4377	2,7238	3,5911
40	1,6839	2,0211	2,4233	2,7045	3,5510
45	1,6794	2,0141	2,4121	2,6896	3,5203
50	1,6759	2,0086	2,4033	2,6778	3,4960
55	1,6730	2,004	2,3961	2,6682	3,4764
60	1,6706	2,0003	2,3901	2,6603	3,4602
70	1,6669	1,9944	2,3808	2,6479	3,4350
80	1,6641	1,9901	2,3739	2,6387	3,4163
90	1,6620	1,9867	2,3685	2,6316	3,4019
100	1,6602	1,9840	2,3642	2,6259	3,3905

Додаток Б

Приклад. Низька правостороння асиметрія

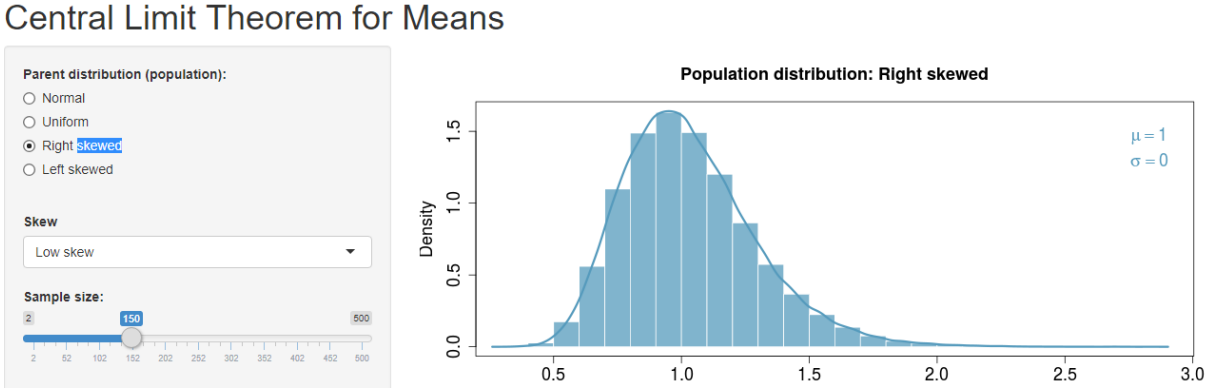
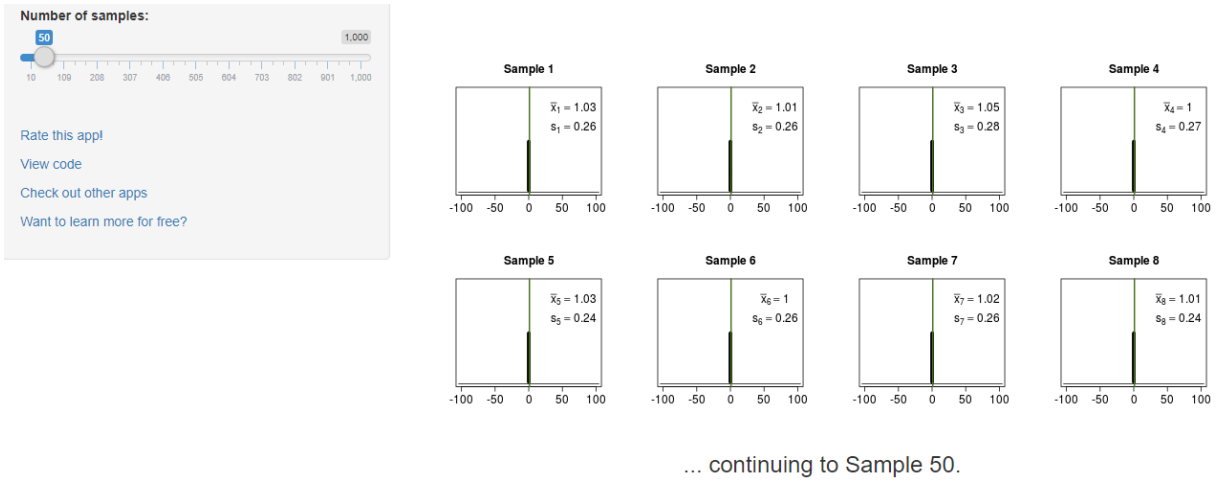
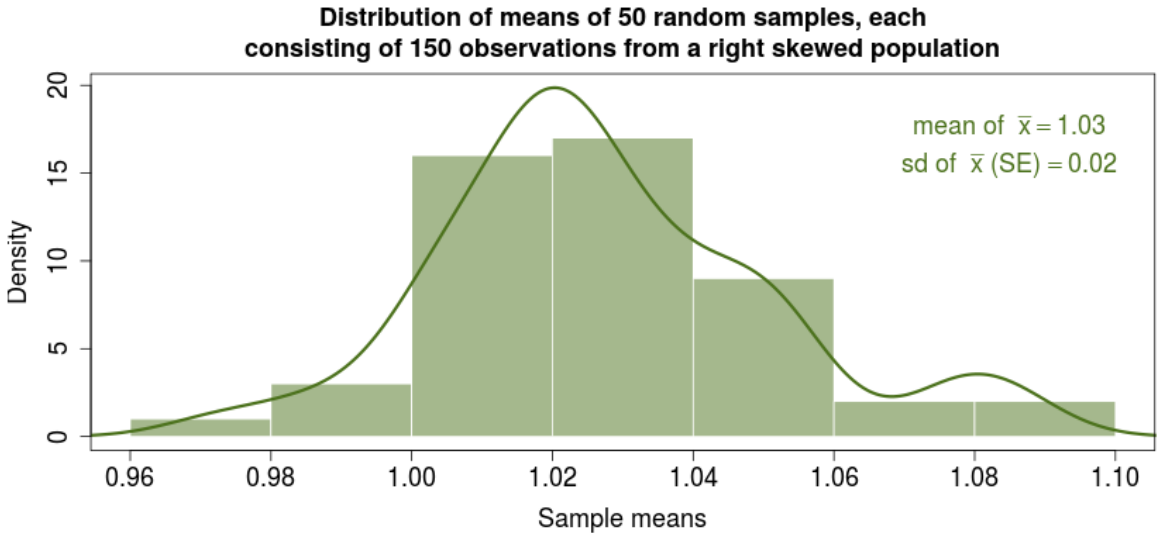


Рис. Низьке (low) правостороннє відхилення (skew) генеральної сукупності





Приклад. Середня правостороння асиметрія

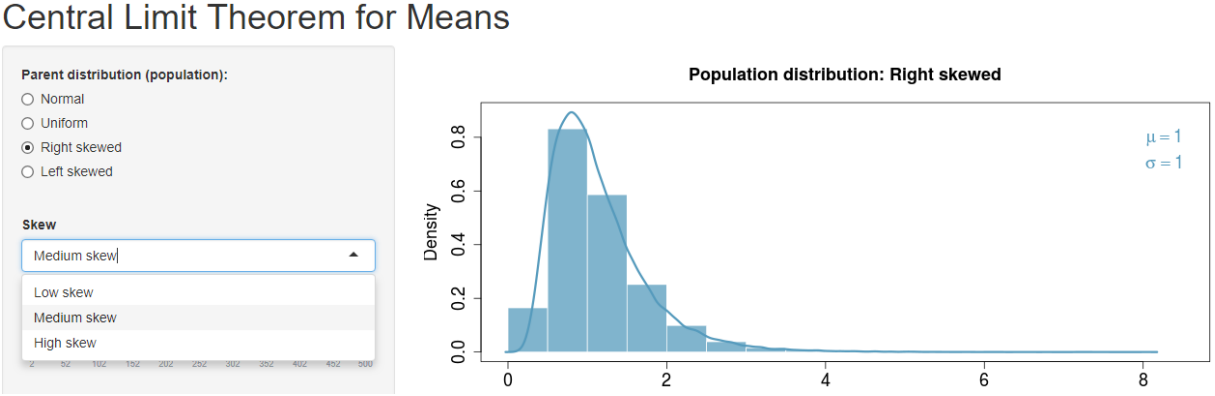
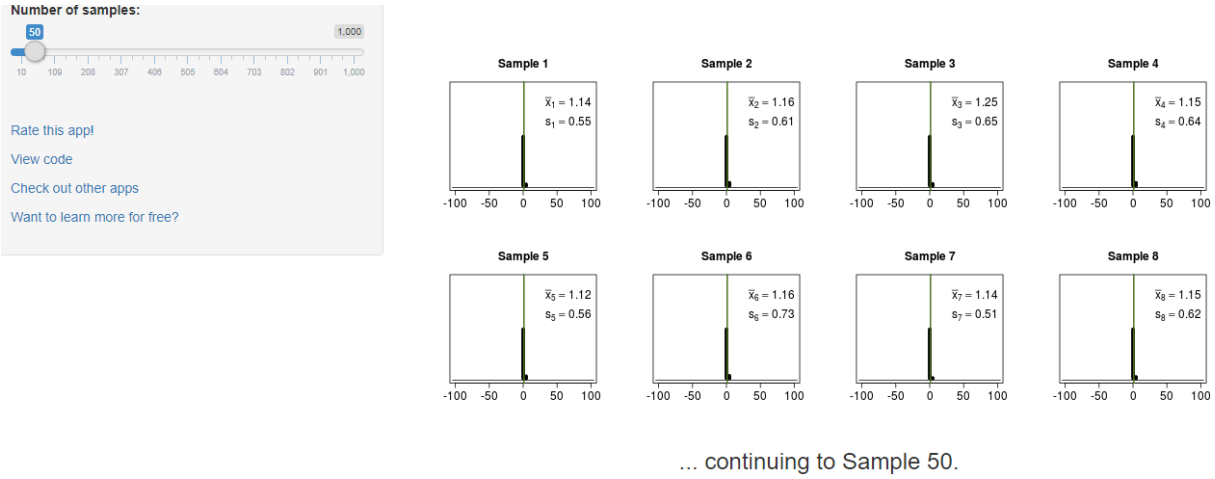
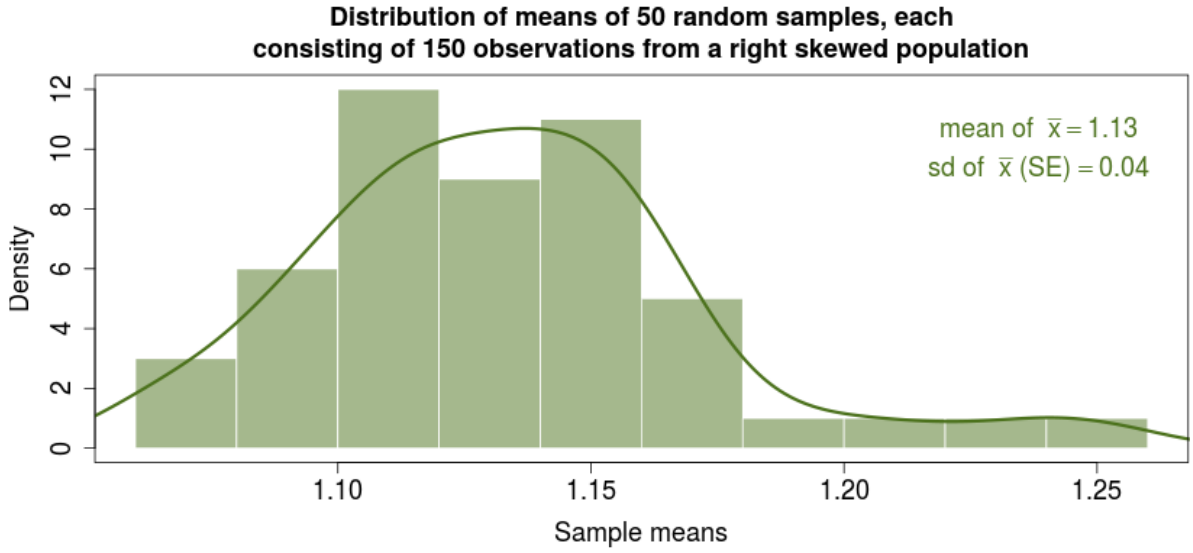


Рис. Середнє (medium) правостороннє відхилення (skew) генеральної сукупності





Приклад. Висока правостороння асиметрія

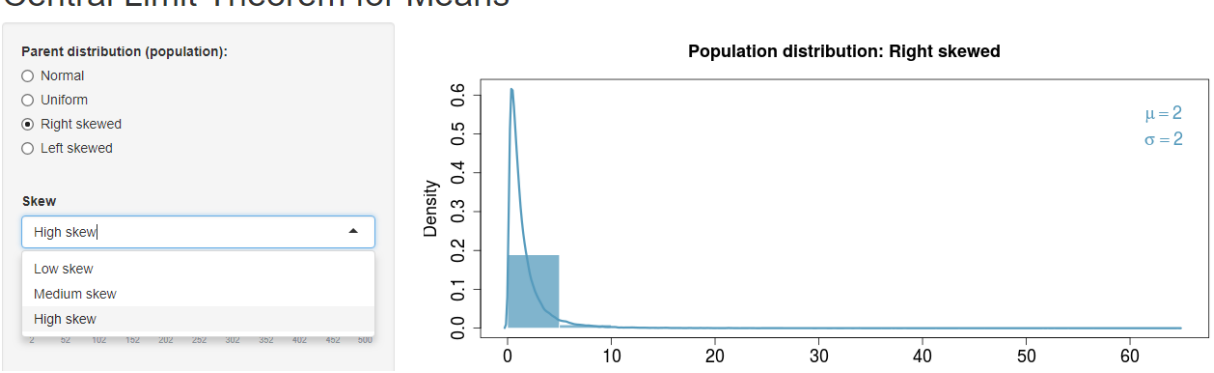
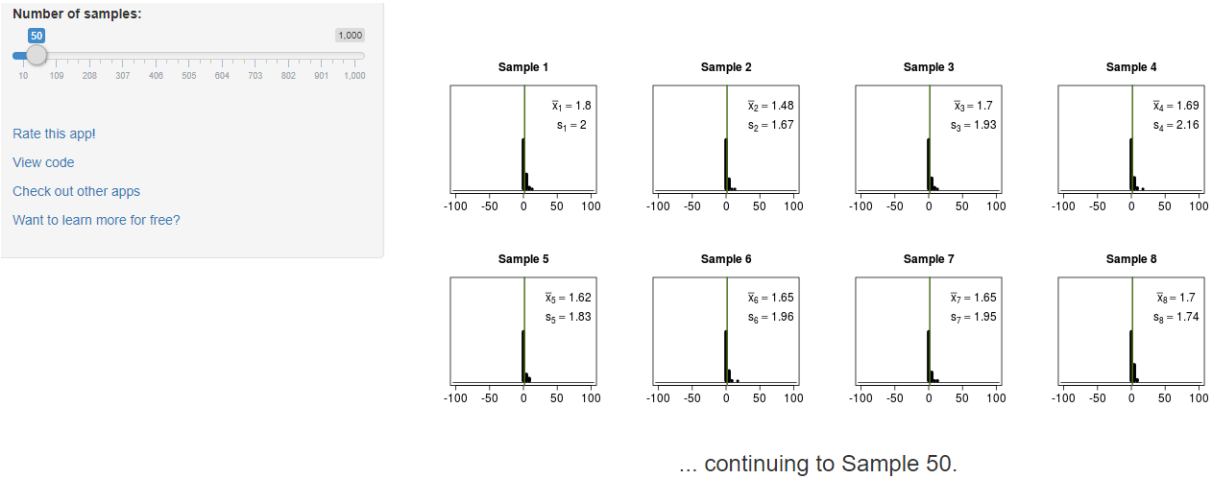
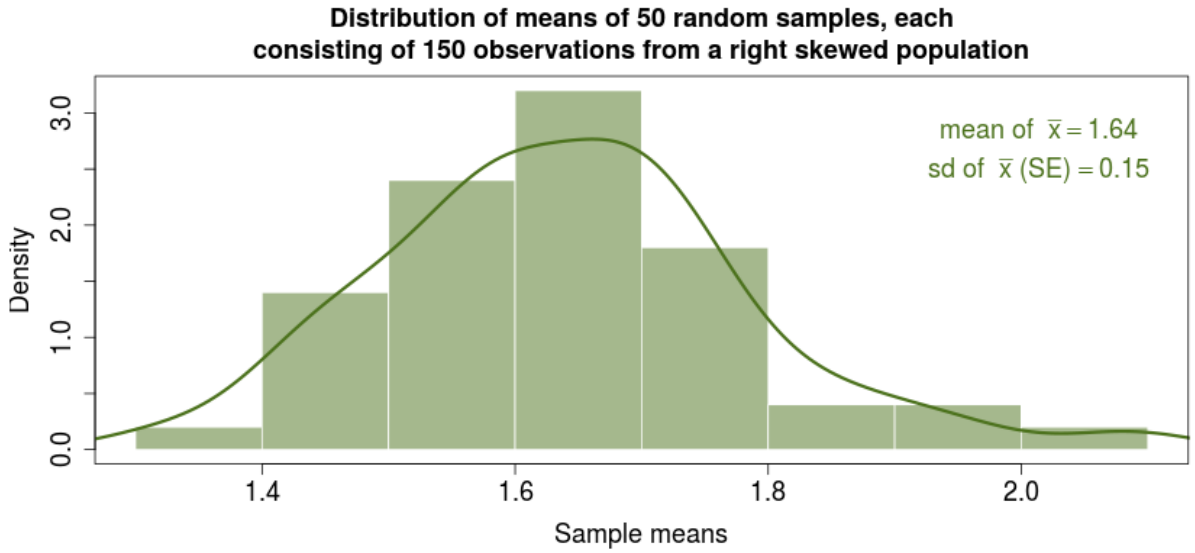


Рис. Високе (high) правостороннє відхилення (skew) генеральної сукупності





Приклад. Низька лівостороння асиметрія

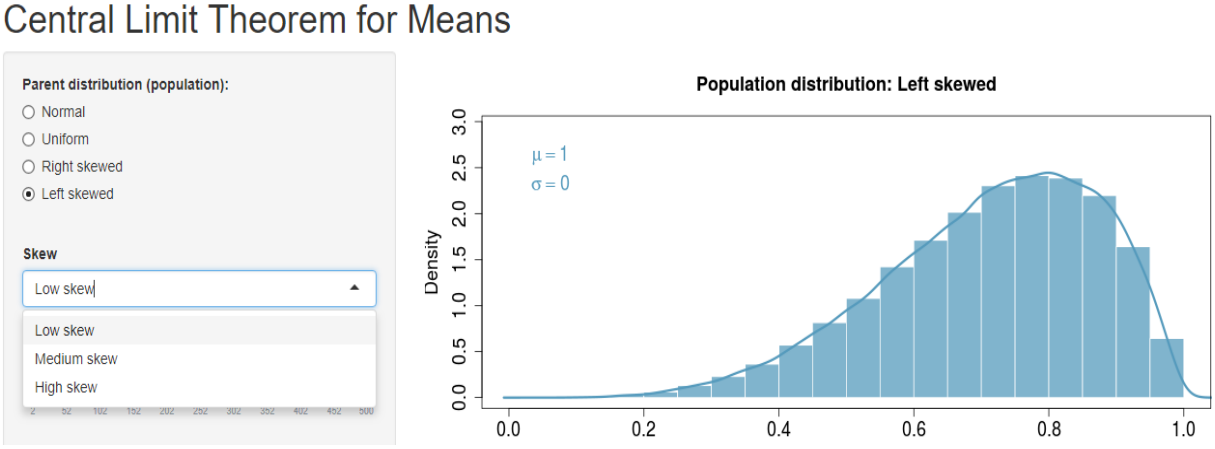
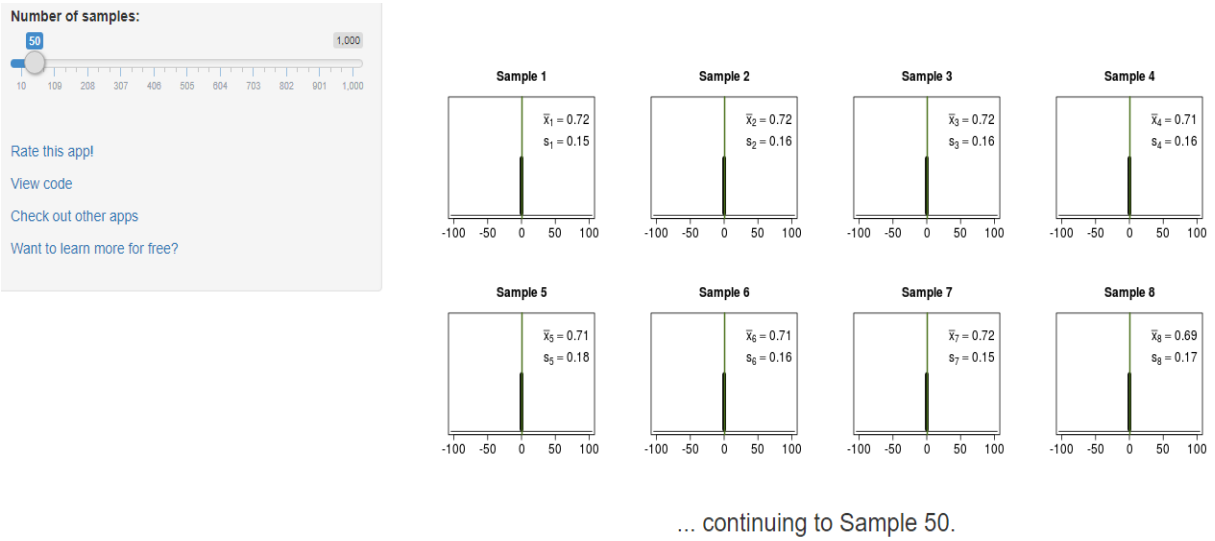
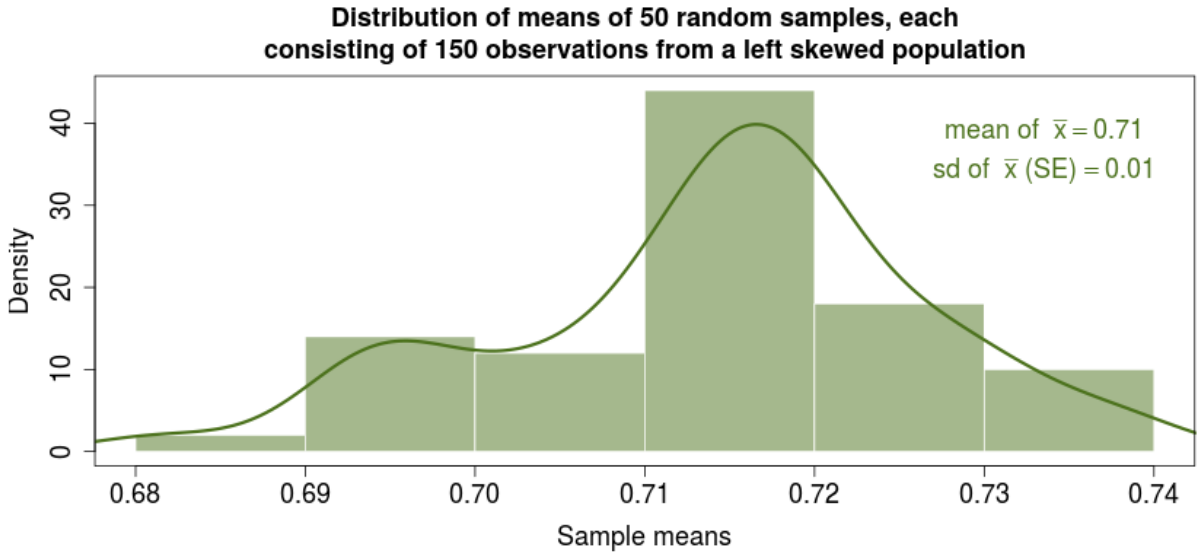
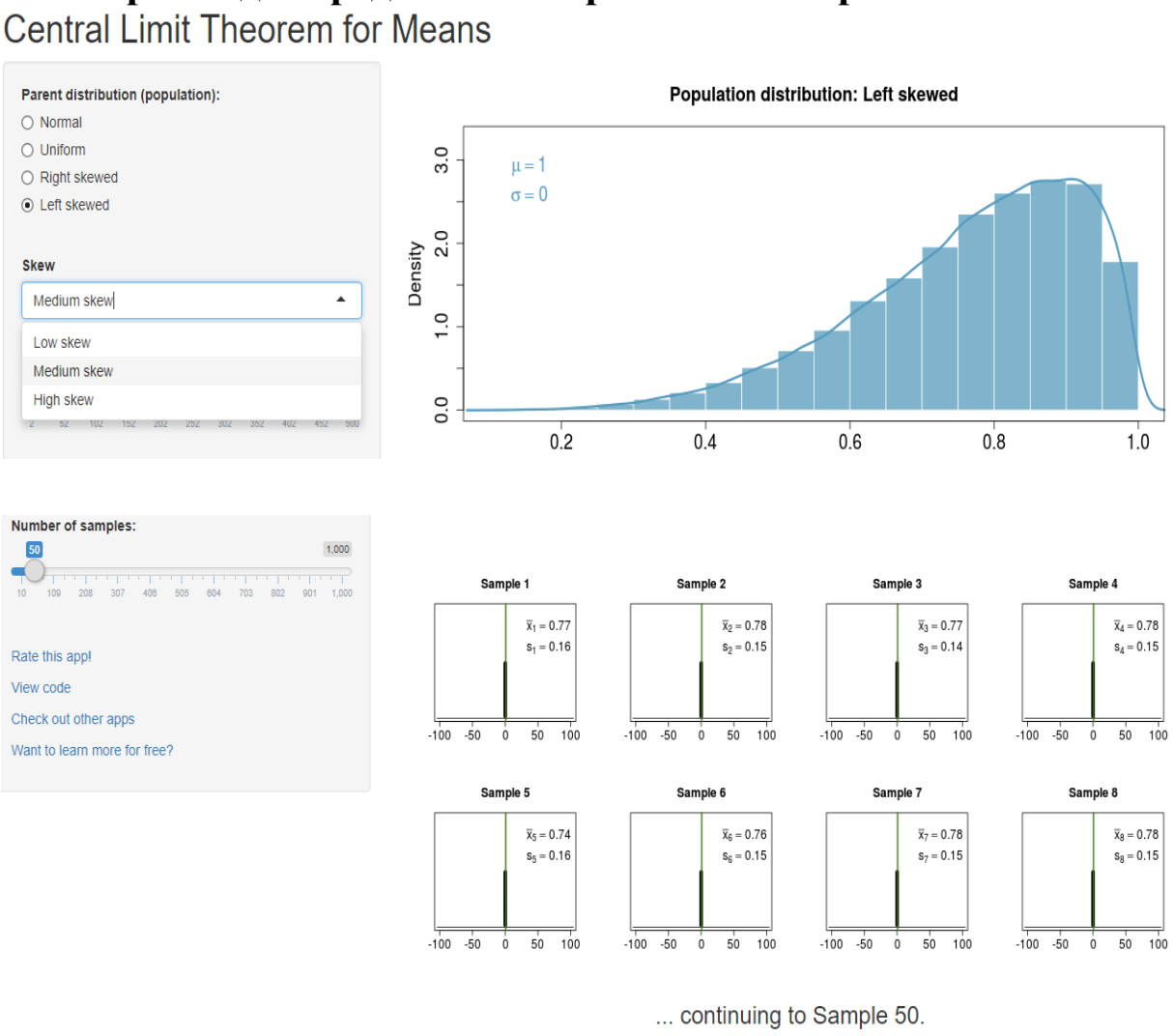


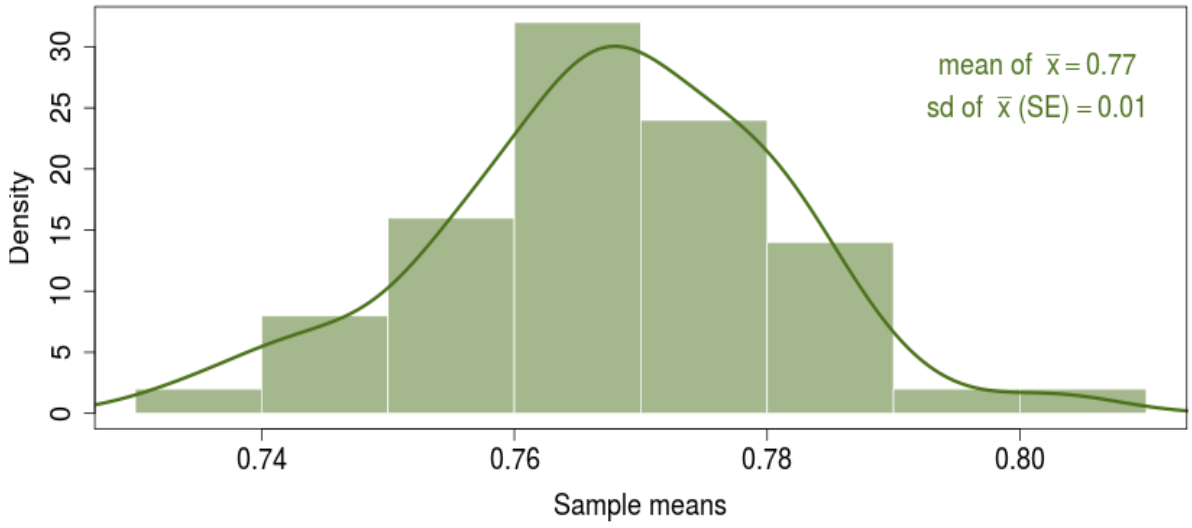
Рис.



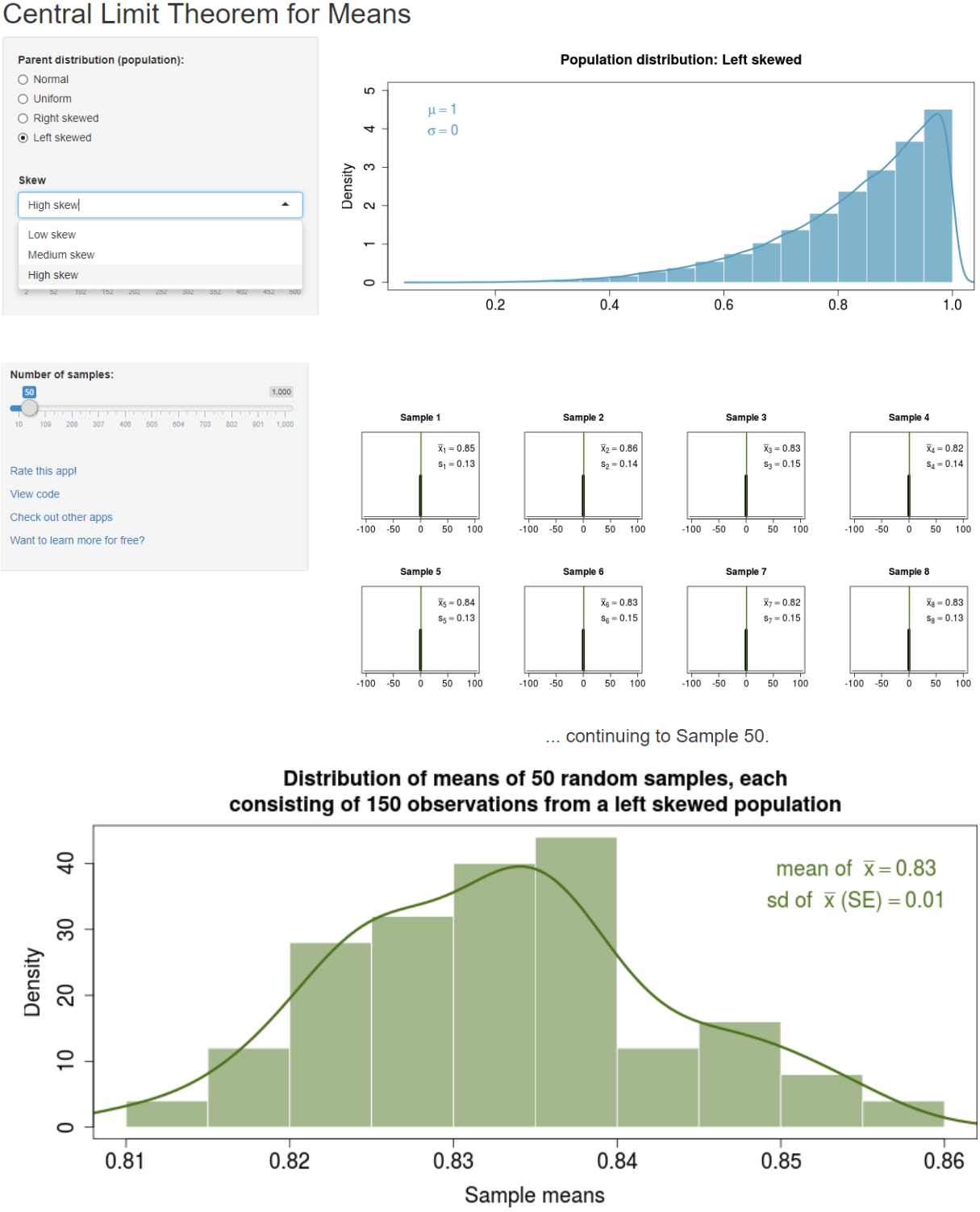


Приклад. Середня лівостороння асиметрія





Приклад. Висока лівостороння асиметрія



Додаток В

Критичне значення χ^2
На рівні значимості 0,10; 0,05; 0,01

Число ступенів свободи	Рівень значимості			Число ступенів свободи	Рівень значимості		
	0,01	0,05	0,1		0,01	0,05	0,1
1	6,63	3,84	2,71	24	42,98	36,42	33,20
2	9,21	5,99	4,61	25	44,31	37,65	34,38
3	11,34	7,81	6,25	26	45,64	38,89	35,56
4	13,28	9,49	7,78	27	46,96	40,11	36,74
5	15,09	11,07	9,24	28	48,28	41,34	37,92
6	16,081	12,59	10,64	29	49,59	42,56	39,09
7	18,48	14,07	12,02	30	50,89	43,77	40,26
8	20,09	15,51	13,36	35	57,34	49,80	46,06
9	21,67	16,92	14,68	40	63,69	55,76	51,81
10	23,21	18,31	15,99	45	69,96	61,66	57,51
11	24,73	19,68	17,28	50	76,15	67,50	63,17
12	26,22	21,03	18,55	60	88,38	79,08	74,40
13	27,69	22,36	19,81	70	100,43	90,53	85,53
14	29,14	23,68	21,06	80	112,33	101,88	96,58
15	30,58	25,00	22,31	90	124,12	113,15	107,57
16	32,00	26,29	23,54	100	135,81	124,34	118,50
17	33,41	27,59	24,77	125	164,69	152,09	145,64
18	34,81	28,87	25,99	150	193,21	179,58	172,58
19	36,19	30,14	27,20	200	249,45	233,99	226,02
20	37,57	31,41	28,41	300	359,91	341,40	331,79
21	38,93	32,67	29,62	400	468,72	447,63	436,65
22	40,29	33,92	30,81	500	576,49	553,12	540,93
23	41,64	35,17	32,01	1000	1106,97	1074,68	1057,72

Додаток Г
(до глави 3, 6)

Значення критерію Фішера (*F*-критерію) для рівня значимості *p = 0,05*

***f_1* – число ступенів свободи більшої дисперсії,**
***f_2* – число ступенів свободи меншої дисперсії**

	f_1										
f_2	1	2	3	4	5	6	7	8	9	10	15
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.18
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.13
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.09
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.06
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.03
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.83

Додаток Д
(до розділу 3)

Прохідний бал абітурієнтів вступної кампанії

№	Стать	Пріоритет	Бал	Українська мова і література	Математика	Географія	Середній бал документа про освіту	Іноземна мова
1	0	1	182,631	176	180	179	10,5	
2	0	4	179,775	179	172	175	10,6	
3	0	1	179,163	186	163	180	9,8	
4	0	3	179,163	175	184	163	10	
5	1	1	175,134	176	186		10,8	140
6	0	4	174,879	187	151	180	9,5	
7	0	2	174,675	179	166		9,7	167
8	0	6	174,522	170	178	158	10,3	
9	0	4	173,808	173	155	183	10,5	
10	0	1	173,451	192	173		10,4	134
11	0	1	171,666	181	166		10,1	138
12	0	1	170,678	176	157	154	9,8	
13	0	1	170,544	177	157		10,4	163
14	0	4	170,187	188	175		10,2	124
15	0	2	170,034	175	166		10,6	150
16	0	3	169,575	178	161		10	154
17	0	1	167,382	178	155	155	9,7	
18	1		167,127	153	166	169	9,6	
19	0	1	166,974	167	151	173	9,5	
20	0	1	166,26	172	153		9,1	163
21	0	7	165,087	169	137	180	10,2	
22	0	1	164,577	193	130		10,7	157
23	0	1	164,539	165	140	167	9,9	
24	0	1	164,424	176	149	155	9,5	
25	0	1	164,373	170	175		9,9	124
26	0	1	164,019	175	155		9,9	132
27	0	1	163,659	170	149	158	9,8	
28	1	2	163,098	183	142	150	9,8	
29	0	1	162,639	164	142	173	9,3	
30	0	1	162,588	157	155	161	9,8	
31	1	2	162,435	147	153	178	9,1	
32	0	2	162,333	176	133	166	10,3	
33	0	2	162,282	165	151	159	9	
34	0	5	161,772	154	153	169	8,6	
35	1	1	161,67	174	133		10	167
36	1	1	161,211	162	165		10,2	134
37	0	1	160,242	182	140	144	9,5	

38	0	5	159,285	170	133	155	8,8	
39	0	1	159,069	160	144	163	8,8	
40	0	7	159,018	178	144	136	10,1	
41	1	3	157,131	170	114	179	10,4	
42	0	1	156,264	175	144	134	8,8	
43	1	2	155,856	138	144	180	8	
44	0	2	155,754	153	149	147	9,9	
45	1	2	155,193	150	149	152	9	
46	1	1	154,916	152	114	182	9,9	
47	0	2	154,135	150	142	145	9,2	
48	0	2	153,927	163	125	150	9,8	
49	0	5	152,592	147	137	163	8,8	
50	1	1	152,031	156	149	134	8,6	
51	1	1	151,572	161	128		9	154
52	1	6	151,47	166	114	166	9,3	
53	0	3	151,326	162	140	123	9,1	
54	0	4	151,317	164	137	134	9,7	
55	0	1	150,858	172	119	149	9,4	
56	0	1	150,654	167	142	120	9,9	
57	1	1	150,39	166	110	157	9	
58	0	3	150,338	160	130	136	9	
59	0	6	150,338	162	114	158	8,5	
60	1	1	148,767	148	137	152	7,5	
61	1	2	148,206	161	128	134	10,7	
62	0	4	148,053	132	153		9	140
63	0	2	147,951	164	149	108	8,7	
64	0	7	146,268	163	117	147	8,8	
65	1	1	145,758	155	107	171	8,2	
66	0	1	143,871	119	155	138	8,6	
67	0		143,259	143	151	112	8,7	
68	0		142,847	155	110	136	10,3	
69	1	1	140,556	140	128	132	10	
70	0	4	140,038	126	130	142	7,8	
71	1	1	138,669	140	117	144	9	
72	1	2	137,853	138	133	126	7,7	
73	0	2	135,772	155	100	128	9	
74	0	1	129,081	147	119		7,8	100
75	1	4	128,724	109	122	138	8,3	
76	0	1	128,622	119	110	142	8,4	
77	1	2	123,216	111	119	121	7,6	
78	1	3	118,728	117	114	106	6,9	
	0 – жінка							
	1 – чоловік							

Додаток Е
(до розділів 3, 4)

Результати анкетування школярів

№	Стать*	Клас	ЗВО	Хобі**	Рахунок в умі	Числові ряди	Словник	Обізнаність	Короткочасна пам'ять	Українська мова	Математика
1	2	2	4	3	6	7	13	10	14	3,9	4,2
2	2	1	4	1	8	9	10	11	11	3,55	3,95
3	2	3	3	2	10	6	10	8	9	3,75	4,65
4	1	3	1	2	13	9	10	12	6	3,85	3,95
5	2	2	3	3	12	8	12	18	12	4,2	3,9
6	1	3	2	3	12	15	17	11	11	4,25	4,25
7	1	3	2	3	6	7	11	16	13	4,45	4,35
8	1	1	1	2	13	11	10	10	10	3,8	3,9
9	1	2	4	3	9	12	14	9	15	3,9	4
10	1	3	2	3	5	9	13	13	12	4,25	3,75
11	2	3	2	2	14	12	8	8	6	4,25	4,25
12	1	2	2	2	12	9	11	8	10	3,8	3,8
13	1	2	2	3	8	10	11	13	12	4,1	4,3
14	1	3	3	3	10	10	11	10	12	3,95	4,55
15	1	3	2	2	10	8	12	11	11	4,25	4,55
16	1	2	4	2	14	14	13	10	15	4,3	4,3
17	2	1	4	1	13	8	10	7	16	3,65	3,55
18	2	3	4	1	10	12	13	13	12	4,1	4
19	2	2	4	1	14	15	11	11	16	3,55	3,95
20	2	1	4	1	13	8	13	14	10	3,45	4,15
21	2	1	4	1	13	10	8	11	13	3,85	4,55
22	2	2	4	1	10	10	17	15	18	3,75	4,15
23	1	3	2	2	11	12	12	12	11	3,95	4,25
24	2	1	2	1	8	9	4	8	5	4,15	3,95
25	1	1	1	3	10	9	9	13	10	3,8	3,9
26	1	3	1	1	9	14	15	10	11	4	4
27	1	1	4	3	15	9	15	10	14	3,8	4
28	2	1	2	1	7	10	12	8	12	3,55	3,95
29	2	2	4	2	12	16	7	4	7	3,9	4,4
30	2	1	1	2	15	15	9	10	7	4,3	4,2
31	1	2	1	3	10	9	15	14	11	4	3,9
32	2	3	1	2	9	8	8	13	8	4,25	4,45
33	1	1	1	1	4	5	11	11	10	3,3	3,85
34	1	3	4	3	13	14	11	13	14	3,75	4,45

35	1	3	2	2	12	4	16	13	14	4,75	4,75
36	2	1	3	3	13	13	13	12	13	4	4,2
37	2	1	2	1	14	6	14	9	8	4,05	3,75
38	1	3	3	3	13	11	14	14	12	4,15	4,55
39	1	2	2	3	6	6	13	21	16	4,1	4,5
40	1	1	4	1	8	6	10	9	9	3,75	4,15
41	1	3	1	3	13	14	15	16	19	4,85	4,45
42	1	2	2	1	6	4	12	12	12	3,95	3,95
43	2	3	4	1	9	10	9	8	8	4,2	4,2
44	1	2	4	3	8	13	17	15	17	4	4,5
45	1	1	1	2	7	9	14	14	13	4	4,2
46	2	3	2	2	13	12	10	13	13	4,55	4,65
47	1	3	2	3	10	10	10	6	10	3,85	4,35
48	1	3	2	3	9	11	13	13	13	4,15	4,55
49	1	2	2	2	10	11	11	12	13	3,6	4
50	2	2	1	2	10	13	10	6	10	3,9	4,2
51	1	3	2	3	13	12	11	11	14	4,25	4,65
52	2	1	1	2	11	11	13	13	11	3,9	3,8
53	1	2	1	1	7	11	9	8	9	4,05	4,25
54	2	1	1	2	11	13	12	8	11	4,2	4,3
55	1	3	1	2	15	10	16	11	18	4,35	4,85
56	2	1	4	2	9	17	8	13	14	3,8	4,4
57	1	1	4	1	10	11	10	13	13	3,45	4,65
58	1	3	2	3	11	13	8	7	13	4,45	4,55
59	2	2	2	2	10	10	12	8	12	4	4,3
60	2	2	2	2	17	14	11	12	7	3,7	4,2
61	1	2	2	3	9	12	13	11	10	4,2	4,1
62	2	1	3	2	9	11	16	13	14	3,9	4,3
63	1	3	1	3	11	8	14	14	12	4,35	4,85
64	1	3	2	2	10	13	12	14	13	4,25	4,55
65	1	3	2	2	10	13	10	8	12	3,95	4,25
66	1	2	4	2	13	13	13	12	14	4,1	4,3
67	2	2	4	2	10	12	11	11	11	3,9	4,3
68	1	3	2	3	11	9	14	16	14	4,25	4,55
69	1	3	4	3	8	11	11	14	18	4,35	4,65
70	1	2	1	2	11	5	17	10	13	3,9	4
71	2	2	4	2	10	14	4	7	10	3,6	4,1
72	1	3	2	1	10	8	17	15	16	4,5	4,1
73	2	2	1	1	14	11	11	11	13	3,45	3,85
74	1	1	4	1	7	10	14	11	11	3,55	3,95
75	2	1	4	2	14	12	12	10	12	4,1	3,8
76	1	1	3	2	13	11	11	14	15	3,6	4
77	1	3	3	2	10	12	10	17	9	4,25	4,45

78	1	2	1	3	5	7	11	13	12	3,8	4,2
79	1	1	4	2	10	9	11	16	17	4,3	4,2
80	1	2	4	2	13	11	12	12	12	4,6	4,3
81	2	2	4	1	9	12	13	6	6	3,55	3,55
82	1	2	4	1	10	11	15	17	13	3,95	4,55
83	1	3	2	1	13	11	15	13	15	4,2	4,6
84	2	2	4	2	9	6	10	10	8	3,8	4,3
85	1	2	4	3	10	13	13	13	7	3,7	4,2
86	2	1	1	1	13	14	10	12	10	3,85	3,95
87	2	1	2	1	6	6	8	9	9	3,55	4,05
88	1	1	3	1	9	8	11	6	9	3,55	4,25
89	1	1	1	1	11	9	18	12	16	3,55	4,35
90	1	2	3	1	9	10	10	12	14	3,85	4,15
91	1	2	4	3	6	8	15	18	14	4	4,4
92	2	2	4	1	10	13	11	12	10	3,75	4,15
93	2	1	4	2	11	9	9	8	8	3,9	4,2
94	1	2	3	1	9	9	12	12	14	3,75	4,25
95	1	3	4	3	6	7	22	16	13	4,15	4,05
96	1	1	2	2	12	12	12	11	9	4,1	4,1
97	1	2	4	1	9	13	13	10	11	3,45	4,35
98	1	1	2	3	5	7	15	15	12	3,7	4,3
99	2	1	4	1	10	14	9	10	12	3,75	4,45
100	2	1	1	1	7	7	14	8	10	3,65	3,85

*Стать: 1 – жінка, 2 – чоловік.
**Хобі: 1 – спорт, 2 – комп’ютер, 3 – мистецтво.

Додаток Ж

Статистика Дарбина-Уотсона: d_L та d_U, рівень значимості 5 %

n	k = 1		k = 2		k = 3		k = 4		k = 5	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
8	0,763	1,332	0,559	1,777	0,368	2,287	-	-	-	-
9	0,824	1,320	0,624	1,699	0,455	2,128	0,296	2,588	-	-
10	0,879	1,320	0,697	1,641	0,525	2,016	0,376	2,814	0,243	2,822
11	0,927	1,324	0,758	1,604	0,595	1,928	0,444	2,283	0,316	2,645
12	0,971	1,331	0,812	1,579	0,658	1,864	0,512	2,177	0,379	2,506
13	1,010	1,340	0,861	1,562	0,715	1,816	0,574	2,094	0,445	2,390
14	1,045	1,350	0,905	1,551	0,767	1,779	0,632	2,030	0,505	2,296
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82
33	1,38	1,51	1,32	1,58	1,26	1,65	1,19	1,73	1,13	1,81
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79

40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
85	1,62	1,67	1,60	1,70	1,57	1,72	1,55	1,75	1,52	1,77
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78

Додаток З

Основні показники трудового потенціалу за регіонами

Показники	Зайняте населення, тис. осіб	Попит на робочу силу на кінець періоду, тис. осіб	Постійне населення, середня чисельність, тис. осіб	Середньомісячна заробітна плата, грн	Рівень зареєстрованого безробіття на кінець року у % до населення працездатного віку	Наявний дохід у розрахунку на одну особу, грн	Індекси споживчих цін	Доходи населення, млн грн
Вінницька	652,7	0,9	1561,016	7801	2,7	54992	109	112916
Волинська	371,1	2,6	1034,165	7324	2	46475,1	109,9	63741
Дніпропетровська	1402,3	6,4	3215,499	8862	1,7	72883,4	109,2	305510
Донецька	741	0,8	4170,296	9686	1,4	31888	112,3	174157
Житомирська	516,7	2,2	1226,485	7372	2,6	52135,9	109,1	83714
Закарпатська	502,4	1,2	1254,645	8070	0,9	40471,6	112,2	67323
Запорізька	732,2	0,8	1713,714	8726	2,4	67982,5	109,2	149083
Івано-Франківська	656,8	1,4	1372,648	7551	1,5	48367,7	109,1	86956
Київська	755,7	4,9	1755,333	9097	1,6	63498,4	110	145715
Кіровоградська	380,5	1,8	944,484	7191	3,4	51018	109	64523
Луганська	298,2	0,5	2155,221	7365	2,1	20618,6	109,3	60086
Львівська	1061,2	6,2	2507,445	8001	1,3	55510,7	110,1	185963
Миколаївська	496,2	1,4	1135,495	8160	2,8	55543,9	109,4	81497
Одеська	1001,9	3,3	2370,631	8011	1	61165,6	109,3	188312

Полтавська	580,6	3,7	1399,296	8375	3	60217,5	109,3	112856
Рівненська	473,6	1,3	1157,914	7469	2,5	47729,1	109,3	72819
Сумська	485,1	1,4	1085,659	7324	2,9	55934,4	109,7	80348
Тернопільська	410,8	1,2	1045,845	6969	1,9	43512,5	109,7	61684
Харківська	1258,9	3,3	2669,167	7657	1,6	60117,7	111,2	216333
Херсонська	448,2	0,6	1040,878	7058	1,9	50109,4	109,5	67894
Хмельницька	522	1,7	1266,394	7346	2,1	52487,6	109,2	87254
Черкаська	522,6	0,5	1209,728	7478	2,9	50292,6	109,7	82043
Чернівецька	382,9	1,3	902,473	6991	1,6	42850,4	108,7	51288
Чернігівська	429,7	1,2	1004,37	6995	2,4	50895,4	109,6	68630
м. Київ	1368,6	7,8	2901,364	13542	0,7	141173,8	108,8	548873

Джерело: за даними[25].

[25] Офіційний сайт Державної служби статистики України. URL : http://www.ukrstat.gov.ua/operativ/gdn/dvn/-arh_dvn2001.html

Навчальне видання

СЕРІЯ «НА ДОПОМОГУ СТУДЕНТУ УДФСУ»

Паянок Тетяна Миколаївна,
Задорожня Тетяна Миколаївна

СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ

Навчальний посібник

Відповідальний за випуск *Д. Ф. Салахова*

Редактор *М. М. Грабарчук*

Форматування та
комп'ютерна верстка *Н. М. Шамардак*

Здано до друку 14.02.2020. Формат 60×84/16.
Папір офсетний № 1. Гарнітура "Times New Roman".
Друк. арк. 18.01.
Наклад 300 прим. Замовлення № 854.

Підготовлено до друку Видавничо-поліграфічним центром
Університету ДФС України
08205, вул. Університетська, 31, м. Ірпінь, Київська обл., Україна

Свідоцтво про внесення суб'єкта видавничої справи
до державного реєстру видавців, виготовлювачів і
розповсюджувачів видавничої продукції
Серія ДК № 5104 від 20.05.2016