Розділ 9

Кореляційний аналіз

9.1. Поняття про кореляційний аналіз

Вивчення реальної дійсності показує, що практично кожне суспільне явище знаходиться в тісному зв'язку й взаємодії з іншими явищами, якими б випадковими вони не здавалися на перший погляд. Так, наприклад, рівень продуктивності праці (виробництво продукції на одного працівника), урожайності сільськогосподарських культур залежить від множини природних і економічних факторів, рівень злочинності – від багатьох соціальних та економічних факторів, тісно пов'язаних між собою.

Дослідження і вимірювання взаємозв'язків і взаємозалежностей соціально-економічних явищ є одним із найважливіших завдань статистики.

Для дослідження взаємозв'язків між явищами статистика використовує ряд методів і прийомів: статистичні групування (прості і комбінаційні). індексний, кореляційний і дисперсійний аналіз, балансовий, табличний, графічний та ін. Зміст, специфіка і можливості застосування деяких з перелічених методів уже були розглянуті в попередніх розділах посібника. Індексний і графічний методи розглядаються відповідно в 11 і 12 розділах.

Поряд з уже розглянутими методами вивчення взаємозв'язків особливе місце займає метод кореляції, який є логічним продовженням таких методів як аналітичне групування, дисперсійний аналіз і зіставлення паралельних рядів. В поєднанні з цими методами він надає статистичному аналізу закінчений, завершений характер.

Засновниками теорії кореляції є англійські статистики Ф.Гальтон (1822-1911 рр.) і К.Пірсон (1857-1936 рр.).

Термін кореляція походить від англійського слова correlation – співвідношення, відповідність (взаємозв'язок, взаємозалежність) між ознаками, що виявляється при масовому спостереженні зміни середньої величини однієї ознаки залежно від значення іншої. Ознаки, що пов'язані між собою кореляційним зв'язком, називають **корельованими**.

Кореляційний аналіз дає змогу виміряти ступінь впливу факторних ознак на результативні, встановити єдину міру тісноти зв'язку й роль досліджуваного фактора (факторів) у загальній зміні результативної ознаки. Кореляційний метод дозволяє одержати кількісні характеристики ступеня зв'язку між двома і більшим числом ознак, а тому на відміну від розглянутих вище методів, дає більш широке уявлення про зв'язок між ними.

Зв'язки між факторами досить різноманітні. При цьому одні ознаки виступають в ролі факторів, що діють на інші, зумовлюючи їх зміну, другі - в ролі дії цих факторів. Перші з них називають **факторними** ознаками, другі - **результативними.**

У правовій статистиці в аналізі злочинності та її факторів і причин важливе значення має виділення **криміногенних факторів**, які вказують на прямий зв'язок зі злочинністю (наприклад, безробіття, пияцтво, відсутність постійного джерела доходів, рівень освіти і виховання злочинців тощо) і **антикриміногенних факторів**, які вказують на обернений зв'язок із злочинністю (наприклад, рівень соціального контролю за злочинністю, адміністративна практика стосовно неповнолітніх правопорушників, навантаження кримінальних справ на суддю, слідчого тощо). Отже, чим вищий рівень криміногенних факторів, тим вище рівень злочинності і чим вищий рівень антикриміногенних факторів, тим рівень злочинності нижче, що відповідно вказує на прямий і обернений зв'язок між соціально-правовими явищами.

Досліджуючи зв'язки між ознаками, необхідно виділити насамперед два види зв'язків: 1) функціональний (повний) і 2) кореляційний (статистичний) зв'язок.

Функціональним називають такий зв'язок між ознаками, при якому кожному значенню однієї змінної (аргумента) відповідає строго визначене значення другої змінної (функції). Такі зв'язки спостерігаються в математиці, фізиці, хімії, астрономії та інших науках.

Наприклад, площа круга ($S = \pi R^2$) і довжина ($C = 2\pi R$) кола повністю визначається величиною радіуса, площі трикутника і прямокутника – довжиною їх сторін тощо. Так, із збільшенням радіуса кола на 1 см його довжина збільшується на 6,28 см, на 2 см - на 12,56 см і т.д.

У виробничій сфері прикладом функціонального зв'язку може бути зв'язок між виручкою від продажу продукції, ціною реалізації 1 т і кількістю реалізованої продукції; валовим збором, урожайністю і розміром посівної площі; фондовіддачею, вартістю валової продукції і основних фондів; заробітною платою й кількістю відпрацьованого часу при погодинній оплаті тощо.

Функціональний зв'язок виявляється як у сукупності в цілому, так і в кожній її одиниці абсолютно точно і виражається за допомогою аналітичних формул.

В соціально-економічних і соціально-правових явищах функціональні зв'язки між ознаками трапляються рідко. Тут дуже часто мають місце такі зв'язки між змінними величинами, при яких чисельному значенню однієї з них відповідає кілька значень інших. Такий зв'язок між ознаками дістав назву **кореляційного (статистичного) зв'язку.** Наприклад, відомо, що із збільшенням доз мінеральних добрив і поліпшенням їхньої структури (співвідношення), як правило, урожайність сільськогосподарських культур підвищується, але добре відомо, що приріст урожайності у кожному окремому випадку буде різним при однакових нормах внесення добрив. Крім того, одні і ті самі норми добрив, навіть при дуже вирівняних умовах, часто по-різному впливають на урожайність. Крім самих добрив на величину формування урожайності впливають також інші фактори, насамперед, такі як якість ґрунту, опади, строки і способи сівби та збирання тощо. Відома закономірність між урожайністю і добривами проявиться при досить великій кількості спостережень і при порівнянні досить великої кількості середніх значень результативної і факторної ознак.

Суспільні явища, в тому числі й соціально-правові взаємопов'язані між собою, залежать одне від одного і зумовлюють одне одного. Юридичні науки мають справу, головним чином, із соціально-правовими явищами, де не має жорстких, функціональних зв'язків. Злочинність, як масове соціальне явище, пов'язане з великою кількістю факторів (за даними науковців таких факторів налічується понад 450), які зі зміною дії хоча б одного з них можуть змінити характер взаємодії у цілому.

Прикладом кореляційного зв'язку в сфері соціально-правових явищ є зв'язок між рівнем злочинності (кількістю зареєстрованих злочинців на 1000 чоловік населення) і рівнем безробіття, споживанням алкогольних напоїв на душу населення, матеріальним положенням, зайнятістю, рівнем освіти та вихованням порушників; між навантаженням кримінальних справ та одного працівника міліції (слідчого, дізнавача) і відсотком розкриття злочинів; навантаженням справ на одного суддю і відсотком скасованих, змінених вироків, скасованих з направленням на новий судовий розгляд, на додаткове розслідування; між кількістю адміністративних правопорушень і кількістю злочинів; між кількістю вчинених адміністративних правопорушень і рівнем споживання алкоголю; між зайнятістю навчанням неповнолітніх і злочинні-

стю неповнолітніх; між злочинністю неповнолітніх і адміністративною практикою стосовно неповнолітніх правопорушників; між судимістю і злочинністю; між кількістю судимостей в розрахунку на одного злочинця і рівнем освіти та виховання злочинців; між судимістю і кількістю ув'язнених; між стажем роботи водіїв і аварійністю на транспорті (кількістю дорожньо-транспортних пригод); між кількістю пожежних команд у населеному пункті і сумою збитків за рік в населеному пункті від пожеж тощо.

Яскравим прикладом кореляційного зв'язку в соціально-правовій сфері є зв'язок між таким криміногенним, що статистично відслідковується фактором, як стан сп'яніння і злочинністю. В Україні в стані алкогольного сп'яніння в 2003 р. вчинено 37,9 тис. злочинів, у тому числі: 1807 умисних убивств, або 45,7 % від числа розслідуваних злочинів цього виду; 2068 умисних тяжких тілесних ушкоджень, або 38,6 %; 479 зґвалтувань або 48,4 %; 4,8 тис. випадків хуліганства, або 37,0 %. Кожний четвертий засуджений вчинив злочин у стані алкогольного чи наркотичного сп'яніння (засуджено майже 50 тис. осіб). Аналогічна картина спостерігалася і в попередні роки.

Наведені відсотки свідчать про прямий кореляційний зв'язок злочинів із пияцтвом. Оскільки ці цифри повторюються з року в рік, вони свідчать не тільки про наявність даного зв'язку, а й певною мірою і про ступінь впливу пияцтва на різні види злочинних діянь.

Істотно впливає на стан законності й правопорядку в країні безробіття. У 2003 р. в Україні майже дві третини (65,6 %) від загальної кількості працездатних громадян, винних у вчиненні злочинів, ніде не працювали, не навчалися та не мали постійного джерела доходів. Кількість засуджених цієї категорії осіб у 2003 р. в Україні становила 126,7 тис., або 63,0 % від усіх засуджених.

Індикатором неблагополуччя в державі є таке ганебне явище як невиплата зарплат. Згідно зі статтею 43 Конституції України право на своєчасне одержання винагороди за працю захищається законом. За невиплату зарплати, стипендії, пенсії та інших установлених законом виплат у 2002 р. в Україні було порушено 898 кримінальних справ, засуджено 72 особи. У 2003 р. ці показники відповідно становили 1265 кримінальних справ і 210 осіб, або на 40,9 і 191,7 % більше, ніж у 2002 р.

Кореляційний зв'язок є неповним, він проявляється при великій кількості спостережень, при порівнянні середніх значень результативної і факторної ознак. У цьому відношенні виявлення кореляційних залежностей пов'язано з дією закону великих чисел: тільки при досить великій кількості спостережень індивідуальні особливості і другорядні

фактори згладяться і залежність між результативною і факторною ознаками, якщо вона має місце, виявиться досить виразно.

За допомогою кореляційного аналізу вирішують такі основні завдання:

а) визначення середньої зміни результативної ознаки під впливом одного або кількох факторів (в абсолютному або відносному вимірі);

б) характеристика ступеня залежності результативної ознаки від одного з факторів при фіксованому значенні інших факторів, включених до кореляційної моделі;

в) визначення тісноти зв'язку між результативними і факторними ознаками (як з усіма факторами, так і з кожним фактором окремо при виключенні впливу інших);

г) визначення і розкладання загального обсягу варіації результативної ознаки на відповідні частини і встановлення ролі кожного окремого фактора в цій варіації;

д) статистична оцінка вибіркових показників кореляційного зв'язку.

Кореляційний зв'язок виражається відповідними математичними рівняннями. За **напрямом** зв'язок між корелюючими ознаками може бути прямим і оберненим. При **прямому зв'язку** обидві ознаки змінюються в одному напрямі, тобто із збільшенням факторної ознаки зростає результативна і навпаки (наприклад, зв'язок між безробіттям і злочинністю, пияцтвом і злочинністю, стажем роботи і продуктивністю праці). При **оберненому зв'язку** обидві ознаки змінюються в різних напрямах (наприклад, зв'язок між навантаженням на слідчого, суддю і якістю розслідування та судочинства, між кількістю пожежних команд і кількістю збитків від пожеж).

За **формою** або **аналітичним вираженням** розрізняють зв'язки прямолінійні (або просто лінійні) і нелінійні (або криволінійні). Якщо зв'язок між ознаками виражається рівнянням прямої лінії, то його називають **лінійним зв'язком,** якщо ж він виражається рівнянням будь-якої кривої (параболи, гіперболи, показникової, степеневої і т.д.), то такий зв'язок називають **нелінійним** або **криволінійним.**

Прикладом криволінійного зв'язку в соціально-правових явищах може бути зв'язок між злочинністю і віком правопорушників. Багаторічні дослідження, як у світі, так і в Україні, свідчать про те , що спочатку кримінальна активність осіб зростає прямо пропорційно зростанню віку правопорушників (приблизно) до 30 років, а потім із збільшенням віку злочинна активність знижується. При цьому, як свідчать дослідження, вершина кривої розподілу правопорушників за віком зсунута від середньої вліво (до більш молодого віку) і є асиметричною.

Залежно від **кількості досліджуваних ознак** розрізняють парну (просту) і множинну кореляцію. При **парній кореляції** вивчають зв'язок між двома ознаками (результативною і факторною), при **множинній кореляції** – зв'язок між трьома і більшим числом ознак (результативною і двома і більшим числом факторів).

За допомогою методу кореляційного аналізу вирішується два головних завдання: 1) визначення форми і параметрів рівняння зв'язку; 2) вимірювання тісноти зв'язку.

Перше завдання вирішується знаходженням рівняння зв'язку і визначенням його параметрів. Друге – за допомогою розрахунку різних показників тісноти зв'язку (коефіцієнта кореляції, кореляційного відношення, індексу кореляції та ін.).

Схематично кореляційний аналіз можна поділити на п'ять етапів:

1) постановка завдання, встановлення наявності зв'язку між досліджуваними ознаками:

2) відбір найістотніших факторів для аналізу;

3) визначення характеру зв'язку, його напряму і форми, вибір математичного рівняння для вираження існуючих зв'язків;

4) розрахунок числових характеристик кореляційного зв'язку (визначення параметрів рівняння і показників тісноти зв'язку;

5) статистична оцінка вибіркових показників зв'язку.

Науково обґрунтоване застосування кореляційного методу потребує передусім глибокого розуміння суті взаємозв'язків соціально-економічних і соціально-правових явищ. Сам метод не встановлює наявність і причин виникнення зв'язків між досліджуваними явищами, його призначення полягає в їх кількісному вимірюванні. На першому етапі кореляційного аналізу здійснюється загальне ознайомлення з досліджуваним об'єктом і явищами, уточнюються мета і завдання дослідження, встановлюється теоретична можливість причинно-наслідкового зв'язку між ознаками.

Встановлення причинних залежностей в досліджуваному явищі передує власне кореляційному аналізу. Тому застосуванню методів кореляції повинен передувати глибокий теоретичний аналіз, який охарактеризує основний процес, що протікає в досліджуваному явищі, визначить суттєві зв'язки між окремими його сторонами і характер їх взаємодії.

Простим, але разом з тим і вельми ефективним і широко застосовуваним методом виявлення зв'язків між суспільними, в тому числі й соціально-правовими, явищами є **метод паралельних рядів**, який дає можливість порівнювати зміну двох або кількох пов'язаних між собою

ознак (наприклад, безробіття і злочинність, стаж роботи водіїв і аварійність на транспорті тощо).

Метод паралельних рядів дає змогу порівнювати пов'язані між собою окремі явища як у часі (за допомогою побудови системи динамічних рядів), так і у просторі (по окремих територіях, регіонах, містах і т.д.).

При побудові паралельних рядів показники, що характеризують одну із ознак, необхідно розташувати у зростаючому або спадаючому порядку (як правило, у такому порядку розташовують результативну ознаку) і установити як змінюються в зв'язку з цим інші, що нас цікавлять показники, - зростають вони чи зменшуються і в якому степені.

Паралельні ряди дають змогу не тільки порівнювати показники двох або кількох пов'язаних між собою ознак, але й уловлювати тенденції їх спряженої зміни. Візуальний погляд на паралельні ряди вже дає можливість виявити наявність або відсутність зв'язку між показниками порівнюваних рядів. Тому перед розв'язуванням однофакторних або багатофакторних кореляційних моделей доцільно побудувати паралельні ряди і впевнитися в наявності кореляційного зв'язку між досліджуваними ознаками.

Наведемо приклади паралельних рядів. Порівняємо у просторі два паралельних ряда: коефіцієнт (рівень) злочинності (на 1000 чоловік населення) – результативна ознака (y) і рівень безробіття – факторна ознака (x) по 10 населених пунктах. Результативну ознаку розташуємо в зростаючому порядку:

№ населеного пункту	1	2	3	4	5	6	7	8	9	10
Коефіцієнт злочинності, злочинів (y)	5,1	6,2	6,3	7,4	7,5	8,2	9,7	10,8	11,8	12,0
Рівень безробіття, % (x)	3,9	3,8	4,7	4,6	5,5	5,4	6,3	6,2	7,2	8,4

Аналіз двох паралельних рядів показує, що із збільшенням рівня факторної ознаки (рівня безробіття) збільшується величина результативної ознаки (рівня злочинності), що свідчить про наявність прямого кореляційного зв'язку між цими ознаками.

Якщо побудувати динамічний ряд двох пов'язаних між собою показників (рівня злочинності і кількості вжитого алкоголю в літрах на душу населення в країні) за останні роки, починаючи з 1986 р., можна виявити наявність прямого кореляційного зв'язку між ними: із зрос-

танням вживання алкоголю зростала й злочинність в країні. Свого піку рівень злочинності досяг у 1995-1996 рр., коли в країні практично був знятий контроль за виробництвом та обігом алкогольних напоїв.

Попередній аналіз даних створює основу для формулювання конкретного завдання дослідження зв'язків, відбору найважливіших факторів, встановлення можливої форми взаємозв'язку ознак і тим самим приводить до математичної формалізації – до вибору математичного рівняння, яке найбільш повно відтворить існуючі зв'язки.

Одним із найважливіших питань кореляційного аналізу є відбір результативної і факторної (факторних) ознак. Факторні і результативні ознаки, що відбираються для кореляційного аналізу, повинні бути суттєвими, перші повинні безпосередньо впливати на другі. Відбір факторів для включення їх в кореляційну модель повинен базуватися перед усім на теоретичних основах і практичному досвіді аналізу досліджуваного соціально-економічного явища. Велику допомогу в розв'язанні цього завдання можуть надати такі статистичні прийоми і методи, як зіставлення паралельних рядів, побудова таблиць розподілу чисельностей за двома ознаками (кореляційних таблиць), побудова статистичних групувань як за результативною ознакою з аналізом взаємопов'язаних з нею факторів, так і за факторною ознакою (або комбінацією факторних ознак) з аналізом їх впливу на результативну ознаку.

Відбір факторів для включення їх в кореляційну модель повинен базуватися перед усім на теоретичних основах і практичному досвіді аналізу досліджуваного соціально-економічного і соціально-правового явища. Велику допомогу в розв'язанні цього завдання можуть надати такі статистичні прийоми і методи, як зіставлення паралельних рядів, побудова таблиць розподілу чисельностей за двома ознаками (кореляційних таблиць), побудова статистичних групувань як за результативною ознакою з аналізом взаємопов'язаних з нею факторів, так і за факторною ознакою (або комбінацією факторних ознак) з аналізом їх впливу на результативну ознаку.

Відбір факторів для парних кореляційних моделей не складний: з множини факторів, що впливають на результативну ознаку, відбирається один з найважливіших факторів, який в основному визначає варіацію результативної ознаки або ж фактор, істотність впливу якого на результативну ознаку передбачається вивчити або перевірити. Відбір факторів для множинних кореляційних моделей має ряд особливостей і обмежень. Вони будуть розглянуті при викладенні питань множинної кореляції.

Однією з головних проблем побудови кореляційної моделі є визначення форми зв'язку і на цій основі встановлення типу аналітичної функції, що відображає механізм зв'язку результативної ознаки з факторною (факторними). Під **формою кореляційного зв'язку** розуміють тип аналітичного рівняння, що виражає залежність між досліджуваними ознаками.

Вибір того або іншого рівняння для дослідження зв'язків між ознаками є найбільш важким і відповідальним завданням, від якого залежать результати кореляційного аналізу. Всі подальші найретельніші розрахунки можуть бути обезцінені, якщо форма зв'язку вибрана невірно. Важливість цього етапу полягає в тому, що правильно встановлена форма зв'язку дає змогу підібрати і побудувати найбільш адекватну модель і на основі її розв'язання одержати статистично вірогідні і надійні характеристики.

Встановлення форми зв'язку між ознаками в більшості випадків обґрунтовується теорією або практичним досвідом попередніх досліджень. Якщо форма зв'язку невідома, то при парній кореляції математичне рівняння може бути встановлено за допомогою складання кореляційних таблиць, побудови статистичних групувань, перегляду різних функцій на ЕОМ і вибір такого рівняння, яке дає найменшу суму квадратів відхилень фактичних даних від вирівняних (теоретичних) значень та ін.

Залежно від вихідних даних теоретичною лінією регресії можуть бути різні типи кривих або пряма лінія. Так, якщо зміна результативної ознаки під впливом фактора характеризується постійними приростами, то це вказує на лінійний характер зв'язку, якщо ж зміна результативної ознаки під впливом фактора характеризується постійними коефіцієнтами зростання, то є підстава припустити криволінійний зв'язок.

Особливе місце в обґрунтуванні форми зв'язку при проведенні кореляційного аналізу належить графікам, побудованих у системі прямокутних координат на основі емпіричних даних. Графічне зображення фактичних даних дає наочне уявлення про наявність і форму зв'язку між досліджуваними ознаками.

Згідно з правилами математики при побудові графіка на осі абсцис відкладають значення факторної ознаки, а на осі ординат – значення результативної ознаки. Відклавши на перетині відповідних значень двох ознак точки, одержимо точковий графік, який називають **кореляційним полем.** За характером розміщення точок на кореляційному полі роблять висновок про напрям і форму зв'язку. Достатньо поглянути

на графік, щоб прийти до висновку про наявність і форму зв'язку між ознаками. Якщо точки концентруються навколо уявної осі напрямленої зліва, знизу, направо, вгору, то зв'язок прямий, якщо к навпаки зліва, зверху, направо, вниз – зв'язок обернений. Якщо точки розкидані по всьому полю, то це свідчить про те, що зв'язок між ознаками відсутній або дуже слабкий. Характер розміщення точок на кореляційному полі вказує також і на наявність прямолінійного або криволінійного зв'язку між досліджуваними ознаками.

За допомогою графіка добирають відповідне математичне рівняння для кількісної оцінки зв'язку між результативною і факторною ознаками. Рівняння, що відображає зв'язок між ознаками, називають **рівнянням регресії** або **кореляційним рівнянням.** Якщо рівняння регресії зв'язує лише дві ознаки, то воно називається **рівнянням парної регресії.** Якщо рівняння зв'язку відображає залежність результативної ознаки від двох і більше факторних ознак, воно називається **рівнянням множинної регресії.** Криві, побудовані на основі рівнянь регресії, називають **кривими регресії** або **лініями регресії**.

Розрізняють емпіричну і теоретичну лінії регресії. Якщо на кореляційному полі з'єднати точки відрізками прямої лінії, то одержимо ламану лінію з деякою тенденцією, яка називається **емпіричною лінією регресії**. **Теоретичною лінією регресії** називається та лінія, навколо якої концентруються точки кореляційного поля і яка вказує основний напрям, основну тенденцію зв'язку. Теоретична лінія регресії повинна відображати зміну середніх величин результативної ознаки в міру зміни величин факторної ознаки при умові повного взаємопогашення всіх інших – випадкових по відношенню до фактора-причин. Отже, пошук, побудова, аналіз і практичне застосування теоретичної лінії регресії називають **регресійним аналізом.**

За емпіричною лінією регресії не завжди вдається встановити форму зв'язку і добрати рівняння регресії. В таких випадках будують і розв'язують різні рівняння регресії. Потім оцінюють їх адекватність і добирають таке рівняння, яке забезпечує найкращу апроксимацію (наближення) фактичних даних до теоретичних і достатню статистичну вірогідність і надійність.

Якщо підходити строго, регресійно-кореляційний аналіз слід розчленувати на регресійний і кореляційний. Регресійний аналіз вирішує питання побудови, розв'язання і оцінки рівнянь регресії, а при кореляційному аналізі до цих питань приєднується ще коло питань пов'язаних із визначенням тісноти зв'язку між результативною і факторною (факторними) ознаками. В подальшому викладенні регресійно-

кореляційний аналіз розглядається як єдине ціле і називається просто кореляційний аналіз.

Щоб результати кореляційного аналізу знайшли практичне застосування і дали науково обґрунтовані результати, повинні виконуватись певні вимоги відносно об'єкта дослідження і якості вихідної статистичної інформації. Основні з цих вимог такі:

- якісна однорідність досліджуваної сукупності, що передбачає близькість формування результативних і факторних ознак. Необхідність виконання цієї умови випливає із змісту параметрів рівняння зв'язку. З математичної статистики відомо, що параметри є середніми величинами. В якісно однорідній сукупності вони будуть типовими характеристиками, в якісно різнорідній спотвореними, що перекручують характер зв'язку. Кількісна однорідність сукупності полягає у відсутності одиниць спостереження, які за своїми числовими характеристиками суттєво відрізняються від основної маси даних. Такі одиниці спостереження слід виключати із сукупності і вивчати окремо;
- досить велике число спостережень, оскільки зв'язки між ознаками виявляються тільки внаслідок дії закону великих чисел. Кількість одиниць спостереження повинна в 6 – 8 разів перевищувати кількість включених у модель факторів;
- випадковість і незалежність окремих одиниць сукупності одна від одної. Це означає, що значення ознак у одних одиниць сукупності не повинні залежати від значень у інших одиниць даної сукупності;
- стійкість і незалежність дії окремих факторів;
- сталість дисперсії результативної ознаки при зміні факторних ознак;
- нормальний розподіл ознак.

9.2. Парна (проста) лінійна кореляція

Найпростішим видом кореляційного зв'язку є зв'язок між двома ознаками: результативною і факторною. Такий зв'язок називають **парною кореляцією** або **простою кореляцією.**

В дослідженнях соціально-правових явищ при взаємозв'язку двох факторів серед множини функцій часто розглядається прямолінійна форма зв'язку, яка виражається рівнянням прямої лінії:

$$\tilde{y}_x = a + bx,$$

де $\tilde{y}_x$ – вирівняне значення результативної ознаки (залежна змінна); x – значення факторної ознаки (незалежна змінна); a – початок відліку, або значення $\tilde{y}_x$ при $b = 0$ (економічного змісту не має); b - **коефіцієнт**

регресії, який показує середню змінну залежної змінної при зміні незалежної змінної на одиницю (одне своє значення).

Коефіцієнти регресії є величинами іменованими і мають одиниці вимірювання, що відповідають змінним, між якими вони характеризують зв'язок.

Якщо $b > 0$, то зв'язок прямий, якщо $b < 0$, то зв'язок обернений, якщо $b = 0$, то зв'язок відсутній.

Параметри рівняння ***a*** і ***b*** визначають способом найменших квадратів. Він дає можливість знайти ту криву, яка порівняно з іншими кривими проходить найближче до точок кореляційного поля, що відображають фактичні дані, тобто дає найменшу суму квадратів відхилень фактичних значень результативної ознаки від вирівняних (теоретичних) значень:

$$\sum (y_i - \tilde{y}_x)^2 = \min.$$

Порядок одержання системи нормальних рівнянь при парній кореляції такий. Для одержання першого рівняння системи необхідно всі члени вихідного рівняння кореляційного зв'язку помножити на коефіцієнти при першому невідомому (***a***) і одержані добутки підсумувати. Потім для отримання другого рівняння необхідно всі члени вихідного рівняння помножити на коефіцієнт при другому невідомому (***b***) і також всі добутки підсумувати.

Техніка одержання системи нормальних рівнянь залишається аналогічною і для побудови системи рівнянь з більшим числом змінних. Так, для парного лінійного зв'язку система нормальних рівнянь має вигляд:

$$\begin{cases} \Sigma y = an + b\Sigma x; \\ \Sigma yx = a\Sigma x + b\Sigma x^2. \end{cases}$$

Параметри ***a*** і ***b*** рівняння прямої лінії можна визначити за іншими робочими формулами:

$$a = \frac{\Sigma y \Sigma x^2 - \Sigma yx \Sigma x}{n\Sigma x^2 - \Sigma x \Sigma x}; \qquad b = \frac{n\Sigma xy - \Sigma y \Sigma x}{n\Sigma x^2 - \Sigma x \Sigma x};$$

або

$$a = \bar{y} - b\bar{x}; \qquad b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}.$$

Рівняння кореляційного зв'язку мають як пізнавальне, так і практичне значення. Їх використовують для обчислення теоретичної лінії

регресії, очікуваних (теоретичних, вирівняних) і прогнозованих значень залежної змінної при тих або інших значеннях фактора (факторів). При цьому слід мати на увазі, що рівняння дає середнє співвідношення між результативною і факторною ознаками, тому найбільшу точність збігання мають розрахункові значення результативної ознаки при величині фактора, близького до середнього його рівня.

Ступінь наближення розрахункових значень результативної ознаки до її фактичного значення залежить від того, наскільки досконала кореляційна модель. Якщо вона включає всі основні фактори, що визначають варіацію результативної ознаки, то точність буде досить високою.

Розглянемо приклад аналізу кореляційного зв'язку між двома ознаками (парна кореляція): рівнем (коефіцієнтом) злочинності (на 1000 чоловік населення), злочинів і рівнем безробіття, % (табл. 9.1).

Таблиця 9.1

Визначення даних для визначення показників кореляційного зв'язку

№ п/п	Коефіцієнт злочинності (на 1000 чол. населення), злочинів	Рівень безробіття, %	Розрахункові дані			
						Очікуване значення коефіцієнта злочинності, злочинів
	y	x	yx	y^2	x^2	$\tilde{y}_x$
1	7,5	5,5	41,25	56,25	30,25	8,34
2	12,0	8,4	100,80	144,0	70,56	12,93
3	5,1	3,9	19,89	26,01	15,21	5,81
4	9,7	6,3	61,11	94,09	39,69	9,61
5	6,2	3,8	23,56	38,44	14,44	5,65
6	10,8	6,2	66,96	116,64	38,44	9,45
7	6,3	4,7	29,61	39,69	22,09	7,08
8	7,4	4,6	34,04	54,76	21,16	6,92
9	11,8	7,2	84,96	139,24	51,84	11,03
10	8,2	5,4	44,28	67,24	29,16	8,18
Разом	**85,0**	**56,0**	**506,46**	**776,36**	**332,84**	**85,00**
У середньому	**8,5**	**5,6**	**50,646**	**77,636**	**33,284**	**8,5**

Результативною ознакою в даному прикладі є коефіцієнт злочинності (y), а факторною – рівень безробіття (x).

Для визначення форми зв'язку між коефіцієнтом злочинності і рівнем безробіття побудуємо графік – кореляційне поле (рис. 9.1.). На осі абсцис відкладемо значення факторної ознаки (незалежної змінної –

рівня безробіття, а на осі ординат – результативної ознаки (залежної змінної – коефіцієнт злочинності).

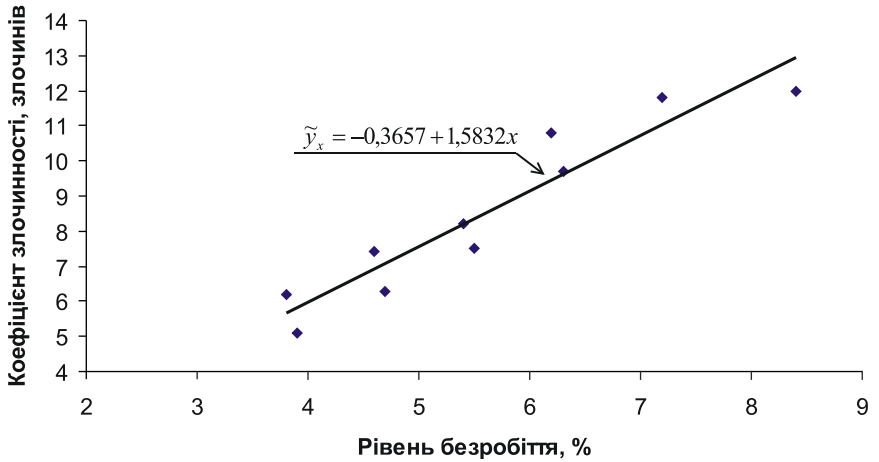

Рис. 9.1. Кореляційне поле залежності коефіцієнта злочинності від рівня безробіття.

Графік показує, що в даному випадку зв'язок близький до прямолінійного і його можна виразити рівнянням прямої лінії

$$\tilde{y}_x = a + bx.$$

Розв'язання цього рівняння регресії покаже зміну коефіцієнта злочинності під впливом рівня безробіття при виключенні випадкових коливань ознаки.

Параметри рівняння прямої лінії ***a*** і ***b*** знайдемо із системи нормальних рівнянь:

$$\begin{cases} \Sigma y = an + b\Sigma x; \\ \Sigma yx = a\Sigma x + b\Sigma x^2. \end{cases}$$

Усі потрібні для розв'язання системи рівнянь дані визначимо в таблиці 9.1.

Одержані дані підставимо в систему рівнянь:

$$\begin{cases} 85{,}0 = 10a + 56b & \Big| :10; \\ 506{,}46 = 56a + 332{,}84b & \Big| :56. \end{cases}$$

Поділимо рівняння на коефіцієнти при ***a***, тобто перше рівняння на 10, а друге – на 56:

$$\begin{cases} 8{,}5000 = a + 5{,}6000b; \\ 9{,}0439 = a + 5{,}9436b. \end{cases}$$

Віднімемо перше рівняння із другого:

$$0{,}5439 = 0{,}3436b.$$

Звідси $b = \dfrac{0{,}5439}{0{,}3436} = 1{,}5832 \approx 1{,}58$ злочину.

Підставимо значення $b = 1{,}5832$ в перше рівняння і знайдемо a:

$$85{,}0 = 10a + 56 \cdot 1{,}5832\ ;$$

$$a = \frac{85 - 88{,}657}{10} = \frac{-3{,}657}{10} = -0{,}3657.$$

Рівняння регресії (кореляційне рівняння), яке виражає зв'язок між коефіцієнтом злочинності і рівнем безробіття буде мати вигляд:

$$\tilde{y}_x = a + bx = -0{,}3657 + 1{,}5832x.$$

Коефіцієнт регресії b =1,5832 показує, що із збільшенням рівня безробіття на один процент коефіцієнт злочинності у середньому для даної сукупності населених пунктів зростає на 1,5832 злочину.

Параметри рівняння регресії можна визначити і за іншими формулами:

$$b = \frac{n\sum yx - \sum y \sum x}{n\sum x^2 - \sum x \sum x} = \frac{10 \cdot 506{,}46 - 85 \cdot 56}{10 \cdot 332{,}84 - 56 \cdot 56} = 1{,}5832 \text{ злочину.}$$

$$a = \frac{\sum y \sum x^2 - \sum yx \sum x}{n\sum x^2 - \sum x \sum x} = \frac{85 \cdot 332{,}84 - 506{,}46 \cdot 56}{10 \cdot 332{,}84 - 56 \cdot 56} = -0{,}3657.$$

або

$$a = \bar{y} - b\bar{x} = 8{,}5 - 1{,}5832 \cdot 5{,}6 = -0{,}3657\ .$$

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\bar{x}^2 - (\bar{x})^2} = \frac{50{,}646 - 8{,}5 \cdot 5{,}6}{33{,}284 - 5{,}6^2} = \frac{3{,}046}{1{,}924} = 1{,}5832 \text{ злочину.}$$

$$a = \bar{y} - b\bar{x} = 8{,}5 - 1{,}5832 \cdot 5{,}6 = -0{,}3657$$

Перевіримо правильність розв'язання системи рівнянь, виходячи із рівності

$$\bar{y} = a + b\bar{x};$$

$$8{,}5 = -0{,}3657 + 1{,}5832 \cdot 5{,}6\ ;\ 8{,}5 = 8{,}5\ .$$

За рівнянням регресії можна визначити очікувані (розрахункові або теоретичні) значення коефіцієнта злочинності ($\tilde{y}_x$) при різних значеннях рівня безробіття (x). Для цього замість x підставимо його конкретні значення:

$$\tilde{y}_{x=5,5} = -0,3657 + 1,5832 \cdot 5,5 = 8,34 \text{ злочину;}$$

$$\tilde{y}_{x=8,4} = -0,3657 + 1,5832 \cdot 8,4 = 12,93 \text{ злочину і т.д.}$$

Усі обчислені дані запишемо в останню графу табл. 9.1. За цими даними на рис. 9.1 побудуємо теоретичну лінію регресії.

5. Перевіримо правильність усіх розрахунків, зіставивши суми фактичного і розрахункового коефіцієнта злочинності:

$$\sum y = \sum \tilde{y}_x ; \quad 85,0 = 85,0 .$$

9.3. Показники тісноти зв'язку

При кореляційному зв'язку разом з досліджуваним фактором або кількома факторами при множинній кореляції на результативну ознаку впливають і інші фактори, які не враховуються або не можуть бути точно враховані. При цьому дія їх може бути направлена як в сторону підвищення результативної ознаки, так і в сторону її зниження. Отже, дослідження зв'язку відбувається в умовах, коли цей зв'язок у більшій або меншій мірі затушовується суперечливою дією інших причин. Тому одне із завдань кореляційного аналізу полягає у визначенні тісноти зв'язку між ознаками, у визначенні сили дії досліджуваного фактора (факторів) на результативну ознаку.

Тіснота зв'язку у кореляційному аналізі характеризується за допомогою спеціального відносного показника, який отримав назву **коефіцієнта кореляції.**

При парній лінійній залежності тіснота зв'язку визначається за допомогою **лінійного коефіцієнта кореляції**

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y},$$

де $$\overline{xy} = \frac{\Sigma xy}{n}; \qquad \bar{y} = \frac{\Sigma y}{n}; \qquad \bar{x} = \frac{\Sigma x}{n};$$

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{n} - (\bar{x})^2}; \qquad \sigma_y = \sqrt{\frac{\Sigma y^2}{n} - (\bar{y})^2}.$$

Він може бути обчислений за іншими формулами:

$$r = \frac{\Sigma(y - \bar{y})(x - \bar{x})}{n\sigma_x \cdot \sigma_y}; \qquad r = \frac{n\Sigma yx - \Sigma y \Sigma x}{\sqrt{[n\Sigma y^2 - (\Sigma y)^2][n\Sigma x^2 - (\Sigma x)^2]}};$$

$$r = \frac{\Sigma(y - \bar{y})(x - \bar{x})}{\sqrt{\Sigma(y - \bar{y})^2 (x - \bar{x})^2}}; \qquad r = b\frac{\sigma_x}{\sigma_y}.$$

Коефіцієнт кореляції знаходиться в межах від 0 до ±1. Якщо коефіцієнт кореляції дорівнює нулю, то зв'язок відсутній, а якщо одиниці, то зв'язок функціональний. Знак при коефіцієнті кореляції вказує на напрям зв'язку ("+" – прямий, "-" – обернений). Чим ближче коефіцієнт кореляції до одиниці, тим зв'язок між ознаками тісніший.

Квадрат коефіцієнта кореляції називається **коефіцієнтом детермінації (r^2).** Він показує, яка частка загальної варіації результативної ознаки визначається досліджуваним фактором. Якщо коефіцієнт детермінації виражений в процентах, то його слід читати так: варіація (коливання) залежної змінної на стільки-то процентів зумовлена варіацією фактора.

Між лінійним коефіцієнтом кореляції (r) і коефіцієнтом повної регресії (b) є такий зв'язок:

$$r = b\frac{\sigma_x}{\sigma_y}. \text{ Звідси } b = r\frac{\sigma_y}{\sigma_x}.$$

Отже, знаючи коефіцієнт кореляції (r) і значення середніх квадратичних відхилень по x і y, можна визначити коефіцієнт регресії (b) і навпаки, знаючи коефіцієнт регресії (b) і відповідні середні квадратичні відхилення можна обчислити коефіцієнт кореляції (r).

При парній лінійній залежності коефіцієнт кореляції і коефіцієнт повної регресії мають однакові знаки (плюс, мінус).

Лінійний коефіцієнт кореляції призначений для оцінки ступеня тісноти зв'язку при лінійній залежності. Для випадків нелінійного зв'язку між ознаками використовується інша формула коефіцієнта кореляції, яка випливає з правила додавання дисперсій:

$$\sigma^2_{заг} = \sigma^2_{м.гр} + \sigma^2_{в.гр}.$$

Із наведеної рівності видно, що чим більшим є вплив фактора на результативну ознаку, тим більшою мірою значення її дисперсії ($\sigma^2_{м.гр}$) наближається до значення загальної дисперсії результативної ознаки. Відповідно, чим більше $\sigma^2_{м.гр}$ і менша $\sigma^2_{в.гр}$, тим зв'язок між ознаками

буде тіснішим і навпаки. Відтак, відношення міжгрупової (факторної) і загальної дисперсій використовується для оцінки тісноти зв'язку між ознаками. Формула коефіцієнта кореляції має вигляд:

$$r = \sqrt{\frac{\sigma^2_{м.гр}}{\sigma^2_{заг}}}.$$

Враховуючи, що $\sigma^2_{м.гр} = \sigma^2_{заг} + \sigma^2_{в.гр}$, формулу коефіцієнта кореляції можна подати як

$$r = \sqrt{1 - \frac{\sigma^2_{в.гр}}{\sigma^2_{заг}}}.$$

Обидві формули коефіцієнта кореляції застосовуються для обчислення тісноти зв'язку при будь-якій формі зв'язку.

Із правила додавання дисперсій видно, що значення коефіцієнта кореляції перебуває в межах від 0 до 1. Знак коефіцієнта кореляції з формули не виводиться. Якщо вивчається зв'язок між двома ознаками (парна проста кореляція), то напрямок зв'язку (знак перед r) визначається безпосередньо за знаком перед коефіцієнтом регресії лінійного рівняння.

При парній криволінійній залежності, тіснота зв'язку як і при лінійній залежності, визначається за допомогою спеціального показника, аналогічного розглянутому вище коефіцієнту кореляції r.

Цей показник (щоб підкреслити його належність до криволінійного зв'язку) позначають символом i_r і називають індексом кореляції:

$$i_r = \sqrt{\frac{\sigma^2_{м.гр}}{\sigma^2_{заг}}}, \text{ або } i_r = \sqrt{1 - \frac{\sigma^2_{в.гр}}{\sigma^2_{заг}}}.$$

Числове значення індексу кореляції аналогічне коефіцієнту кореляції: якщо $i_r = 1$ – зв'язок функціональний, якщо $i_r = 0$ – зв'язок відсутній; чим i_r ближче до одиниці, тим зв'язок між ознаками тісніший.

Якщо відомі коефіцієнти регресії рівняння зв'язку, то індекс кореляції можна визначити за іншою, простішою формулою. Так, при параболічній залежності формула індексу кореляції може бути подана як

$$i_r = \sqrt{\frac{\sigma^2_{м.гр}}{\sigma^2_{заг}}} = \sqrt{\frac{a\Sigma y + b\Sigma yx + c\Sigma yx^2 - n\bar{y}^2}{\Sigma y^2 - n\bar{y}^2}}.$$

або

$$i_r = \sqrt{1 - \frac{\sigma^2_{в.гр}}{\sigma^2_{заг}}} = \sqrt{1 - \frac{\Sigma y^2 - a\Sigma y - b\Sigma yx - c\Sigma yx^2}{\Sigma y^2 - n\bar{y}^2}}.$$

Тіснота зв'язку при множинній кореляції визначається за допомогою **коефіцієнта множинної кореляції (R)** і **коефіцієнта множинної детермінації (R^2).** За змістом вони аналогічні коефіцієнтам кореляції і детермінації при парному зв'язку. Їх обчислення ґрунтується на порівнянні міжгрупової (факторної) і загальної дисперсій:

$$R = \sqrt{\frac{\sigma^2_{м.гр}}{\sigma^2_{заг}}}, \text{ або } R = \sqrt{1 - \frac{\sigma^2_{в.гр}}{\sigma^2_{заг}}}.$$

Ця формула може бути застосована для визначення тісноти зв'язку при будь-якій формі зв'язку.

Величина R змінюється від 0 до 1 і розглядається як додатна, оскільки при множинних залежностях зв'язок результативної ознаки з одними факторами може бути додатнім, а з іншими – від'ємним.

Для випадку залежності результативної ознаки від двох факторів формула коефіцієнта множинної кореляції має вигляд

$$R = \sqrt{\frac{r^2_{yx_1} + r^2_{yx_2} - 2r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r^2_{x_1x_2}}},$$

де r_i – парні лінійні коефіцієнти кореляції.

Наведена формула застосовується для визначення тісноти зв'язку при лінійній залежності.

Для визначення тісноти зв'язку між результативною ознакою і кожним фактором при виключені впливу інших факторів визначають **часткові коефіцієнти кореляції,** які характеризують "чистий" вплив фактора на результативну ознаку. Для їх розрахунку використовують парні коефіцієнти кореляції.

У випадку залежності результативної ознаки від двох факторів (x_1 і x_2) можна розрахувати три коефіцієнта часткової кореляції:

1) між y і x_1 при виключенні впливу x_2:

$$r_{yx_1(x_2)} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{(1 - r^2_{yx_2})(1 - r^2_{x_1x_2})}};$$

2) між y і x_2 при виключенні впливу x_1:

$$r_{yx_2(x_1)} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{(1-r_{yx_1}^2)(1-r_{x_1x_2}^2)}};$$

3) між x_1 і x_2 при виключенні впливу y:

$$r_{x_1x_2(y)} = \frac{r_{x_1x_2} - r_{yx_1} \cdot r_{yx_2}}{\sqrt{(1-r_{yx_1}^2)(1-r_{yx_2}^2)}}.$$

Коефіцієнти кореляції при парних і множинних зв'язках, а також індекс кореляції – це відносні величини, тому вони можуть бути використані для зіставлення тісноти зв'язку по кількох аналізуємих явищах.

Слід мати на увазі, що показники тісноти зв'язку залежать від розмаху варіювання досліджуваних ознак. Чим більшою є варіація змінних, тим вищою буде величина показників тісноти зв'язку.

Визначимо тісноту зв'язку між досліджуваними ознаками для нашого прикладу. Оскільки між коефіцієнтом злочинності і рівнем безробіття має місце лінійний зв'язок, тісноту зв'язку визначимо за допомогою лінійного коефіцієнта кореляції

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y},$$

де $\overline{xy} = \frac{\sum xy}{n} = \frac{506,46}{10} = 50,646$; $\bar{x} = \frac{\sum x}{n} = \frac{56}{10} = 5,6\%$;

$$\bar{y} = \frac{\sum y}{n} = \frac{85}{10} = 8,5 \text{ злочину};$$

$$\sigma_x = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2} = \sqrt{\frac{332,84}{10} - 5,6^2} = \sqrt{1,924} = 1,3871\%;$$

$$\sigma_y = \sqrt{\frac{\sum y^2}{n} - \bar{y}^2} = \sqrt{\frac{776,36}{10} - 8,5^2} = \sqrt{4,886} = 2,2104 \text{ злочину};$$

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y} = \frac{50,646 - 5,6 \cdot 8,5}{1,3871 \cdot 2,2104} = \frac{3,046}{3,066} = 0,9935.$$

Коефіцієнт кореляції показує, що між коефіцієнтом злочинності і рівнем безробіття спостерігається тісний (сильний) зв'язок.

Коефіцієнт детермінації $r^2 = 0,9935^2 = 0,9870$ показує, що 98,7 % загального варіювання коефіцієнта злочинності зумовлено відмінно-

стями в рівні безробіття, а решта 1,3 % (100-98,7) – іншими факторами, які в даному випадку не було враховано.

Коефіцієнт кореляції можна знайти і за іншими формулами:

$$1)\ r = \frac{n\sum yx - \sum y \sum x}{\sqrt{\left[n\sum y^2 - \left(\sum y\right)^2\right]\left[n\sum x^2 - \left(\sum x\right)^2\right]}} =$$

$$= \frac{10 \cdot 506,46 - 85 \cdot 56}{\sqrt{\left(10 \cdot 776,36 - 85^2\right)\left(10 \cdot 332,84 - 56^2\right)}} = 0,9935.$$

$$2)\ r = b\frac{\sigma_x}{\sigma_y} = 1,5832 \cdot \frac{1,3871}{2,2104} = 0,9935.$$

Таким чином, одержано такі самі результати, що й за основною формулою.

9.4. Криволінійна кореляція

Дослідження форми зв'язку інколи зумовлює потребу використання нелінійних (криволінійних) рівнянь регресії. Це пояснюється тим, що взаємодія між ознаками, що характеризують окремі явища і процеси, нерідко має більш складний характер, ніж просто пропорційні залежності.

Характерною особливістю цього зв'язку є те, що рівномірна зміна однієї ознаки супроводжується нерівномірною зміною (збільшенням або зменшенням) значення іншої ознаки.

Прикладом криволінійного зв'язку в соціально-правових явищах є зв'язок між злочинністю і віком правопорушників, про що згадувалось вище.

При дослідженні криволінійних зв'язків, так само як і при вивченні лінійних зв'язків, принципове значення має вибір форми і рівняння зв'язку, яке найточніше відобразить наявний зв'язок. Для розв'язання цього завдання використовуються ті самі прийоми, що й при обґрунтуванні лінійного зв'язку. При цьому особлива увага належить графічному методу.

Криволінійні форми зв'язку досить різноманітні. В статистичному аналізі найчастіше використовують параболу другого порядку, гіперболу і степеневу функцію.

При криволінійній залежності система рівнянь будується так само, як і для лінійного зв'язку: вихідне рівняння множиться на коефіцієнти при невідомих і добутки підсумовуються почленно. Так, система рівнянь для параболи другого порядку

$$\tilde{y}_x = a + bc + cx^2$$

має вигляд:

$$\Sigma y = an + b\Sigma x + c\Sigma x^2;$$

$$\Sigma yx = a\Sigma x + b\Sigma x^2 + c\Sigma x^3;$$

$$\Sigma yx^2 = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4.$$

Однією з особливостей параболи другого порядку є те, що вона завжди має точку перегину (критичну точку), яка характеризує оптимальний варіант розміру величини результативної ознаки і змінює свій напрям тільки один раз. Якщо в рівнянні параметр a_1 виражений додатним числом, а параметр – a_2 від'ємним, то крива змінює напрям із зростання на зниження.

Система рівнянь для гіперболи

$$\tilde{y}_x = a + b\frac{1}{x}$$

має вигляд

$$\begin{cases} \Sigma y = an + b\Sigma\frac{1}{x}; \\ \Sigma y\frac{1}{x} = a\Sigma\frac{1}{x} + b\Sigma\frac{1}{x^2}. \end{cases}$$

Формули, які випливають із розв'язання цієї системи рівнянь, для визначення параметрів гіперболи мають вигляд:

$$a = \frac{\Sigma y\Sigma\frac{1}{x^2} - \Sigma\frac{y}{x}\Sigma\frac{1}{x}}{n\Sigma\frac{1}{x^2} - \Sigma\frac{1}{x}\Sigma\frac{1}{x}}; \qquad b = \frac{n\Sigma\frac{y}{x} - \Sigma y\Sigma\frac{1}{x}}{n\Sigma\frac{1}{x^2} - \Sigma\frac{1}{x}\Sigma\frac{1}{x}}.$$

Щоб полегшити обчислення параметрів рівнянь регресії способом найменших квадратів при криволінійній залежності вибране рівняння регресії доцільно звести до лінійного вигляду відповідними перетвореннями.

Процес перетворень нелінійних рівнянь регресії в лінійні називають **лінеаризацією.**

Покажемо на прикладі трьох нелінійних функцій, найчастіше застосовуваних при вивченні взаємозв'язків, перетворення до лінійного вигляду.

1. Гіперболу $\tilde{y}_x = a + \frac{b}{x}$ зводять до лінійного вигляду замінивши x новою змінною (її зворотним значенням $z = \frac{1}{x}$);

$$\tilde{y}_x = a + bz;$$

2. Параболу другого порядку $\tilde{y}_x = a + bx + cx^2$ перетворюють замінивши квадрат значень факторної ознаки ($z = x^2$). Одержимо лінійну функцію з двох змінних:

$$\tilde{y}_x = a + bx + cz;$$

3. Степеневу $\tilde{y}_x = ax^b$ зводять до лінійного вигляду логарифмуванням

$$\lg y = \lg a + b\Sigma\lg x.$$

Подальші розрахунки аналогічні розрахункам лінійної функції.

Система рівнянь має вигляд :

$$\begin{cases} \Sigma\lg y = n\lg a + b\Sigma\lg x; \\ \Sigma\lg y\lg x = \lg a\Sigma\lg x + b\Sigma(\lg x)^2. \end{cases}$$

Формули для визначення параметрів степеневої функції

$$\lg a = \frac{\Sigma\lg y(\Sigma\lg x)^2 - \Sigma\lg y\lg x\Sigma\lg x}{n\Sigma(\lg x)^2 - \Sigma\lg x\Sigma\lg x};$$

$$\lg b = \frac{n\Sigma\lg y\lg x - \Sigma\lg y\Sigma\lg x}{n\Sigma(\lg x)^2 - \Sigma\lg x\Sigma\lg x}.$$

На відміну від прямолінійної залежності коефіцієнти регресії криволінійної регресії не можна інтерпретувати однозначно, так як швидкість зміни результативної ознаки при різному значенні фактора буде неоднаковою. Наприклад, якщо залежність злочинності від віку правопорушників, яка характеризується тим, що із зміною віку спочатку злочинність зростає, а потім поступово знижується, виразити рівнянням параболи другого порядку $\tilde{y}_x = a + bx + cx^2$, то коефіцієнт a_1 покаже швидкість приросту злочинності, а a_2 – її уповільнення.

Порядок визначення показників зв'язку при криволінійній залежності розглянемо на такому прикладі.

В області вивчалась ураженість злочинністю окремих вікових груп населення. В результаті дослідження одержані такі дані про коефіцієнт ураженості злочинністю окремих вікових груп населення (число злочинів на 1000 чоловік населення даного віку) і вік злочинців (табл., 9.2).

Таблиця 9.2

Дані для розрахунку показників кореляційного зв'язку

№ п/п	Коефіцієнт злочинності, злочинів	Вік злочинців	Розрахункові дані						Очікуване значення коефіцієнта злочинності, злочинів
	y	x	x^2	x^3	x^4	yx	yx^2	y^2	$\tilde{y}_x$
1	2,1	14	196	2744	38416	29,4	411,6	4,41	2,77
2	3,0	15	225	3375	50625	45,0	675,0	9,00	3,94
3	4,5	16	256	4096	65536	72,0	1152,0	20,25	5,04
4	5,6	17	289	4913	83251	95,2	1618,4	31,36	6,08
5	7,0	18	324	5832	104976	126,0	2268,0	49,00	7,06
6	10,3	20	400	8000	160000	206,0	4120,0	106,09	8,81
7	12,6	24	576	13824	331776	302,4	7257,6	158,76	11,52
8	14,8	27	729	19683	531441	399,6	10789,2	219,04	12,95
9	16,7	30	900	27000	810000	501,0	15030,0	278,89	13,58
10	13,0	34	1156	39304	1336336	442,0	15028,0	169,0	13,67
11	10,2	36	1296	46656	1679616	367,2	13219,2	104,04	13,34
12	9,3	40	1600	64000	2560000	372,0	14880,0	86,49	11,85
13	7,4	45	2025	91125	4100625	333,0	14985,0	54,76	8,60
14	5,9	49	2401	117649	5764801	289,1	14165,0	34,81	4,66
15	5,0	старше 50	2500	125000	6250000	250,0	12500,0	25,00	3,53
Разом	**127,4**	**435**	**14873**	**573201**	**23867399**	**3829,9**	**128099,9**	**1350,90**	**127,4**

Потрібно встановити форму зв'язку між двома ознаками, визначити параметри рівняння регресії, очікувані значення коефіцієнта злочинності для різного віку злочинців і тісноту зв'язку.

Для визначення форми зв'язку між коефіцієнтом злочинності (y) і віком злочинців (x) побудуємо графік – кореляційне поле (рис. 9.2).

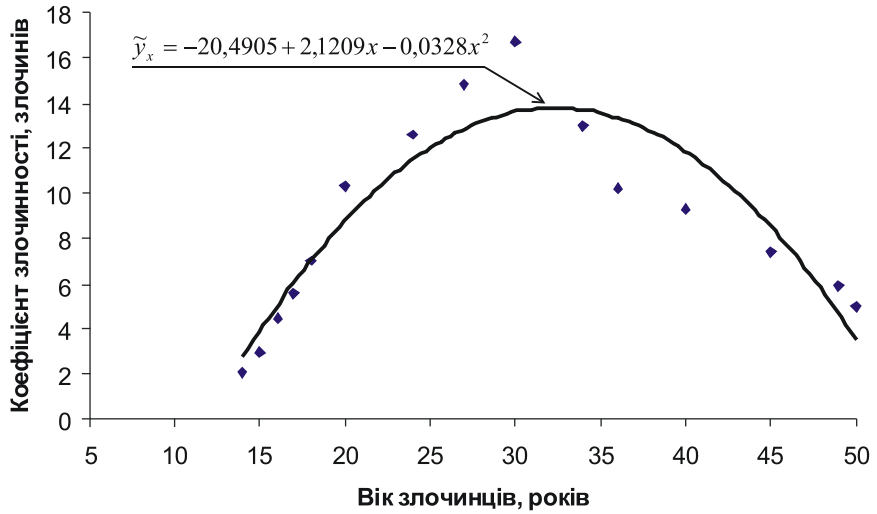

Рис. 9.2. Кореляційне поле залежності коефіцієнта злочинності від віку злочинців

З графіка видно, що між коефіцієнтом злочинності і віком злочинців зв'язок нелінійний. Коефіцієнт злочинності зростає у міру зростання віку злочинців до 30 років, а потім знижується. Розташування точок на кореляційному полі показує, що зв'язок між коефіцієнтом злочинності і віком злочинців можна виразити рівнянням параболи другого порядку:

$$\tilde{y}_x = a + bx + cx^2,$$

де $\tilde{y}_x$ - коефіцієнт злочинності, злочинів; x – вік злочинців; a, b, c – параметри рівняння.

Для визначення параметрів рівняння регресії a, b, c складемо систему рівнянь, для цього послідовно перемножимо всі члени вихідного рівняння на коефіцієнти при невідомих, а знайдені добутки підсумуємо:

$$\begin{cases}\sum y = an + b\sum x + c\sum x^2;\\ \sum yx = a\sum x + b\sum x^2 + c\sum x^3;\\ \sum yx^2 = a\sum x^2 + b\sum x^3 + c\sum x^4.\end{cases}$$

Усі потрібні для розв'язку системи нормальних рівнянь дані (Σy; Σx; Σx^2; Σx^3; Σx^4; Σyx; Σyx^2; Σy^2) розрахуємо в таблиці 9.2.

Підставимо одержані дані в систему рівнянь:

$$\begin{cases}127{,}4 = 15a + 435b + 14873c; & (1)\\ 3829{,}9 = 435a + 14873b + 573201c; & (2)\\ 128099{,}9 = 14873a + 573201b + 23867399c; & (3)\end{cases}$$

Розв'яжемо систему рівнянь і знайдемо коефіцієнти регресії ***a, b, c***:

а) поділимо всі члени рівняння на коефіцієнти при а (перше на 15, друге – на 435, третє – на 14873):

$$\begin{cases}8{,}4933 = a + 29{,}0000b + 991{,}5333c; & (4)\\ 8{,}8044 = a + 34{,}1908b + 1317{,}7034c; & (5)\\ 8{,}6129 = a + 38{,}5397b + 1604{,}7468c; & (6)\end{cases}$$

б) віднімемо з 5-го рівняння 4-е і із 6-го рівняння 5-е, в результаті одержимо систему рівнянь з двома невідомими:

$$\begin{cases}0{,}3111 = 5{,}1908b + 326{,}1701c; & (7)\\ -0{,}1915 = 4{,}3489b + 287{,}0434. & (8)\end{cases}$$

в) поділимо обидва рівняння на коефіцієнти при ***b***:

$$\begin{cases}0{,}0599 = b + 62{,}8362c; & (9)\\ -0{,}0440 = b + 66{,}0037c. & (10)\end{cases}$$

г) віднімемо з 9-го рівняння 10-е:

$$0{,}1039 = -3{,}1675c$$

звідси $\boldsymbol{c} = -0{,}0328$;

д) підставимо значення ***c*** в рівняння 9 і знайдемо коефіцієнт регресії ***b***:

$$0{,}0599 = b + 62{,}8362(-0{,}0328);$$

$$0{,}0599 = b - 2{,}06103;$$

$$b = 2{,}1209 \text{ злочину.}$$

е) визначимо коефіцієнт регресії ***a***, підставивши значення ***b*** і ***c*** в перше рівняння:

$$127{,}4 = 15a + 435 \cdot 2{,}1209 + 14{,}873(-0{,}0328);$$

$$a = -20{,}4905.$$

Перевіримо правильність обчислення коефіцієнтів регресії за такою формулою:

$$\bar{y} = a + bx + c\bar{x}^2,$$

де $\bar{y} = \dfrac{\sum y}{n} = \dfrac{127{,}4}{15} = 8{,}4933$; $\bar{x} = \dfrac{\sum x}{n} = \dfrac{435}{15} = 29{,}0$;

$$\bar{x}^2 = \frac{\sum x^2}{n} = \frac{14873}{15} = 991{,}533.$$

$$8{,}4933 = -20{,}4905 + 2{,}1209 \cdot 29{,}0 + (-0{,}0328) \cdot 991{,}5333;$$

$$8{,}4933 = 8{,}4933.$$

Отже, рівняння регресії, яке характеризує зв'язок між коефіцієнтом злочинності і віком злочинців, має вигляд:

$$\tilde{y}_x = -20{,}4905 + 2{,}1209x - 0{,}0328x^2$$

Коефіцієнт регресії $b = 2{,}1209$ показує, що у міру зростання віку злочинців до 30 років (див. графік і очікуване значення коефіцієнта злочинців - $\tilde{y}_x$) коефіцієнт злочинності збільшується на 2,1209 злочину потім із збільшенням віку злочинців рівень злочинності зменшується. Про це свідчить коефіцієнт регресії $c = -0{,}0328$ злочину, який показує уповільнення приростів коефіцієнта злочинності.

Оптимальне значення фактора визначимо за формулою:

$$-\frac{b}{2c} = \frac{2{,}1209}{2(-0{,}0328)} = 32 \text{ роки.}$$

Визначимо очікувані (розрахункові) значення коефіцієнту злочинності для різного віку злочинців ($\tilde{y}_x$).

Для цього до рівняння регресії замість x (вік злочинців) підставимо його конкретні значення $x = 14, 15, 16, \ldots, 50$ років. Так, очікуване значення коефіцієнта злочинності для злочинців у віці 14 років становить:

$$\tilde{y}_{x=14} = -20{,}4905 + 2{,}1209 \cdot 14 - 0{,}0328 \cdot 14^2 = 2{,}77 \text{ злочину;}$$

для злочинців у віці 15 років

$$\tilde{y}_{x=15} = -20{,}4905 + 2{,}1209 \cdot 15 - 0{,}0328 \cdot 15^2 = 3{,}94 \text{ злочину і т.д.}$$

Результати розрахунків запишемо в останню колонку таблиці 9.2.

Перевіримо правильність розрахунків:

$$\sum y = \sum \tilde{y}_x \; ; \; 127{,}4 = 127{,}4 \, .$$

За очікуваними значеннями коефіцієнту злочинності на рис. 9.2 побудуємо теоретичну лінів регресії.

Визначимо тісноту зв'язку між коефіцієнтом злочинності і віком злочинців, для чого розрахуємо індекс кореляції

$$i_r = \sqrt{\frac{\sigma^2_{м.гр}}{\sigma^2_{заг}}} = \sqrt{\frac{a\sum y + b\sum yx + c\sum yx^2 - n\bar{y}^2}{\sum y^2 - n\bar{y}^2}} =$$

$$= \sqrt{\frac{-20{,}4905 \cdot 127{,}4 + 2{,}1209 \cdot 3829{,}9 + (-0{,}0328) \cdot 128099{,}9 - 15 \cdot 8{,}493^2}{1350{,}9 - 15 \cdot 8{,}493^2}} =$$

$$= \sqrt{\frac{228{,}61}{268{,}85}} = \sqrt{0{,}8503} = 0{,}9221$$

Такий самий результат може бути одержаний і за іншою формулою (через залишкову дисперсію):

$$i_r = \sqrt{1 - \frac{\sigma^2_{в.гр}}{\sigma^2_{заг}}} = \sqrt{1 - \frac{\sum y^2 - a\sum y - b\sum yx - c\sum yx^2}{\sum y^2 - n\bar{y}^2}} =$$

$$= \sqrt{1 - \frac{1350{,}9 - (-20{,}4905) \cdot 127{,}4 - 2{,}1209 \cdot 3829{,}9 - (-0{,}0328) \cdot 128099{,}9}{1350{,}9 - 15 \cdot 8{,}493^2}} =$$

$$= \sqrt{1 - \frac{40{,}24}{268{,}85}} = \sqrt{1 - 0{,}1497} = \sqrt{0{,}8503} = 0{,}9221$$

Коефіцієнт кореляції показує, що між коефіцієнтом злочинності і віком злочинців є тісний зв'язок. Коефіцієнт детермінації ($i_r^2 = 0{,}9221 = 0{,}8503$) показує, що 85,03% відмінностей у коефіцієнтах злочинності пов'язані з віком злочинців, а решта 14,97% - з іншими факторами, дію яких у даному випадку не було враховано.

9.5. Множинна кореляція

Рівень результативних показників соціально-правових явищ формується під впливом цілого комплексу взаємопов'язаних між собою факторів, які діють з різною силою і з різною спрямованістю. Тому на

практиці найчастіше доводиться вивчати взаємозв'язки між кількома ознаками одночасно.

Особливе значення у вивченні взаємозв'язків між ознаками в дослідженні соціально-правових явищ належить багатофакторному кореляційно-регресійному аналізу, при якому визначається залежність результативної ознаки від кількох факторів одночасно.

Використання ЕОМ і типових програм кореляційно-регресійного аналізу дає змогу розв'язувати кореляційні моделі різних залежностей і вибрати з цієї множини таке рівняння, яке найточніше описує ступінь наближення фактичних даних до теоретичних і відповідно дає найменшу суму квадратів відхилень фактичних даних від розрахованих за рівнянням зв'язку.

Багатофакторний кореляційно-регресійний аналіз може бути застосований для:

1) розрахунку очікуваних (теоретичних) значень результативної ознаки;

2) зіставлення і оцінки фактичного і розрахункового значень результативної ознаки;

3) порівняльного аналізу різних сукупностей;

4) об'єктивної оцінки результатів роботи установ, організацій і підприємств;

5) виявлення резервів виробництва;

6) розроблення нормативів;

7) прогнозування суспільних явищ тощо.

Парна кореляція, в силу того, що разом з досліджуваним фактором на результативну ознаку впливають й інші фактори не завжди дає правильне уявлення про зв'язок між результативною і факторною ознакою (перебільшує або применшує міру залежності). Перевага багатофакторного кореляційно-регресійного аналізу порівняно з простою кореляцією полягає в тому, що він дає змогу оцінити ступінь впливу на результативну ознаку кожного з включених у модель (рівняння) факторів при фіксованому положенні (звичайно на середньому рівні) решти факторів.

Методологія множинної кореляції ґрунтується на загальних принципах кореляційного аналізу. Водночас в ній ускладнюється змістовний аналіз, зростає складність математичного апарату.

При формуванні множинної кореляційної моделі необхідно враховувати ряд обмежень, пов'язаних з відбором, кількістю і взаємозв'язком факторів, вибором форми зв'язку (рівняння регресії).

Відбір найістотніших факторів до кореляційної моделі є одним з найбільш важливіших і принципових завдань багатофакторного коре-

ляційно-регресійного аналізу. Природно, що всі фактори, які впливають на досліджувану результативну ознаку, до рівняння регресії включити не можна. З усього комплексу таких факторів необхідно відібрати найбільш важливі, істотні. Захоплення великою кількістю факторів при відносно невеликій чисельності сукупності може призвести до неякісних результатів. Крім того, із збільшенням в рівнянні регресії кількості параметрів значно утруднюється інтерпретація одержаних результатів.

Велику роль у відборі факторів відіграють завчасно побудовані і проаналізовані факторні групування. Дуже важливого значення тут набувають комбінаційні групування, які дозволяють визначити вплив на результативну ознаку фактора, що цікавить дослідника, при фіксованих значеннях інших факторів. **Можна зробити безперечний висновок про те, що статистичні групування становлять основу для кореляційного і дисперсійного аналізу і найбільшої ефективності останні досягають в поєднанні з методом групувань.**

Практичні розрахунки показують, що для забезпечення стійкості параметрів рівняння зв'язку, кількість факторів. включених до моделі, має бути в 6 – 8 разів меншою від чисельності досліджуваної сукупності. При цьому сукупність, з якої відбирають фактори, повинна бути якісно однорідною.

Відбираючи фактори, потрібно виключати ті, що взаємно дублюють один одного і перебувають у функціональному зв'язку. Функціональний або близький до нього зв'язок між самими факторами вказує на мультиколінеарність (для двох – колінеарність). Наявність мультиколінеарності свідчить про те, що ці фактори відображають ту саму сторону впливу на результативну ознаку.

При високій корельованості факторів (тіснота зв'язку між двома факторами перевищує $r > 0{,}8$) вплив одного з них акумулює і вплив другого. Одержані при цьому кореляційні моделі стають нестійкими.

При формуванні кореляційної моделі до неї потрібно включити один з цих факторів, який істотніше впливає на результативну ознаку. При мультіколінеарності включення до кореляційної моделі взаємопов'язаних факторів можливе тоді, коли тіснота зв'язку між ними менша, ніж тіснота зв'язку результативної ознаки з кожним фактором. Потрібно, щоб кореляційна модель містила незалежні і такі, що не дублюють один одного, фактори. Небажаним є включення до однієї моделі часткових і загальних факторів. Повністю слід виключити фактори, функціонально пов'язані з результативною ознакою.

Важкою і складною проблемою побудови рівняння множинної регресії є також вибір функції зв'язку, тобто вибір математичного рівняння, яке найповніше проявляє характер взаємозв'язку між результативною ознакою і включеними до рівняння регресії факторами.

Одна із складностей полягає у взаємозв'язку і взаємодії факторів між собою та з результативною ознакою. Тому звичайні прийоми, використовувані при виборі форми зв'язку при парній кореляції (графічний та ін.) тут мало прийнятні.

Вибір рівняння регресії може спиратися на положення теорії досліджуваного явища або практичний досвід попередніх досліджень. Якщо таких даних немає, то допомогти у вирішенні цього питання може побудова комбінаційних групувань, таблиць розподілу чисельностей, експертні оцінки, вивчення парних зв'язків між результативною ознакою і кожним фактором, графіки, перебирання функцій різних типів (при розв'язанні задач на ЕОМ), послідовний перехід від лінійних рівнянь зв'язку до більш складних видів тощо.

Виконання усіх цих прийомів пов'язане із значною кількістю зайвих підрахунків. Тому, приймаючи до уваги, що кореляційні зв'язки в більшості випадків відображаються функціями лінійного типу або степеневими, які шляхом логарифмування або заміни змінних можна звести до лінійного вигляду, рівняння множинної регресії можна будувати у лінійній формі.

При n змінних лінійне рівняння має вигляд:

$$\tilde{y}_x = a_0 + a_1 x_1 + a_2 x_2 + ... + a_n x_n,$$

де $\tilde{y}_x$ – залежна змінна (результативна ознака);

x_i – незалежні змінні (фактори);

a_1 – незалежні змінні (фактори);

a_0 – початок відліку, який економічного смислу немає;

$a_1, a_2, ..., a_n$ – коефіцієнти регресії.

Рівняння, за допомогою якого виражається кореляційний зв'язок між кількома ознаками називають **рівнянням множинної регресії**. Параметри рівняння регресії, так само як і у випадку парної кореляції, знаходять способом найменших квадратів.

Коефіцієнти множинної регресії показують ступінь середньої зміни результативної ознаки при зміні відповідної факторної ознаки на одиницю (одне своє значення) за умови, що всі інші фактори, які включені до рівняння регресії, залишаються постійними (фіксованими) на одному (звичайно середньому) рівні.

Коефіцієнти множинної регресії, які характеризують зв'язок між результативною ознакою і фактором при фіксованому значенні інших

факторів, називаються **коефіцієнтами чистої регресії,** а коефіцієнти парної регресії – **коефіцієнтами повної регресії.**

Коефіцієнти чистої регресії, що мають різний фізичний смисл і одиниці вимірювання не дають чіткого уявлення про те, які саме фактори найістотніше впливають на результативну ознаку. Крім того, величина коефіцієнтів регресії залежить від ступеня варіації ознаки.

Щоб привести коефіцієнти чистої регресії до порівнянного вигляду, їх виражають у стандартизованій формі у вигляді коефіцієнтів еластичності (E) і бета-коефіцієнтів (β).

Коефіцієнти еластичності показують, на скільки процентів змінюється величина результативної ознаки при зміні відповідного фактора на один процент при фіксованому значенні інших факторів.

Коефіцієнти еластичності і коефіцієнти чистої регресії зв'язані між собою таким відношенням:

$$E_i = a_i \frac{\bar{x}_i}{\bar{y}},$$

де a_i – коефіцієнт чистої регресії при i-му факторі;

$\bar{x}_i$ і $\bar{y}$ – середні значення відповідно i-го фактора і результативної ознаки.

Бета-коефіцієнти показують, на скільки середньоквадратичних відхилень σ_y зміниться результативна ознака при зміні відповідного фактора на одне значення середньоквадратичного відхилення σ_x (при постійності інших факторів, включених до рівняння регресії).

Бета-коефіцієнти обчислюються за формулою:

$$\beta_i = a_i \frac{\sigma_{x_i}}{\sigma_y},$$

де a_i – коефіцієнт чистої регресії при i-му факторі;

σ_{x_i} і σ_y – середні квадратичні відхилення відповідно по i-му фактору і результативній ознаці.

З наведеної формули випливає, що бета-коефіцієнти мають той самий знак (плюс, мінус), що й коефіцієнти чистої регресії.

По суті бета-коефіцієнти характеризують фактори, у розвитку яких приховуються найбільші резерви поліпшення результативної ознаки.

При парному лінійному зв'язку коефіцієнт кореляції являє собою бета-коефіцієнт:

$$r = b\frac{\sigma_x}{\sigma_y} = \beta.$$

Як зазначалося вище, коефіцієнт множинної детермінації (R^2) показує, яка частина загальної варіації результативної ознаки визначається варіацією факторів, включених до кореляційної моделі. Щоб визначити частку впливу кожного фактора у загальній варіації, треба знайти добуток парних коефіцієнтів кореляції (r_{yx_i}) на відповідні бета-коефіцієнти (β_i), а одержані по всіх факторах результати підсумувати ($\sum r_{yx_i}\beta_i$).

Якщо потрібно частку впливу кожного фактора визначити у процентах, то знайдені коефіцієнти множать на сто процентів.

Порядок визначення і аналізу показників зв'язку при множинній кореляції розглянемо на такому прикладі.

При побудові системи взаємопов'язаних факторних і результативних групувань у розділі 5 було використано статистичний матеріал щодо рівня злочинності та її факторів у 30 містах. За допомогою статистичних групувань було встановлено, що відмінності в рівні злочинності в основному пов'язані з різним рівнем факторів, які детермінують злочинність.

Для розв'язання задачі до кореляційної моделі включимо такі ознаки: 1) y – коефіцієнт злочинності (на 1000 чол. населення); 2) x_1– рівень безробіття (%); x_2– вжито алкоголю на душу населення, л/рік.

Вихідні дані подамо у вигляді матриці (табл. 9.3).

Попереднє вивчення форми залежності між вказаними ознаками показало, що зв'язок може бути виражений за допомогою лінійного рівняння регресії:

$$\tilde{y}_x = a_0 + a_1x_1 + a_2x_2.$$

Таблиця 9.3

Матриця вихідних даних для багатофакторного кореляційно-регресійного аналізу рівня злочинності

№ п/п	Коефіцієнт злочинності (на 1000 чол. населення), злочинів	Рівень безробіття, %	Вжито алкоголю на душу населення, л/рік
	y	x_1	x_2
1	2	3	4
1	8,1	5,3	4,3
2	8,2	5,5	4,4
3	8,3	6,0	4,0
4	8,4	6,1	4,9

Продовження таблиці 9.3

1	2	3	4
5	8,5	5,8	5,1
6	8,5	5,7	5,2
7	9,0	6,2	5,5
8	9,1	6,4	6,0
9	9,3	6,5	6,3
10	9,5	6,0	6,6
11	9,7	7,1	6,0
12	9,8	7,2	7,1
13	10,2	7,0	7,5
14	10,8	8,1	8,0
15	11,1	8,2	8,1
16	12,0	9,0	7,2
17	12,1	7,6	8,3
18	13,0	7,7	8,4
19	13,5	8,1	8,0
20	14,0	8,2	8,1
21	14,1	8,5	9,0
22	14,8	8,6	8,8
23	15,0	9,1	8,7
24	15,7	10,2	9,1
25	15,9	11,2	10,3
26	16,0	9,6	10,0
27	17,1	10,3	9,9
28	18,7	10,5	9,8
29	19,0	10,0	8,9
30	20,4	11,3	10,2

Розв'язавши рівняння множинної регресії і обчисливши інші показники кореляційного зв'язку на ЕОМ, одержимо таку машинограму:

1. Коефіцієнт рівняння регресії $\mathbf{a_i}$	***2. Середня помилка коефіцієнта регресії*** μ_{a_i}	***3.* t *– коефіцієнт***
-1,3062		
1,1260	*0,3083*	*3,6518*
0,6291	*0,2950*	*2,1320*
4. Парний коефіцієнт кореляції r_{yx_i}	***5.* β *- коефіцієнт*** β_i	***6. Коефіцієнт еластичності*** $\mathbf{E}_i$
0,9554	*0,6145*	*0,7263*
0,9427	*0,3588*	*0,3853*

7. Середні значення $\overline{y}$ і $\overline{x}$	***8. Середнє квадратичне відхилення*** σ_i	***9. Кореляційна матриця***
11,6933	*4,0142*	*1 0,9554 0,9427*
7,5433	*2,1911*	*1 0,9500*
7,1633	*2,2897*	*1*
10. Загальний коефіцієнт кореляції **R**	***11. Коефіцієнт детермінації*** $\mathbf{R^2}$	***12. Різниця*** $y - \widetilde{y}_x$
0,9620	*0,9255*	*1. 8,10 7,366 0,734* *2. 8,20 7,654 0,546* *3. 8,30 7,965 0,335* *... ... ...* *30. 20,4 11,300 9,100*

У добутій машинограмі пронумеруємо її стовпці з 1-го по 12-й і проаналізуємо знайдені результати.

На ЕОМ матимемо таку кореляційну залежність рівня злочинності від включених до моделі факторів (1-й стовпець машинограми):

$$\widetilde{y}_x = -1{,}3062 + 1{,}1260x_1 + 0{,}6291x_2.$$

Оцінка коефіцієнтів регресії. Подальший аналіз пов'язаний з перевіркою значущості коефіцієнтів регресії. Для цього визначимо табличне значення *t* - критерію нормального розподілу (n 30) і порівняємо його з фактичними значеннями (3-й стовпець машинограми). Табличне значення *t* – критерію нормального розподілу при заданому рівні надійної ймовірності $P = 0{,}95$ становитиме $t = 1{,}96$ (дод. 16).

Відповідні фактичні значення нормованих відхилень для коефіцієнтів регресії такі:

$$t_{a_1} = 3{,}6518;\ \ t_2 = 2{,}1320.$$

Фактичні значення коефіцієнтів *t* вище табличного значення ($t = 1{,}96$). Тому наведене вище рівняння регресії можна використати для подальшого аналізу.

Коефіцієнти регресії показують на скільки зміниться рівень (коефіцієнт) злочинності у разі зміни кожного фактора на одиницю його виміру при фіксованих значеннях інших факторів, включених до рівняння. Так, збільшення рівня безробіття на 1% підвищує рівень злочинності на 1,1260 злочину, збільшення споживання алкоголю на душу населення на 1 літр в рік – на 0,6291 злочину.

Аналіз коефіцієнтів кореляції. Коефіцієнт множинної кореляції (10-й стовпець машинограми), який характеризує тісноту зв'язку між рівнем

злочинності та її факторами дорівнює 0,9620. Коефіцієнт множинної детермінації (11-й стовпець машинограми) $R^2 = 0{,}9620^2 = 0{,}9255$ показує, що варіація рівня злочинності в зв'язку зі зміною розглядуваних факторів, становить 92,55%.

Тісноту зв'язку між ознаками, включеними до рівняння регресії характеризують 4-й і 9-й стовпці машинограми і складена на їх основі така матриця парних коефіцієнтів кореляції:

	y	x_1	x_2
y	1,0000	0,9554	0,9427
x_1		1,0000	0,9500
x_2			1,0000

З даних таблиці видно, що рівень злочинності перебуває в досить тісному зв'язку з включеними до моделі факторами. Так, тіснота зв'язку між рівнем злочинності і безробіттям становить $r_{yx_1} = 0{,}9554$, між кількістю вжитого алкоголю $r_{yx_2} = 0{,}9427$.

Аналіз коефіцієнтів еластичності і β-коефіцієнтів. Найбільше впливає на рівень злочинності, якщо робити висновок за наведеним рівнянням регресії, рівень безробіття, тому що коефіцієнт регресії при цьому коефіцієнті найбільший ($a_1 = 1{,}1260$) і менше впливає кількість вжитого алкоголю ($a_2 = 0{,}6291$).

Однак, коефіцієнти регресії, що мають різний фізичний смисл і одиниці вимірювання, не дають чіткого уявлення про те, які фактори найістотніше впливають на злочинність. Для проведення такого аналізу на ЕОМ визначено коефіцієнти еластичності, які показують на скільки процентів зміниться величина результативної ознаки у разі зміни величини фактора на 1% при фіксованому значенні інших факторів (6-й стовпець машинограми).

На підставі обчислених коефіцієнтів еластичності $E_1 = 0{,}7263$; $E_2 = 0{,}3853$ можна зробити висновок, що збільшення на 1% рівня безробіття веде до збільшення рівня злочинності на 0,7263%, на 1% вжитого алкоголю – на 0,3853%. Таким чином, найбільший вплив на рівень злочинності має рівень безробіття.

Проте і цих даних недостатньо, щоб скласти об'єктивне уявлення проте, як по групі досліджуваних міст розподіляються фактори впливу на зростання злочинності.

З цією метою на ЕОМ обчислено β - коефіцієнти, які показують, на скільки середньоквадратичних відхилень σ_y зміниться результативна ознака (злочинність) при зміні відповідного фактора на одне значення свого середньоквадратичного відхилення. По суті β – коефіцієнти характеризують фактори, в розвитку яких приховано найбільші причини збільшення результативної ознаки (злочинності).

Фактичні значення β - коефіцієнтів (5-й стовпець машинограми) такі:

$$\beta_1 = 0{,}6145;\ \beta_2 = 0{,}3588.$$

У розрахованій нами моделі найбільші можливості збільшення рівня злочинності закладено в рівні безробіття ($\beta_1 = 0{,}6145$), тому що при зміні на одне середнє квадратичне відхилення рівня безробіття рівень злочинності змінюється на 0,6145 свого середнього квадратичного відхилення. Далі за ступенем впливу йде такий фактор як вжита кількість алкоголю на душу населення ($\beta_2 = 0{,}3588$).

Розкладання загальної варіації. Коефіцієнт множинної детермінації, який дорівнює $R^2 = 0{,}9255$, свідчить про те, що коливання злочинності пояснюване варіацією включених до рівняння регресії факторів, дорівнює 92,55%.

Становить інтерес розкладання загального обсягу варіації коефіцієнта злочинності за рахунок кожного включеного до рівняння регресії фактора. Для цього визначимо **коефіцієнти детермінації,** які розраховують як добуток парних коефіцієнтів кореляції на β - коефіцієнти за відповідними факторами (4-й і 5-й стовпці машинограми).

Усі розрахунки зведемо в табл. 9.4.

Таблиця 9.4

Розкладання загального обсягу варіації за факторами

№ п/п	Фактор x_i	Парний коефіцієнт кореляції r_{yx_i}	β-коефіцієнт β_i	Добуток, % $r_{yx_i}\beta_i$
1	x_1 – рівень безробіття, %	0,9554	0,6145	58,72
2	x_2 – вжито алкоголю, л/рік	0,9427	0,3588	33,83
Разом	–	–	–	**92,55**

Таким чином, із 92,55% загального коливання рівня злочинності 58,72% пояснюється варіацією рівня безробіття і 33,83% – кількістю

вжитого алкоголю. Найвпливовішим фактором, як показали розрахунки, є рівень безробіття.

9.6. Статистична оцінка вибіркових показників зв'язку

У тих випадках, коли вивчення кореляційної залежності базується на вибіркових даних, виникає потреба оцінки вибіркових показників кореляції (коефіцієнтів регресії і кореляції).

Статистична оцінка вибіркових показників кореляції дає змогу зробити висновок про те, наскільки вибіркові статистичні показники відповідають показникам генеральної сукупності. Однак така оцінка проводиться у випадках, коли: 1) вибірка сформована у випадковому порядку; 2) вибірка зроблена з нормально розподіленої сукупності; 3) відхилення фактичних значень результативної ознаки від її теоретичних значень, обчислених за рівнянням, також розподілені нормально.

Розглянемо порядок статистичної оцінки вибіркових показників зв'язку при парній лінійній регресії.

В кореляційному аналізі середня помилка вибірки обчислюється на основі залишкової дисперсії, оскільки ця величина характеризує точність підбору кривої до фактичних даних. Проте залишкова дисперсія, розрахована за вибірковими даними, зменшує дійсну величину залишкової дисперсії в генеральній сукупності, тобто є зміщеною оцінкою. Це зміщення коригується внесенням в знаменник формули залишкової дисперсії поправки на втрату ступенів свободи. При парній лінійній залежності втрачаються відповідно числу параметрів рівняння (*a* і *b*) дві ступені свободи, при кореляції трьох змінних з параметрами *a*, *b* і *c* - три ступені свободи і т.д.

Квадрат середньої помилки вибіркового коефіцієнта регресії являє собою відношення залишкової дисперсії, скоригованої на втрату числа ступенів свободи варіації, до суми квадратів відхилень незалежної змінної.

Позначаючи залишкову дисперсію через S^2_{yx}, а квадрат середньої помилки вибіркового коефіцієнта регресії через $\mu^2_в$, одержимо

$$\mu^2_в = \frac{S^2_{yx}}{\Sigma(x_i - \tilde{x})^2},$$

де $S^2_{yx} = S^2_{зал}\frac{n}{n-m} = \frac{\Sigma(y_i - \tilde{y}_x)^2}{n-m}$,

де m – кількість параметрів рівняння регресії, яке дорівнює двом при парній лінійній залежності; n – чисельність вибірки.

Відповідно середня помилка коефіцієнта регресії:

$$\mu_в = \sqrt{\frac{S^2_{yx}}{\Sigma(x_i - \tilde{x})^2}}.$$

Гранична помилка вибіркового коефіцієнта регресії визначається за формулою:

$$\varepsilon_в = t\mu_в,$$

де t – значення нормованого відхилення, величина якого встановлюється за таблицями. Для великих вибірок (n > 30) значення t знаходять за дод. 16, для малих вибірок (n < 30) – за дод. 17.

Довірчі межі коефіцієнта регресії у генеральній сукупності ($\boldsymbol{b_0}$) становитимуть:

$$в_0 = в \pm t\mu_в.$$

Вірогідність вибіркового коефіцієнта регресії визначається як відношення:

$$t = \frac{b}{\mu_в}.$$

Якщо $t_{факт} > t_{табл}$ при заданому рівні значущості і відповідному числі ступенів свободи варіації, то нульова гіпотеза про рівність коефіцієнта регресії у генеральній сукупності нулю ($b_0 = 0$) відхиляється і робиться висновок про те, що вибірковий коефіцієнт регресії є вірогідним. Якщо ж $t_{факт} < t_{табл}$, то нульова гіпотеза приймається і робиться висновок про те, що значення $\boldsymbol{b}$ у вибірці є неістотним, випадковим.

Обчислимо середню і граничну помилку для коефіцієнта регресії, що характеризує залежність коефіцієнта злочинності від рівня безробіття ($b = 1{,}5832$ злочину).

Визначимо залишкову дисперсію, використовуючи коефіцієнти рівняння регресії, а також дані табл. 9.1.

$$\sigma^2_{зал} = \frac{\Sigma y^2 - a\Sigma y - в\Sigma yx}{n} = \frac{776{,}36 - (-0{,}3657)85 - 1{,}5832 \cdot 506{,}46}{10} = \frac{5{,}61}{10} = 0{,}561.$$

Обчислимо скориговану залишкову дисперсію:

$$S^2_{yx} = \sigma^2_{зал}\frac{n}{n-m} = 0{,}561 \cdot \frac{10}{10-2} = 0{,}701.$$

де m – кількість параметрів рівняння регресії ($m = 2$).

Визначимо середню помилку параметра *b*:

$$\mu_в = \sqrt{\frac{S^2_{yx}}{\Sigma(x_i - \tilde{x})^2}} = \sqrt{\frac{0{,}701}{19{,}21}} = \sqrt{0{,}03649} = 0{,}1910,$$

де $\Sigma(x_i - \tilde{x})^2 = n\sigma_x^2 = 10 \cdot 1{,}3871^2 = 19{,}21,$

або $\Sigma(x_i - \tilde{x})^2 = \Sigma x^2 - n\tilde{x}^2 = 332{,}84 - 10 \cdot 5{,}6^2 = 19{,}21.$

Визначимо фактичне значення *t* – критерію Стьюдента:

$$t_{факт} = \frac{b}{\mu_в} = \frac{1{,}5832}{0{,}1910} = 8{,}332.$$

За таблицею (дод. 17) при $\alpha = 0{,}05$ і числі ступенів свободи $k = n - m = 10 - 2 = 8$ знайдемо $t_{0{,}05} = 2{,}307.$

Оскільки $t_{факт} > t_{0{,}05}$ (8,332 > 2,307), від нульової гіпотези, яка передбачає відсутність зв'язку між коефіцієнтом злочинності і рівнем безробіття в генеральній сукупності ($b_0 = 0$), слід відмовитись. Вибірковий коефіцієнт регресії $b = 1{,}5832$ є вірогідним, істотним.

Обчислимо граничну помилку вибіркового коефіцієнта регресії:

$$\varepsilon_в = t_{0{,}05}\mu_в = 2{,}307 \cdot 0{,}1910 = 0{,}4406 \text{ злочину.}$$

Визначимо інтервал, в якому із заданим рівнем значущості знаходиться коефіцієнт регресії в генеральній сукупності:

$$b_0 = b \pm \varepsilon_в = 1{,}5832 \pm 0{,}4406, \text{ або } 1{,}1426 \le b_0 \le 2{,}0238.$$

Отже з рівнем значущості $\alpha = 0{,}05$ (з імовірністю помилитись в 5 випадках із 100) можна стверджувати, що величина коефіцієнта регресії, який характеризує зв'язок між коефіцієнтом злочинності і рівнем безробіття в генеральній сукупності, перебуває в інтервалі від 1,1426 до 2,0238 злочинів на 1% безробіття.

Проведемо перевірку гіпотези та інтервальну оцінку вибіркового коефіцієнта кореляції.

Сформулюємо нульову гіпотезу про відсутність зв'язку між коефіцієнтом злочинності і рівнем безробіття, а також рівності нулю коефіцієнта кореляції в генеральній сукупності

$$H_0 : r_0 = 0; \; H_a : r_0 \neq 0.$$

Оскільки в задачі чисельність вибірки є невеликою ($n = 10$), а вибірковий коефіцієнт кореляції близький до одиниці ($r = 0{,}9935$), оцінку його істотності здійснимо за допомогою методу Р. Фішера, який дістав назву перетвореної кореляції.

Р. Фішер довів, що розподіл логарифмічної функції вибіркового лінійного коефіцієнта кореляції (z) наближується до кривої нормального розподілу навіть при невеликому обсязі вибірки і високому значенні *r*.

Величину z визначають за формулою:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r}$$

Перехід від *r* до z і назад здійснюють за допомогою спеціальних таблиць, що виключають потребу логарифмування.

Середня квадратична помилка z – розподілу залежить тільки від обсягу вибірки і визначається за формулою:

$$\mu_z = \frac{1}{\sqrt{n-3}}.$$

Обчислимо середню помилку z – розподілу для нашої задачі:

$$\mu_z = \frac{1}{\sqrt{10-3}} = \frac{1}{\sqrt{7}} = 0{,}378.$$

За таблицею (дод. 22) знайдемо, що коефіцієнту кореляції $r = 0{,}9935$ відповідає $z = 2{,}826$.

Визначимо відношення z до середньої помилки вибіркового коефіцієнта кореляції:

$$t_{факт} = \frac{z}{\mu_z} = \frac{2{,}826}{0{,}378} = 7{,}48.$$

Знайдемо табличне значення *t*-критерію Стьюдента (дод. 17) при $a = 0{,}05$ і $k = 10 - 2 = 8$; $t_{0{,}05} = 2{,}307$. Оскільки фактичне відношення *t* виявилося більше табличного $t_{0{,}05}$ (7,48 > 2,307), то можна зробити висновок про те, що висунута гіпотеза про рівність нулю коефіцієнта кореляції у генеральній сукупності не узгоджується з фактичними даними, відтак її потрібно відхилити. Вибірковий коефіцієнт кореляції є вірогідним, істотним.

Побудуємо надійний інтервал, в якому із заданим рівнем значущості знаходиться коефіцієнт кореляції в генеральній сукупності:

$r_0 = z \pm t\mu_z = 2{,}826\ 2{,}307 \cdot 0{,}378 = 2{,}826 \pm 0{,}872$, тобто від 1,954 до 3,698.

Користуючись таблицею значень z у зворотному порядку, знайдемо границі довірчого інтервалу для коефіцієнта кореляції в генеральній сукупності:

$$0{,}96 < r_0 < 0{,}99.$$

Отже, із заданим рівнем значущості $\alpha = 0{,}05$ можна стверджувати, що тіснота зв'язку між коефіцієнтом злочинності і рівнем безробіття в генеральній сукупності знаходиться в межах від 0,96 до 0,99.

Вірогідність вибіркового коефіцієнта кореляції може бути встановлена і без обчислень за таблицею Р. Фішера (дод. 23).

Для нашої задачі табличне значення коефіцієнта кореляції при $\alpha = 0{,}05$ і $k = 8$ становитиме $r_{0,05} = 0{,}632$. Оскільки $r_{факт} > r_{0,05}$ (0,9935 > 0,632), можна підтвердити попередній висновок про те, що вибірковий коефіцієнт кореляції є вірогідним. Це дає підставу для висновку про дійсний зв'язок між коефіцієнтом злочинності і рівнем безробіття в генеральній сукупності.

9.7. Непараметричні критерії оцінки кореляційного зв'язку

Наведені вище формули для визначення тісноти зв'язку між ознаками передбачають, що сукупності, до яких вони застосовуються, мають нормальний, або близький до нормального розподіл. Якщо ж характер розподілу досліджуваної сукупності навіть передбачувано невідомий, то тісноту зв'язку можна обчислити за допомогою **непараметричних критеріїв** визначення тісноти зв'язку.

Особливістю цих критеріїв є те, що тіснота зв'язку між ознаками визначається не за кількісними значеннями варіантів, а за допомогою порівняння їх рангів. Під **рангом** розуміють порядковий номер одиниці сукупності в ранжированому ряду розподілу. Чим менші розбіжності між рангами, тим тісніший зв'язок між ознаками.

До непараметричних критеріїв показників тісноти зв'язку відносяться коефіцієнти: кореляції рангів Фехнера, асоціації, контингенції та ін.

Коефіцієнт кореляції рангів – це один з найпростіших показників тісноти зв'язку (його же називають ранговим коефіцієнтом кореляції Спірмена). Суть його розрахунку полягає в такому. Парні спостереження двох взаємопов'язаних ознак (результативної і факторної) ранжируються, а потім відповідно величині ознаки їм надається ранг від 1 до n. Тіснота зв'язку визначається на основі близькості рангів і формула коефіцієнта кореляції рангів буде мати вигляд:

$$r_p = 1 - \frac{6\Sigma d^2}{n(n-1)}.$$

де d – різниці між величинами рангів в порівнюваних рядах;
n – число спостережень.

Суть його така сама як і лінійного коефіцієнта кореляції. Коефіцієнт кореляції рангів, як і лінійний коефіцієнт кореляції, може приймати значення від – 1 до + 1. Якщо ранги двох паралельних рядів повністю співпадають, то $\Sigma d^2 = 0$ і тоді має місце прямий функціональний зв'язок, а $r_p = 1$. При повному зворотному зв'язку (ранги розміщуються в зворотному порядку) $r_p = -1$. Ранжирувати обидві ознаки потрібно в одному і тому самому порядку: або від менших значень ознаки до більших, або навпаки.

Методику розрахунку коефіцієнта кореляції рангів покажемо на прикладі визначення тісноти зв'язку між суспільно корисною зайнятістю підлітків і рівнем злочинності неповнолітніх (табл. 9.5).

Таблиця 9.5

Дані для розрахунку коефіцієнта кореляції рангів

№ п/п	Коефіцієнт злочинності неповнолітніх, злочинів	Чисельність підлітків, які не навчаються і не працюють, чол.	Ранги		Різниця рангів	Квадрат різниці рангів
			за коефіцієнтом злочинності	за зайнятістю		
	y	x	R_y	R_x	$d = R_y - R_x$	d^2
1	2,0	9	1	5	-4	16
2	2,5	8	2	4	-2	4
3	2,8	6	3	2	1	1
4	3,0	5	4	1	3	9
5	3,1	7	5	3	2	4
6	3,5	11	6	7	1	1
7	3,6	10	7	6	1	1
8	3,8	13	8	9	-1	1
9	4,0	16	9	10	-1	1
10	4,5	12	10	8	2	4
11	5,0	18	11	12	-1	1
12	5,5	19	12	13	1	1
13	6,0	17	13	11	2	4
14	7,2	21	14	15	-1	1
15	8,3	20	15	14	1	1
Разом	–	–	–	–	0	50

Визначимо тісноту зв'язку між коефіцієнтом злочинності і зайнятістю підлітків, обчисливши коефіцієнт кореляції рангів:

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \cdot 50}{15(15^2-1)} = 1 - \frac{300}{3360} = 0{,}911.$$

Величина коефіцієнта кореляції рангів свідчить про тісний зв'язок між рівнем злочинності і суспільно корисною зайнятістю неповнолітніх.

Вірогідність коефіцієнта кореляції рангів можна перевірити за таблицею Фішера (дод. 23). Табличне значення коефіцієнта кореляції при $\alpha = 0{,}05$ і $k = n - m = 15 - 2 = 13$ становить $r_p = 0{,}514$. Оскільки $r_{\text{факт}} > r_{0{,}05}$ (0,911 > 0,514), можна зробити висновок про те, що вибірковий коефіцієнт кореляції рангів є вірогідним.

Недоліком коефіцієнта кореляції рангів є те, що однаковим різницям можуть відповідати зовсім відмінні різниці значень ознак (у випадку кількісних ознак). Тому для останніх слід вважати кореляцію рангів приблизною мірою оцінки тісноти зв'язку.

Коефіцієнт кореляції рангів може бути також використаний для визначення тісноти зв'язку між якісними (атрибутивними) ознаками, яким може бути надана рангова оцінка.

Коефіцієнт Фехнера застосовується для оцінки тісноти зв'язку на основі порівнянь знаків відхилень значень результативної і факторної ознак від їх середніх, його обчислюють за формулою:

$$k = \frac{\Sigma a - \Sigma b}{\Sigma a + \Sigma b},$$

де Σa – сума збігів знаків; Σb – сума незбігів знаків.

Коефіцієнт Фехнера змінюється від 0 до ±1. Якщо знаки всіх відхилень збігаються, то $\Sigma b = 0$, а коефіцієнт Фехнера дорівнює одиниці, що свідчить про наявність прямого зв'язку. Якщо знаки всіх відхилень будуть різними, то $\Sigma a = 0$, а коефіцієнт Фехнера дорівнює -1, що вказує на наявність оберненого зв'язку.

Розглянемо порядок обчислення коефіцієнта Фехнера на прикладі табл. 9.6.

Знак мінус означає, що значення ознаки менше середньої, знак плюс – більше середньої. Збіг знаків по обох ознаках означає узгоджену варіацію, незбіг - порушення узгодженості.

Коефіцієнт Фехнера для нашого прикладу становитиме

$$k_\phi = \frac{\Sigma a - \Sigma b}{\Sigma a + \Sigma b} = \frac{5-2}{5+2} = \frac{3}{7} = 0{,}4286.$$

Одержана додатна величина коефіцієнта Фехнера свідчить про те, що між судимістю і злочинністю є прямий кореляційний зв'язок.

Таблиця 9.6

Дані для розрахунку коефіцієнта Фехнера

№ п/п	Коефіцієнт злочинності (на 1000 чол. населення), злочинів	Коефіцієнт судимості (на 1000 чол. населення), судимостей	Різниця щодо середньої (+,–)		Збіг (a), незбіг (b) знаків
			для *x*	для *y*	
	x	*y*			
1	8,1	2,8	–	–	a
2	8,6	3,4	–	–	a
3	9,0	3,9	–	–	a
4	9,5	4,8	–	+	b
5	10,1	4,1	+	–	b
6	11,0	5,0	+	+	a
7	10,9	5,4	+	+	a
Разом	**67,2**	**29,4**	**–**	**–**	**–**
У середньому	**9,6**	**4,2**	**–**	**–**	**–**

Слід мати на увазі, що коефіцієнт Фехнера тільки констатує наявність і напрям кореляційного зв'язку і не залежить від величини відхилень результативної і факторної ознак від відповідних середніх, у зв'язку з чим оцінка тісноти зв'язку є наближеною. Коефіцієнт Фехнера може бути деяким орієнтиром в оцінці інтенсивності зв'язку.

Тісноту зв'язку між атрибутивними (якісними) ознаками можна виміряти за допомогою спеціальних коефіцієнтів **асоціації і контингенції,** запропонованих відповідно Д. Юлом і К. Пірсоном.

Для їх обчислення будується чотириклітинна таблиця, яка показує зв'язок між двома ознаками, кожна з яких повинна бути альтернативною, тобто такою, що складається з двох якісно відмінних один від одного значень (наприклад, особа засуджена або не засуджена, вживала або не вживала наркотики тощо).

Загальна схема чотириклітинної таблиці має вигляд (табл. 9.7).

В цій таблиці А і В ознаки, між якими вивчається зв'язок; не А і не В – протилежні (альтернативні) ознаки: *a*, *b*, *c*, *d* – частоти відповідних комбінацій ознак; *N* – загальне число спостережень.

Коефіцієнти обчислюються за формулами:

асоціації $r = \frac{ad - bc}{ad + bc};$

Таблиця 9.7

Чотириклітинна таблиця для розрахунку коефіцієнтів асоціації і контингенції

Ознаки	А	Не А	ΣВ
В	a	b	$a + b$
Не В	c	d	$c + d$
ΣА	$a + c$	$b + d$	N

контингенції $r = \dfrac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$.

Методику розрахунку коефіцієнтів асоціації і контингенції розглянемо на такому прикладі визначення тісноти зв'язку між двома якісними ознаками (табл. 9.8).

Таблиця 9.8

Розподіл засуджених чоловіків і жінок за вживанням накротиків

Стать	Кількість засуджених, які		Разом
	вживали наркотики	не вживали наркотики	
Чоловіки	120	41	161
Жінки	8	20	28
Разом	**128**	**61**	**189**

Оскільки дані подано у вигляді читириклітинної таблиці розподілу чисельностей за якісними ознаками, тісноту зв'язку визначимо, обчисливши коефіцієнти асоціації і контингенції.

За таблицею $a = 120$, $b = 41$, $c = 8$, $d = 20$.

1. Визначимо коефіцієнт асоціації:

$$r_a = \frac{ad - bc}{ad + bc} = \frac{120 \cdot 20 - 41 \cdot 8}{120 \cdot 120 + 41 \cdot 8} = \frac{2072}{2728} = 0{,}76.$$

2. Обчислимо коефіцієнт контингенції:

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

$$= \frac{120 \cdot 20 - 41 \cdot 8}{\sqrt{(120+41)(8+20)(120+8)(41+20)}} = \frac{2072}{5932{,}8} = 0{,}35.$$

Знайдені коефіцієнти асоціації і контингенції вказують на досить тісний зв'язок між вживанням наркотиків і статтю засуджених. При цьому коефіцієнт контингенції дає більш обережну оцінку тісноти зв'язку між ознаками, він завжди менший від коефіцієнта асоціації.

Коефіцієнти асоціації і контингенції можуть приймати будь-які значення від - 1 до + 1. Коефіцієнт контингенції завжди менше коефіцієнта асоціації. Для великих вибірок ($n \geq 30$) зв'язок практично вважається значущим, якщо $r_a \geq 0{,}5$, або $r_k \geq 0{,}3$. Величини коефіцієнтів асоціації і контингенції, як показників тісноти зв'язку, тлумачаться так само як і величина коефіцієнта кореляції.

Питання для самоконтролю

1. Дайте поняття функціонального і кореляційного зв'язку. Як проявляється кореляційний зв'язок?
2. Які Ви знаєте форми кореляційного зв'язку? Наведіть приклади.
3. Які задачі розв'язуються за допомогою кореляційно-регресійного аналізу?
4. Назвіть основні етапи кореляційного аналізу і розкрийте їх суть.
5. Які основні прийоми встановлення форми зв'язку між ознаками? Що таке кореляційне поле?
6. Як визначаються параметри рівняння регресії при лінійній і криволінійній залежності?
7. Що характеризує коефіцієнт регресії?
8. Охарактеризуйте рівняння регресії.
9. Які показники використовують для вимірювання тісноти зв'язку між ознаками при лінійній і криволінійній залежності?
10. Що характеризує коефіцієнт кореляції і коефіцієнт детермінації?
11. На що вказує знак лінійного коефіцієнта кореляції?
12. Як розраховується кореляційне відношення і що воно показує?
13. Дайте визначення множинної кореляції.
14. Що характеризують коефіцієнти регресії в рівнянні множинної регресії?
15. Який зміст коефіцієнтів повної і чистої регресії?

16. У чому суть коефіцієнтів еластичності і β-коефіцієнтів?

17. Яка методика перевірки істотності коефіцієнта регресії?

18. Як оцінюється істотність лінійного коефіцієнта кореляції?

19. Охарактеризуйте коефіцієнт кореляції рангів. Наведіть алгоритм його розрахунку.

20. За допомогою яких показників вивчається і вимірюється кореляційна залежність між якісними показниками?

21. Охарактеризуйте коефіцієнти асоціації і контингенції.

22. За наведеними даними розрахуйте коефіцієнти регресії, кореляції і детермінації. Зробіть висновки.

№ п/п	1	2	3	4	5	6	7	8	9	10
Кількість засуджених за хуліганство (на 100 тис. чол. населення)	83	87	90	82	91	89	103	115	100	99
Душове вживання алкоголю населення, літрів/рік	3,5	3,4	3,6	3,8	3,7	4,1	4,5	5,1	5,7	6,5