DOI: https://doi.org/10.36486/978-966-2310-66-5-19

Думай своєю головою, користуючись об'єктивними законами логіки. Висувай гіпотези, перевіряй докази, ніколи не приймай слова на віру й ніколи не вір, що хтось за тебе відповість на поставлене тобою ж запитання

Майкл Монро Льюїс

РОЗДІЛ 19
Статистичні методи в кримінальному аналізі

Олександр Користін

Багатьох людей лякає математика і статистика. Однак роль статистики дуже важлива для розуміння звітів, публікацій, політики та загальної інформації, яку ми обробляємо щодня. Знання статистики сьогодні стає необхідним для успішної діяльності у будь-якій сфері науки й практики.

> *Фундаментальне розуміння статистики має важливе значення для кримінального аналізу (IACA) [1].*

Аналітикам підрозділів кримінального аналізу необхідно володіти основами статистичного аналізу та принципами досліджень, щоб здійснювати глибокий аналіз даних та інформації, доступних правоохоронним органам. Аналітики повинні вміти використовувати відносно не складні, на перший погляд, але потужні, саме з точки зору аналітики, методи статистичного аналізу, для опису даних і робити висновки на їх основі.

Статистичні методи допомагають описувати та інтерпретувати дані, знаходити закономірності й робити висновки на основі великих масивів даних та вивчати причинно-наслідкові зв'язки.

ОПИСОВА СТАТИСТИКА

Описова статистика узагальнює великі обсяги даних в ефективний і зрозумілий спосіб. Ці дані можуть бути або кількісними, як, наприклад, вимірювання зросту та ваги, або якісними (категоріальними), як, наприклад, стать чи тип особистості. Величезні масиви даних, як правило, повинні бути узагальненими перш ніж вони будуть інтерпретовані людиною. Тобто, описова статистика є інструментом, що описує, узагальнює або зводить до бажаного вигляду розподіл значень окремих характеристик у сукупності.

Вимірювання, його рівні (шкали) та статистика

Одне з важливих понять, яке застосовується під час вибору статистичних показників й інтерпретації отриманих результатів – це «шкали вимірювання». Дійсно, вимірювання – це приписування чисел властивостям об'єктів для кількісного представлення відповідно до певних правил. Числа присвоюються для того, щоб зробити дані придатними для статистичного аналізу. Традиційно виділяють чотири шкали вимірювання: *номінальна, порядкова, інтервальна* та *відношення*. Ці відмінності важливі, оскільки спосіб аналізу даних залежить від того, як вони були зібрані. Кожен рівень передає різну кількість інформації.

Номінальна шкала

Номінальна шкала вимірювання – це класифікація об'єктів за набором категорій. Тобто, це процес групування предметів у класи, коли об'єкти, що належать до одного класу, є ідентичними відносно певної ознаки або властивості. Для розрізнення спостережень класам присвоюють назви або ярлики, але класам для ідентифікації також можуть присвоювати і числові коди.

> ***Наприклад**, ми можемо класифікувати респондентів опитування за сімейним станом або статтю. Сімейний стан кожного респондента може бути закодований як «одружений» або «неодружений». За статтю кожен респондент кодується як «чоловік» або «жінка». Кожен респондент підпадає лише під одну класифікацію, і для кожного респондента існує відповідна категорія відповідно до його раси та статі. Хоча ми можемо присвоїти числовий код кожній категорії (наприклад, «1» для чоловіків і «2» для жінок), цей числовий код все одно сприймається як номінальні дані – не існує логічного способу впорядкувати їх або виконати з ними обчислення [1].*

Порядкова шкала

Порядкове вимірювання має місце тоді, коли числа, які позначають категорії об'єктів, можливо порівнювати, використовуючи відношення «більше», «такий самий» та «менше». Цей ступінь різниці вказує на порядок або ранжування між категоріями. Категорії певним чином впорядковані, але фактична відстань між цими порядками невідома і не має значення.

> ***Прикладами*** *порядкової шкали можуть бути: думка про поліцію, серйозність злочинів або рівень страху. Шкали відповідей, такі як «добре, краще, найкраще», «згоден, нейтрально, не згоден» або «дуже малоймовірно, малоймовірно, не визначився, ймовірно, дуже ймовірно», є поширеними показниками порядкової шкали [1].*

Інтервальна шкала

Інтервальне вимірювання складається з усіх характеристик номінального та порядкового вимірювання. Крім того, інтервальний рівень передбачає, що осмисленою є операція порівняння інтервалів (тобто, різниць) між парами значень на шкалі вимірювання. Для цього має існувати одиниця вимірювання тієї властивості, яка нас цікавить. На відміну від порядкових шкал, відстань між категоріями має значення.

> ***Прикладами*** *інтервального вимірювання є показники температури та IQ [1].*

Шкала відношень

На додаток до всіх характеристик попередніх трьох шкал вимірювання, шкала відношень має істинну нульову точку, в якій властивість зникає. Істинна нульова точка дозволяє виміряти повну відсутність поняття, що вимірюється. Завдяки цьому можливо порівнювати між собою відношення між парами значень на шкалі.

> *Дохід, вага, час і вік є прикладами вимірювань на цій шкалі [1].*

Багато властивостей можна виміряти за допомогою різних шкал вимірювання. Наприклад, **температуру** можна виміряти за допомогою відповідних наборів відповідей для кожної шкали вимірювання:

> *Номінальна шкала*: «Холодний» або «Гарячий»
>
> *Порядкова шкала*: 1 – дуже холодний, 2 – холодний, 3 – теплий, 4 – гарячий, 5 –дуже гарячий
>
> *Інтервальна шкала*: 20 градусів Цельсія, 20 градусів Фаренгейта
>
> *Шкала відношення*: 20 градусів за шкалою Кельвіна

Міри центральної тенденції

Міри центральної тенденції є найпоширенішими різновидом описових статистик. Вони використовуються для опису властивостей «типового значення» певної сукупності даних, яким можна представити всі інші. Наприклад, «типове значення» може бути описане показником «середнє арифметичне». Ці описові індекси використовуються для відповіді на запитання, як-то «Який вік типового квартирного крадія у місті за минулий рік?». Якщо уявити собі, що множина результатів розміщена на числовій прямій віку людини, то ця властивість описуваної сукупності відображається в орієнтації результатів відносно цієї прямої. Групуються вони навколо значення 28 років чи центром є 31 рік? Різні міри центральної тенденції сукупності даних спираються на різні визначення «центрального значення на шкалі». Існує порівняно невелика кількість таких мір центральної тенденції, але основними є: *мода, медіана та середнє арифметичне*.

Мода

Мода – це таке значення у множині спостережень, яке зустрічається найбільш часто. Мода визначається шляхом безпосереднього спостереження за розподілом частот. Водночас, не будь-яка сукупність має єдину моду, тому остаточне розуміння моди має особливості, зокрема:

- *у випадку, коли усі значення в групі зустрічаються з однаковою частотою, то прийнято вважати, що група оцінок не має моди, тобто в групі (0,5; 0,5; 1,6; 1,6; 3,9; 3,9) мода не можна визначити;*
- *коли два сусідніх значення мають однакову частоту і вона є більшою частоти будь-якого іншого значення, мода є середнім цих двох значень, тобто в групі (0, 1, 1, 2, 2, 2, 3, 3, 3, 4) мода 2,5;*
- *якщо два несуміжних значення у групі мають рівні частоти і вони більші від частот будь-якого значення, то існує дві моди, тобто в групі (10, 11, 11, 11, 12, 13, 14, 14, 14, 17) модами є і 11, і 14; в такому випадку кажуть, що група оцінок є бімодальною.*

Аналогічно, якщо є більше двох однакових оцінок, розподіл є мультимодальним.

Мода вказує на бал, який найбільш часто зустрічається у розподілі частот, або інтервал у згрупованому розподілі частот. Вона надає обмежену інформацію і часто є дуже нестійкою у невеликих вибірках, тому використовується рідше за інші міри центральної тенденції. Хоча мода може засто-

совуватися до всіх чотирьох шкал вимірювання, оскільки вона використовує дуже мало інформації та є найбільш придатною для даних номінальної шкали.

Медіана

Медіана (Md) – це середина або середній бал впорядкованої за величиною сукупності значень. Це значення, яке ділить упорядковану множину даних навпіл, так що одна половина значень виявляється більше медіани, а інша – меншою, тобто, точка на лінії, в якій 50% оцінок лежать вище, а 50% нижче. Однією з сильних сторін медіани є стабільність. На відміну від середнього значення, медіана не зазнає значного впливу екстремальних значень. Екстремальні значення або «викиди» можуть траплятися на обох кінцях розподілу, але вони не впливають на медіану.

Мають місце також особливості визначення медіани:

- якщо дані містять непарну кількість різних значень, наприклад 11, 13, 18, 19, 20, то медіана є серединним (центральним) значенням для випадку, коли вони упорядковані, тобто Md = 18;

- якщо дані містять парну кількість значень, наприклад 4, 9, 13, 14, то медіана є точкою, що лежить посередині між двома центральними значеннями, коли вони впорядковані, тобто Md = (9+13)/2 = 11.

Процедура знаходження медіани в таблиці частот подібна до процедури знаходження медіани в масиві. Визначається положення медіани, використовуючи кумулятивну частоту. Потім визначається значення або інтервал пов'язаного з нею класу або інтервального групування.

Середнє

Середнє значення – це середнє арифметичне. Середнє обчислюється шляхом ділення суми значень на кількість цих значень. Розподіл даних може мати лише одне середнє значення.

Середнє корисне для даних на рівні шкали інтервалів або шкали відношень. Обчислення середнього для номінальної змінної не має сенсу. Наприклад, припустимо, що набір даних містить номінальну змінну, таку як стать, і кодується «1» для чоловіків і «2» для жінок. Якщо набір даних містить десять чоловіків і три жінки, середній бал можна теоретично обчислити за наведеною вище формулою, отримавши в результаті середній бал 1,23. Однак 1,23 не має сенсу, оскільки цифри «1» і «2» використовуються як мітки, а не як арифметичні числа.

> *Кожна міра центральної тенденції наділена характеристиками, які роблять її цінною за певних умов.*

Мода найбільш просто вираховується – її можливо визначити «на око» на графіку чи в таблиці розподілу. Крім того, для надто великих груп даних це достатньо стабільна міра центру розподілу. У багатьох симетричних розподілах значної кількості вимірювань, мода є близькою і до медіани, і до середнього.

Медіана займає проміжне положення між модою та середнім з точки зору її обрахування, якщо останнє виконується вручну. У великих масивах дані спочатку групуються, а потім визначається медіана.

Середнє множини даних передбачає переважно арифметичні операції. На величину середнього впливають значення усіх результатів. Медіана й мода не потребують для визначення усіх значень.

***Наприклад**, що відбудеться з середнім, медіаною та модою, якщо подвоїться максимальне значення у наступній множині:*

Множина	Середнє	Медіана	Мода
1: 1, 3, 3, 5, 6, 7, 8	33/7	5	3
2: 1, 3, 3, 5, 6, 7, 16	41/7	5	3

На величину середнього особливо впливають результати, які називають «викидами», тобто дані, які розміщуються далеко від більшості інших значень. Перевага це чи ні – залежить від конкретних питань, що вирішуються.

Міри розсіювання

Міри центральної тенденції дають нам уявлення про концентрацію групи значень на числовій шкалі, або про «типовий» випадок у великому наборі даних. Кожна міра дає таке значення, яке «представляє», репрезентує собою усі оцінки сукупності. У цьому випадку нехтують відмінностями, що мають місце між окремими значеннями. Але варіація є основою статистики. Для вимірювання варіації оцінок всередині сукупності необхідно використовувати інші описові статистики, зокрема, окремі статистичні характеристики, які по-різному виконують роль міри розсіювання (розмах, дисперсія, стандартне відхилення).

Розмах

Розмах вимірює на числовій шкалі відстань, в межах якої змінюються оцінки. Подібно до цього, діапазон використовується для створення групувань даних за інтервалами класів, і він також є простою мірою варіації. Діапазон показує відстань (різницю) між найвищим і найнижчим значеннями в розподілі. Оскільки він використовує крайні значення, він мало що говорить про інші значення в розподілі. Ця статистика може бути нестабільною, оскільки спирається лише на два крайні значення. Між оцінками на крайніх кінцях розподілу може бути велика відстань, але невелика варіація всередині крайніх значень [1].

***Наприклад**, розглянемо два набори даних:*

Набір даних №1: 1, 7, 7, 7, 7, 8, 8, 8, 10

Набір даних №2: 1, 3, 4, 6, 9, 9, 9, 10

У наведених вище наборах даних обидва мають діапазон 9, але очевидно, що набір даних №2 має набагато більшу варіацію значень, ніж набір даних №1.

Дисперсія

Дисперсія (у популяції позначається як σ^2 – грецька літера «сигма») – це сума квадратів відхилень кожного значення сукупності від середнього, поділена на загальну кількість значень. Інакше кажучи, це середнє значення суми всіх квадратів відхилень від середнього значення будь-якого розподілу. По суті, вона підсумовує розсіювання оцінок навколо середнього значення.

Стандартне відхилення

Мірою розсіювання, тісно пов'язаною з дисперсією, є *стандартне відхилення*, яке визначається як позитивне значення квадратного кореня з дисперсії. Стандартне відхилення вимірює середню відстань, на яку кожен елемент даних віддалений від середнього значення всіх елементів даних у розподілі. Таким чином, воно вказує на середню величину варіації щодо середнього значення. Стандартне відхилення є важливим, оскільки воно дає уявлення про те, як бали в розподілі співвідносяться з типовим, центральним. Воно також дозволяє порівнювати декілька розподілів.

Через потенційний вплив викидів середнє значення може вводити в оману. Таким чином, оцінка середніх значень разом зі стандартними відхиленнями дає краще розуміння всіх оцінок разом. Іншими словами, якщо середнє значення – це типове значення сукупності всіх спостережень, то стандартне відхилення відображає середню величину відхилення кожного окремого спостереження від середнього значення. Використовуються всі бали в розподілі. Оскільки вони базуються на середньому значенні, дисперсія та стандартне відхилення потребують даних про інтервали або рівні відношення.

***Наприклад**, визначимо дисперсію та стандартне відхилення для невеликого набору даних з шести значень: 1, 6, 12, 17, 18, 24:*

Значення	Середнє	Відхилення	Відхилення в квадраті	Дисперсія	Стандартне відхилення
1	13	-12	144	71,2	8.438009
6		-7	49		
12		-1	1		
17		4	16		
18		5	25		
24		11	121		

Почнемо з обчислення середнього значення (13), потім віднімаємо кожне значення від середнього. Це дає нам відхилення кожного значення. Зводимо кожне відхилення в квадрат і підсумовуємо квадрати (356). Потім ділимо суму квадратичних відхилень на кількість значень мінус один (у нашому випадку 6 - 1 = 5) і беремо квадратний корінь з результату. Таким чином, 71,2 є дисперсією для цього набору даних, а 8.438009 – стандартним відхиленням.

Існує три міркування щодо інтерпретації стандартного відхилення:

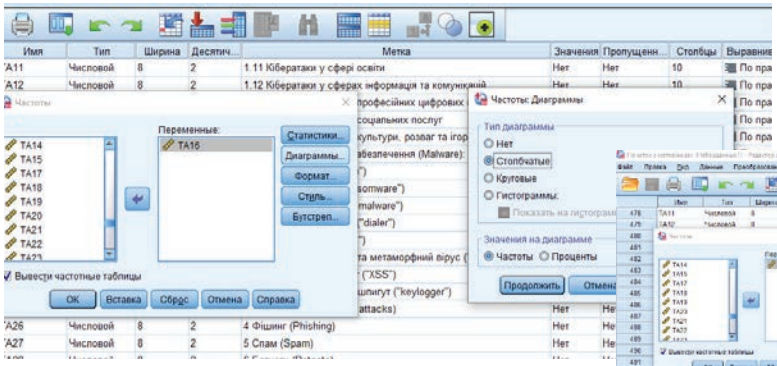
будь-яке конкретне значення стандартного відхилення не має реального інтуїтивного значення;

воно є найбільш корисним у порівняльному сенсі;

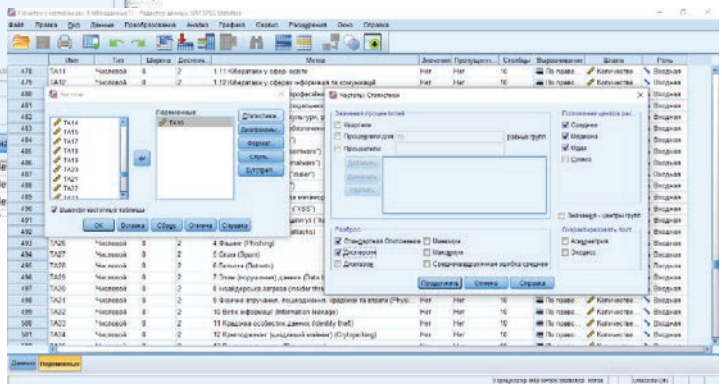
порівняння відносних значень стандартного відхилення та середнього значення показує, наскільки велика варіація в групі випадків відносно середнього значення.

> *Сучасні комп'ютерні програми, такі як Excel, SPSS, або програмне середовище R чи Python швидко обчислюють як дисперсію, так і стандартне відхилення, значною мірою усувають необхідність проходити всі ці кроки і значно спрощують процес розрахунку, залишаючи аналітику можливість та необхідність змістовної інтерпретації отриманих результатів.*

Приклад обчислення описових статистик з використанням пакету **IBM SPSS Statistics:**



1. Замовлення описових статистик: середнє, медіана, мода, дисперсія та стандартна відхилення за змінною TA16



2. Замовлення стовпчастої діаграми

```
DATASET ACTIVATE Наборданных1.
FREQUENCIES VARIABLES=TA16
 /STATISTICS=STDDEV VARIANCE MEAN MEDIAN MODE
 /BARCHART FREQ
 /ORDER=ANALYSIS.
```

3. Синтаксис за зробленим запитом

4. Отримані описові статистики з розподілом частот

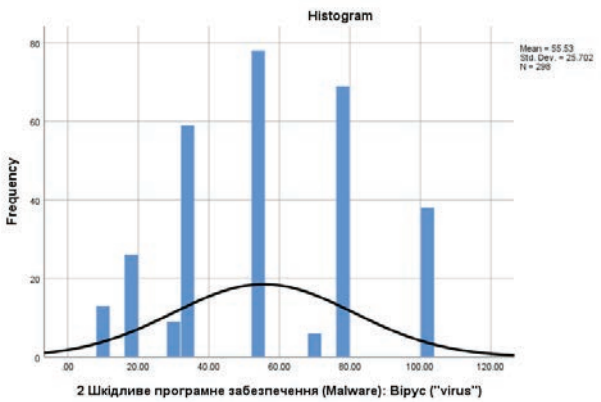
Statistics

2 Шкідливе програмне забезпечення (Malware): Вірус («virus»)

N	Valid	298
	Missing	500
Mean		55.5294
Median		52.2388
Mode		52.24
Std. Deviation		25.70201
Variance		660.594
Range		89.55

2 Шкідливе програмне забезпечення (Malware): Вірус («virus»)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	10.45	13	1.6	4.4	4.4
	19.40	26	3.3	8.7	13.1
	28.36	9	1.1	3.0	16.1
	35.82	59	7.4	19.8	35.9
	52.24	78	9.8	26.2	62.1
	68.66	6	.8	2.0	64.1
	76.12	69	8.6	23.2	87.2
	100.00	38	4.8	12.8	100.0
	Total	298	37.3	100.0	
Missing	System	500	62.7		
Total		798	100.0		



5. Гістограма з описовими статистиками та розподілом частот

СТАТИСТИЧНИЙ ВИСНОВОК

У той час як описова статистика описує характеристики конкретного набору даних, статистичний висновок використовуються для того, щоб зробити узагальнення про більшу сукупність на основі вибірки з цієї сукупності.

За допомогою описової статистики узагальнюються дані та описуються характеристики, такі як міри центральної тенденції та розсіювання. Ці методи використовуються, щоб допомогти зрозуміти конкретні характеристики певного набору даних. На відміну від них, процедури статистичного висновку використовуються для узагальнення характеристик генеральної сукупності на основі вибірки, взятої з цієї популяції (генеральної сукупності).

Щоб краще зрозуміти цю концепцію, корисно визначити деякі терміни, що за таких обставин використовуються, зокрема [1]:

> *- **Популяція (генеральна сукупність)** – це весь загальний можливий набір елементів (наприклад, випадків, людей, адрес), який цікавить дослідника. Розмір сукупності може варіюватися залежно від групи, що вивчається.*

***Прикладом** генеральної сукупності, яка часто використовується аналітиками кримінального аналізу, може бути група автомобільних крадіжок. Сукупність таких крадіжок може включати всі зареєстровані випадки у певній юрисдикції, районі чи групі, або може зосереджуватися на сукупності викликів поліції в межах однієї конкретної території у певний час.*

> *- **Вибірка** означає підмножину спостережень, взятих з генеральної сукупності.*
>
> *- **Статистичний висновок** – це перехід від результатів, отриманих на підмножині (вибірці) даних, взятих з генеральної сукупності, до оцінок характеристик цієї сукупності.*

Метою статистичного висновку є визначення характеристик генеральної сукупності, коли все, що ми маємо, – це вибірка з цієї сукупності. У наведеному вище прикладі аналітика можуть попросити визначити певну характеристику автомобільних крадіжок, яку неможливо отримати з бази даних масово (наприклад, змінну, яка не може бути надійно зафіксована під час процесу звітності).

***Наприклад**, може виникнути потреба у визначенні відсотка крадіжок, в яких не було насильницьких дій. Якщо ця змінна не фіксується в базі даних, аналітик може зробити вибірку зі спеціальних звітів, щоб визначити, як часто відбуваються крадіжки без ознак насильницьких дій. Список, з якого формується вибірка, називається основою вибірки.*

Хоча вибірка може розповісти нам щось про генеральну сукупність, не всі вибірки «створені однаковими». Найпотужніший метод формування вибірки (і метод, на якому ґрунтується більшість статистичних процедур) називається *випадковим відбором*. Як випливає з назви, вибірка є випадковою, коли кожен елемент генеральної сукупності має однакову ймовірність потрапити до вибірки. Вибірку, зібрану таким методом, часто називають простою випадковою вибіркою.

За умови достатнього розміру вибірки, випадковий відбір є достатньо потужним. Хоча аналітики кримінального аналізу часто вивчають цілі групи населення, можливість використовувати підгрупу населення для визначення характеристик всієї популяції може заощадити час і дозволити дослідити ті сфери, які інакше були б неможливими.

Випадковий відбір, однак, не гарантує, що ваша вибірка та генеральна сукупність матимуть ідентичні характеристики. Насправді, випадковість може призвести до того, що вибірка не збігатиметься з генеральною сукупністю. На ймовірність того, що статистика вашої вибірки буде схожою на параметри генеральної сукупності, впливають як розмір вибірки, так і варіативність досліджуваного об'єкта.

> ***Розмір вибірки** в цьому контексті означає кількість елементів, взятих з генеральної сукупності. Чим більшою є випадкова вибірка, тим більша ймовірність того, що за інших рівних умов вона буде схожою на генеральну сукупність.*
>
> ***Варіативність** – це кількість відмінностей між значеннями у популяції. За інших рівних умов, більша варіативність вимагає більшого розміру вибірки.*

Багато з найпоширеніших статистичних процедур розроблені для даних, які в популяції характеризуються так званим «нормальним» (гаусівським) розподілом.

> ***Нормальний розподіл** – це розподіл ймовірностей, який графічно зображується класичною симетричною дзвоновидною кривою.*

Нормальний розподіл визначається кількома характеристиками.

По-перше, нормальний розподіл є унімодальним (лише один пік), водночас середнє, медіана та мода є рівними і падають точно в центрі розподілу.

По-друге, нормальний розподіл є симетричним (тобто кожна половина, поділена центральною лінією, є дзеркальним відображенням іншої половини) і не має асиметрії.

Загальна площа нормального розподілу дорівнює одиниці або 100%. Властивістю нормального розподілу є те, що певні відсотки значень потрапляють у встановлену кількість стандартних відхилень від середнього значення сукупності. Приблизно 68% значень у розподілі знаходяться в межах одного стандартного відхилення від середнього (в позитивний і негативний бік) і приблизно 95% значень знаходяться в межах двох стандартних відхилень від середнього значення. Хоча форма нормальної кривої завжди однакова, висота й ширина кривої визначаються середнім значенням і стандартним відхиленням досліджуваних даних [1].

Нормальний розподіл часто використовується в аналізі даних, оскільки великі вибірки (зазвичай 100 або більше випадків) часто мають нормальний розподіл. Ця концепція демонструється в центральній граничній теоремі:

***центральна гранична теорема** стверджує, що розподіл середніх значень безкінечної кількості незалежних випадкових вибірок однакового обсягу буде наближатися до нормального розподілу незалежно від основного розподілу генеральної сукупності.*

Це важливо, оскільки багато статистичних тестів працюють належним чином лише тоді, коли дані мають приблизно нормальний розподіл. Одним з наслідків цієї теореми є те, що аналітик може припустити, що розподіл вибіркових середніх є нормальним, якщо вибірка є достатньо великою.

Аналітик може використовувати статистичні висновки, щоб отримати цінну інформацію про одну популяцію або декілька популяцій. Так, використовуючи приклад з автомобільними крадіжками, аналітик може розробити точкову оцінку відсотка автомобільних крадіжок зі зломом.

> ***Точкова оцінка** – це одне значення на основі вибірки, яке найкраще оцінює або наближається до істинного значення в генеральній сукупності.*

У цьому випадку ми використовуємо вибірку автомобільних крадіжок для оцінки загального відсотка автомобільних крадіжок з насильницькими діями.

Статистичний висновок реалізується шляхом побудови довірчих інтервалів та перевірки статистичних гіпотез.

> ***Перевірка гіпотез** – це покроковий процес визначення дослідницького питання, визначення статистичних методів, які можуть бути використані для відповіді на нього, проведення тестів, а потім прийняття рішення щодо даних на основі результатів, отриманих на основі вибірки.*

Вибірки та популяції

Існує низка показників, які можна використовувати для опису набору даних. Часто ми не маємо доступу до всієї сукупності даних, що ускладнює обчислення описових статистик, таких як міри центральної тенденції (наприклад, середнє, медіана і мода), міри дисперсії (наприклад, розмах, дисперсія, стандартне відхилення) та інших статистичних показників, що представляють інтерес. Під час використання для опису всієї сукупності будь-яка з цих статистик називається *параметром*. І навпаки, той самий показник, розрахований на основі вибірки, називається *статистикою*. Статистика потрібна для того, щоб оцінити значення параметру. Прикладом статистики може бути середнє значення випадково відібраної вибірки (тобто, вибіркове середнє). Параметром буде середнє значення, яке можна було б розрахувати, якби ми мали доступ до всіх спостережень у популяції.

Правильно зібрані вибірки дозволяють аналітику оцінити параметри генеральної сукупності. Однак важливо пам'ятати, що ця оцінка може бути неточною, не зважаючи на випадковий характер відбору спостережень. Основним обмеженням вибірки (і узагальнюючої статистики загалом) є те, що ми не повністю виміряли генеральну сукупність, що призводить до можливості того, що наша вибірка не буде репрезентативною для генеральної сукупності [1].

Імовірнісні та неімовірнісні вибірки

Стратегії вибірки можна умовно поділити на дві великі категорії: імовірнісні та неімовірнісні вибірки. Проста випадкова вибірка, про яку ми говорили на початку цього розділу, належить до категорії імовірнісних вибірок. Імовірнісні вибірки повинні використовувати певну форму випадкового відбору. Це означає, що всі елементи генеральної сукупності або основи вибірки повинні мати відомі шанси потрапити у вибірку.

Неімовірнісні вибірки не використовують випадковий відбір під час формування вибірки. Хоча неімовірнісні вибірки можуть бути репрезентативними для генеральної сукупності, з якої вони отримані, до них не застосовуються інструменти статистичного висновку.

Найпоширеніші методи збору імовірнісних вибірок включають [1]:

- проста випадкова вибірка,

- систематична випадкова вибірка,

- стратифікована випадкова вибірка,

- кластерна вибірка.

Проста випадкова вибірка

Проста випадкова вибірка формується шляхом надання рівних шансів бути відібраним кожному елементу генеральної сукупності або вибіркової сукупності. Елементи, відібрані за допомогою цієї процедури, становлять вибірку. Випадкова вибірка може бути із заміщенням або без заміщення. У разі вибірки із заміщенням кожен елемент повертається до основи вибірки, з якої він був відібраний, після того, як він був використаний у вибірці. Це створює ситуацію, коли один і той самий елемент може бути використаний кілька разів в одній і тій самій вибірці.

Часто вибірка відбирається без заміщення. Таким чином, кожен елемент спочатку має однакові шанси бути відібраним. Однак, по мірі того, як процес формування вибірки продовжується, шанси на те, що будь-який елемент (ще не відібраний) буде відібраний, зростають.

***Наприклад**, припустимо, що аналітик хоче визначити характеристики, не зафіксовані в базі даних свого відомства, у вибірці зі 100 випадків викрадення автомобілів, що сталися в певному регіоні. Замість того, щоб переглянути всі 100 випадків, аналітик вирішує випадковим чином відібрати 30 крадіжок без заміщення. Спочатку будь-який автомобіль, що потрапив до вибірки, мав би шанс 1/100 бути відібраним. Після відбору цей автомобіль буде вилучений з вибірки, що дасть наступному автомобілю шанс потрапити до вибірки з імовірністю 1/99. Цей процес продовжиться, коли третій об'єкт матиме шанс 1/98 і так далі [1].*

Чому це важливо? Основна відмінність між вибіркою із заміщенням та без заміщення полягає в тому, що елементи у вибірці із заміщенням не є незалежними. Зазвичай це не має особливого значення для великих груп населення. Однак для невеликих груп населення вибірка із заміщенням може створити проблеми, пов'язані з коваріацією. У цьому контексті коваріація просто означає ступінь, до якого дві величини змінюються разом. На практиці проблеми коваріації значно зменшуються у великих вибірках .

Систематична випадкова вибірка

Систематична випадкова вибірка – це різновид простої випадкової вибірки, яка може бути корисною в ситуаціях, коли аналітик не має основи для вибірки (тобто, списку всіх спостережень, які належать до популяції).

***Прикладом** цього можуть бути паперові звіти, які зберігаються в шафі й не доступні в електронному вигляді. У такому випадку аналітик може вирішити взяти перший елемент (у цьому випадку – звіт), а потім відібрати кожен n-й звіт після першого. Число n зазвичай обирається шляхом ділення загальної кількості випадків на бажану кількість для вибірки. Добуток цього рівняння називається інтервалом вибірки. Наприклад, якщо аналітик хоче відібрати 50 звітів про крадіжки автомобілів з файлів, що містять 1000 звітів, він ділить 1000 на 50. Це дасть інтервал вибірки, що дорівнює 20.*

Наступним кроком у цьому процесі буде випадковий вибір числа з інтервалу вибірки (у цьому випадку від 1 до 20) і початок вибірки з цього випадково вибраного значення. Якщо наш аналітик випадково вибере 15, то він зарахує 15 справ до файлу для першого елемента вибірки. Потім аналітик продовжить, відбираючи кожен 20-й звіт у картотеці [1].

У деяких випадках інтервал вибірки не дорівнює цілому числу. У такому випадку аналітик буде чергувати цілі числа, відбираючи кейси. Наприклад, якщо інтервал вибірки становитиме 20,5 замість 20, аналітик буде чергувати 20-й і 21-й звіти.

Стратифікована випадкова вибірка

Стратифікована випадкова вибірка може бути використана для більш ефективного отримання ймовірнісної вибірки. Це досягається шляхом використання інформації, відомої про населення до формування вибірки, для створення страт (тобто, підгруп), з яких випадковим чином відбираються елементи вибірки. Страти створюються шляхом поділу частини населення на групи відповідно до характеристик, що становлять інтерес. Потім з кожної страти випадковим чином відбирається вибірка заздалегідь визначеного розміру.

Існує низка причин для використання цього методу. Наприклад, якщо аналітику потрібна імовірнісна вибірка частини першої злочинів і він хоче переконатися, що кожен тип злочину має

достатню кількість зразків, проста випадкова вибірка повинна бути надзвичайно великою, щоб охопити злочини, які рідко трапляються: такі як вбивства, зґвалтування і пограбування. Аналітик не може бути впевненим, що всі злочини частини 1 будуть представлені у вибірці. Більш ефективною стратегією може бути стратифікація генеральної сукупності (усіх злочинів за частиною 1) за типом злочину, а потім відбір достатньої кількості елементів з кожної страти для вивчення явищ, що цікавлять дослідника. Такий тип стратифікованої випадкової вибірки називається непропорційною стратифікованою вибіркою. Цей метод використовується для того, щоб забезпечити достатню репрезентативність випадків з менших страт, і має потенціал для підвищення точності вибірки за використання фіксованих вхідних даних.

Інша форма стратифікованої вибірки називається пропорційною стратифікованою вибіркою. У пропорційній стратифікованій вибірці обсяг вибірки кожної страти визначається її часткою в генеральній сукупності. Хоча це може виглядати складно, розрахунок є дуже простим.

Розрахунок полягає в тому, що кількість випадків у стратах (тобто кількість крадіжок) ділиться на загальну кількість випадків (тобто загальну кількість злочинів за ч. 1), помножену на загальний обсяг вибірки. Потім аналітик повинен повторити цей розрахунок для кожного типу злочинів частини 1.

Пропорційна стратифікована вибірка допомагає зменшити упередженість відбору і знизити похибку, пов'язану з рандомізацією. Однак, як пропорційна, так і непропорційна стратифікована вибірка вимагає, щоб населення було поділене на страти і щоб кожен елемент сукупності перебував в одній і тільки в одній страті (тобто, елементи є взаємовиключними). Жоден елемент сукупності не може належати до двох страт. У разі стратифікації за кількома характеристиками (наприклад, за статтю, расою та етнічною приналежністю) кількість необхідних страт може швидко стати некерованою.

Кластерна вибірка

Тоді як стратифікована вибірка вимагає більше інформації про досліджувану сукупність, кластерна вибірка може бути використана для подолання обмежень у разі спроби відібрати великі групи населення. Зокрема, вона корисна, коли відсутня основа вибірки. Це може статися під час вивчення груп населення, розміщених на великих географічних територіях або розпорошених між багатьма різними організаціями.

> ***Кластерна вибірка** – це стратегія вибірки, яка використовується, коли необхідно відібрати природньо існуючі групи з високим ступенем однорідності, і аналітик хоче використати ці групи для більш ефективного проведення аналізу.*

Процес починається з визначення кластерів, що підлягають вивченню, а потім випадкового відбору елементів для вибірки. Кластери визначаються як природні об'єднання елементів сукупності. Кожен елемент може з'явитися лише в одному кластері.

> ***Прикладом** кластера, що має відношення до аналізу злочинності, може бути вибірка поліцейських через кластери поліцейських відомств або вибірка ув'язнених у групах в'язниць. Якщо аналітик опитує всі елементи (поліцейських та ув'язнених відповідно), він проводить одноступінчасту кластерну вибірку. Як варіант, аналітик може спочатку випадковим чином відібрати управління поліції або в'язниці, а потім випадковим чином відібрати працівників поліції або ув'язнених для вибірки. Це приклад багатоступеневої кластерної вибірки. Багатоступеневі кластерні вибірки можуть включати ще більше ступенів. Наприклад, дослідник або аналітик може спочатку випадковим чином відібрати кластери регіонів, потім з цих регіонів випадковим чином відібрати кластери поліцейських управлінь, а з цих управлінь випадковим чином відібрати окремих поліцейських.*

Хоча кластерна вибірка є ефективним методом опитування великих груп, які існують і природно об'єднані в кластери, вона може збільшити похибку вибірки через складність процесу.

> ***Похибка вибірки** – це ймовірна розбіжність вибіркової статистики з параметрами генеральної сукупності.*

Оскільки вибірки не включають усіх членів генеральної сукупності, існує ймовірність того, що між статистикою та параметром буде існувати певна різниця. Ця різниця і є похибкою вибірки.

Перевірка гіпотез

Часто дані опитувань збираються з метою проведення цільового аналізу. Нехай завданням, яке визначило керівництво для аналітика, є з'ясування взаємозв'язку ефективності розслідування кримінальних проваджень із співвідношенням аналітиків і детективів у підрозділі чи органі.

Для відповіді на це питання аналітик має визначити: *покращує чи ні співвідношення між аналітиками та детективами якість розслідувань?*

Відповідь на це дослідницьке питання потребує перевірки відповідних гіпотез. Технічно, перевірка гіпотези – це форма статистичного висновку, коли аналітик використовує набори даних для перевірки гіпотези як твердження про значення параметру. Хороша гіпотеза є дуже конкретною щодо того, що станеться або не станеться.

Припустимо, що аналітик, який взяв початкову вибірку, вважає, що більша кількість і співвідношення аналітиків підрозділів кримінального аналізу з детективами покращує якість розслідування. Однак це дуже широка гіпотеза, і її важко перевірити. Щоб допомогти звузити питання, аналітик може розглянути, підходи вимірювання «покращення розслідувань». Відповідь на це друге питання може полягати в тому, що детективи, які отримують більше допомоги в аналізі злочинів, з більшою ймовірністю розкривають справи. Цю концепцію можна перетворити на конкретне твердження, яке можна перевірити, та сформулювати гіпотезу так:

> *Збільшення співвідношення аналітиків до детективів покращує розслідування (за рахунок покращення показників розкриття злочинів).*

У цьому випадку залежна змінна – це показники розкриття справ, а незалежна змінна – співвідношення кількості аналітиків і детективів у відомстві. А це вже можливо перевірити, оскільки маємо залежну змінну, тобто змінну, на яку впливають, і незалежну змінну, тобто змінну, яка впливає на залежну змінну.

Водночас це також спрямоване твердження, яким стверджується, що вищий відсоток аналітиків до детективів покращить результати розслідувань, на відміну від неспрямованого твердження, яке лише вказує на різницю, але не визначає напрям впливу.

> *Спрямовані твердження називають односпрямованими тестами, тоді як неспрямовані – двоспрямованими.*

Перевіряючи гіпотезу, аналітик не «доводить» її. Це неможливо. Замість цього аналітик перевіряє нульову гіпотезу, тобто твердження, яке передбачає відсутність різниці між змінними, на відміну від дослідницької гіпотези, яка стверджує, що деяка різниця існує.

> *Наразі нульова гіпотеза буде такою: співвідношення аналітиків та детективів не впливає на розслідування (через покращення показників розкриття злочинів).*

Якщо дані демонструють хибність нульової гіпотези з достатньо високим рівнем достовірності, аналітик може відкинути нульову гіпотезу і прийняти альтернативну гіпотезу як можливу.

Перевірка нульової гіпотези базується на тому, що можливим є доведення за допомогою спостереження, аналізу або логічного аргументу, що вона є некоректною (тобто неправильною або неправдивою).

> ***Правило**: ми не можемо явно довести дослідницьку гіпотезу. Все, на що ми можемо сподіватися, – це виявити інші можливі пояснення і виключити їх.*

Багато аналітиків використовують цитату сера Артура Конан Дойля (1890) як девіз:

> *«Після того, як ви відкинете неможливе, те, що залишиться, незалежно від того, наскільки воно неймовірне, має бути правдою» (IACA) [1].*

Це те, чого намагається досягти перевірка гіпотези. Перевірка гіпотези – це аргумент зведення до абсурду, який намагається продемонструвати правдивість дослідницької гіпотези шляхом спростування нульової гіпотези. Якщо нульову гіпотезу можна спростувати, це залишає дослідницьку гіпотезу як єдину можливу альтернативу. Однак часто неможливо виключити всі потенційні конкуруючі пояснення певного явища. Через це обмеження ми не можемо довести правдивість будь-якого твердження-гіпотези. Замість цього аналітик намагається збалансувати ймовірність різних видів помилок.

Часто аналітик стикається з вибором між збільшенням ймовірності помилки першого або другого типу. Перш ніж приймати рішення щодо відхилення чи не відхилення певної нульової гіпотези, доцільно зважити відносні витрати, пов'язані з помилкою першого або другого типу. Потім аналітик може врахувати ці витрати під час визначення конкретного рівня *статистичної значущості*.

P-значення та статистична значущість

Вивчення (перевірка) *p-значення* є одним із запобіжників у прагненні зменшити похибку першого типу – спростування нульової гіпотези, коли вона є фактично вірною для популяції. Розрахунок *p* враховує розмір вибірки, варіативність даних і розмір ефекту, щоб отримати число, яке вказує, наскільки спостережувані дані узгоджуються з нульовою гіпотезою.

> ***Розмір ефекту** – це статистика, яка оцінює величину або вплив змінної, що нас цікавить.*

Для визначення статистичної значущості використовується p-значення.

> *Результат є статистично значущим, коли p-значення є меншим за встановлений рівень значущості, який позначається символом α (грецька літера «альфа»). У соціальних науках рівень значущості часто встановлюється на рівні α = 0.05, що означає, що p-значення, менше 0.05, є статистично значущим на рівні 0.05 [1].*

Слід зазначити, що *p-значення* не вимірює розмір ефекту та не надає жодної інформації про практичну значущість результату (наприклад, чи є ефект достатньо великим, щоб бути значущим у практичному плані).

За гіпотезою, яку визначено вище, розмір ефекту – це вплив різниці у співвідношенні аналітиків і детективів на показники розкриття злочинів. Тому краще використовувати p-значення лише як один з інструментів в арсеналі аналітика.

Перевірка гіпотез є важливим статистичним інструментом, але ним можна легко зловживати. Це пов'язано з тим, що перевірка гіпотез просто оцінює два взаємовиключних твердження щодо сукупності, використовуючи вибірку, взяту з цієї сукупності. Статистично значущий результат 1) не доводить альтернативну гіпотезу і 2) стосується лише вибірки, взятої випадковим чином з генеральної сукупності. Враховуючи те, що аналітики злочинності часто мають доступ до повної генеральної сукупності (тобто до всіх зареєстрованих випадків викрадення автомобілів), корисність цих методів може бути обмеженою.

Іншою частою помилкою є припущення, що незначущий результат вказує на те, що нульова гіпотеза, є правильною або що значущий результат показує з 95% ступенем впевненості, що альтернативна гіпотеза правильна. Важливо пам'ятати, що незначущий результат просто вказує на те, що розміру вибірки недостатньо, щоб відкинути нульову гіпотезу. Пам'ятайте, що загальним консенсусом є використання $p < .05$. Такий висновок про значущість можна перекласти так, що з 95% ступенем впевненості можна стверджувати, що це не пов'язано з випадковістю, спричиненою процесом вибірки [1].

І навпаки, значущий результат підтверджує альтернативну гіпотезу лише в тій мірі, в якій існує лише одна альтернативна можливість після того, як нульова гіпотеза відкинута. Він не робить різниці між альтернативними гіпотезами, якщо існує більше однієї альтернативної гіпотези.

У правоохоронній діяльності може виявитися, що значущість не настільки важлива, як розмір ефекту. Якщо розмір ефекту є значним, може виявитися, що навіть відносно високе значення *p* (більше за 0.05, наприклад) є достатнім приводом для інформування осіб, які приймають рішення. Важливо пам'ятати про ці міркування під час перевірки гіпотез.

СТАТИСТИЧНІ МЕТОДИ ВИЯВЛЕННЯ ТА ПЕРЕВІРКИ СТАТИСТИЧНИХ ЗВ'ЯЗКІВ

Більш складні статистичні методи, такі як кореляційний або регресійний аналіз, досліджують взаємозв'язок між двома або більшою кількістю змінних, щоб краще зрозуміти, як ці змінні пов'язані одна з одною. Використовуючи такі інструменти, як дескриптивний, кореляційний і регресійний аналіз, аналітики підрозділів кримінального аналізу можуть краще зрозуміти, як взаємодіють дані про злочинність та інші дані про громадську безпеку.

Міри тісноти зв'язку – кореляція

У наборах даних, що підлягають одно- або багатовимірному аналізу, неявно передбачається, що значення, або числа, будь-якої змінної представляють послідовно визначені міри деякого поняття або об'єкта, і що вони логічно і дефінітивно пов'язані один з одним у значущий спосіб. В іншому випадку статистичні вимірювання будуть безглуздими. Хоча цей момент може здатися очевидним, варто поставити низку запитань до будь-якого набору даних, на основі якого планується проведення аналізу, щоб переконатися, що все в порядку, перш ніж продовжувати.

Перед початком аналізу необхідно розглянути певні питання, що стосуються будь-якого набору даних.

> *Перше – чи пов'язані між собою логічно значення, що представляють змінну сутність або поняття.*
>
> *Друге – який тип даних, маючи на увазі, що існують статистичні операції, які підходять для даних кожного типу, і що використання невідповідних процедур призводить до недостовірних результатів.*

Третє питання полягає в тому, чи можна належним чином порівнювати значення безпосередньо одне з одним, тобто чи базуються вони на одному стандарті вимірювання.

Для розуміння зазначених вимог IACA у своєму посібнику [1] наводить кілька прикладів відповідного набору даних.

Набір даних 1. *Номери телефонів місцевих мешканців, щодо яких є невиконані ордери на арешт. Почнемо з найпростішого. Незважаючи на те, що номери телефонів виглядають як числа, насправді вони є номінальними даними – кодовими назвами з використанням цифр – і тому абсолютно непридатні для будь-яких математичних обчислень або статистики. (Впорядкувати їх для полегшення пошуку може бути корисно, але це не буде статистикою).*

Набір даних 2. *Кількість років формальної шкільної освіти, яку закінчили дорослі, заарештовані за будь-які або окремі види правопорушень. Можна сказати, що такі вимірювання «операціоналізують» концепцію освітнього рівня, тобто переводять абстракцію освіти в більш відчутний і вимірюваний формат. Незважаючи на певні обмеження, такі як труднощі з обліком різних видів навчальних програм поза межами шкільних систем або коледжів, можна сказати, що таке вимірювання репрезентативно відображає рівень освіти, а дані будуть мати тип шкали відношень. Однак може виникнути проблема з третім критерієм – критерієм «рівних умов». Якщо деякі заарештовані були надто молодими, вони ще не встигли закінчити стільки років формальної освіти, скільки старші за віком заарештовані. (Саме тому перепис населення США надає ці дані лише для осіб віком 25 років і старше). З поправкою на врахування лише заарештованих мінімального віку, ця змінна виступає як належним чином визначена єдина змінна для використання в багатовимірному аналізі з різними цілями.*

Набір даних 3. *Кількість крадіжок автотранспорту протягом певного року в містах з населенням понад 100 тис. осіб. Підрахунки будь-якого повторюваного явища, такого як тип або категорія злочину, можна вважати достовірними даними типу відношення, за умови, що злочин визначається та реєструється у досить схожий спосіб у кожному місті. Однак, очевидною проблемою тут є те, що розміри міст варіюються в широкому діапазоні, так що кількість крадіжок транспортних засобів в одному місті не може бути справедливо порівняна з кількістю крадіжок в іншому місті, яке значно відрізняється за розміром. Для того, щоб поставити випадки (міста) в рівні умови, необхідно стандартизувати набір даних шляхом перетворення кількості злочинів у коефіцієнти злочинності. Зазвичай прийнято переводити кількість злочинів у показники на 100 000 осіб. Використання таких показників дає змогу порівнювати міста з різною кількістю населення (оминаючи питання відмінностей у процедурах збору та кодування даних у різних містах).*

Набір даних 4. *Кількість скарг на порушення громадського порядку (викликів поліції або інцидентів, залежно від місцевих термінів та процедур) щомісяця в юрисдикції протягом певного періоду часу. Оскільки географічна територія залишається незмінною, кількість скарг на порушення громадського порядку (шкала відношень) є відповідними значеннями єдиного визначеного поняття (якими б нечіткими не були його межі на практиці). Ця змінна містить характеристику, яка не зустрічалася в жодному з попередніх випадків. Вона складається з двох змінних: кількості порушень та інтервалів часу.*

З перерахованих прикладів, лише набір даних 4 відповідає умовам багатовимірного аналізу, оскільки визначено дві (або більше) змінних і увага приділяється спостережуваним зв'язкам між змінними.

Таким чином, отримавши 52 щотижневих підрахунки викликів про порушення громадського порядку (або будь-яке інше повторюване явище, що нас цікавить), ми можемо визначити середнє значення, дисперсію та стандартне відхилення, а також максимальну й мінімальну кількість викликів за рік. Також ми можемо повідомити або використати в подальшому аналізі, що ми «бачили» близько 25 квартирних крадіжок на тиждень, «в середньому», плюс-мінус кілька (на основі дисперсії). Такі дескриптори можуть дати нам краще уявлення про частоту виникнення проблеми або, принаймні, полегшити її пояснення іншим. Але чи буде таке середнє значення «найкращим припущенням» щодо найбільш імовірної приблизної кількості скарг на порушення громадського порядку, які можна очікувати протягом наступного, 53-го тижня?

Використання середнього значення як центральної тенденції прогнозу (а також дисперсії) на кілька тижнів вперед є, більш за все, процедурною помилкою, бо базується на припущенні про лінійність даних, тобто, що значення даних, розподілені в часі, утворюють більш-менш пряму лінію. Але це не так, і часто не так. Суспільні явища, зокрема і злочинність та порушення громадського порядку, найчастіше характеризуються різними видами нелінійних моделей зв'язку.

Кореляцію можна описати як ступінь, до якого два або більше значення (змінні) змінюються в унісон, й зв'язок описується прямою лінією. Коли дві характеристики мають позитивну кореляцію, зі збільшенням однієї збільшується і друга. Якщо кореляція негативна, то зі збільшенням значення однієї змінної значення іншої змінної буде зменшуватися. Це називається парною кореляцією.

> *Цей зв'язок не обов'язково має бути причинно-наслідковим, але за визначенням вказує на залежність.*

Сила (тіснота) цього зв'язку вимірюється коефіцієнтом кореляції Пірсона зі значеннями від +1 (значення зростають разом) до від'ємного (-1) (зі збільшенням одного значення інше зменшується).

Значення коефіцієнту кореляції, близькі до нуля, вказують на те, що лінійна залежність між досліджуваними величинами незначна або відсутня.

Таким чином, застосування кореляції до вимірювання причинновості зв'язку може вводити в оману, виходячи з того, що:

- навіть в тих випадках коли можливим є уявити наявність причиновного зв'язку між двома змінними, що корелюють, сам по собі коефіцієнт кореляції нічого не говорить про те чи впливає одна змінна на іншу, чи навпаки;

- нерідко виявлений зв'язок існує завдяки наявності та певного зв'язку з іншими змінними;

- взаємозв'язки соціального характеру, що і характеризують злочинність, завжди надзвичайно складні й ніяким чином не можуть бути пояснені виключно однією причиною.

Хоча кореляція прямо не вказує на причиновний зв'язок, вона може слугувати ключем до розгадки причини. За певних умов на її основі можливим є сформулювати гіпотези, що перевіряються експериментально, коли можливим є контроль впливу інших змінних, що виходять за наявні межі.

Існує багато різних випадків, коли аналітики підрозділів кримінального аналізу можуть бути зацікавлені у вивченні взаємозв'язку між двома або більше змінними в поліцейських даних.

***Наприклад**, кореляцію можна дослідити між змінними, як пов'язаними, так і не пов'язаними зі злочинністю, наприклад, даними перепису населення. Хоча існує буквально тисячі змінних, доступних для аналізу на основі даних перепису населення, деякі змінні, які аналітики можуть розглянути для включення в аналіз злочинності (окрім кількості населення), включають: житло (наприклад, вільне, орендоване, багатоквартирне), транспорт/власність на транспортний засіб, дохід та зайнятість.*

Залежно від рівня шкали вимірювання та характеру розподілу даних (нормального чи ненормального) можуть використовуватися різні типи коефіцієнтів кореляції. Ці обмеження називаються припущеннями щодо природи досліджуваних даних. Порушення цих припущень може поставити під сумнів результати аналізу.

Коефіцієнт кореляції Пірсона

Метод, який зазвичай використовується для вимірювання тісноти лінійного зв'язку відносних або інтервальних змінних, відомий, на честь його автора, як коефіцієнт кореляції Пірсона, і позначається літерою «*r*».

Цей коефіцієнт кореляції, який часто називають *r* Пірсона, використовується для визначення сили та напрямку лінійного зв'язку між двома змінними. Це означає, що за умови дотримання припущень цей коефіцієнт може допомогти визначити ступінь і напрямок, в якому одна змінна змінюється по відношенню до іншої змінної. Наприклад, коли $r = 1$, існує досконала і позитивна асоціація, що означає, що зі збільшенням однієї змінної, інша зростає в унісон (створюючи пряму лінію). Коефіцієнт Пірсона матиме від'ємне значення, коли одна змінна збільшується, а інша зменшується.

Нуль означає відсутність лінійного зв'язку, що дуже малоймовірно, оскільки майже кожна можлива пара змінних демонструє принаймні незначну, випадкову кореляцію.

Як зазначалося вище, кореляція не доводить причинново-наслідковий зв'язок між змінними.

Загалом, кореляція, що наближається до 1.0 (або -1.0), вказує на сильний зв'язок, але наскільки сильним є цей зв'язок? Нам потрібен спосіб пов'язати r-бали з чимось більш відчутним. Відповідь лежить у квадраті r Пірсона, показнику r-квадрат.

> *Відомий як коефіцієнт детермінації, r-квадрат показує частку дисперсії однієї змінної, що пояснюється іншою змінною.*

Таким чином, якщо r = 0.8, r-квадрат дорівнює 0,64, що означає «64% дисперсії Y пояснюється X». Інші 36% дисперсії залишаються під впливом інших факторів.

У прикладі щодо квартирних крадіжок коефіцієнт кореляції, r Пірсона, наприклад, становив 0,775, що дає коефіцієнт детермінації (r-квадрат) 0,601, який можна інтерпретувати як «близько 60% дисперсії пояснюється незалежною змінною». Але, очевидно, не можна сказати, що плин часу (незалежна змінна) «спричинив» зміну кількості квартирних крадіжок; що б не спричиняло

збільшення або зменшення кількості квартирних крадіжок, це точно не сам час, а якісь треті змінні, з ним пов'язані.

Існує кілька важливих припущень, яких необхідно дотримуватися за використання r Пірсона:

- *змінні, що вимірюються, повинні бути неперервними (тобто, виміряні в інтервальній шкалі або шкалі відношень);*
- *обидві змінні формують двовимірний нормальний розподіл;*
- *цей коефіцієнт є чутливим до викидів, а тому їх не повинно бути;*
- *зв'язок між двома змінними є лінійним.*

Цей приклад ілюструє два важливих міркування для аналітиків кримінального аналізу, які використовують більш досконалі статистичні методи.

По-перше, важливо вивчити дані про злочинність до застосування статистичного тесту, щоб переконатися, що вони відповідають необхідним припущенням. З даними про злочинність рідко буває легко працювати, і вони часто потребують або трансформації для забезпечення відповідності припущенню про нормальність, або непараметричного тесту.

По-друге, непараметричні дані не є нормально розподіленими.

Але існують міри кореляції, які підходять для не нормально розподілених даних.

Коефіцієнт рангової кореляції Спірмена

Кореляція Спірмена використовує ранжування змінних для мінімізації впливу викидів. Цей тест можна використовувати з двома неперервними змінними, двома порядковими змінними або однією порядковою змінною та однією неперервною змінною.

На додаток до припущень щодо r Пірсона, для використання цього тесту необхідно виконати ще два припущення:

- *змінні, що аналізуються, повинні бути парними спостереженнями;*
- *між змінними має бути монотонний зв'язок (зі збільшенням однієї змінної інша змінна або збільшується, або зменшується).*

Перевагою монотонного зв'язку є те, що залежність між даними не обов'язково має бути лінійною (дані повинні послідовно зростати та/або зменшуватися по відношенню один до одного, але не обов'язково описуватися прямою лінією).

Критерій хі-квадрат на незалежність категоріальних змінних

У наведених вище прикладах ми використовували тести зв'язків, які підходять для відносних, інтервальних та порядкових даних. Тест хі-квадрат на незалежність (зазвичай називається хі-квадрат (χ^2)) має перевагу в тому, що він підходить для номінальних/дихотомічних змінних. Під час використання цього тесту метою є визначення того, чи є дві номінальні категоріальні змінні статистично незалежними. Це просто означає, що знання однієї змінної не дає жодних знань про властивості іншої змінної.

Наприклад, вираз $\chi^2 = 596{,}983$, $p > 0{,}001$ вказує на наявність зв'язку між двома змінними, що тестуються. Але для визначення сили та напряму цього зв'язку необхідно вже застосувати відповідне вимірювання цього зв'язку.

Важливо пам'ятати, що жоден з описаних вище тестів не визначає причинно-наслідковий зв'язок. Навіть якщо зв'язок (або кореляцію) виявлено, все ще існує ймовірність того, що неідентифікована третя змінна корелює з обома факторами і спричиняє цей зв'язок.

Дисперсійний аналіз (ANOVA)

Дисперсійний аналіз корисний, коли аналітик хоче порівняти середні значення з трьох або більше незалежних вибірок, які походять з однієї або більше категорійних груп. Існує кілька типів ANOVA, які використовуються в різних ситуаціях.

Однофакторний ANOVA можна використовувати для більш ніж двох груп. Важливо, що однофакторний дисперсійний аналіз – це всеохоплююча тестова статистика. Це означає, що він лише перевіряє, чи існують значущі відмінності середніх між досліджуваними групами. Він не скаже аналітику, в чому полягають ці відмінності. Враховуючи, що аналітик може досліджувати кілька груп, може виникнути необхідність у проведенні додаткових тестів.

Існує кілька припущень, яких слід дотримуватися у разі проведення однофакторного ANOVA:

- ✓ *залежна змінна повинна бути неперервною;*
- ✓ *спостереження повинні бути незалежними одне від одного;*
- ✓ *незалежна змінна є номінальною і два або більше значень.*

Оскільки це параметричний тест, розподіл залежної змінної повинен наближатися до нормального у всіх популяціях. Нарешті, в даних не повинно бути значних викидів, а дисперсії залежної змінної мають бути однорідними в досліджуваних популяціях.

В однофакторному дисперсійному аналізі аналітик прагне порівняти кілька рівнів одного фактора. Однак можуть бути випадки, коли аналітик хоче розглянути два фактори, кожен з яких має кілька рівнів.

> ***Наприклад**, аналітик хоче дослідити вплив освіти (вимірюваної категоріально як середня школа, неповний коледж або закінчений коледж) і звання (офіцер, сержант або капітан) на моральний дух.*

Двофакторний ANOVA може визначити, чи існує ефект статистичної взаємодії між двома незалежними номінальними змінними (факторами) на залежну змінну. Використовуючи наш приклад, цей тест дозволить нам побачити, чи впливає освіта на бойовий дух і чи відрізняється вплив освіти залежно від звання.

Однак такий тест зазвичай проводять лише тоді, коли аналітик вважає, що обидві незалежні змінні (освіта і звання) певним чином взаємодіють між собою. Наприклад, аналітик може припустити, що звання впливає на моральний дух, але цей вплив залежить від рівня освіти.

Це називається ефектом взаємодії. Як і у випадку з однофакторним ANOVA, цей тест визначить статистично значущі ефекти, але також потребує подальшого аналізу.

Прогноз та оцінювання – регресійний аналіз

Кореляція вимірює відносну «відповідність» або ступінь зв'язку між двома змінними. Також цей метод дає змогу оцінити надійність будь-яких прогнозів якщо у доповнення використати якісну характеристику лінії регресії.

> *Регресійний аналіз – це метод моделювання зв'язку між залежною змінною (змінною, на яку впливають) та однією, двома або більше незалежними змінними.*

Регресійний аналіз часто використовується в процесі прогнозування, що має очевидну цінність для аналітика в кримінальному аналізі, але він також може допомогти визначити конкретні незалежні змінні, пов'язані з залежною змінною.

Регресійний аналіз використовується для встановлення причинно-наслідкового зв'язку, коли кореляція вже виявлена. Однак через низьку якість поліцейських даних, мультиколінеарність та потенційну можливість упередженості щодо пропущених змінних, до спроб виявлення причинно-наслідкового зв'язку слід підходити з обережністю.

Під час аналізу часових рядів основне питання полягає в тому, чи існує, і якщо так, то в якій мірі, основна тенденція, що лежить в основі роботи. Тобто, чи свідчать відомі значення даних про наявність якоїсь лінії тренду, що проходить через розкид точок даних, навіть якщо вона рідко торкається або проходить безпосередньо через будь-яку з окремих точок на графіку. Якщо таку лінію тренду можна виявити й побудувати, і якщо ця лінія виявиться достатньо якісним відображенням структури точок, то можна сказати, що вона має певний прогностичний потенціал [1].

> *По суті, регресійний аналіз - це процедура розпізнавання закономірностей, що є однією з головних цілей кримінального аналізу (IACA) [1].*

Навіть якщо аналітик рідко стикається з можливістю проведення повномасштабного регресійного аналізу, елементарне розуміння його концепцій і місця у світі наукового аналізу має виявитися цінним хоча б для того, щоб допомогти відточити свої навички мислення.

Регресійний аналіз вирішує завдання пошуку загальної закономірності, що проявляється в спостереженнях за явищами, включаючи тенденції злочинності або порушень громадського порядку, взаємозв'язки між соціально-економічними явищами і поведінкою, а також інші різноманітні й складні аспекти сфери діяльності аналітика підрозділу кримінального аналізу.

Регресія передбачає розпізнавання закономірностей і чітке, хоча й не зовсім досконале, розміщення закономірності на лінії тренду, прямій чи кривій відповідно до якоїсь іншої математичної функції. Її партнер, кореляційний аналіз, показує нам, наскільки сильним є зв'язок і до якої міри ми можемо вважати його значущим, корисним і надійним.

Таким чином, *регресійний і кореляційний аналіз* – це дві сторони однієї медалі:

> ***регресія** знаходить найкращу лінію, що проходить через скупчення точок даних, і показує, як вона виглядає;*
>
> ***кореляція** показує, наскільки сильним є лінійний зв'язок і до якої міри ви можете робити з нього висновки.*

> *За наявності достатньо сильного зв'язку лінія регресії дає можливість передбачити майбутні значення залежної змінної.*

Умовно кажучи, метою регресійного аналізу є пошук найкращої прямої лінії, що проходить через хмарку точок (рис. 19.1). Такою лінією є та, яка порівняно з усіма іншими можливими розташуваннями має найменшу дисперсію, що відображає розбіжності між фактичним розміщенням точок даних на графіку та відповідними розташуваннями на найкращій прямій лінії. Точніше кажучи, розташування лінії регресії є таким, що мінімізує суму квадратів відхилень – тобто дисперсію похибок передбачення.

Алгебраїчна формула прямої має вигляд: $y = mX+b$

де:

- m – константа, відома як коефіцієнт нахилу;

- b – константа, що представляє місце на осі y, де лінія перетинається, тобто значення y при $X = 0$.

Рисунок 19.1. Діаграма регресії (*з використанням SPSS Statistics*)

> *У Microsoft Excel додавання лінії тренду до діаграми розсіювання – це просте завдання: натиснути правою кнопкою миші на одній з точок даних і вибрати «Додати лінію тренду». Програма дозволяє створювати численні типи ліній тренду (є низка варіантів нелінійної функції), в тому числі лінійну лінію тренду, яку продемонстровано (рис. 19.1). Також є можливість переглянути рівняння та його результати (а також значення R-квадрат).*

Таким чином, лінія регресії працює для знаходження набору точок (у цьому випадку лінійно вирівняних), які мінімізують квадратичні відхилення (дисперсію навколо середнього значення) фактичних значень від прогнозованих значень Y.

У випадку з лінією регресії дисперсія – це, по суті, те саме, за винятком того, що відхилення, які підносяться в квадрат, – це відхилення від фактичних точок до прогнозованих точок вздовж осі y для кожного значення X.

Таким чином, лінія регресії замінює середнє значення в одновимірному аналізі. Зауважимо, що стандартне відхилення було квадратним коренем з дисперсії у випадку однієї змінної, стандартна похибка, або стандартна похибка оцінки, є квадратним коренем з дисперсії у випадку регресії. Очевидно, що чим більша стандартна похибка оцінки, тим менша точність, або якість, лінії регресії, і тим менша надійність прогнозів майбутніх значень.

> *Це свідчить про здатність лінії регресії проектуватися в майбутнє і відображає той факт, що сутність математичного прогнозування полягає в регресійному аналізі (IACA) [1].*

Регресійний аналіз жодним чином не обмежується даними часових рядів. Насправді, дуже часто дві (або більше) змінні досліджуються разом, щоб визначити не тільки їхню лінію тренду (лінійну або іншу), але й ступінь зв'язку між ними, тобто їхню кореляцію. Ці процедури можна проводити з трьома або більше змінними одночасно. Коли задіяно три і більше змінних, процедури дещо ускладнюються, хоча програмне забезпечення може легко впоратися з цією проблемою.

Зазвичай, в аналізах, які використовують регресію та кореляцію, одна змінна розглядається як залежна від значень іншої, незалежної змінної, як така, що випливає з неї, з можливою, і часто оманливою, видимістю причинно-наслідкового зв'язку. Тут є певна аномалія, яку слід врахувати:

у регресійному аналізі ми будуємо дослідження таким чином, щоб припустити або шукати можливість причинно-наслідкового зв'язку – як мінімум, щоб сказати, що *«y, здається, випливає з x»* в тій чи іншій мірі. Проте неможливо визначити причинно-наслідковий зв'язок лише на основі спостережуваного патерну зв'язків [1].

Головною причиною цього є те, що можуть існувати інші фактори «під поверхнею» або поза межами безпосереднього аналізу. Ці невідомі фактори можуть впливати на значення залежної змінної, окрім тієї, що визначена як незалежна змінна. Крім того, можливо, що якась латентна змінна насправді опосередковано виявляє себе в обох досліджуваних змінних (або спричиняє їх) [1].

Звичайний хід подій у дослідницькому процесі полягає у зборі необхідних даних (достатньої кількості для обґрунтування достовірних висновків) та висуненні гіпотези, яку можна перевірити, який взаємозв'язок очікується, чому він очікується та його значущість. Таким чином встановлюється мета і сенс аналізу.

Види регресійного аналізу

Лінійна регресія

Лінійна регресія використовує існування лінійного зв'язку для прогнозування значення однієї змінної (залежної змінної) на основі значення іншої змінної (незалежної змінної). Для використання лінійної регресії необхідно, щоб обидві змінні були безперервними. Як випливає з назви, зв'язок між залежною і незалежною змінними повинен бути лінійним. Як і у випадку з кореляцією Пірсона, викиди можуть впливати на результати.

На додаток до основних критеріїв, лінійна регресія також вимагає двовимірної нормальності, гомоскедастичності та відсутності автокореляції (це кореляція змінної з самою собою в часі). Нарешті, лінійна регресія чутлива до мультиколінеарності.

> *Важливо пам'ятати, що порушення цих припущень може призвести до неточних результатів (IACA) [1].*

Існує кілька способів перевірити ці умови, але це завдання більш поглибленого додаткового вивчення, що виходить за межі цього посібника.

Лінійна регресія може бути корисною в окремих випадках, але часто наші дані не описуються лінійними закономірностями. У таких випадках використання лінійної регресії може виявитися неінформативним або навіть ввести аналітика в оману.

Логістична регресія

Логістичну регресію можна використовувати для моделювання зв'язку категоріальної залежної змінної. Це часто буває корисно, оскільки дані, які використовують аналітики, часто мають категоріальний характер (наприклад, типи арештів або демографічні змінні, такі як раса чи стать).

> ***Наприклад**, під час моделювання бінарного результату – арешт або відсутність арешту, застосування сили за поліцейським радіо-викликом або відсутність застосування сили за поліцейським радіо-викликом, аналітик може розрахувати співвідношення шансів настання події (або відсутності події), контролюючи при цьому низку незалежних змінних. Це може бути надзвичайно корисним для складання певних видів прогнозів [1].*

Пуассонівська регресія

Для моделювання даних підрахунку частоти якоїсь події можна використовувати пуасонівську регресію. З даними підрахунку може бути важко працювати через наявність нулів і відсутність від'ємних значень. Це, як правило, створює позитивно асиметричний набір даних, який часто не може бути перетворений на щось, що наближається до нормального розподілу. Пуассонівська регресія є одним з методів аналізу таких даних.

Пуассонівська регресія припускає, що наявні помилки розподілені за пуассонівським (на відміну від нормального) розподілом. За визначенням, це означає, що середнє значення й дисперсія помилок рівні. У даних про злочинність це часто не так.

> ***Наприклад**, якщо розглядати арешти за викрадення автомобілів, більшість людей не мають жодного арешту, але набагато менша група може мати десятки арештів. Це може призвести до того, що дисперсія буде більшою за середнє значення [1].*

У таких випадках доцільніше використовувати негативну біноміальну регресію.

Негативна біноміальна регресія

Негативна біноміальна регресія може бути використана, коли залежна змінна є числом (наприклад, кількість арештів за незаконне заволодіння автомобілем) і є надмірно розпорошеною.

***Прикладом** такого типу аналізу може бути ситуація, коли аналітик прогнозує кількість арештів за крадіжки з магазинів на основі характеристик підприємства (наприклад, наявність охорони під прикриттям або охорони у формі, тип магазину (біг-бокс, розмір або обсяг продажів) та певних показників соціально-економічного стану району, в якому розташоване підприємство [1].*

Аналітик може продемонструвати, що наявність охорони під прикриттям у порівнянні з охороною у формі (в такому випадку – предиктор або залежна змінна) призводить до більшої або меншої кількості арештів, незважаючи на інші фактори. За умови правильного проведення, такі аналізи можуть бути дуже корисними для запобігання злочинності або обговорення політики. Неправильно виконаний аналіз, у кращому випадку, може виставити аналітика з питань злочинності дурнем і/або некомпетентним. Використовуйте їх на свій страх і ризик [1].

Зазначене лише дуже коротко характеризує різні типи регресійного аналізу. Якщо аналітик бажає дізнатися більше про ці методи, доцільно прослухати кілька курсів зі статистики, особливо ті, що присвячені соціальним наукам.

> *Таким чином, оскільки дані стають все більш доступними, на аналітиків кримінального аналізу зростатиме тиск щодо вдосконалення способів проведення аналізу. Популярність нових «прогностичних» методів лише посилить цей тиск. Тому вивчення нових форм аналізу дає аналітикам неймовірну можливість підвищити свою цінність.*

Як зазначено у посібнику IACA «*Exploring crime analysis: Readings on essential skills*» [1], незалежно від того, чи це простий опис центральної тенденції різних факторів у ряді (наприклад, «середній час інцидентів становить 15:33; 70% злочинів сталися між 15:00 і 16:00»), чи використання регресійного аналізу для прогнозування рівня викликів на наступний рік, описова і багатовимірна статистика відіграє величезну роль в аналізі злочинності.

Кримінальні бюлетені та щорічні звіти аналітиків переповнені середніми, медіанами, модами, темпами, відсотковими змінами та кореляціями. Найчастіше ці інструменти використовуються правильно, але іноді їх застосовують некоректно, що призводить до сумнівних висновків і рекомендацій. Глибоке розуміння як описової, так і вивідної статистики має вирішальне значення для якісного аналізу та прийняття правильних рішень.

> *Важливою є потреба в осмисленні, аналізі та інтерпретації статистичних даних. Завжди пам'ятайте, що аналіз злочинності – це соціологічна сфера, а в соціології всі дані певною мірою є якісними. Статистика - це основа, а не продукт аналізу (IACA) [1].*

ПРАКТИЧНІ КЕЙСИ
ВИКОРИСТАННЯ СТАТИСТИЧНИХ МЕТОДІВ В КРИМІНАЛЬНОМУ АНАЛІЗІ

Приклад 1.
Аналіз інформаційних загроз в контексті гібридної війни: оцінка ризиків та прогнозування

Гібридні загрози у системі сучасної безпеки суспільства займають одне із надзвичайно важливих місць. Особливо це важливо для країн, де зовнішня агресія або інші форми збройних конфліктів знаходяться в активній фазі. Безпекова ситуація вимагає підвищення рівня обізнаності щодо інформаційних гібридних загроз та спроможності ефективної протидії їх поширенню [2]. Сучасний ризик-менеджмент передбачає аналіз та моніторинг як ключових загроз так і факторів, що формують механізми ефективної протидії таким загрозам, формування системи стійкості до загроз, прийняття управлінських рішень до підвищення спроможності протидії гібридним загрозам, уникнення вразливих компонентів у діяльності суб'єктів безпеки тощо.

***Аналітичним завданням визначено** – на основі ризик-орієнтованого підходу:*

- *оцінити інформаційні гібридні загрози у секторі цивільної безпеки,*
- *проаналізувати [3] фактори, що визначають ефективність протидії гібридним загрозам,*
- *сформувати систему індикаторів спроможності|вразливості [4], що корелюються з інформаційними гібридними загрозами та впливають на рівень зниження ризику їх поширення;*
- *на основі регресійного аналізу [5] побудувати прогнозну модель поширення інформаційних гібридних загроз за умови підвищення ключових спроможностей та зниження рівня ризику вразливості системи протидії інформаційним гібридним загрозам.*

На основі відповідної методології [2], під час оцінювання як загроз, так й індикаторів спроможності/вразливості системи протидії цим гібридним загрозам, застосовано ризик-орієнтований підхід. Крім

того, виділено узагальнюючі групи гібридних загроз у секторі цивільної безпеки, бо гібридизація може поширюватися на усі сфери суспільних відносин. Однією із ключових груп гібридних загроз у сфері цивільної безпеки є «*гібридні загрози пов'язані з інформаційним середовищем*» [6].

З усього масиву [2] ідентифікованих гібридних загроз у сфері цивільної безпеки, з інформаційним середовищем пов'язано 25 гібридних загроз, частина з яких (11) відображені в *табл. 19.2* з визначеним рівнем ризику.

Таблиця. 19.2. Оцінка ризиків поширення гібридних загроз

Гібридні загрози пов'язані з інформаційним середовищем	**№ТА**	**Risk, %**
Створення та підтримка (в т.ч. фінансова) різноманітних політичних проектів з метою розхитування соціально-політичного стану, дискредитації керівництва держави, збудження населення до громадянської непокори тощо	ТА3	**51,36**
Стимулювання та заохочення громадян України, що мешкають в прикордонних регіонах, до отримання 2-го громадянства всупереч закону	ТА4	**53,33**
Підбурювання окремих верств населення до громадянської непокори й висунення економічних та політичних вимог	ТА28	**41,10**
Контроль агресором контенту у регіонах близьких до бойових дій та прикордонних районів	ТА39	**51,04**
Провокаційна діяльність контрольованого агресором медіа ресурсу через приховування кінечного бенефіціара	ТА40	**48,14**
Просування агресивного деструктивного наративу щодо органів і працівників системи МВС через інформаційні засоби	ТА41	**41,37**
Вплив на свідомість працівників органів системи МВС, шляхом цілеспрямованої дискредитації та зниження довіри до керівництва	ТА42	**38,60**
Вплив на свідомість працівників органів системи МВС, шляхом цілеспрямованої дискредитації, спрямованої на зниження довіри до політичної еліти в державі	ТА42.4	**43,59**
Вплив на свідомість населення з метою дискредитації влади, органів і працівників системи МВС	ТА44	**49,15**
Вплив на свідомість населення з метою дискредитації влади, органів і працівників системи МВС шляхом використання цільових інформаційно-психологічних операцій	ТА44.1	**43,38**
Вплив на свідомість населення з метою дискредитації влади, органів і працівників системи МВС шляхом проведення компл інформкампаній із штучної поляризації та радикалізації суспільства, використовуючи збудженість громадян до проблеми	ТА44.4	**44,62**

Використовуючи підходи, що базуються на відображенні граничних розмірів ризиків, значні загрози, в межах 50-60%, виділено оранжевим кольором, а менш значні (40-50%) – жовтим та (менше 40%) зеленим кольорами.

Таким чином, за рівнем ризику поширення значними гібридними загрозами, пов'язаними з інформаційним середовищем, є такі:

- *стимулювання та заохочення громадян України, що мешкають в прикордонних регіонах, до отримання 2-го громадянства всупереч закону* – 53,33%;
- *створення та підтримка (в т.ч. фінансова) різноманітних політичних проєктів з метою розхитування соціально-політичного стану, дискредитації керівництва держави, збудження населення до громадянської непокори тощо* – 51,36%;
- *контроль агресором контенту в близьких до бойових дій регіонах та прикордонних районах* – 51,04%.

Незважаючи на те, що за середнім значенням, під час оцінювання ризику поширення гібридної загрози: в*плив на свідомість працівників органів системи МВС, шляхом цілеспрямованої дискредитації та зниження довіри до керівництва*, – одержано 38,6%, експертна оцінка, згідно з розподілом за усією шкалою, у 30% експертів вбачає оцінювання ризику на рівні вище 50% (*рис. 19.2*). Зазначене вказує на необхідність додаткового контролю за визначеною загрозою та певну її актуальність у сучасних умовах.

Продовжуючи аналіз, важливим є оцінювання спроможностей системи цивільної безпеки щодо зниження ризиків поширення гібридних загроз, пов'язаних з інформаційним середовищем. Для оцінювання спроможностей використано індикатори, що характеризують ефективність діяльності органів системи цивільної безпеки щодо протидії гібридним загрозам [2]. Оцінювання здійснено на основі ризик-орієнтованого підходу та враховує як ймовірність рівня ефективності, так і вплив щодо протидії гібридним загрозам.

Всього використано 152 індикатори, що характеризують ефективність (рівень можливостей, спроможностей та вразливостей) системи цивільної безпеки у протидії гібридним загрозам [2]. Гіпотетично частина з них характеризуються певним статистичним зв'язком з окремими гібридними загрозами та можуть прямо або протилежно впливати на зниження рівня ризику поширення загрози.

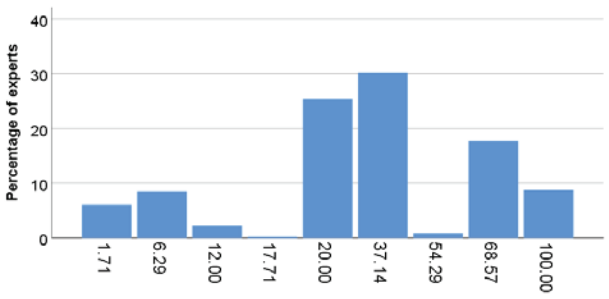


Рисунок 19.2. Вплив на свідомість працівників органів системи МВС, шляхом цілеспрямованої дискредитації та зниження довіри до керівництва

З метою підтвердження зазначеної гіпотези, на основі лінійної регресії здійснено аналіз (з *використанням SPSS Statistics*) за методом STEPWISE (відповідний синтаксис відображено на рис. 19.3), де залежною змінною використано гібридну загрозу TA42 – в*плив на свідомість прац. органів системи МВС, шляхом цілеспрямованої дискредитації та зниження довіри до керівництва*.

```
REGRESSION
 /DESCRIPTIVES MEAN STDDEV CORR SIG N
/MISSING LISTWISE
 /STATISTICS COEFF OUTS R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT TA42
 /METHOD=STEPWISE Rs1 Rs2 Rs3 Rs4 Rs5 Rs6 Rs7
Rs8 Rs9 Rs10 Rs11 Rs12 Rs13 Rs14 Rs15 Rs16 Rs17
Rs18 Rs19 Rs20 Rs21 Rs22 Rs23 Rs24 Rs25 Rs26 Rs27
Rs28 Rs29 Rs30 Rs31 Rs32 Rs33 Rs34 Rs35 Rs36 Rs37
Rs38 Rs39 Rs40 Rs41 Rs42 Rs43 Rs44 Rs45 Rs46 Rs47
Rs48 Rs49 Rs50 Rs51 Rs52 Rs53 Rs54 Rs55 Rs56 Rs57
Rs58 Rs59 Rs60 Rs61 Rs62 Rs63 Rs64 Rs65 Rs66 Rs67
Rs68 Rs69 Rs70 Rs71 Rw1 Rw2 Rw3 Rw4 Rw5 Rw6
Rw7 Rw8 Rw9 Rw10 Rw11 Rw12 Rw13 Rw14 Rw15
Rw16 Rw17 Rw18 Rw19 Rw20 Rw21 Rw22 Rw23 Rw24
Rw25 Ro1 Ro2 Ro3 Ro4 Ro5 Ro6 Ro7 Ro8 Ro9 Ro10
Ro11 Ro12 Ro13 Ro14 Ro15 Ro16 Ro17 Ro18 Ro19
Ro20 Ro21 Ro22 Ro23 Ro24 Ro25 Ro26 Ro27 Ro28
Ro30 Ro31 Ro32 Ro33 Ro34 Ro35 Ro36 Ro37 Ro38
Ro39 Ro40 Ro41 Ro42 Ro43 Ro44 Ro45 Ro46 Ro47
Ro48 Ro49 Ro50 Ro51 Ro53 Ro54 Ro55 Ro56 Ro57.
```

Рисунок 19.3. Синтаксис лінійної регресії із залежною змінною TA42.

Використовуючи результати застосування лінійної регресії, була визначена статистична залежність гібридної загрози – «*вплив на свідомість працівників органів системи МВС, шляхом цілеспрямованої дискредитації та зниження довіри до керівництва*» від 5 з переліку 152 змінних, що характеризують ефективність діяльності органів системи цивільної безпеки щодо протидії гібридним загрозам (*табл. 19.3*).

Таблиця 19.3. Модель лінійної регресії із залежною змінною TA42

Factor[a]					
Model	Unstand factors		Standard factors	т	P value
	B	Std. Err.	Beta		
(Constant)	13,886	3,333		4,188	0,000
RS44	-0,154	0,051	-0,160	-3,110	0,000
RS51	0,290	0,055	0,277	5,396	0,000
RW1	0,105	0,037	0,128	2,776	0,006
RW2	0,131	0,034	0,177	3,830	0,000
RW9	0,141	0,035	0,180	4,020	0,000

Таким чином, ключовими предикторами за результатами регресійного аналізу є:

- *RS44 - Рівень особистої стійкості співробітників органів МВС до сприйняття дезінформації;*
- *RS51 - Відповідність професійної підготовки вимогам практики на рівні початкової (базової) підготовки;*
- *RW1 - Недосконале нормативно-правове забезпечення діяльності системи МВС;*
- *RW2 - Дискредитація реформи органів МВС окремими працівниками;*
- *RW9 - Недостатній рівень професійної підготовки управлінського рівня.*

Один предиктор RS44 має негативне значення, а тому з підвищенням його рівня, ризик поширення загрози буде мати тенденцію до зниження. Інші ж предиктори з однаковим знаком у моделі лінійної регресії, а тому впливатимуть на загрозу з прямою залежністю: підвищення рівня предиктора впливатиме на підвищення означеної загрози.

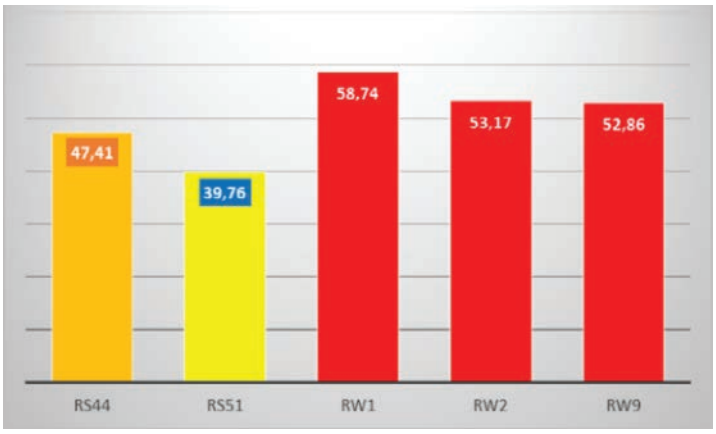


Рисунок 19.4. Оцінка ризику найзначніших предикторів.

Водночас надзвичайно важливим є врахування у моделі реального стану. У нашому варіанті це оцінювання вразливості або спроможності системи цивільної безпеки щодо зниження ризиків поширення інформаційних гібридних загроз. За даними дослідження

[2] рівень ризику щодо зазначених у моделі (*табл. 19.3*) предикторів характеризується переважно вразливістю (*рис. 19.4*).

Предиктори RW1, RW2, RW9 за своїм змістом вже є вразливостями оскільки оцінюються як «слабкі сторони».

У свою чергу, предиктори RS44 та RS51, хоча за своїм змістом є явищами позитивного характеру, за оцінкою рівня ризику, яка є меншою 50%, сприймаються більше як вразливість. Така думка нами доводиться експериментально із застосуванням граничних рівнів на основі матричного аналізу [2].

Із врахуванням рівня вразливості за ключовими предикторами та використовуючи модель лінійної регресії визначено рівень ризику поширення загрози «*вплив на свідомість прац. органів системи МВС, шляхом цілеспрямованої дискредитації та зниження довіри до керівництва*» (рис. 19.5)

Model and forecast	Coeff	P value	Rating %	
(Constant)	13,886	0,000		
RS44	-0,154	0,000	47,41	-7,280920591
RS51	0,290	0,000	39,76	11,5234602
RW1	0,105	0,006	58,74	6,144196434
RW2	0,131	0,000	53,17	6,98138524
RW9	0,141	0,000	52,86	7,438466821
TA42			**RA = 38,64 %**	

Рисунок 19.5. Оцінка ризику за залежною змінною TA42.

Враховуючи сучасний стан вразливості системи цивільної безпеки у протидії визначеній гібридній загрозі, на основі зазначених у моделі лінійної регресії предикторів, реальний рівень ризику поширення загрози *«вплив на свідомість прац. органів системи МВС, шляхом цілеспрямованої дискредитації та зниження довіри до керівництва»* (38,69 %) є дещо вищим ніж його оцінювання на основі інтегрованого показника значень імовірності та наслідків (38,6 %). У свою чергу, розглянута модель лінійної регресії формує базові засади для прогнозу зниження рівня ризику загрози *«вплив на свідомість працівників органів системи МВС, шляхом цілеспрямованої дискредитації та зниження довіри до керівництва»* за умови підвищення спроможності системи цивільної безпеки щодо протидії інформаційним гібридним загрозам (рис. 19.6).

Coefficient	P value	Rating +/-20%	
13,886	0,000		
-0,154	0,000	67,41	-10,35239099
0,290	0,000	39,76	11,5234602
0,105	0,006	38,74	4,05219901
0,131	0,000	33,17	4,355323461
0,141	0,000	32,86	4,624063938
			RA = 28,09 %

Рисунок 19.6. Прогноз впливу на ризик поширення гібридної загрози TA42 шляхом підвищення спроможності за ключовими предикторами на 20%

Таким чином, на основі оцінювання ризиків поширення *гібридних загроз, пов'язаних з інформаційним середовищем*, визначено найзначніші загрози:

- *стимулювання та заохочення громадян України, що мешкають в прикордонних регіонах, до отримання 2-го громадянства всупереч закону* – 53,33%;

- *створення та підтримка (в т.ч. фінансова) різноманітних політичних проектів з метою розхитування соціально-політичного стану, дискредитації керівництва держави, збудження населення до громадянської непокори тощо* – 51,36%;

- *контроль агресором контенту в близьких до бойових дій регіонах та прикордонних районах* – 51,04%.

Із загального переліку індикаторів (152), що характеризують ефективність правоохоронної системи у протидії гібридним загрозам, 5 з них характеризуються найбільшим впливом на загрозу «*вплив на свідомість працівників органів системи МВС, шляхом цілеспрямованої дискредитації та зниження довіри до керівництва*». Так усі предиктори за своїм рівнем оцінювання ризику характеризуються вразливістю системи, але по-різному впливають на рівень ризику визначеної загрози.

На основі моделі лінійної регресії побудовано прогноз зниження ризику поширення гібридної загрози *«вплив на свідомість прац. органів системи МВС, шляхом цілеспрямованої дискредитації та зниження довіри до керівництва»*, із врахуванням зміни рівня вразливості на 20 % (тобто підвищення спроможності) системи цивільної безпеки. За прогнозною моделлю, такі зміни ключових предикторів призводять до зменшення ризику поширення гібридної загрози *«вплив на свідомість працівників органів системи МВС, шляхом цілеспрямованої дискредитації та зниження довіри до керівництва»* з 38,64 % до 28, 09 %, більше ніж на 10 %.

Приклад 2.
Аналіз кримінального компонента гібридних загроз та прогнозування зниження ризику діяльності незаконних збройних формувань

В умовах гібридної війни у системі суспільної безпеки важливе місце займає розуміння того, що гібридизація може поширюватись на усі сфери соціальної поведінки, зокрема, не є винятком і злочинність. Враховуючи зовнішній характер загроз в умовах гібридної війни, тобто поширення зазначених загроз за безпосередньої чи опосередкованої ініціативи агресора, або третьою стороною в інтересах агресора, класичний академічний підхід до аналізу злочинності, а тим більше щодо запобігання її поширенню, є малоефективним.

За таких умов причини поширення злочинності завжди будуть мати зовнішню характеристику детермінації та найбільш імовірно, що їх провокація буде актуальною увесь період активності воєнної агресії. За таких обставин важко говорити про упереджувальний вплив на причини поширення злочинності без ліквідації факту агресії.

> *В умовах гібридної війни важливим є розуміння такого підходу, а тому можливим є виділення гіпотези щодо поширення гібридних загроз, пов'язаних з поширенням злочинності.*

Стратегії гібридної війни, крім суто збройної складової, включають принципи «рефлексивного контролю», інформаційної пропаганди, психологічного тиску, заперечення факту прямої причетності до конфлікту, заохочення дезертирства, кібератаки та різноманітні шпигунські технології [7]. На думку F. Hoffman, поширення злочинності також є елементом гібридної війни, бо криміналітет достатньо ефективно може впливати на дестабілізацію політичної ситуації та сприяти незаконним збройним формуванням [8]. А професор соціології Рущенко І.П., говорячи про гібридну війну, зазначає: соціальна дестабілізація одночасно планується й здійснюється у трьох сферах: «хаос у головах», «хаос в інституціях», «хаос на вулицях» [9].

> *Таким чином, достатньо **обґрунтованим є висунення і відповідної гіпотези** щодо необхідності врахування особливостей гібридної війни в системі кримінологічного аналізу злочинності в Україні та подальшого формування державної політики у сфері протидії злочинності із врахуванням гібридності.*

На основі факторного аналізу (*рис. 19.7*) в межах статистичної перевірки нашої гіпотези, виділено групи з найбільшою часткою поясненої відмінності окремих груп гібридних загроз.

Використовуючи додатково експертне оцінювання змісту гібридних загроз й середовища їх поширення, сформовано сутнісне сприйняття гібридних загроз з подальшим їх розподілом у 6 груп, що характеризуються так:

- *гібридні загрози, що пов'язані з інформаційним середовищем;*
- *гібридні загрози, що пов'язані з використанням кібероперацій;*
- *гібридні загрози, що пов'язані з поширенням злочинності;*
- *гібридні загрози, що пов'язані з провокуванням громадянської непокори, порушенням громадського порядку;*
- *гібридні загрози, що пов'язані з діяльністю ОКІ (об'єкти критичної інфраструктури);*
- *гібридні загрози, що пов'язані з корупцією.*

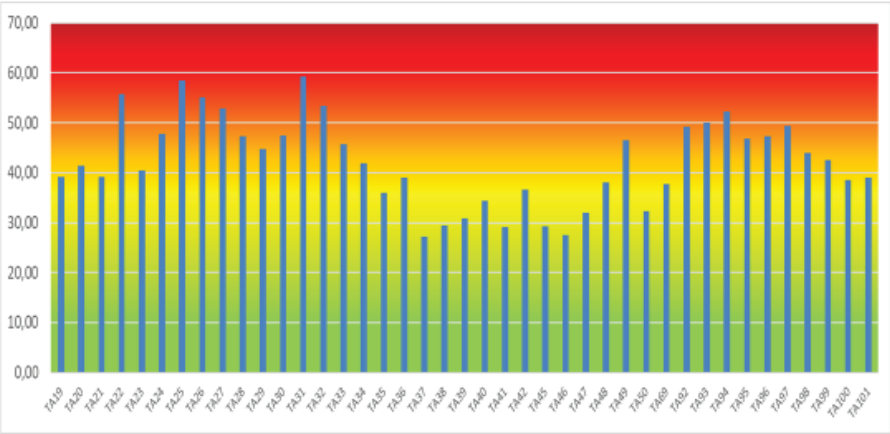
Рисунок 19.7. Результати факторного аналізу та частка поясненої відмінності окремих груп гібридних загроз

Таким чином, висунута гіпотеза щодо наявності групи гібридних загроз, пов'язаних з поширенням злочинності, має своє підтвердження і цей напрям потребує більш ґрунтовного дослідження.

Емпірична база та система індикаторів, за якою було ідентифіковано та оцінено гібридні загрози вказує на 41 вид гібридних загроз, пов'язаних з поширенням злочинності (*рис. 19.8*).

До зазначеної групи гібридних загроз увійшли: *незаконний обіг ЗБВ; діяльність незаконних збройних формувань; провокування сепаратизму; поява нових злочинних схем в економічній сфері; фінансування тероризму; терористичні акти; незаконне втручання у функціонування транспорту; діяльність ОЗУ, що спрямована на поглиблення криміногенного стану; відмивання коштів та інші.*

TA19	TA20	TA21	TA22	TA23	TA24	TA25	TA26	TA27	TA28	TA29	TA30	TA31	TA32	TA33	TA34	TA35	TA36	TA37	TA38	TA39	TA40	TA41	TA42	TA45	TA46	TA47	TA48	TA49	TA50	TA69	TA92	TA93	TA94	TA95	TA96	TA97	TA98	TA99	TA100	TA101
39,25	41,38	39,25	55,74	40,52	47,75	58,53	55,11	52,85	47,27	44,74	47,46	59,20	53,41	45,70	41,96	35,98	38,95	27,17	29,38	30,97	34,34	29,14	36,56	29,30	27,52	32,08	38,00	46,45	32,26	37,68	49,29	50,03	52,25	46,88	47,36	49,36	44,03	42,58	38,51	39,03



Найбільш значні гібридні загрози, повязані зі злочинністю	Оцінка ризику, %
TA21 Цілеспрямована діяльність організованої злочинності в інтересах агресора	39,25
TA22 Діяльність незаконних збройних формувань	55,74
TA24 Терористичні акти	47,75
TA25 Провокація сепаратизму	58,53
TA26 Незаконний обіг зброї, боєприпасів та вибухових речовин	55,11
TA28 Контрабанта озброєння	47,27
TA31 Диверсії на складах, арсеналах боєприпасів	59,20
TA35 Поширення транскордонної організованої злочинності	35,98
TA48 Ініційоване агресором поширення відмивання коштів на користь сепаратистів, інформаційних дезінформаторів тощо	38,00
TA49 Ініційоване агресором поширення фінансування тероризму	46,45

Рисунок 19.8. Гібридні загрози, повязані з поширенням злочинності

До найбільш небезпечних гібридних загроз у сфері цивільної безпеки, що пов'язані з поширенням злочинності, на основі оцінювання ризику їх поширення віднесено:

- *диверсії на складах, арсеналах боєприпасів (59,2 %),*
- *провокація сепаратизму (58,53 %),*
- *діяльність незаконних збройних формувань (55,74 %),*
- *незаконний обіг зброї, боєприпасів та вибухових речовин (55,11 %).*

> *Подальший аналіз зосереджено на одній гібридній загрозі, достатньо високого рівня ризику поширення, -* ***діяльність незаконних збройних формувань.***

Використовуючи результати оцінювання ризику щодо діяльності незаконних збройних формувань за географічною ознакою, на основі адміністративно-територіальний поділу України, побудували теплову карту, що характеризує територіальну ознаку за рівнем ризику діяльності незаконних збройних формувань в регіонах (*рис. 19.9*).

Найбільшим рівнем ризику характеризуються Луганська та Київська області (більше 70%). Водночас також високим (в межах 60% – 70%) ризиком діяльності незаконних збройних формувань характеризуються Донецька, Миколаївська, Сумська, Черкаська, Волинська, Закарпатська області та м. Київ. За основною методикою визначення шкали оцінювання ризиків, зазначений рівень ризику в регіонах вказує на загрози надзвичайно високого рівня щодо діяльності незаконних збройних формувань саме в цих регіонах.

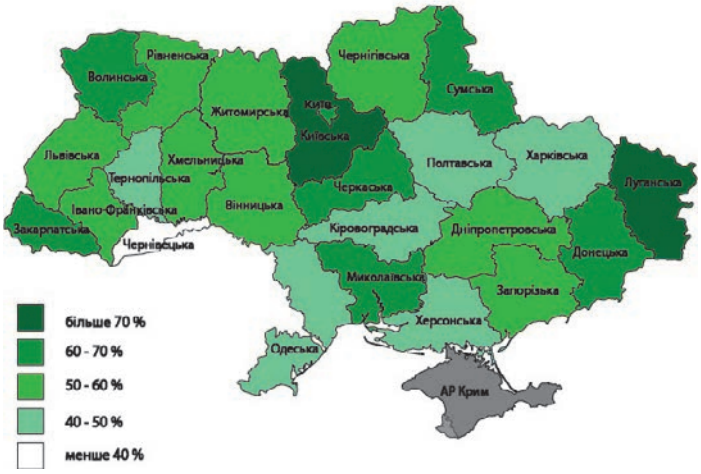


Рисунок 19.9. Теплова карта за рівнем ризику діяльності незаконних збройних формувань в регіонах України

На основі кореляційного аналізу визначено фактори, що мають певний статистичний зв'язок із зазначеною загрозою.

Із загального переліку індикаторів (152), що характеризують ефективність правоохоронної системи у протидії гібридним загрозам, лише 53 характеризуються певним статистичним зв'язком із загрозою «діяльність незаконних збройних формувань». При чому, частина корелятів (43) має негативне значення і потенційно з підвищенням їх оціночного рівня впливатимуть на зниження означеної загрози (*табл. 19.4*). Такі фактори характеризуються позитивним змістом і, за умови рівня їх оцінювання вище 50%, визначають саме спроможності. Тобто це група факторів, що сприймаються як спроможності протидії означеній гібридній загрозі.

Таблиця 19.4. Кореляційний аналіз статистичного зв'язку діяльності незаконних збройних формувань з факторами, що характеризують спроможності

1 Оновлено та удосконалено організаційні структури органів системи МВС.	-,108*	0,017
4 Імплементовано демократичні норми та процедури провідних країн ЄС і НАТО у діяльність ПО	-,110*	0,015
5.1 Досвід МВС щодо протидії гібридним загрозам	-,122**	0,007
32 Рівень фінансового забезпечення розвитку освітньої, наукової та дослідницької діяльності у сфері НБ	-,117**	0,010
33 Рівень спроможностей навчальних та наукових установ щодо стратегічних досліджень у сфері НБ	-,101*	0,026
39 Ефективність системи моніторингу використання в Україні криміналітету в інтересах агресора	-,167**	0,000
41.3.1 Рівень використання аналіт інструментарію на оперативному рівні	-,113*	0,013
41.3.2 Рівень використання аналіт інструментарію на стратегічному рівні	-,102*	0,024

Поряд з цим, частина корелятів (10) має позитивне значення кореляції і характеризується однаковим трендом із загрозою (*табл. 19.5*). За даних умов важливою є констатація того, що фактори, поширення яких підвищують ризик діяльності незаконних збройних формувань, є негативними за своїм змістом явищами та характеризуються як вразливість системи щодо протидії гібридній загрозі «діяльність незаконних збройних формувань».

Таблиця 19.5. Кореляційний аналіз статистичного зв'язку діяльності незаконних збройних формувань з факторами, що характеризують вразливості

1 Недосконале нормативно-правове забезпечення діяльності органів системи МВС	,104*	0,022
2 Дискредитація реформи органів системи МВС окремими працівниками	,099*	0,028
3 Низькі темпи процесів реформування органів системи МВС	,130**	0,004
4 Невисока заробітна плата працівників нижчої ланки	,175**	0,000
5 Незадоволені потреби (передбачені законодавством) працівників органів системи МВС – учасників АТО	,131**	0,004
6 Окремі випадки безкарності працівників органів системи МВС	,162**	0,000
8.1 Недостатній рівень професійної підготовки управлінського (начальницького) рівня	,139**	0,002
8.2 Недостатній рівень професійної підготовки виконавського рівня (персоналу)	,103*	0,024
12 Ослаблений внутрішній контроль	,109*	0,016
23 Рівень корупційних проявів серед працівників (усіх категорій) органів системи МВС	,138**	0,002

Поряд із зазначеними результатами дослідження, важливим залишається питання визначення ключових факторів, що впливають на визначену нами гібридну загрозу та є пріоритетними щодо зменшення ризику її поширення. Зазначене аналітичне завдання базується на тому, що з переліку 152 факторів, що впливають на гібридні загрози та характеризують ефективність діяльності суб'єктів протидії гібридним загрозам, їхній вплив на конкретну загрозу щодо діяльності незаконних збройних формувань є різним. Зазначене потребує пошуку відповідного статистичного зв'язку та змістовного осмислення. З цією метою нами побудована модель лінійної регресії за методом поетапного включення, поетапного виключення та пошуку найкращих підмножин (*рис. 19.10*).

```
REGRESSION
/DESCRIPTIVES MEAN STDDEV CORR SIG N
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT TA22
/METHOD=STEPWISE Rs1 Rs2 Rs3 Rs4 Rs6 Rs7 Rs8 Rs9 Rs10 Rs11 Rs12 Rs13 Rs15 Rs16 Rs17 Rs18 Rs19 Rs20 Rs21 Rs22
Rs23 Rs24 Rs25 Rs26 Rs27 Rs28 Rs29 Rs30 Rs31 Rs32 Rs33 Rs34 Rs35 Rs36 Rs37 Rs38 Rs39 Rs40 Rs41 Rs42 Rs43 Rs44 Rs45
Rs46 Rs47 Rs48 Rs49 Rs50 Rs51 Rs52 Rs53 Rs54 Rs55 Rs56 Rs57 Rs58 Rs59 Rs60 Rs61 Rs62 Rs63 Rs64 Rs65 Rs66 Rs68 Rs69
Rs70 Rs71 Ro1 Ro2 Ro3 Ro4 Ro5 Ro6 Ro7 Ro8 Ro9 Ro10 Ro11 Ro12 Ro13 Ro14 Ro15 Ro16 Ro17 Ro18 Ro19 Ro20 Ro21 Ro22
Ro23 Ro24 Ro25 Ro26 Ro27 Ro28 Ro30 Ro31 Ro32 Ro33 Ro34 Ro35 Ro36 Ro37 Ro38 Ro39 Ro40 Ro41 Ro42 Ro43 Ro44 Ro45
Ro46 Ro47 Ro48 Ro49 Ro50 Ro51 Ro52 Ro53 Ro54 Ro55 Ro56 Ro57.
```

Рисунок 19.10. Синтаксис лінійної регресії із залежною змінною «Діяльність незаконних збройних формувань» (TA22)

Найбільш оптимальна модель лінійної регресії визначає два найбільш значущі кореляти діяльності незаконних збройних формувань (*табл. 19.6*):

- *розгорнуто на прохання Уряду України КМЄС для реформування в рамках мандату щодо розбудови спроможності органів влади,*
- *низька заробітна плата працівників нижчого рівня.*

У свою чергу, спроможність за фактором «*розгорнуто на прохання Уряду України КМЄС для реформування в рамках мандату щодо розбудови спроможності органів влади*» визначена на рівні 53,57 %, а вразливість за фактором «*низька заробітна плата працівників нижчого рівня*» – 73,34 %.

Таблиця 19.6. Модель лінійної регресії на основі залежної змінної «Діяльність незаконних збройних формувань».

Factora					
Model	Unstand factors		Standard factors	т	P value
	B	Std. Err.	Beta		
(Constant)	51,156	2,717		18,829	0,000
розгорнуто на прохання Уряду України КМЄС для реформування в рамках мандату щодо розбудови спроможності органів влади	-0,169	0,038	-0,195	-4,435	0,000
низька заробітна плата працівників нижчого рівня	0,169	0,042	0,175	3,995	0,000

Зазначене вказує на високі статистичні зв'язки залежної змінної та встановлених предикторів. Водночас змістовне осмислення такого статистичного зв'язку вказує на зниження ризику діяльності незаконних збройних формувань за умови ефективної діяльності Консультативної місії ЄС щодо розбудови спроможностей правоохоронних органів, але очікуваним є підвищення рівня досліджуваної загрози за умови зниження рівня заробітної плати працівників нижчого рівня. Зазначене має зрозумілу логіку й статистичний зв'язок закріплює змістовним зв'язком.

Враховуючи оціночні величини та застосувавши модель лінійної регресії, був визначений ризик поширення діяльності незаконних збройних формувань за умови наявного рівня спроможностей системи протидії зазначеній гібридній загрозі (*табл. 19.11*)

Потрібно зазначити, що рівень ризику діяльності незаконних збройних формувань за наявного рівня спроможностей (за встановленими предикторами) (54,49 %) є дещо нижчим ніж його оцінювання лише за інтегрованим показником двох оціночних значень – ймовірність та наслідки – (55,74 %) в статичному стані.

Таблиця 19.11. Оцінка ризику діяльності незаконних збройних формувань із врахуванням реального рівня предикторів

Модельта прогноз	**Coeff**	**P value**	**Rating %**	
(Constant)	51,156	0,000		
розгорнуто на прохання Уряду України КМЄС для реформування в рамках мандату щодо розбудови спроможності органів влади	-0,169	0,000	53,57	-9,05
низька заробітна плата працівників нижчого рівня	0,169	0,000	73,34	12,39
Поточний рівень ризику (із врахуванням спроможностей) **RA = 54,49 %**				

Поряд з цим, побудована модель лінійної регресії дає можливість розрахувати прогноз щодо зниження рівня ризику діяльності незаконних збройних формувань як гібридної загрози у секторі цивільної безпеки за умови підвищення спроможності (*розгорнуто на прохання Уряду України КМЄС для реформування в рамках мандату щодо розбудови спроможності органів влади*) – та зниження вразливості (*низька заробітна плата працівників нижчого рівня*) на 20% (*табл.* 19.12):

Coefficient	P value	Rating	
51,156	0,000		
0,169	0,000	73,57	-12,43
-0,169	0,000	53,34	9,01
RA = 47,74 %			

Таблиця 19.12. Прогноз рівня ризику діяльності незаконних збройних формувань із врахуванням підвищення спроможностей на 20%

Таким чином, на основі факторного аналізу визначено ключові групи гібридних загроз у секторі цивільної безпеки, серед яких обґрунтованою є необхідність розглядати групу гібридних загроз, пов'язаних з поширенням злочинності, з яких 41 ідентифікована, оцінений ризик їх поширення та на основі ризику встановлено найзначніші загрози.

На основі кореляційного аналізу визначено фактори, що мають певний статистичний зв'язок із зазначеною загрозою. Із загального переліку індикаторів (152), що характеризують ефективність правоохоронної системи у протидії гібридним загрозам, лише 53 характеризуються певним статистичним зв'язком із загрозою «діяльність незаконних збройних формувань».

На основі моделі лінійної регресії оцінено ризик поширення незаконних збройних формувань за умови існуючого рівня спроможності та вразливості, а також визначено пріоритетні фактори зниження цього ризику: «розгорнуто на прохання Уряду України КМЄС для реформування в рамках мандату щодо розбудови спроможності органів влади» та «низька заробітна плата працівників нижчого рівня», а також побудовано відповідну прогнозну модель.

> *За умови підвищення спроможності на 20% лише за встановленими двома індикаторами, ризик поширення діяльності незаконних збройних формувань зменшиться майже на 7%: з рівня 54,49% до 47,74%.*

ВИКОРИСТАНІ ДЖЕРЕЛА

1. International Association of Crime Analysts (*2017). Exploring crime analysis: Readings on essential skills* (3rd ed.). K. Gallagher, J. Wartell, S. Gwinn, G. Jones, & G. Stewart (Eds.). Overland Park.
2. Ковальчук Т.І., Користн О.Є., Свиридюк Н.П. Hybrid threats in the civil security sector in Ukraine. *Наука і правоохоронна*. 2019. № 3 (45). С. 69-79. DOI: doi.org/10.36486/np.2019308
3. Hua-Gang Yu, Gao-Ming Huang & Jun Gao (2010). Nonlinear Blind Source Separation Using Kernel Multi-set Canonical Correlation Analysis. *International Journal of Computer Network and Information Security*, 2 (1), 1-8. DOI: 10.5815/ijcnis.2010.01.01
4. Sonali, Sharma & Shilpa, Mahajan (2017). Design and Implementation of a Security Scheme for Detecting System Vulnerabilities. *International Journal of Computer Network and Information Security*, 9 (10), 24-32. DOI: 10.5815/ijcnis.2017.10.03
5. Mohamed Zaim, Shahrel, Sofianita, Mutalib & Shuzlina, Abdul-Rahman (2021). PriceCop–Price Monitor and Prediction Using Linear Regression and LSVM-ABC Methods for E-commerce Platform. *International Journal of Information Engineering and Electronic Business*, 13 (1), 1-14. DOI: 10.5815/ijieeb.2021.01.01
6. Oleksandr, Korystin & Nataliia, Svyrydiuk (22 March 2021). Activities of Illegal Weapons Criminal Component of Hybrid Threats. *Proceedings of the International Conference on Economics, Law and Education Research* (ELER 2021), 170, 86-91. DOI:10.2991/aebmr.k.210320.016
7. Рущенко І.П., Рущенко Ю.І. Гібридна агресія та громадянський спротив у Харкові 2014 р.: уроки першої фази російсько-української війни. *Український соціум*. 2016. № 3(58). С. 88-99.
8. Hoffman, F. (2009). Hybrid Warfare and Challenges. *JFQ*. 52, 34-39.
9. Рущенко І.П. Тактика керованого хаосу як інструмент гібридної війни. *Вісник ЛНУ імені Тараса Шевченка.* Червень 2017. № 5 (310). С. 5-29.