ЛЕКЦІЯ.
КЛАСИФІКАЦІЯ НАБОРІВ ДАНИХ

Дані можуть бути представлені в різній формі. Корисно мати базову класифікацію різних типів даних, щоб відразу ж визначати тип нових даних і використовувати відповідний метод аналізу. Набір даних складається з результатів спостережень об’єктів, які зазвичай включають реєстрацію однієї і тієї ж інформації для кожного об'єкта. Під **елементарними одиницями** розуміються самі об’єкти (наприклад, компанії, люди, домогосподарства, міста, телевізори), щоб відрізняти їх від **результатів вимірювань або спостережень** (наприклад, обсяги продажів, вага, дохід, населення, розмір). Можна визначити чотири основних способи класифікації наборів даних.

Перший. За кількістю порцій інформації (змінних) для кожної елементарної одиниці.

Другий. За типом вимірювання (числа або категорії) для кожного спостереження.

Третій. По тому, важлива чи ні впорядкованість в часі записів про результати вимірювань.

Четвертий. По тому, чи збиралася інформація спеціально для цього аналізу або дані збиралися раніше кимось іншим для своїх потреб.

2.1. Скільки змінних?

Порція інформації, що реєструється для кожного об'єкта (наприклад, вартість), називається **змінною**. Кількість змінних, або порцій інформації, що реєструються для кожного об'єкта, вказує на складність набору даних і визначає відповідний тип аналізу. Залежно від того, чи маємо ми справу з однією, двома або багатьма змінними, ми отримуємо відповідно одновимірний, двовимірний або багатовимірний набір даних.

Одномірні дані

Одномірні набори даних (одна змінна) містять лише одну ознаку, що реєструється для кожної елементарної одиниці. В цьому випадку статистичні методи використовують для узагальнення основних властивостей цієї єдиної ознаки, відповідаючи на такі питання.

1. Чому дорівнює типове (узагальнене) значення?
2. Наскільки різняться ці об'єкти?
3. Чи є в цьому наборі даних окремі елементи або групи елементів, які потребують особливої уваги?

Зазначена нижче таблиця одновимірних даних містить обсяги доходу 10 компаній в сфері ресторанного та готельного бізнесу (зі списку Fortune 500) у 2019 році:

Назва	Дохід (млн. $)
Starbucks	24 719,5
McDonald's	21 025,2
Marriott International	20 758,0
Las Vegas Sands	13 729,0
MGM Resorts International	11 763,1
Hilton Worldwide Holdings	8 906,0
Yum China Holdings	8 415,0
Caesars Entertainment	8 391,0
Darden Restaurants	8 080,1
Wynn Resorts	6 717,7

Джерело даних: https://fortune.com/fortune500/2019

Наведемо ще кілька прикладів одновимірних наборів даних.

1. Доходи окремих людей, виявлені в ході маркетингового дослідження. Статистичний аналіз виявив би структуру (або розподіл) доходів, виявивши типовий рівень доходу, ступінь варіації доходів і відсоток людей, дохід яких знаходиться в будь-якому заданому діапазоні.

2. Кількість дефектів у кожному телевізорі з вибірки об’ємом 50, взятої з телевізорів, виготовлених сьогодні вранці. Статистичний аналіз можна використовувати для обліку якості (оцінювання) і спостереження за тим, щоб виробничий процес не вийшов з під контролю (перевірка гіпотез).

3. Кольори, обрані членами фокус-групи. Аналіз допоможе зробити відповідний вибір для нового виду продукції.

4. Оцінки платоспроможності фірм в інвестиційному портфелі. Аналіз показав би ризик інвестиційного портфеля.

Двовимірні дані

Набори двовимірних (дві змінні) даних містять інформацію про дві ознаки для кожного з об'єктів. На додаток до узагальнення властивостей кожної з цих двох змінних, що розглядаються як окремі набори одновимірних даних, статистичні методи можна використовувати для вивчення зв’язку між цими двома факторами, з'ясовуючи при цьому наступне.

1. Чи існує між цими двома змінними простий зв’язок?
2. Наскільки сильно взаємопов’язані змінні?
3. Чи можна передбачити значення однієї змінної на підставі іншого? Якщо так, то з яким ступенем надійності?
4. Чи існують окремі об’єкти або групи, які потребують особливої уваги?

Наведена нижче таблиця містить двовимірні дані про розміри прибутку 10 компаній в сфері ресторанного та готельного бізнесу (зі списку Fortune 500) у 2019 році і зміни прибутку (у відсотках) в порівнянні з попереднім роком:

Name	Прибуток (млн. %)	Зміна прибутку, у % до 2018 р.
Starbucks	4 518,3	56,6%
McDonald's	5 924,3	14,1%
Marriott International	1 907,0	39,0%
Las Vegas Sands	2 413,0	-14,0%
MGM Resorts International	466,8	-76,2%
Hilton Worldwide Holdings	764,0	-39,3%
Yum China Holdings	708,0	75,7%
Caesars Entertainment	303,0	-
Darden Restaurants	596,0	24,4%
Wynn Resorts	572,4	-23,4%

Джерело даних: https://fortune.com/fortune500/2019

Розглянемо ще кілька прикладів двовимірних наборів даних.

1. Дані за минулий квартал про витрати на виробництво продукції (перша змінна) і кількості вироблених виробів (друга змінна) для кожної з семи фабрик (об'єкти або елементарні одиниці), що випускають телефонні навушники. Двовимірний статистичний аналіз показав би взаємозв'язок між витратами і кількістю вироблених телефонних навушників. Зокрема, аналіз визначив би постійні витрати, пов’язані з використанням виробничого обладнання, і змінні витрати, що характеризують виробництво однієї пари інтегральних навушників. Аналітик, подивившись на дані семи фабрик, міг би порівняти їх ефективність.

2. Ціна однієї звичайної акції вашої фірми (перша змінна) і дата (друга змінна), зареєстровані кожен день протягом останніх шести місяців. Зв'язок між ціною і часом дозволяє побачити тенденції в зміні вартості інвестицій. Однак важко сказати, чи можна на підставі таких даних передбачити майбутню вартість інвестицій (це, зокрема, залежить від того, чи є зміна вартості непередбачуваним «випадковим блуканням», або існує деяка реальна закономірність).

3. Дані опитування 100 чоловік в торговому центрі: купувати чи не купувати деякий товар (перша змінна, записується відповідь «так / ні», або 1/0), і можливість згадати рекламу цього товару (друга змінна, записана аналогічним чином). Такі дані (і, звичайно ж, дані більш докладних досліджень) допомагають прояснити ефективність реклами, тобто вивчити зв’язок між рекламою і покупкою.

Багатовимірні дані

Набори багатовимірних (багато змінних) даних містять інформацію про три або більше ознаках для кожного об’єкта. На додаток до узагальнення властивостей кожної з цих змінних (що розглядаються як окремі набори одновимірних даних) і встановленню залежності між парами змінних (як при аналізі набору двовимірних даних) статистичні методи можна використовувати для вивчення взаємозв'язків між усіма цими змінними, з’ясовуючи при цьому такі питання.

1. Чи існує проста залежність між цими ознаками?

2. Наскільки сильно вони взаємопов’язані?

3. Чи можна передбачити значення однієї («виділеної») змінної виходячи з значень інших? З яким ступенем надійності?

4. Чи існують окремі об’єкти або групи, які потребують особливої уваги?

У наведеній нижче таблиці містяться багатовимірні дані про розміри прибутку 10 компаній в сфері ресторанного та готельного бізнесу (зі списку Fortune 500) разом з відсотком зміни прибутку по відношенню до попереднього року, кількістю працівників і розмірами доходу:

Name	Дохід (млн. $)	Прибуток (млн. %)	Зміна прибутку, у % до 2018 р.	Кількість працівників
Starbucks	24 719,5	4 518,3	56,6%	291000
McDonald's	21 025,2	5 924,3	14,1%	210000
Marriott International	20 758,0	1 907,0	39,0%	176000
Las Vegas Sands	13 729,0	2 413,0	-14,0%	51500
MGM Resorts International	11 763,1	466,8	-76,2%	74500
Hilton Worldwide Holdings	8 906,0	764,0	-39,3%	169000
Yum China Holdings	8 415,0	708,0	75,7%	450000
Caesars Entertainment	8 391,0	303,0	-	66000
Darden Restaurants	8 080,1	596,0	24,4%	180656
Wynn Resorts	6 717,7	572,4	-23,4%	26000

Джерело даних: https://fortune.com/fortune500/2019

Розглянемо ще кілька прикладів наборів багатовимірних даних.

1. Темп зростання (виділена змінна) і набір характеристик стратегії (інші змінні), таких як тип обладнання, обсяг інвестицій, стиль керівництва для кожної з декількох нових підприємницьких фірм. Аналіз міг би показати, яке поєднання призводить до успіху, а яке – ні.

2. Заробітна плата (виділена змінна) а також стать (реєструється як «чоловіча / жіноча», або 1/0), стаж роботи, категорія роботи і продуктивність для кожного працівника. Такі дані можуть розглядатися в судовому процесі про дискримінацію (з точки зору більш низької середньої оплати праці) жінок. Ключове питання, на яке може відповісти багатовимірний аналіз, полягає в наступному. Чи можна пояснити розбіжність у розмірі заробітної плати

іншими чинниками, окрім статі працівника? Статистичні методи можуть виключити вплив цих інших факторів і таким чином виміряти середнє розходження заробітної плати між чоловіками і жінками, які однакові в інших відносинах.

3. Для кожного з будинків в районі ціна цього будинку (виділена змінна) і ряд змінних, від яких залежить вартість нерухомості, а саме кількість будинків такого типу, площа будинку, кількість кімнат, наявність або відсутність басейну, вік будинку. Аналіз показав би, як оцінюється нерухомість в цьому районі. Такий результат можна було б використовувати для визначення реальної ринкової вартості будинку в цьому районі або при будівництві, щоб визначити, яка комбінація характеристик нового будинку підвищує його ціну.

2.2. Кількісні дані: числа

Числа, які мають змістовну інтерпретацію, – це числа, які безпосередньо представляють виміряний або спостережуваний обсяг певної ознаки або кількість елементарних одиниць. До чисел, які мають змістовну інтерпретацію, можна віднести, наприклад, кількість гривень, частоти, розміри, кількість службовців або число кілометрів на літр. До них не належать ті числа, які використовують для кодування або нумерації чого-небудь, як, наприклад, номер на футбольній спортивній формі або кодування угод виду 1 = покупка акції, 2 = продаж акції, 3 = покупка зобов'язань, 4 = продаж зобов'язань. Якщо дані представляють собою числа, що мають змістовну інтерпретацію, то ми маємо справу з кількісними даними (тобто вони представляють кількість чого-небудь). З кількісними даними можна виконувати всі звичайні операції над числами, такі як обчислення середнього та оцінку мінливості. З такими даними можна проводити безпосередні обчислення. Залежно від того, які значення може потенційно приймати змінна, виділяють два типи кількісних даних: дискретні і безперервні.

Дискретні кількісні дані

Дискретна змінна – це така змінна, яка може набувати значень тільки з деякого списку певних чисел. Наприклад, число дітей в сім'ї є дискретною змінною. Оскільки можливі значення змінної можна перерахувати, то з наборами дискретних даних працювати відносно легко. Розглянемо кілька прикладів дискретних змінних.

1. Скільки разів за останні 24 години на підприємстві вимикали комп'ютер.
2. Кількість дійсно укладених контрактів з 18 підготовлених вами пропозицій.
3. Число іноземних танкерів, які пришвартувалися сьогодні в певному порту.
4. Стать службовця, записана за допомогою числа 0 або 1.

Безперервні кількісні дані

Безперервною будемо вважати будь-яку числову змінну, яка не є дискретною. Термін «безперервна» використовують, оскільки можливі значення змінної утворюють «континуум», як, наприклад, безліч всіх позитивних чисел, безліч всіх чисел або всі значення між 0 і 100%. Наприклад, фактична вага льодяника на паличці, записана як «нетто вага 1,7 грамів», є безперервною випадковою змінною, оскільки фактична вага може дорівнювати 1,70235 або 1,69481 грамів, а не точно 1,7 грамів. Якщо ви все ще не володієте статистичними мисленням, можете вважати, що фактична вага точно дорівнює 1,7 грамів; в дійсності в будь-яких реальних вимірах завжди є невеликі (а іноді великі) відхилення від очікуваних значень.

Розглянемо кілька прикладів безперервних змінних.

1. Ціна унції золота в доларах зараз. Здавалося б, що можна розглядати цю величину як дискретну (і технічно це вірно, оскільки число виду 390,79 долара належить списку дискретних значень, записаних з точністю до центів: 0,00; 0,01; 0,02; ...). Однак краще розглядати такі числа як безперервні дані, оскільки розмір дискретної зміни малий і не важливий для аналізу. Якби

золото продавалося за ціною кілька центів за унцію, то його вартість, можливо, потрібно було б розглядати як дискретні дані. Однак більш імовірно, що в цьому випадку ціна була б вказана з урахуванням тисячних часток цента, що, по суті, теж виступало б як безперервна кількість.

2. Інвестиційні показники і показники бухгалтерського обліку: дохід на одну акцію, ставка доходу на інвестиції, показник перекриття.

3. Кількість енергії, необхідної для роботи одного комп'ютера.

Обережно з числами, які не мають змістовної інтерпретації!

Перш ніж приступити до аналізу кількісних даних, слід зробити одне важливе попередження. Слід переконатися, що числа, які аналізуються, дійсно мають змістовну інтерпретацію. На жаль, числа можна використовувати для запису чого завгодно. Якщо йде справа з довільним кодуванням, то результат аналізу таких даних не матиме сенсу.

2.3. Якісні дані: категорії

Якщо набір даних показує, якій з кількох нечислових категорій належить кожен з об'єктів, то дані є якісними (оскільки вони реєструють певну якість, якою володіє об'єкт). Треба бути уважними і обережними, щоб уникнути спокуси приписати числові значення категоріям (класам) і далі проводити з ними обчислення. За наявності кількох класів, можна оперувати відсотками (частотами) подій в кожному класі (створивши таким чином щось числове з представлених категоріями даних). Якщо є в точності дві категорії, їх можна позначити цифрами 1 і 0, приписати ці значення відповідно до кожного з об'єктів і потім (в досить багатьох випадках) обробляти отримані дані як кількісні. Спочатку розглянемо загальний випадок, коли мова йде про три або більше категорій.

Існують два типи якісних даних: порядкові (ординальні, для яких порядок має змістовний сенс, але немає змістовного числового позначення) і номінальні (для яких немає змістовного інтерпретованого порядку).

Порядкові якісні дані

Набір даних є ординальним, якщо існує порядок, що має змістовний сенс: можна вести мову про перший (наприклад, «кращий»), другий, третій і т.д. Можна ранжувати дані відповідно до цього порядку і використовувати це ранжування при виконанні аналізу, особливо якщо воно має відношення до досліджуваного питання.

Розглянемо деякі приклади порядкових даних.

1. Посада, записана для кожного з групи керівників: президент, віце-президент, начальник відділу, заступник начальника відділу. Хоча класифікатор не містить чисел і не зовсім ясно, яким чином їх можна тут використовувати, об’єкти можна природним чином упорядкувати.

2. Характеристики, такі як AA+, AA, AA–, A+, A, A–, B+, B і B–, зафіксовані для набору боргових зобов'язань. Це чисто порядкові категоріальні дані, оскільки впорядкованість має сенс з точки зору ризику вкладів і використовується в аналізі інвестицій.

3. Відповіді на запитання анкети: «Будь ласка, висловіть свою думку щодо вашої роботи в компанії, використовуючи шкалу від 1 до 5, де 1 означає «насилу чекаю закінчення робочого дня», а 5 – «всі мої думки зайняті роботою». Незважаючи на те, що відповіді виражені числами, ми маємо справу з порядковими даними, оскільки запропонована шкала оцінок носить суб’єктивний характер. Незрозуміло, чи можна вважати, що різниця між оцінками 5 і 4 така ж, як і між оцінками 2 і 1. Крім того, можна вважати, що оцінка 2 в два рази краще оцінки 1. Однак впорядкування і ранжування тут явно мають місце.

Номінальні якісні дані

Номінальні якісні дані визначаються в термінах категорій, які не можна змістовно впорядкувати. Для таких категорій немає чисел, з якими можна робити обчислення, і немає основи для ранжирування. Все, що можна зробити, – це підрахувати відсоток (або кількість) спостережень, що

потрапили в кожну з категорій, і використовувати в якості узагальнюючого показника моди (категорії, що найбільш часто зустрічається).

Розглянемо кілька прикладів номінальних даних.

1. Райони області, в яких проживають студенти навчальної групи. Раніше відзначалося, що це лише категорії. Щоб ранжувати райони, необхідно обов'язково ввести будь-яку іншу змінну (наприклад, чисельність населення району або розмір доходу на душу населення), яку краще використовувати безпосередньо.

2. Головний продукт кожного з кількох виробничих підприємств диверсифікованого бізнесу, як, наприклад, пластмаса, електроніка, деревина. Ці категорії дійсно не впорядковані. Щоб їх упорядкувати, необхідно розглянути додатковий фактор (наприклад, потенціал зростання даної фірми в галузі), який не є внутрішньою властивістю цих категорій.

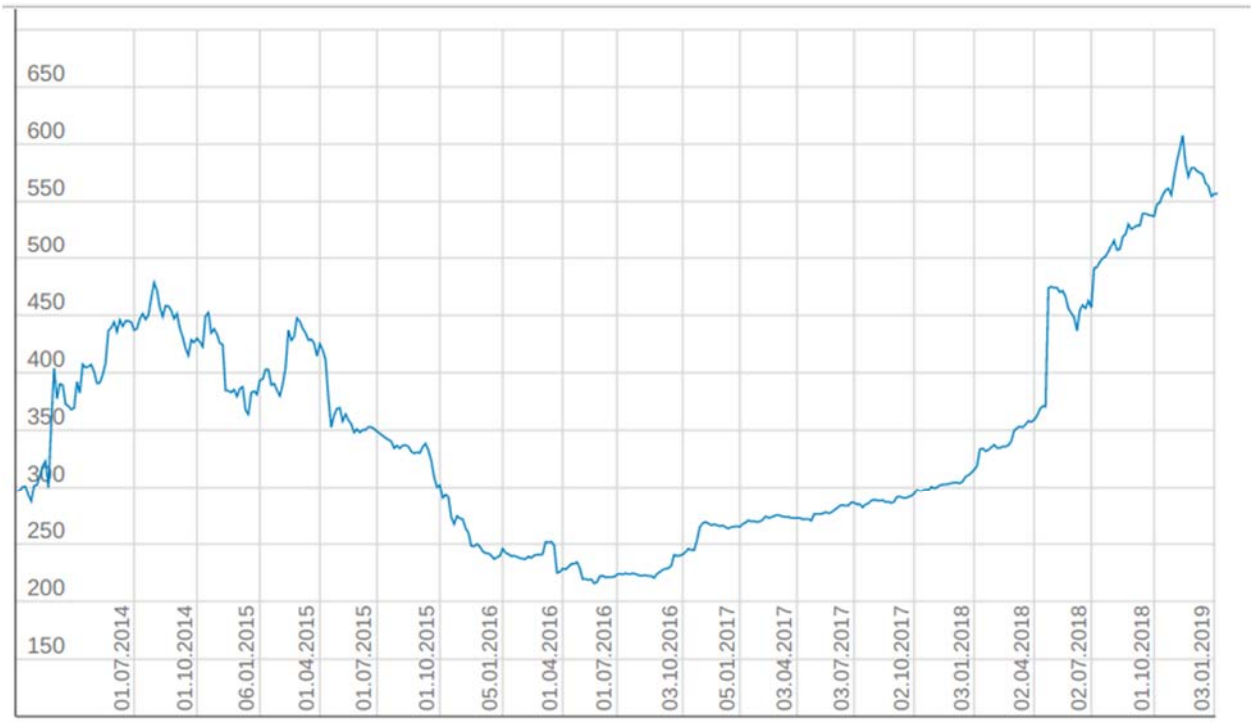
3. Назви всіх фірм, зазначених на першій сторінці сьогоднішнього випуску Україна Сьогодні.

2.4. Часові ряди і дані про один часовий зріз

Якщо порядок запису значень даних має змістовний сенс, наприклад, щоденні ціни на фондовому ринку, то ми маємо справу з часовим рядом. Якщо послідовність, в якій записані дані, не важлива, скажімо, доходи восьми аптечних фірм в першому кварталі 2020 року, то ми маємо дані про один часовий зріз. Слова про один часовий зріз в даному випадку означають лише те, що немає ніякого впорядкування в часі, а є лише інформація про деякі об'єкти в певний момент часу (свого роду «моментальний знімок»).

Аналіз часових рядів в цілому складніший за аналіз даних про один часовий зріз, оскільки вимагає ретельного врахування порядку спостережень. Тому в наступних лекція ми почнемо з даних про один часовому зрізі.

Нижче наведено графік індексу фондового ринку України за останні 5 років, що побудовано на основі часових рядів. Часовий ряд показує зміну стану фондового ринку України з плином часу.



Джерело: https://static.ukrinform.com/photos/2019_01/1546965846-134.png

Розглянемо ще кілька прикладів часових рядів.

1. Ціна на пшеницю за останні 50 років з урахуванням інфляції. Вважаючи, що з трансформаційних змін в майбутньому будуть носити той же характер, що і зміни в минулому, можна використовувати ці тимчасові тренди для довгострокового планування.

2. Обсяги місячних продажів за останні 20 років. Цей набір даних має структуру, що показує зростання продажів з плином часу, а також чітку сезонну особливість з піками в грудневі свята.

3. Результати щохвилинних вимірювань товщини паперу на виході з папероробної машини. Такі дані важливі для контролю якості. Часова послідовність важлива, оскільки невеликі зміни товщини паперу можуть або послідовно «дрейфувати» в сторону неприпустимого рівня, або «коливатися», стаючи ширше або вже в чітко допустимих межах.

Тепер розглянемо кілька прикладів наборів даних про одному часовому зрізу.

1. Виміряна для 30 осіб тривалість сну в останню ніч, яка використовується для оцінки ефективності нових ліків, що продається без рецепта.

2. Сьогоднішня балансова вартість випадкової вибірки банківських ощадних сертифікатів.

3. Кількість телефонних дзвінків, оброблених вчора кожним з працюючих співробітників колл-центру.

2.5. Джерела даних, включаючи Internet

Звідки беруть дані? Існує багато джерел, вибір яких здійснюють виходячи з їх вартості, доступності та потреб економічної діяльності. Якщо план збору даних розробляється самостійно (навіть якщо дані збирають інші), то результатом будуть первинні дані. Якщо ж використовуються дані, раніше зібрані іншими людьми і для інших цілей, то використовуються вторинні дані.

Головна перевага первинних даних полягає в тому, що в цьому випадку більше можливостей зібрати інформацію, яка відповідає меті дослідження, оскільки можна керувати процесом отримання даних шляхом планування питань або вимірювань, а також шляхом визначення вибірки елементарних одиниць для вимірювання. На жаль, часто отримання первинних даних занадто дороге і займає багато часу. З іншого боку, вторинні дані дешевше (або взагалі безкоштовні), і можна знайти саме те (або майже те), що потрібно. Це передбачає таку стратегію отримання даних: пошук вторинних даних, які швидко задовольняють потреби дослідника за прийнятну ціну. Якщо це неможливо, необхідно оцінити вартість збору первинних даних і вирішити, яке джерело (первинне або вторинне) використовувати, виходячи зі співвідношення витрат і переваг кожного з підходів.

Розглянемо кілька прикладів джерел первинних даних.

1. Інформація про продуктивність обладнання заводу, включаючи обсяг і якість (наприклад, рівень браку) продукції, що випускається щодня. Такі дані може автоматично збирати інформаційна система компанії.

2. Дані опитування, проведеного службовцями маркетингової фірми, найнятими з метою вивчення впливу можливої рекламної кампанії на поведінку споживачів.

3. Зібрані в ході політичної кампанії дані про проблеми, якими стурбовані виборці, які збираються голосувати на майбутніх виборах.

А тепер розглянемо приклади джерел вторинних даних.

1. Зібрані і зведені в таблицю урядом України економічні та демографічні дані, які доступні безкоштовно в бібліотеці або через Internet.

2. Дані із спеціалізованих журналів (наприклад, реклама, обсяги виробництва, фінанси і т.п.), які допомагають фірмам, що працюють в цьому секторі ринку, оцінити ситуацію на ринку і успіх окремих продуктів.

3. Дані, зібрані компаніями, що спеціалізуються на зборі даних і продають їх іншим компаніям. Наприклад, Nielsen Media Research продає телевізійні рейтинги (виходячи зі спостережень за тим, які телешоу дивилася вибірка людей) телевізійним кабельним мережам, незалежним станціям, рекламодавцям, рекламним агентствам і ін. Більшість публікацій, які можна знайти в бібліотеці, посилаються саме на спеціалізовані дані, отримані такими компаніями.

2.6. Додатковий матеріал

Резюме

Набір даних містить одне або кілька значень для кожного з окремих об’єктів, які називаються елементарними одиницями. В якості таких об'єктів можуть виступати люди, домогосподарства, міста, телевізійні приймачі або що завгодно, що представляє інтерес для вивчення. Для кожного з об’єктів реєструються одна й та сама ознака (або ознаки). Ознака, що реєструються для кожного з об’єктів (наприклад, вартість), називається змінною.

Існують три основні способи класифікації наборів даних:

– за кількістю змінних (одновимірний, двовимірний і багатовимірний);

– за типом представлених кожною зі змінних інформацією (числа або категорії);

– за часовим характером: чи є набір даних часовим рядом або це дані про один часовий зріз.

Одномірні набори даних (одна змінна) містять інформацію тільки про одну ознаку, зареєстровану для кожного об'єкта. Одновимірний набір даних дозволяє визначити типове значення і характеристику мінливості даних, а також виділити специфічні особливості або проблеми в даних.

Двовимірні набори даних (дві змінні) містять дві ознаки, значення яких реєструються для кожного об’єкта. Двовимірні дані на додаток до інформації про кожну змінну як набор одновимірних даних дозволяють дослідити зв’язок між двома змінними та передбачити значення однієї змінної на основі значення іншої.

Багатовимірні набори даних (багато змінних) містять три або більше ознак, значення яких реєструються для кожного об'єкта. Багатовимірні дані на додаток до інформації про кожну змінну як набор одновимірних даних дають можливість вивчити зв'язок між змінними і передбачити значення однієї змінної на основі значення інших.

Значення змінних, які реєструються як числа, що мають змістовний сенс, називають кількісними даними. Дискретна кількісна змінна може приймати значення тільки з деякого списку конкретних чисел (наприклад, 0 або 1, або перелік чисел 0, 1, 2, 3, ...). Будь-яку кількісну змінну, яка не є дискретною, будемо називати безперервною. Значення безперервної змінної не обмежені простим переліком можливих значень.

Якщо змінна містить інформацію про те, до якої з кількох нечислових категорій належить об’єкт, то вона називається якісною змінною. Якщо категорії можна природним чином і змістовно впорядкувати, то мова йде про порядкову якісну змінну. Якщо такий порядок відсутній, то мова йде про номінальну якісну змінну. Незважаючи на те, що часто значення якісної змінної можна записати за допомогою чисел, така змінна все одно залишається

якісною, а не кількісною, оскільки ці числа не мають будь-якої інтерпретації, що змістовно властива цій змінній.

До кількісних даних можна застосовувати ті ж операції, що і до звичайних чисел: підрахунок частоти, ранжування, арифметичні дії. З порядковими даними можна виконувати тільки підрахунок частоти і ранжування, з номінальними даними – тільки підрахунок частоти.

Якщо послідовність запису даних має змістовний сенс, то відповідний набір даних представляє собою часовий ряд. Якщо послідовність запису даних не важлива, то відповідний набір містить дані про один часовий зріз. Аналіз часових рядів складніше аналізу даних про один часовий зріз.

Якщо збір даних планується роботи самостійно (навіть якщо власне збір даних роблять інші), то отримуються первинні дані. Якщо використовуються дані, що попередньо зібрані іншими людьми і для інших цілей, то це вторинні дані. Отримання первинних даних часто обходиться дорого і займає багато часу, але лише дослідник отримуєте те, що необхідно. Вторинні дані можна отримати дешевше (або навіть безкоштовно), але можливість отримання необхідних даних під питанням.

Основні терміни

- набір даних (data set)
- елементарні одиниці (elementary units)
- змінна (variable)
- одновимірна (univariate)
- двовимірна (bivariate)
- багатомірна (multivariate)
- кількісна (quantitative)
- дискретна (discrete)
- безперервна (continuous)
- якісна (qualitative)
- порядкова або ординальна (ordinal)
- номінальна (nominal)

– часові ряди (time series)

– один часовий зріз (cross-sectional)

– первинні дані (primary data)

– вторинні дані (secondary data)

Контрольні питання

1. Що таке набір даних?
2. Що таке змінна?
3. Що таке елементарна одиниця?
4. Якими трьома основними способами можна класифікувати набори даних? (розгорнута відповідь)
5. На які основні питання можна відповісти, проаналізувавши:

а) одномірні дані;

б) двовимірні дані;

в) багатовимірні дані.

6. Чому двовимірні дані представляють собою більше, ніж просто два окремих одновимірних набори даних?
7. Що можна робити з багатовимірним набором даних?
8. У чому різниця між якісними та кількісними даними?
9. У чому різниця між дискретними і безперервними кількісними змінними?
10. Що таке якісні дані?
11. У чому різниця між порядковими і номінальними якісними даними?
12. Чим відрізняються часові ряди від даних про один часовий зріз?
13. Що легше аналізувати: часові ряди або дані про один часовий зріз?
14. Визначте різницю між первинними і вторинними даними.