ЛЕКЦІЯ.
СУТНІСТЬ ДАНИХ

Мета цієї лекції полягає в тому, щоб познайомити студентів зі способами обробки і представлення даних у документах, особливо тих, що мають безпосереднє відношення до виконання основних функціональним обов'язків соціального працівника на виробництві – проведенні соціологічних досліджень. У реальних соціологічних дослідженнях зазвичай вибірка буває великою і може включати кілька тисяч респондентів. Правильно розпорядитися з таким обсягом статистичного матеріалу можливо лише при правильній організації роботи і хорошому знанні способів подання результатів спостережень. Ми вивчимо поняття розподілу частот, навчимося обчислювати відносні частоти, використовувати таблиці і графіки для представлення даних.

1.1. Розподіл частот

Соціальні працівники працюють з емпіричними даними. Якщо значення ознаки повторюється кілька разів, можна говорити про частоту появи певного значення ознаки.

Частота (frequency) – кількість спостережень, в яких ознака приймає певне значення або знаходиться в певному інтервалі.

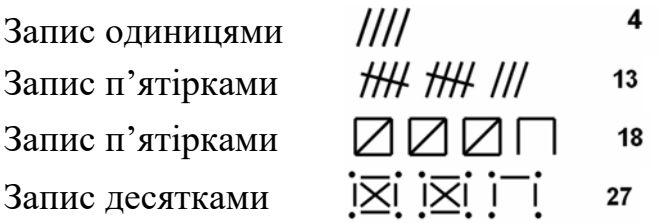
Розподіл частот (frequency distribution) показує частоти у взаємозв'язку з результатами спостережень.

Ми будемо розрізняти два випадки: перший, коли ознака вимірюється номінальною або порядковою шкалою, приймає дискретні значення (будемо називати їх категорії), і другий, коли значення ознаки вимірюються за допомогою числової шкали і безперервні. У першому випадку будемо називати розподіл частот категоріальним, у другому будемо будувати інтервальний розподіл. Як пізніше буде пояснено, інтервальний розподіл можна будувати також і для дискретної ознаки, якщо її значення є рідкісними і можуть бути об'єднані в кілька інтервалів.

Підрахунок частот може виявитися зовсім не простим заняттям. Уявіть, що вам належить опитати понад сто респондентів, а потім дані опитування представити у вигляді таблиці з розподілом частот. Спочатку ваші записи будуть виглядати таким чином:

Номер респондента	Політичні уподобання
1	демократи
2	ліберали
3	ліберали
4	комуністи
5	демократи
6	демократи
і так далі	

Припустимо, 105 респондентів потраплять в ваш довгий список, а ваші записи займуть кілька сторінок. Тепер вас чекає обробка. З давніх-давен відомо кілька графічних способів зручного ручного підрахунку частот: п'ятірками, десятками і т.п. Виглядає це наступним чином:



При виборі одного з наведених вище способів записи після отримання відповіді від чергового респондента потрібно проставляти відповідні мітки, а потім порахувати їх загальне число. Мистецтво ручного підрахунку і обробки даних було поширене зовсім недавно – до появи обчислювальної техніки і її масового використання. Вміти користуватися ручним записом виявляється корисним і тепер, особливо «в польових умовах».

Категоріальний розподіл частот

Що ми хочемо отримати в результаті обробки даних? Якщо у випадковій вибірці з 105 респондентів виявиться, що 45 з них віддали політичні уподобання демократам – це і є частота ознаки у вибірці. Ознакою, яку спостерігають, або змінною, є політичні уподобання виборців.

Таблиця 1–1. Політичні уподобання

Категорії	*f*
Демократи	45
Центристи	43
Ліберали	15
Всього	105

Вимірювання мають номінальну шкалу з трьома категоріями: демократи, центристи і ліберали. Розподіл частот для вибірки 105 виборців показано в таблиці 1–1. Символ *f* позначає частоти, *n* – загальна кількість спостережень. Зверніть увагу, що сума частот повинна дорівнювати *n*.

Побудована нами таблиця являє собою категоріальний розподіл частот для політичних уподобань 105 опитаних респондентів. Дамо також загальне визначення для категоріального розподілу частот ознаки.

Категоріальне розподіл частот складається з категорій, які є значеннями досліджуваної ознаки, і відповідних цим категоріям частот. Категоріальний розподіл будується для даних, вимірюваних номінальною або порядковою шкалою.

Для порядкових шкал розподіл частот будується аналогічно. У першому стовпці розташовуються можливі значення ознаки в порядку зростання. У другому – частоти, що відповідають кожному значенню.

У таблиці 1–2 наведено приклад розподілу частот по вибірці 60 глядачів. Вимірювалося їх ставлення до переглянутого фільму за порядковою шкалою з п'яти значень. Результати показують, що 36 глядачам фільм сподобався або дуже сподобався.

Таблиця 1–2. Відношення до фільму

Категорії	*f*
Дуже сподобався	24
Сподобався	12
Фільм середній	10
Не сподобався	6
Дуже поганий	8
Разом	60

Інтервальний розподіл частот

Для кількісних шкал кількість варіантів потенційно може бути дуже великою, оскільки кількісні шкали можуть включати безперервну кількість значень і ознака може приймати унікальне значення для кожного спостереження. Отже, представлення розподілу змінної, що вимірюється кількісної шкалою, вимагає іншого підходу.

Уявімо собі, що результати вимірів записані у вигляді послідовності чисел, які є значеннями ознаки, вимірюваними для кожного окремого об'єкта. Такі дані можуть бути неупорядкованими або впорядкованими.

Невпорядковані дані складаються з результатів, які записані в довільному порядку, наприклад в тому, в якому були отримані.

Впорядковані дані містять результати спостережень, записані в порядку зростання або зменшення.

В Таблиці 1–3 представлені впорядковані дані вимірювання ваги довільно обраних 77 осіб в кілограмах. Недоліком такого подання є його слабка придатність для подальшого вивчення вибірки.

Таблиця 1–3. Вага тіла (впорядковані дані)

(N = 77)						
46	59	65	69	71	74	79
49	60	65	69	71	75	80
50	60	65	69	72	75	81
50	60	66	70	72	75	81
52	61	67	70	73	76	83
53	62	67	70	73	76	84
54	62	67	70	73	77	84
55	63	68	70	73	77	85
55	64	68	71	74	78	87
56	64	68	71	74	79	89
58	64	69	71	74	79	90

Побудуємо інше, більш зручне представлення. Всі наші спостереження знаходяться в інтервалі від найменшого 46 до найбільшого 90. Згрупуємо наші спостереження і побудуємо таблицю інтервального розподілу частот.

Інтервальний розподіл частот складається з певної кількості інтервалів рівної довжини, на які ділиться весь діапазон зміни ознаки, і відповідних цим інтервалам частот.

Таблиця 1–4. Інтервальний розподіл частот

Інтервали	*f*
45–49	2
50–54	5
55–59	5
60–64	10
65–69	14
70–74	20
75–79	11
80–84	6
85–89	3
90–94	1
Разом	77

Дані про вагу 77 осіб тепер перетворені в таблицю за кількома інтервалами. Правий стовпець, що має *f* в якості заголовка, містить кількість людей, вага яких знаходиться у відповідному інтервалі. Наприклад, вага 10 осіб перебуває між 50 і 59 кг, а вага 14 осіб – між 60 і 69 кг.

Існує різниця між впорядкованими даними і інтервальним розподілом частот. Можна сказати, що в результаті перетворення ми втратили частину інформації, оскільки ми не маємо тепер 77 точних вимірювань, а отримали лише частоти потрапляння результатів спостережень в певний інтервал. Нам довелося пожертвувати деталями заради отримання знань про вид розподілу. Очевидно, що навіть недосвідченим спостерігачам таблиця з частотами більше говорить про особливості ваги людей у вибірці, ніж чим просто упорядкований набір спостережень.

Побудова інтервального розподілу

При побудові інтервального розподілу потрібно дотримуватися кількох правил.

Правило 1. Інтервали не повинні перетинатися, щоб одне значення не потрапило в два різних інтервали.

Правило 2. Інтервали повинні охоплювати всі можливі значення ознаки.

Правило 3. Інтервали повинні мати однакову довжину, щоб не спотворювати вигляд розподілу даних. Якщо в крайні інтервали потрапляє невелика кількість значень ознаки, вони можуть бути об'єднані з сусідніми і

тільки в цьому випадку інтервали будуть мати подвійну довжину в порівнянні зі звичайними.

Правило 4. Інтервали не повинні мати пробілів. Якщо в якийсь період не потрапило жодного значення, його не слід враховувати при розгляді. Винятком можуть бути лише перший і останній інтервали, які виключаються без шкоди для вигляду розподілу.

Правило 5 (не обов'язкове). Також важливо, щоб середини інтервалів були цілим числом. Це зручно для наступних обчислень і побудови графіків. Крім цього, зазвичай визначається від 5 до 20 інтервалів. Інша кількість не буде помилкою, але існує така негласна угоду.

Якщо перевірити вищенаведені умови, таблиця 2–4 не бездоганна. Зокрема, якщо людина має вагу 79,5 кг, то виникає питання, в якому інтервалі знаходиться це значення. Оголошені нами межі не є точними.

Таблиця 1–5. Визначені і точні межі інтервалів ваги тіла

Оголошені межі	**Точні межі**	***f***
45–49	менше 49,5	2
50–54	49,5–54,5	5
55–59	54,5–59,5	5
60–64	59,5–64,5	10
65–69	64,5–69,5	14
70–74	69,5–74,5	20
75–79	74,5–79,5	11
80–84	79,5–84,5	6
85–89	84,5–89,5	3
90–94	89,5 і більше	1
Разом		77

У таблиці 1–5 обчислені точні межі інтервалів зменшенням нижньої межі і збільшенням верхньої межі оголошених інтервалів на 0,5 кг. Точні межі, скажімо, сьомого інтервалу отримані в такий спосіб:

(Нижня межа) = 75 – 0,5 = 74,5

(Верхня межа) = 79 + 0,5 = 79,5

Для остаточного усунення двозначності слід сказати, що значення 74,5 потрапляє в інтервал 74,5 – 79,5 (відповідно до правила округлення – 74,5 округляється до 75, а 74,4 до 74), і не потрапляє в інтервал 69,5 – 74,5. Інтервал

69,5 – 74,5 містить всі значення, що більше або дорівнюють 69,5 і менші 74,5. Тепер ми повністю усунули перетин інтервалів, зробили їх вичерпними, тобто таким, що охоплюють всі можливі значення. Для цього нам знадобилося розширити два крайніх інтервали.

Як вже говорилося, після переходу від упорядкованого масиву до інтервального розподілу частот, ми втратили інформацію про кожне індивідуальне значення ознаки у вибірці. З таблиці 2–5 видно, що 14 осіб важать від 64,5 до 69,5 кг, але точну вагу будь-якого індивідуума при цьому невідома. Найкращою оцінкою ваги кожного з цих людей є середина інтервалу.

Середина інтервалу розраховується за наступною формулою:

(середина інтервалу) = (початок інтервалу) + (довжина інтервалу) / 2

Для прикладу, стосовно інтервалу 69,5–74,5, (середина інтервалу (69,5-74,5)) = 69,5 + 5/2 = 69,5 + 2,5 = 72. Вдало, що середина є цілим числом.

Якщо ми виберемо довжину інтервалу у 20 кг, то отримаємо всього 3 інтервали, які не дадуть нам необхідної інформації. При довжині в 2 кг ми, навпаки, отримаємо 30 інтервалів і невеликі частоти, які також не дадуть нам досить ясної картини. Вибір інтервалу довжиною в 5 кг дозволив побудувати 10 інтервалів і отримати цілком придатний для аналізу розподіл частот.

Розглянемо наступний приклад. Є дані про тривалість 20 розмов по телефону в хвилинах. Отримайте розподіл частот, використовуючи 7 інтервалів.

11	29	6	33	14
21	18	17	22	38
31	22	27	19	22
23	26	39	34	27

Рішення. Побудова інтервального розподілу частот представимо по кроках.

Крок 1. Знайдемо найбільше і найменше значення: 39 і 6.

Крок 2. Визначимо діапазон значень вибірки:

(діапазон) = (найбільше значення) – (найменше значення) = 39 – 6 = 33.

Крок 3. Визначимо кількість інтервалів: (число інтервалів) = 7.

Крок 4. Знайдемо довжину інтервалу, розділивши діапазон даних на кількість інтервалів:

$$\text{Довжина інтервалу} = \frac{\text{діапазон}}{\text{число інтервалів}} = \frac{33}{7} \approx 4{,}7.$$

Округлимо до найближчого цілого: 4,7 ≈ 5.

Таблиця 1-6. Обчислення інтервального розподілу частот

Інтервал	Точні межі	Частота
6-10	5,5-10,5	1
11-15	10,5-15,5	2
16-20	15,5-20,5	3
21-25	20,5-25,5	5
26-30	25,5-30,5	4
31-35	30,5-35,5	3
36-40	35,5-40,5	2

Крок 5. Нижньою межею першого інтервалу є найменше значення 6. Додамо 5 і отримаємо нижню межу другого інтервалу. Продовжимо додавати до тих пір, поки не отримаємо 7 значень: 6, 11, 16, 21, 26, 31, 36 – нижні межі.

Крок 6. Віднімемо одиницю з нижніх меж, щоб отримати верхні межі. Перший інтервал 6-10, другий 11-15 і так далі.

Крок 7. Знайдемо точні межі інтервалів, віднімаючи 0,5 від кожної нижньої межі і додаючи 0,5 до верхньої межі інтервалу.

Крок 8. Підрахуємо кількість даних, що потрапляють в кожен з інтервалів. Зручно вести підрахунок графічними значками.

Крок 9. Запишемо підсумкові чисельні значення в стовпці для частот.

Резюме

У цій частині розглянуті два типи розподілу частот. Перший тип використовується для категоріальних даних і називається категоріальним розподілом. Другий тип розподілу використовується для числових даних, які необхідно групувати в інтервали. Він називається інтервальним розподілом частот. Обидва типи розподілів часто використовуються в статистичних обчисленнях і корисні для тих, хто описує і представляє дані.

Як ми зможемо переконатися, частотні розподілу необхідні, щоб представити дані правильним і зрозумілим чином, візуально визначити характер і форму розподілу, побудувати таблиці, графіки, діаграми даних, спростити обчислення середнього значення та інших параметрів, порівняти різні дані між собою.

2.2. Відносні частоти, частки, відсотки

Існують загальноприйняті способи опису і використання розподілу частот. Найбільш застосовні – відносини частот, частки, відносні частоти, відсотки, накопичені частоти і накопичені відсотки. Будь-яка форма подання частот пов'язана з певним сформованим стандартом.

Відносини частот

Найбільш загальний вигляд відносних частот є звичайне відношення двох чисел. Наприклад, відношення чисельності студентів-соціологів до чисельності студентів-економістів:

$$\frac{\text{Соціологи}}{\text{Економісти}}$$

Якщо серед 60 студентів 40 соціологів і 20 економістів, то відношення соціологів до економістів є 40/20 = 2,0, що означає, що на два соціолога приходить по одному економісту. Ми можемо порахувати також відношення економістів до соціологів: 20/40 = 0,5, і це означає, що на 0,5 економістів приходить один соціолог.

Відношення частот (ratio) – це традиційний поділ однієї частоти на іншу і побудова дроби.

Загальна формула для відносин частот:

$$\text{Відношення} = \frac{f_1}{f_2}$$

де, f_1 – частота першого значення ознаки, f_2 – частота другого значення ознаки.

Відносини дуже корисні, оскільки вони дозволяють порівнювати різні виміри. Наприклад, ви можете порахувати кількість студентів на одного

викладача у закладі вищої освіти, розмір оплати за навчання до кількості аудиторних годин, кількість автомобілів або мобільних телефонів в сім'ї і так далі. Відносини можна обчислювати для всіх типів шкал – числових, порядкових, номінальних, оскільки відносини засновані на кількості спостережень з одним значенням ознаки. Вони є потужним наочним інструментом, тому що легко зрозумілі і завжди під рукою.

Частки і відносні частоти

Будемо розглядати частку як відношення деякої підмножини частот до загальної суми частот. Ми можемо записати це у вигляді такої формули:

$$\text{Частка} = \frac{\text{Частина}}{\text{Ціле}}$$

Частка (proportion) є відношення частини до цілого. Частка в розподілі частот є відношенням однієї з частот до загальної кількості спостережень, яке також прийнято називати **відносною частотою** значення ознаки.

Частка в розподілі частот може бути виражена як:

$$\text{Частка} = \frac{f_i}{n}$$

де f_i – одна з частот в розподілі, n – загальне число спостережень.

Частка соціологів серед 60 студентів (соціологів та економістів) буде обчислена в такий спосіб:

$$\frac{\text{Соціологи}}{\text{Соціологи} + \text{Економісти}}$$

У нашому прикладі, частка соціологів в групі з 60 студентів є 40 / (40 + 20) = 0,67. Частка економістів є 20 / (40 + 20) = 0,33. Зауважимо, що сума 0,67 + 0,33 дорівнює 1,00. Сума часток розподілу частот повинна дорівнювати одиниці. Обчислюючи частку, ми отримуємо, що число соціологів щодо загального числа студентів складає 0,67 і що число економістів щодо загальної кількості – 0,33.

Ми можемо розглядати частки з ще однієї точки зору. Перехід від початкових даних до часток може розглядатися як лінійне перетворення розподілу частот. У нашому прикладі, первинний розподіл соціологів та

економістів було перетворено від виду (40 + 20 = 60) до виду (0,67 + 0,33 = 1,00). Перетворення виявляється корисним, оскільки новий вид більш зрозумілий і наочний, особливо, якщо ми маємо справу з великими числами. Крім того, перетворення дозволяє нам приводити різні виміри до стандартного розподілу значень, яке розташовується від 0,00 до 1,00. Вимірювання, засновані на різних обсягах вибірки, можна порівняти, якщо перетворити їх в частки.

Таблиця 1–7. Політичні уподобання

Категорії	Частоти, f	Відносні частоти	Відсотки, %
Демократи	45	0,429	43%
Центристи	43	0,410	41%
Ліберали	15	0,143	14%
Разом	105	1,000	100%

Відсотки

Відсотки (percentages) – це частка, помножена на 100%:

$$\text{Відсотки} = \frac{f_i}{n} \times 100$$

Відсотки соціологів та економістів в групі 60 студентів: 40/60*100 = 66,7% і 20/60*100 = 33,3%. Сума відсотків всіх частот розподілу завжди дорівнює 100. Тим самим, сума відсотків соціологів та економістів складає 66,7 + 33,3 = 100. Перетворення частот розподілу в відсотки є іншим лінійним перетворенням, яке переводить початкові частоти в нову числову шкалу, що розташовується від 0 до 100. Ми перетворили первинний розподіл частот 40 + 20 = 60 в 66,7 + 33,3 = 100. Переваги використання відсотків такі ж, як використання відносин, їх настільки часто використовують повсякденно, що вони розуміються та використовуються кожним з нас.

Відсотки в прикладі вище округлені до одного десяткового знака. Це загальне правило дозволяє зберігати дані у вигляді, подібному вихідним вимірам. Спочатку ми говорили про соціологів і економістів в термінах індивідів, а потім в термінах відсотків з одним десятковим знаком.

Повернемося до таблиці категоріального розподілу з політичними уподобаннями і продовжимо її двома стовпцями – третім і четвертим (таблиця 1-7). У третьому стовпці ми обчислюємо відносні частоти (частки) для кожного із значень ознаки. Для абсолютної частоти демократів 45 відносна частота обчислюється так: P = 45/105 = 0,429. Відсотки обчислюються за формулою: 45/105*100% = 42,9%. До якого знака доречно округляти? В даному випадку це залежить від мети подальшого використання результатів. Пам'ятайте, що округлення завжди призводить до помилок округлення. У нашому прикладі, після округлення сума по четвертому стовпчику не дорівнює 100%, як би цього нам не хотілося! Будьте уважні!

Накопичені частоти, відносні частоти і відсотки

Відносини, частки і відсотки можуть застосовуватися для всіх типів шкал. Відносини і відсотки, крім цього, можна накопичувати, щоб показати число спостережень вище або нижче будь-якого обраного значення розподілу. Однак, таке накопичення можливо тільки в інтервальних і порядкових шкалах. Категорії в номінальній шкалі не можуть бути впорядковані, і тому не має сенсу думати в термінах «вище» або «нижче» для категорій номінальної шкали.

Таблиця 1–8. Частоти, частки і відсотки результатів іспиту

Категорії	*f*	C*f*	P	CP	%	C%
Відмінно	17	17	0,200	0,200	20,0%	20,0%
Добре	41	58	0,482	0,682	48,2%	68,2%
Задовільно	20	78	0,236	0,918	23,6%	91,8%
Незадовільно	7	85	0,082	1,000	8,2%	100,0%
Разом	85		1,000		100,0%	

Позначення стовпців: *f* – частота, C*f* – накопичена частота, P – відносна частота, CP – накопичена відносна частота, % – відсотки, C% – накопичені відсотки.

Таблиця 1–8 показує результати іспиту для 85 осіб. Накопичена частота для оцінки «Добре» отримана складанням 17 + 41 = 58. Накопичена відносна частота для оцінки «Добре» отримана складанням 0,200 + 0,482 = 0,682.

Накопичені відсотки для оцінки «Добре» отримані складанням 20,0% + 48,2% = 68,2%.

Накопичені відносні частоти, СР, показують, що 0,918 загальної чисельності студентів отримали оцінки вище «Незадовільно». Віднімаючи, ми отримаємо: 1,000 – 0,682 = 0,318 – така частка студентів отримала менше 4 балів.

Зверніть увагу, що сума стовпця відносних частот Р, дорівнює 1,00 і що накопичена частота СР, дорівнює 1,00. Таким чином, первинний розподіл частот *f*, що включало 85 спостережень, було перетворено до нового розподілу зі значеннями в межах від 0,00 до 1,00. Тим самим, частки перетворюють частоти розподілу незалежно від кількості первинних спостережень до відносного масштабу в межах від 0,00 до 1,00, в той час як відсотки перетворюють їх в масштаб від 0,00 до 100,00. Ці перетворення легко інтерпретуються і дозволяють порівнювати групи з різною кількістю спостережень.

Таблиця 1–9. Результати іспиту двох груп

	частоти	
Сума балів	**Група 1**	**Група 2**
100	10	47
95	20	94
90	30	141
85	30	141
80	20	94
75	10	47
РАЗОМ	120	564

У Таблиці 1-9 представлені дані результатів іспиту в двох різних груп студентів. З такими даними досить складно мати справу, оскільки візуально важко розглядати і аналізувати надані частоти.

Складемо іншу таблицю. У Таблиці 1-10 наведені також відносні частоти і відсотки. Тепер можна зробити деякі висновки. Очевидно, що розподіли частот для обох груп однакові. Перетворення початкових розподілів

частот до часток або відсотків дозволило нам побачити, що немає ніяких відмінностей між групами.

Таблиця 1–10. Частоти, частки і відсотки результатів іспиту для двох груп

	ГРУПА 1			**ГРУПА 2**		
СУМА БАЛІВ	*f*	*P*	*%*	*f*	*P*	*%*
100	10	0,083	8,3	47	0,083	8,3
95	20	0,167	16,7	94	0,167	16,7
90	30	0,250	25	141	0,250	25
85	30	0,250	25	141	0,250	25
80	20	0,167	16,7	94	0,167	16,7
75	10	0,083	8,3	47	0,083	8,3
РАЗОМ	120	1	100	564	1	100

Резюме

Відносні частоти, частки і відсотки виключно корисні, оскільки вони представляють частоти в деякому стандартному, широко відомому вигляді. Частки і відсотки, до того ж, є лінійним перетворенням частот в стандартну форму. Далі ми опишемо табличний спосіб представлення даних.

2.3. Таблиці

Таблиці є зручною і популярною формою представлення даних. Слід обговорити основні правила, яким слід дотримуватися при побудові і використанні таблиць.

Стандартний вид таблиці

На малюнку 1–1 представлені компоненти стандартної таблиці. Таблиця завжди має назву, яка ясно повідомляє про її зміст. Назва повинна бути короткою і має вказувати змінні, які містяться в таблиці. Якщо в документі є більше однієї таблиці, вони нумеруються. Якщо документ містить більше одного розділу, таблиці можуть нумеруватися двома числами, наприклад, 1-1, 1-2, і 2-1, де перше число відповідає номеру розділу, а друге число – номеру таблиці в розділі. Статистичні таблиці, крім того, зазвичай містять інформацію про кількість спостережень, дані про які представлені в таблиці. Кількість

спостережень вказується в самій таблиці або її назві. Основне поле (або тіло) таблиці містить дані спостережень.

<table>
<tr><td colspan="2">Номер, Заголовок таблиці</td></tr>
<tr><td rowspan="2">Назва рядків</td><td>Назва стовпців</td></tr>
<tr><td>Заголовки стовпців</td></tr>
<tr><td>Заголовки рядків</td><td>Поле (тіло) таблиці</td></tr>
</table>

Малюнок 1–1. Стандартний вид таблиці

Таблиці спряженості

Розрізняють способи побудови таблиць для однієї, двох або кількох змінних. Таблиці для двох змінних містять одну змінну в рядках, а іншу в стовпці. Таблиці для двох змінних називаються таблицями спряженості, вони показують зв'язок або відносини між ними. Приклад таблиці спряженості наведено в таблиці 1-11.

Таблиця 1–11. Вид діяльності і задоволеність оплатою праці

Задоволеність розміром оплати праці	**Вид діяльності**		**РАЗОМ**
	Робочий	*Службовець*	
Низька	35	11	46
Висока	12	49	61
РАЗОМ	47	60	107

Якщо вивчається зв'язок між незалежною і залежною змінною, то залежна частіше розміщується в рядках, а незалежна в стовбцях. Якщо в таблиці розраховуються відсотки, то вони розміщуються у напрямку незалежної змінної.

Таблиця 1–12 показує зв'язок між місцем проживання і бажаною формою дозвілля. Тіло таблиці показує вимірювання вибірки 469 респондентів з чотирьох різних міст. Місце проживання розглядається як незалежна змінна, і вважається, що форма дозвілля залежить від місця проживання. Кожна клітинка таблиці показує кількість спостережень, для яких збіглося місце проживання і форма дозвілля. Таблиця показує частоту і відсоток для кожної комірки. Відсотки розраховані для незалежної змінної (в стовпці), щоб

показати відсоток людей, що віддають перевагу тій чи іншій формі дозвілля в залежності від місця проживання. Якщо скласти відсотки в будь-якому стовпці, отримаємо 100%.

Можна було ще розрахувати відсотки в рядках і відсоток для кожного осередку по відношенню до загального числа спостережень. Однак, існує правило, що в таблиці повинно бути стільки обчислень, скільки потрібно для дослідження. Слід уникати перевантаження даними.

Таблиця 1–12. Форма дозвілля і місце проживання

(N = 469)					
Форми дозвілля	Місце проживання				**Разом по рядку**
	Київ	Дніпро	Харків	Запоріжжя	
Спорт	94 58,4%	47 29,9%	43 48,3%	26 41,9%	**210** **44,8%**
Автомобілі	49 30,4%	93 59,2%	21 23,6%	31 50,0%	**194** **41,4%**
Комп’ютер	18 11,2%	17 10,8%	25 28,1%	5 8,1%	**65** **13,9%**
Разом по стовбцю	**161** **34,3%**	**157** **33,5%**	**89** **19,0%**	**62** **13,2%**	**469** **100,0%**

Зауважимо, що ми вирахували відсотки і для рядка підсумків за стовпцями. Це показує нам, як вибірка 469 осіб розподілена по відношенню до форм дозвілля або яка частка жителів конкретного міста у вибірці. Ми бачимо, наприклад, що 44,8% вважають кращою формою дозвілля спорт, 41,4% – автомобілі і 13,8% – комп'ютери. Ми також знаємо, що 34,3% – проживають в Києві, 33,5% – Дніпрі, 19% – Харкові і 13,2% – Запоріжжі.

2.4. Візуалізація даних

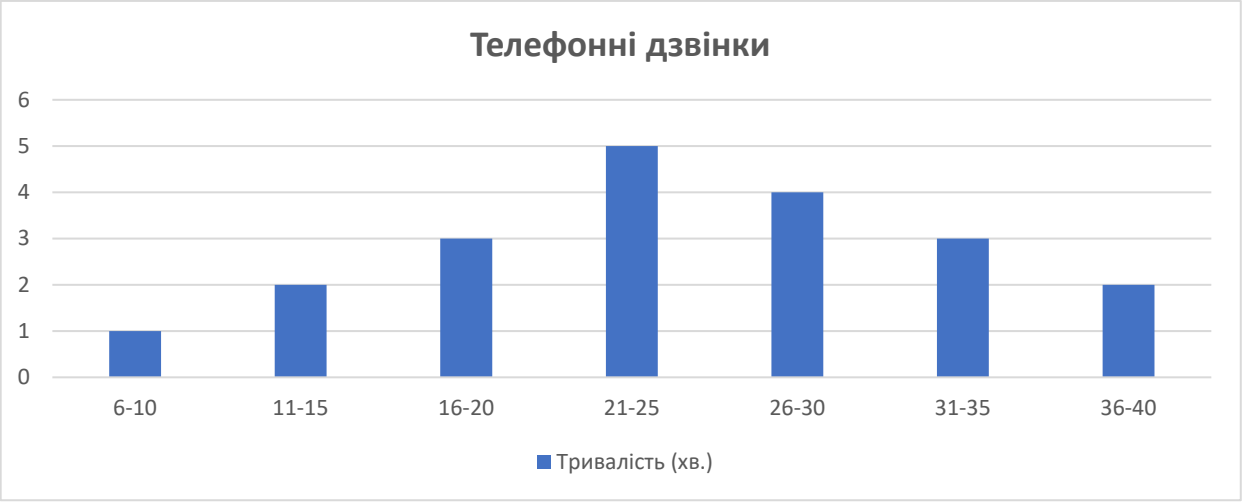
Важливим інструментом представлення (візуалізації) числових даних є використання графічних зображень. Наступний параграф розглядає деякі графічні методи. Ми розглянемо гістограми, полігони частот, кумуляти.

Гістограми частот

Гістограма частот – це графічне представлення, в якому по осі X відкладаються значення змінної, а по осі Y відповідні їм частоти.

Гістограма будується в вигляді прямокутників, висота яких відповідає частоті значення змінної.

Побудуємо гістограму частот для прикладу з телефонними дзвінками. Використовуємо для цього результати, отримані нами в таблиці 2–6 розподілу частот. По осі X відкладаємо відповідні інтервали: 6–10, 11–15 і т.д., разом сім. В інтервалі 6–10 будуємо прямокутник висотою 1, що означає, що лише один дзвінок з 20 потрапив за тривалістю в перший інтервал, в другому інтервалі – прямокутник висотою 2 і так далі. Гістограма частот корисна для візуальної оцінки даних. Можна відзначити, що найбільш популярний час: 21–25 хвилин, в цей інтервал потрапило 5 дзвінків, це більше, ніж в будь-якому іншому інтервалі.



Малюнок 1–2. Гістограма частот для телефонних дзвінків

Малюнок 1–3. Гістограма відносних частот

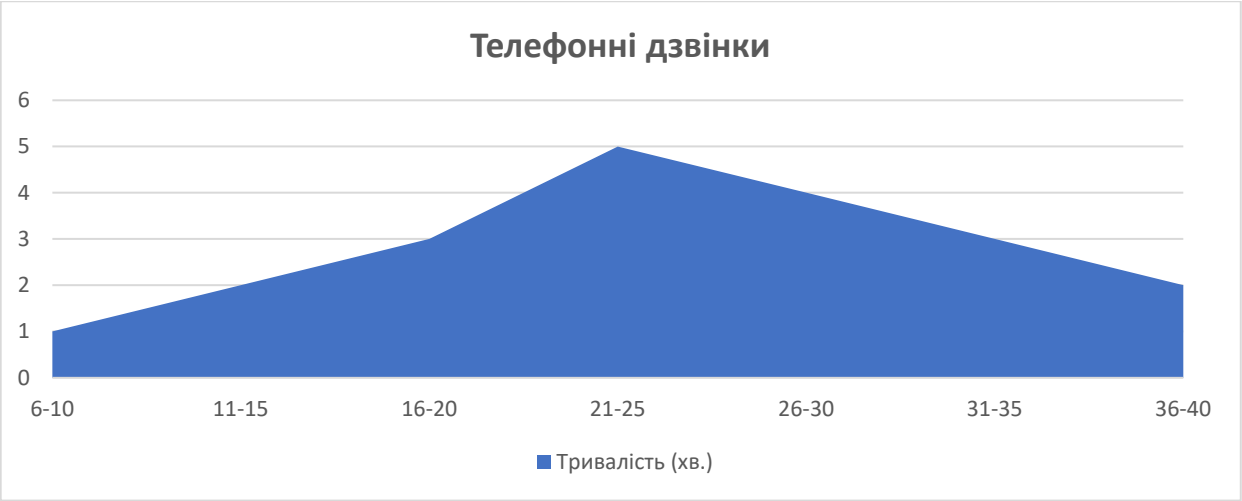
Гістограма відносних частот – це графічне представлення, в якому по осі X відкладаються значення змінної, а по осі Y відповідні їм частки або відсотки. Гістограма будується в вигляді прямокутників, висота яких відповідає частці або відсоткам для значення змінної.

Гістограма відносних частот будується таким чином, що замість частот на осі Y відзначені частки або відсотки для відповідних значень змінної. Для розглянутого нами прикладу гістограма відносних частот побудована на малюнку 1-3. Обидві гістограми ідентичні, відмінність полягає лише в інший шкалі по осі Y.

Спираючись на гістограму, можна сказати, наприклад, що 0,20 всіх дзвінків потрапило за тривалістю в інтервал 26–30 хвилин (або 20%).

Полігони частот

Полігон частот будується подібним до гістограми чином. Замість прямокутників в полігоні будується лінія по точкам, які відповідають серединам інтервалів і частотам. Полігон дає зорове уявлення про розподіл частот, яке сильно відрізняється від гістограми при використанні одних і тих же даних. Не існує правил, які б визначали краще представлення даних. Все залежить від конкретної ситуації і від смаків дослідника, який обирає вид представлення з існуючих альтернатив.



Малюнок 2–4. Полігон частот для телефонних дзвінків

Полігон – це графічне представлення, в якому по осі X відкладаються значення змінної, а по осі Y відповідні їм частоти. Полігон будується в вигляді області, обмеженої лінією, що проходить по точках (середина інтервалу, частота).

Полігони, також як і гістограми, можна будувати для частот, відносних частот і відсотків. У всіх трьох випадках графік залишиться однаковим, за винятком осі Y. На малюнку 1–4 полігон побудований для частот.

Кумуляти

Ще одним часто використовуваним графічним представленням даних є кумулята – від слова «кумулятивний» – той, що накопичується. В цьому випадку по осі Y відкладаються накопичені частоти або накопичені відсотки.

На малюнку 1–5 графік побудований для накопичених відсотків. За графіком можна сказати, наприклад, що близько 70% дзвінків має тривалість менше 30 хвилин, зате менше 20 хвилин розмовляють лише 30%.

Кумуляти можуть бути побудовані як за зростанням значень ознаки, так і за спаданням, в залежності від того, який аналіз ми плануємо проводити з використанням отриманого графічного зображення.

Малюнок 1–5. Кумулята частот для телефонних дзвінків

Аналіз візуалізацій

Ми не ставили за мету назвати всі можливі графічні способи представлення даних. Загальновідомі кругові діаграми, лінійні графіки та інші. Із спектром можливостей можна познайомитися в будь-якій комп’ютерній програмі, що займається обробкою і візуалізацією даних. Ми зупинилися лише на тих, які часто використовуються для візуального аналізу даних з точки зору їх розподілу. Зокрема, вони дозволяють відповісти на наступні питання:

– Які значення є мінімальними і максимальними?

– Який розмах наявних даних?

– Які значення зустрічаються в наборі даних частіше за все? Які значення є найбільш типовими?

– Яка стандартна різниця між значеннями в наборі даних?

– Який вигляд має розподіл? Де зосереджена основна частина даних? Наскільки симетрично вони розташовані навколо типового значення? В який бік зміщені?

– Чи є характерні особливості? Викиди? Чи є пропущені значення ознаки?

Використовуємо комп’ютер

Ця глава надає великий матеріал для комп'ютерних вправ. Застосування програмного забезпечення до матеріалу глави допоможе, одночасно,

навчитися використовувати комп’ютер для представлення даних і глибше зрозуміти зміст глави. Перше завдання з використанням комп’ютера це організація і введення даних. Продумайте список змінних, з якими будете працювати, визначте їх можливі значення. Потім створіть таблиці частот, включіть в них групові частоти. Порахуйте абсолютні значення і відсотки, накопичені частоти і накопичені відсотки.

Статистичний пакет, ймовірно, містить опції для створення різних типів діаграм. Ви можете створити діаграми, які відповідають інтервальним, порядковим та номінальним даними. Ці початкові кроки щодо застосування статистичного пакету дуже важливі для його подальшого використання для вирішення більш складних завдань.

Символи та формули

- X_i – елементи вибірки, варіанти
- f_i – частоти
- $P = f_i / n$ – відносні частоти
- CP – накопичені відносні частоти
- % – відсотки
- C% – накопичені відсотки