Вступ

Самі числа говорити не вміють. За них говоримо ми.
Ми наповнюємо їх сенсом.
— Нейт Сілвер, «Сигнал і шум»[1]

Чому нам потрібна статистика

Гарольд Шіпман був найплодовитішим на жертви з усіх засуджених убивць Британії, хоча він зовсім не підходив під стать типового образу серійного вбивці. В період між 1975 та 1998 роками цей тихий сімейний лікар, що працював у передмісті Манчестера, зробив смертельну ін'єкцію **опіату** не менш ніж 215 своїм пацієнтам переважно похилого віку. Зрештою він припустився помилки, сфальсифікувавши заповіт однієї зі своїх жертв, щоб заволодіти її грошима. Дочка цієї жертви виявилась адвокатом. У неї виникли підозри, і судова експертиза його комп'ютера виявила, що він заднім числом змінював історію хвороби своїх пацієнтів так, щоб вони мали *більш* хворий вигляд, ніж насправді. Шіпман був відомим затятим прихильником нових технологій, але, як виявилось, недостатньо технічно підкутим, щоб усвідомити, що кожна внесена ним зміна мала часову мітку (до речі, хороший приклад даних, що розкривають прихований зміст).

Після ексгумації тих його пацієнтів, що не були кремовані, у п'ятнадцяти виявили смертельний рівень діаморфіну — медичної форми героїну. В 1999 році Шіпмана засудили за п'ятнадцять убивств. Він відмовився від захисту і не вимовив ні слова на суді. Його визнали винним та засудили до довічного ув'язнення, а також розпочали громадське розслідування, щоб з'ясувати, які ще злочини, крім тих, за які його засудили, він міг вчинити і чи **могли** його спіймати раніше.

Я був одним із кількох статистиків, що давав свідчення в рамках цього громадського розслідування, яке дійшло висновку, що він точно вбив 215 своїх пацієнтів, а можливо, й ще 45[2].

Основна увага в цій книзі буде зосереджена на використанні **статистики**[**] для отримання 1945 відповідей на питання, що виникають, коли ми намагаємося краще зрозуміти світ. Деякі з цих питань будуть винесені у текстові врізки. Щоб отримати певне уявлення про мотиви Шіпмана, перш за все хочеться поставити таке запитання:

> Яких саме людей вбивав Гарольд Шіпман, і коли вони помирали?

Громадське розслідування надало подробиці про вік, стать і дату смерті жертв. Рис. 0.1 — це доволі складна візуалізація цих даних, на якій ми бачимо діаграму розсіювання (точкова діаграма), де зображено вік жертви і рік убивства, при цьому відтінок крапок вказує на стать жертви. На осях додані гістограми, які вказують на вікові групи (з інтервалом у п'ять років) та роки.

Роздивившись цей рисунок, можна дійти певних висновків. На ньому набагато більше чорних крапок, ніж сірих, тож жертвами Шіпмана переважно були жінки. Гістограма з правого боку рисунка демонструє, що більшості з його жертв було від 70 до 80 років, але якщо подивитись на те, як розкидані точки, стає очевидно, що хоча спочатку вони всі були похилого віку, з роками до них додалося кілька молодших пацієнтів. Гістограма вгорі чітко показує прогалину приблизно в 1992 році, коли вбивств не було. Виявилось, що перед тим Шіпман вів сумісну практику з іншими лікарями, а потім, можливо, відчувши, що вони почали щось підозрювати, почав працювати сам. Після цього його діяльність активізувалась, як видно на верхній гістограмі.

---

[*] Терміни, виділені **напівжирним** шрифтом, наводяться у «Глосарії» наприкінці книги, де містяться як основні, так і технічні визначення.

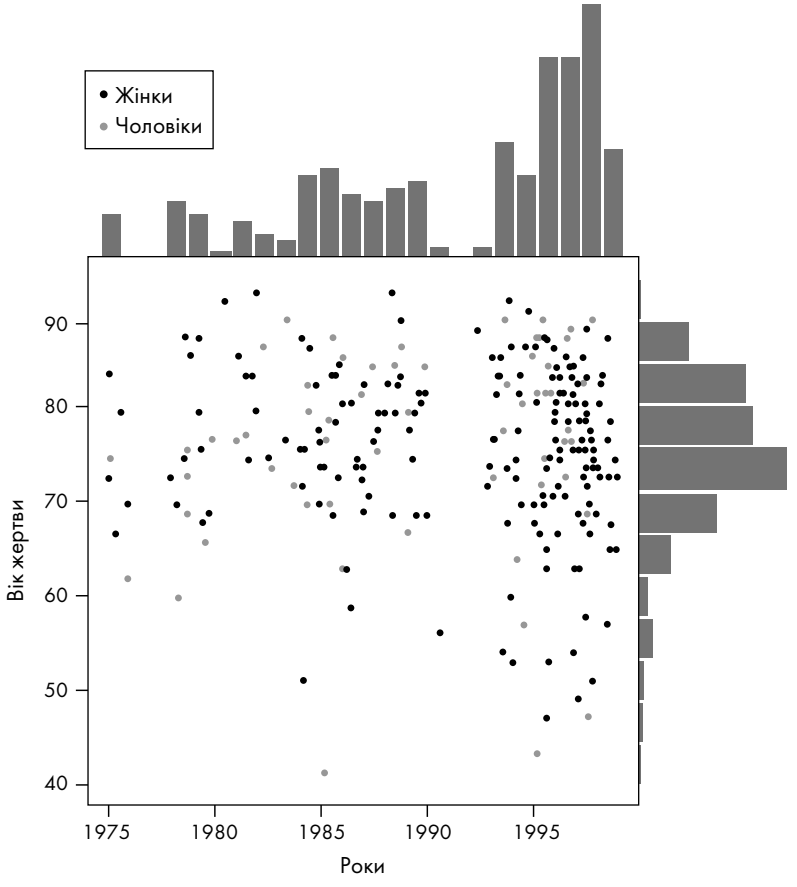

Рис. 0.1. Діаграма розсіювання, де показані вік та рік смерті 215 підтверджених жертв Гарольда Шіпмана. Гістограми на осях, демонструють розподіл за віком та роком скоєння вбивства

Цей аналіз жертв, виявлених під час розслідування, викликає ще більше запитань щодо того, як він скоював убивства. Деякі статистичні докази вдалося отримати з даних про час смерті його ймовірних жертв, указаний у свідоцтві про смерть. Рис. 0.2 — це точкова

діаграма, в якій час смерті пацієнтів Шіпмана порівнюється з часом смерті пацієнтів інших місцевих сімейних лікарів. Щоб помітити закономірність, не потрібно бути досвідченим аналітиком: такі очевидні результати видно неозброєним оком. Пацієнти Шіпмана переважною більшістю помирали вранці.

Ці дані не можуть сказати нам, *чому* вони здебільшого помирали в той час, але подальше дослідження виявило, що він навідувався до своїх літніх пацієнтів додому після обіду, коли зазвичай міг залишитися з ними наодинці. Він пропонував зробити їм ін'єкцію, яка, за його словами, мала покращити їхнє самопочуття, але насправді вколював їм смертельну дозу діаморфіну. Після того як пацієнт тихо помирав у нього перед очима, він змінював запис у його медичній картці, щоб усе виглядало як очікувана природна смерть. Пізніше суддя Джейн Сміт, яка очолювала громадське розслідування, сказала: «Я досі думаю про те, як це невимовно жахливо — просто невимовно, й немислимо, й неймовірно, — те, що він приходив день у день, прикидаючись їхнім чудовим турботливим лікарем, та носив із собою в сумці свою смертельну зброю... і просто діловито діставав її звідти».

Він був свідомий певного ризику, оскільки достатньо було єдиного розтину трупа, щоб викрити його, але, враховуючи вік пацієнтів та очевидні природні причини їх смерті, жодного розтину так і не було зроблено. І він так ніколи й не пояснив причини, через які здійснював ці вбивства: Шіпман не давав свідчень у суді, ніколи не говорив ні з ким про свої лиходійства, навіть зі своєю сім'єю, і вчинив самогубство у в'язниці, вельми доречно, саме в той час, коли його дружина ще мала право на його пенсію.

Ми можемо розглядати подібний тип циклічної дослідницької роботи як «криміналістичну» статистику, і в цьому разі буквально так і було. Ніякої математики, ніякої теорії, просто пошук закономірностей, які можуть навести на ще більше цікавих запитань. Деталі Шіпманових злочинів визначались за допомогою свідчень окремо у кожному конкретному випадку, але цей аналіз даних дав нам загальне розуміння того, як він здійснював свої злочини.

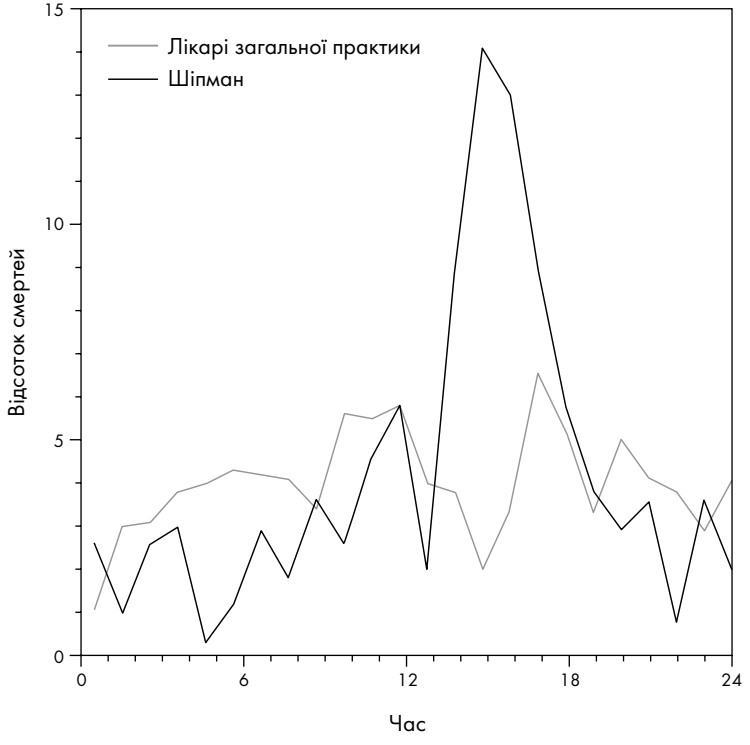

Рис. 0.2. Час смерті пацієнтів Гарольда Шіпмана порівняно з часом смерті пацієнтів інших місцевих лікарів загальної практики. Не треба вдаватися до складного статистичного аналізу, щоб зрозуміти, про що говорить цей графік

Далі в цій книзі, у розд. 10, ми дізнаємося, чи міг би формальний статистичний аналіз допомогти спіймати Шіпмана раніше*. Тим часом історія Шіпмана наочно демонструє великий потенціал вико-

* Увага, спойлер: майже напевно, зміг би.

ристання даних для кращого розуміння світу та ухвалення кращих рішень. У цьому й полягає суть статистики.

Перетворюємо світ на дані

Статистичний підхід до злочинів Гарольда Шіпмана вимагає від нас не враховувати довгий список окремих трагедій, за які він несе відповідальність. Усі ті особисті, унікальні деталі, що стосуються життя і смерті людей, необхідно звести до низки фактів та чисел, які можна порахувати й зобразити у вигляді графіків. На перший погляд такий підхід здається бездушним і нелюдяним, та якщо ми збираємося використовувати статистику, щоб кинути світло на наш світ, тоді наш повсякденний досвід слід перетворити на дані, а для цього треба категоризувати й класифікувати події, записати виміри, проаналізувати результати й сформулювати висновки. Однак уже проста категоризація та класифікація можуть виявитись складною задачею. Візьмемо просте запитання, яке має цікавити кожного, кого хвилюють проблеми довкілля:

> Скільки на нашій планеті дерев?

Перш ніж задуматись над тим, як ми можемо відповісти на це запитання, ми маємо з'ясувати одне просте поняття. Що таке дерево? Ви можете вважати, що впізнаєте дерево, коли побачите його, але ваше судження може кардинально відрізнятися від думки інших, що назвуть його кущем або чагарником. Тож для того щоб перетворити свій досвід на дані, ми маємо почати з чітких визначень.

Виявляється, за офіційним визначенням, дерево — це рослина з дерев'янистим стовбуром, що має доволі великий діаметр на рівні грудей (ДРГ). Лісова служба США вважає, що рослину можна офіційно називати деревом, якщо її ДРГ — понад 12,7 см, але більшість організацій задовольняються значенням ДРГ 10 см.

Проте ми не можемо особисто обійти всю планету, щоб виміряти кожну рослину з дерев'янистим стовбуром і полічити ті, що відповідають цьому критерію. Тому дослідники, які вивчали це питання, вирішили вдатися до більш прагматичного підходу: спочатку вони взяли низку територій з поширеним типом ландшафту, відомим як біом, та підрахували середню кількість дерев на квадратний кілометр. Потім за допомогою супутникових знімків вони оцінили загальну площу планети, вкриту кожним типом біому, виконали складне статистичне моделювання і, зрештою, отримали приблизну суму: 3,04 трильйона (тобто 3 040 000 000 000) дерев на планеті. Ця цифра здається великою, хоча вчені вважають, що колись дерев було вдвічі більше*[3].

Якщо організації не в змозі дійти згоди в тому, що таке дерево, то не дивно, що визначення більш розпливчастих понять викликає ще більші труднощі. Ось вам яскравий приклад: в період з 1979 по 1996 рік офіційне визначення слова «безробіття» у Великій Британії змінювалося щонайменше тридцять один раз[4]. Визначення валового внутрішнього продукту (ВВП) постійно переглядається; так, у 2014 році до ВВП Великої Британії були додані нелегальна торгівля наркотиками та проституція, а для попередніх розрахунків використовувались незвичні джерела даних, наприклад, розцінки на різні види послуг проституток взяли з вебсторінки Punternet, на якій оцінювались послуги повій[5].

Навіть наші особисті почуття можна систематизувати й піддати статистичному аналізу. В рамках річного опитування, яке закінчилось у вересні 2017 року, 150 000 жителям Великої Британії поставили таке запитання: «Наскільки щасливими ви почувалися вчора?»[6] За шкалою від нуля до десяти опитані, в середньому, почувалися щасливими на 7,5 балів — краще, ніж у 2012 році, коли цей середній

* Похибка для цієї величини 0,1 трильйона. Це означає, що, на думку дослідників, справжня кількість дерев лежить в діапазоні від 2,94 до 3,14 трильйона (я припускаю, що ця цифра занадто точна, враховуючи багато припущень, зроблених під час моделювання). За оцінками фахівців, щороку вирубується 15 мільярдів (15 000 000 000) дерев, і з часу виникнення людської цивілізації планета втратила 46 % своїх дерев.

бал становив 7,3, що може бути пов'язано з відновленням економіки після фінансової кризи 2008 року. Найнижчі бали поставили люди віком від 50 до 54 років, а найвищі — віком від 70 до 74 років, що цілком типово для Великої Британії.*

Виміряти щастя важко, натомість із визначенням жива людина чи мертва не має бути жодних проблем. Як можна побачити з наведених у цій книзі прикладів, питання народжуваності та смертності цікаві для статистики. Але в Сполучених Штатах кожен штат має своє власне законне визначення смерті, і хоча для того щоб досягти хоч якоїсь стандартизації, в 1981 році був прийнятий Закон про єдине визначення смерті, деякі незначні розбіжності все ж залишились. Так, людина, яку оголосили мертвою в Алабамі, теоретично могла перестати вважатись мертвою з юридичної точки зору після перетину державного кордону з Флоридою, тому що там мертвою вважається лише та людина, смерть якої засвідчили два дипломовані лікарі[7].

Ці приклади свідчать, що статистичні дані певною мірою завжди базуються на судженнях, і було б очевидною оманою думати, що особистий досвід у всій його складності можна закодувати без жодних двозначностей та оформити у вигляді електронних таблиць чи комп'ютерних програм. Однак, незважаючи на складності з визначенням, підрахуванням та вимірюванням наших власних характеристик та характеристик нашого оточення, вони все ж таки залишаються інформацією та єдиною вихідною точкою до реального розуміння світу.

Як джерело подібних знань дані мають два основних обмеження. По-перше, вони завжди є недосконалим мірилом того, що нас по-справжньому цікавить: запитуючи про те, наскільки щасливими люди почувалися минулого тижня за шкалою від нуля до десяти, навряд чи можна дізнатися про емоційний стан всієї нації. По-друге, все, що б ми не вирішили вимірювати, відрізнятиметься у різних людей залежно від місця та часу, і здобути корисну інформацію

* Якби я був пересічною людиною, то принаймні мав би на що з нетерпінням очікувати.

з усієї цієї очевидно випадкової **мінливості** може бути досить проблематично.

Упродовж століть статистика намагалась подолати ці дві проблеми та відігравала провідну роль у спробах науковців пізнати світ. Вона надала основу для інтерпретації даних — завжди недосконалих, — щоб відрізнити важливі взаємозв'язки від індивідуальних відмінностей, які роблять нас усіх унікальними. Але світ постійно змінюється, виникають нові запитання, стають доступними нові джерела даних, і статистика також була змушена змінитись.

Люди рахували та вимірювали споконвіку, але сучасна статистика як дисципліна виникла в 1650-х роках, коли, як ми дізнаємося з розд. 8, Блез Паскаль та П'єр де Ферма вперше представили поняття ймовірності. З огляду на появу такого міцного математичного підґрунтя під час роботи з мінливістю прогрес значно пришвидшився. В поєднанні з даними про вік смерті людей теорія ймовірності дала змогу розраховувати пенсії та щорічні виплати. В астрономії відбулася революція, коли вчені усвідомили, що теорія ймовірності може впоратись із розбіжністю результатів вимірювань. Ентузіасти часів Вікторіанської епохи одержимо збирали дані про людське тіло (та все інше) і встановили міцний зв'язок між статистичним аналізом і генетикою, біологією та медициною. Потім у XX столітті статистика стала більш математичною і, на жаль для багатьох студентів та практиків, почала асоціюватись із механізованим застосуванням ряду статистичних інструментів, багато з яких названі на честь ексцентричних статистиків, любителів суперечок, з якими ми познайомимося далі у цій книзі.

Таке поширене сприйняття статистики як базового «набору інструментів» тепер зіткнулося з великими труднощами. По-перше, ми живемо в епоху **науки про дані** — великі та складні набори даних, зібраних із таких повсякденних джерел, як монітори дорожнього руху, публікації в соціальних мережах та інтернет-покупки, формують базу для технологічних інновацій, як-от оптимізація транспортних маршрутів, адресна реклама чи системи рекомендації

покупок. Ми розглянемо **алгоритми**, що базуються на **великих масивах даних**, у розд. 6. Щоб стати фахівцем з обробки даних, потрібно не тільки вивчати статистику та володіти навичками управління даними, програмування й розробки алгоритмів, але й добре розбиратися в цій галузі.

Ще одна проблема, пов'язана з традиційним уявленням про статистику, викликана неймовірним зростанням кількості наукових досліджень, зокрема в галузях біомедичних та соціальних наук, в поєднанні з вимогою публікацій у високорейтингових журналах. Це призвело до сумнівів у достовірності значної частини наукової літератури і заяв інших дослідників про неможливість відтворити багато так званих відкриттів. Як приклад можна навести нескінченні суперечки про те, чи здатна так звана владна поза, якою людина транслює впевненість у собі, спровокувати гормональні та інші зміни[8]. Неналежне використання стандартних статистичних методів значною мірою винне у виникненні явища, відомого в науці як криза відтворюваності, чи криза реплікації.

На тлі зростання доступу до масивів даних та зручного у використанні програмного забезпечення для їх аналізу може здатися, що вже немає потреби у вивченні методів статистики. Ця думка вкрай наївна. Через збільшення кількості даних і зростання числа та складності наукових досліджень зробити належні висновки стає дедалі складніше. Більше даних означає, що ми маємо ще краще усвідомлювати, чого насправді варті такі докази.

Наприклад, ретельний аналіз масивів даних, отриманих із поточних даних, може підвищити ймовірність помилкових відкриттів. Це пов'язано з тим, що джерелам даних притаманні систематичні помилки, а також із тим, що з безлічі виконаних аналізів звітують тільки про ті, які здаються найцікавішими, — практика, відома під назвою «сліпе прочісування даних». Для того, щоб мати змогу критично оцінити опубліковану наукову працю, не кажучи вже про публікації у ЗМІ, з якими ми стикаємось на повсякденній основі, ми повинні чітко усвідомлювати небезпеку вибіркової звітності, необхідність

відтворення наукових даних незалежними дослідниками та ризик хибної інтерпретації результатів єдиного дослідження, висмикнутих із контексту.

Усі ці висновки можна об'єднати під терміном «**грамотність використання даних**», який описує здатність не лише проводити статистичний аналіз проблем реального світу, але й розуміти й критично оцінювати будь-які висновки, зроблені іншими на основі статистики. Але для підвищення цієї грамотності необхідно змінити спосіб навчання статистики.

Навчання статистики

Покоління студентів були змушені проходити сухі курси зі статистики, на яких вони вивчали набір технік для застосування в різних ситуаціях. На цих курсах більше уваги приділялось математичній теорії, ніж розумінню, чому саме використовуються конкретні формули, та які складнощі виникають під час спроби використати дані, щоб отримати відповідь на запитання.

На щастя, все змінюється. Наука про дані та грамотність їх використання вимагає підходу, орієнтованого на розв'язання проблем, коли застосування специфічних статистичних інструментів вважається лише одним із компонентів повного циклу досліджень. Як модель розв'язання проблем був запропонований цикл **PPDAC** (Problem, Plan, Data, Analysis, Conclusion), який ми будемо використовувати в цій книзі[9]. Рис. 0.3 базується на прикладі з Нової Зеландії — країни, що є світовим лідером з викладання статистики у школах.

Перша стадія циклу — це визначення Проблеми; статистичний запит завжди починається із запитання, на кшталт нашого запитання про закономірність вбивств Гарольда Шіпмана чи кількість дерев у світі. Далі в цій книзі ми зосередимось на найрізноманітніших проблемах, починаючи з очікуваних переваг різних методів терапії відразу після хірургічного лікування раку грудей і закінчуючи питанням, чому в літніх людей великі вуха.

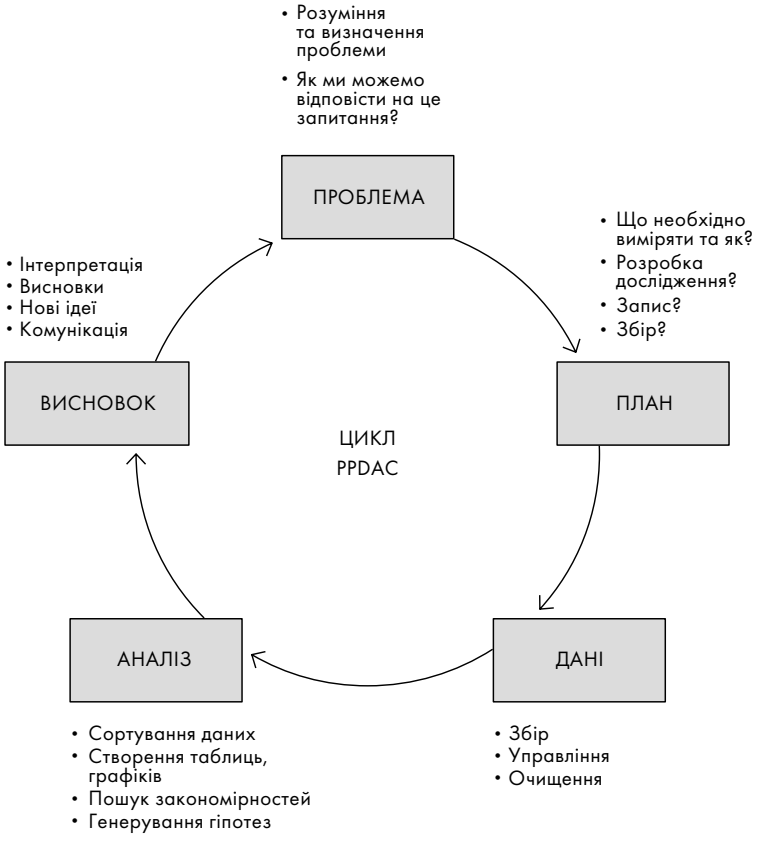

Рис. 0.3. Цикл розв'язання проблем PPDAC — від проблеми, плану, даних, аналізу до висновків та комунікації, і знову початок нового циклу

Не слід піддаватися спокусі знехтувати розробкою ретельного Плану. В історії з Шіпманом необхідно було просто зібрати якомога більше даних про його жертви. Але люди, що рахували дерева, приділяли пильну увагу точним визначенням і методам вимірювання, оскільки впевнені висновки можна зробити тільки на основі належно спланованого дослідження. На жаль, поспішаючи отримати дані та розпочати аналіз, про це часто забувають.

Збір хороших Даних вимагає організаційних навичок і навичок кодування. Значення цих навичок у галузі науки про дані все більше зростає, зокрема тому, що дані зі звичайних джерел часом потрібно ретельно очистити, щоб підготувати їх до аналізування. Системи збору даних з часом можуть змінюватись, там можуть з'являтися очевидні помилки і так далі — сам вислів «зібрані дані» чітко вказує на те, що ці дані можуть бути досить брудними, як щось, підібране на вулиці.

Етапу Аналізу на курсах статистики традиційно приділяється найбільше уваги, і в цій книзі ми розглянемо цілу низку аналітичних прийомів, хоча іноді для аналізу потрібна лише зручна візуалізація, як на рис. 0.1. Нарешті, запорукою хорошої статистики є здатність робити відповідні Висновки, які повною мірою визнають обмеження в доказах, а також чітко їх формулюють, як на графічних ілюстраціях з даними про Шіпмана. Будь-який висновок зазвичай викликає ще більше запитань, тож цикл знову починається від самого початку, як у ситуації, коли ми почали досліджувати час смерті пацієнтів Шіпмана.

Хоча на практиці цикл PPDAC, зображений на рис. 0.3, може не виконуватись з абсолютною точністю, він підкреслює, що офіційні методи статистичного аналізу є лише частиною роботи статистика або спеціаліста з обробки та аналізу даних. Статистика — це набагато більше, ніж галузь математики із заплутаними формулами, над якими (часто проти своєї волі) б'ються покоління студентів.

Про цю книгу

Коли я був британським студентом у 1970-х роках, у нас було всього три телевізійних канали, комп'ютери були розміром із подвійну шафу, а найбільш схожим на Вікіпедію був вигаданий портативний пристрій (дивовижно пророчий) з кінострічки Дугласа Адамса «Автостопом по галактиці». Тому для самовдосконалення ми зверталися до книг видавництва Pelican, і їхні культові блакитні корінці були звичайним явищем на полиці кожного студента.

Оскільки я вивчав статистику, в моїй добірці видань від Pelican були «Факти з цифр» М. Дж. Мороні («Facts from Figures», 1951) та «Як брехати за допомогою статистики» Даррела Хаффа («How to Lie with Statistics», 1954). Ці авторитетні видання були випущені накладом у сотні тисяч примірників, що свідчило як про рівень зацікавленості статистикою, так і про гнітючу відсутність вибору на той час. Ці класики навдивовижу добре протрималися впродовж шістдесяти п'яти років, проте сучасна епоха вимагає іншого підходу до навчання статистики, який би базувався на викладених вище принципах.

Тому в цій книзі розв'язання проблем є відправною точкою для впровадження ідей статистики. Деякі з цих ідей можуть видатись очевидними, але інші — більш делікатні й вимагають певних розумових зусиль, хоча математичні навички вам не знадобляться. На відміну від традиційних текстів у цій книзі більше уваги приділяється концептуальним питанням, а не технічним деталям і міститься лише кілька досить безневинних рівнянь та глосарій. Програмне забезпечення — це важлива частина будь-якої роботи, що стосується науки про дані та статистики, але в цій книзі ми на ньому не фокусуємось. Ви й самі можете легко знайти навчальні матеріали з таких мов програмування, як R та Python.

На питання у текстових вставках певною мірою також можна відповісти за допомогою статистичного аналізу, хоча вони мають зовсім різну проблематику. Деякі з них — це важливі наукові гіпотези (на-

приклад, чи існує бозон Хіггса, або чи насправді є переконливі докази екстрасенсорних здібностей). Інші — це питання, що стосуються охорони здоров'я, наприклад, чи правда, що у більш завантажених лікарнях вищі показники виживання, та чи є користь зі скринінгів на рак яєчників. Часом ми просто хочемо оцінити величини, такі як ризик захворіти раком через споживання сандвічів з беконом, кількість статевих партнерів у британців упродовж життя та користь щоденного споживання статинів.

А деякі питання просто цікаві, наприклад: кому із вцілілих на «Титаніку» пощастило найбільше, чи можна було спіймати Гарольда Шіпмена раніше та яка ймовірність того, що скелет, який було знайдено на автомобільній стоянці в місті Лестер, насправді належить Річарду III.

Ця книга стане в пригоді як студентам, що вивчають статистику і хочуть ознайомитись із предметом, не занурюючись у технічні деталі, так і звичайним читачам, яким цікаво більше дізнатися про статистику, з якою вони стикаються і на роботі, і в повсякденному житті. Я наголошую на необхідності використовувати статистику вправно та обережно: числа можуть здаватись сухими, суворими фактами, але спроби виміряти дерева, щастя та смерть уже показали, що з ними треба поводитись з великою делікатністю.

Статистика може принести ясність та розуміння в проблеми, з якими ми стикаємося, але ми всі знаємо, як нею можна зловживати, часто для просування потрібної думки або просто щоб привернути увагу. Вміння оцінити достовірність статистичних тверджень залишається ключовою навичкою в сучасному світі, і я сподіваюсь, що ця книга зможе допомогти людям поставити під сумнів цифри, з якими вони стикаються в своєму повсякденному житті.