Розділ 10

Статистичні методи обробки інформації

Математична статистика — це розділ математики, присвячений методам збирання, аналізу й обробки статистичних даних для нукових і практичних цілей.

Сучасна математична статистика поділяється на описову та аналітичну. Описова статистика охоплює методи збирання статистичних даних, представлення їх у формі таблиць, розподілів, графіків тощо. Аналітична статистика називається також теорією статистичних висновків, які мають прикладне значення для медичної практики.

Основою наукового медичного дослідження є спостереження, в результаті якого дослідик робить вимірювання, одержуючи кількісні величини, що характеризують ту чи іншу ознаку. Це первинна медична статистична інформація.

Розрізняють два види спостережень. При спостереженнях першого типу робиться N вимірювань досліджуваної ознаки у даного об'єкта, сигналу чи процесу (наприклад, вимірюють концентрацію цукру в крові в одного пацієнта новим методом). При спостереженнях другого типу проводяться одиничні

вимірювання досліджуваної озаки у кожного з n однорідних об'єктів (наприклад, вимірюють концентрацію цукру в крові у n пацієнтів).

Можна довести, що ці типи спостережень принципово різні. Перший тип охоплює тривалий процес і застосовується у випадках, коли необхідно простежити, як досліджувана ознака змінюється з часом. При проведенні спостережень другого типу потрібно, щоб умови, в яких відбуваються спостереження різних об'єктів, залишилися незмінними від спостереження до спостереження, від об'єкта до об'єкта.

Експериментальні дані в галузі біотехнічної науки, як правило, являють собою результати вимірювання деяких параметрів (наприклад, значення мінімальної та максимальної амплітуди пульсового сигналу, температури, тиску і т.д.). Отримані значення випадкової величини — це проста статистична сукупність, або простий статистичний ряд, що підлягає обробці й науковому аналізу.

10.1 Генеральна та вибіркова сукупності

Множина об'єктів дослідження, що об'єднані загальними, суттєвими для цього дослідження властивостями (ознаками) називається *сукупністю*.

Обсяг сукупності (n) — це кількість об'єктів (елементів) сукупності.

Генеральна сукупність — найбільша сукупність обсягом n, яка об'єднує всі об'єкти дослідження із загальними, суттєвими для цього дослідження ознаками. Звичайно, краще провести дослідження для всієї генеральної сукупності, однак здебільшого це зробити неможливо через надто велику кількість об'єктів або їх недоступність. Тому з генеральної суку-

пності обирають для вивчення частину об'єктів, які утворюють *вибірку*.

Вибірка (або *вибіркова сукупність*) — це група елементів, яка вибрана для дослідження з усієї генеральної сукупності елементів. Вибірка повинна мати ті самі загальні суттєві для дослідження ознаки, що й генеральна сукупність.

Для того, щоб властивості вибірки досить добре відбивали властивості генеральної сукупності, вибірка має бути репрезентативною. Це можливо у тому випадку, коли вибірка формується випадковим чином. За одними ознаками елементи вибірки можуть збігатися разом (наприклад, стать, вік), значення інших ознак змінюються від одного до іншого (наприклад, вага, зріст). Предметом вивчення у статистиці є саме ті ознаки, що змінюються. Вони поділяються на якісні та кількісні.

Якісні ознаки не піддаються безпосередній кількісній оцінці, однак мають ряд якісних градацій, що дозволяє порівнювати між собою окремі об'єкти за ступненем виразності даної ознаки (наприклад, інтенсивність болю, відсоток шуму в сигналі тощо).

Кількісні ознаки являють собою результати підрахунку або вимірювання. Вони поділяються на *безперервні* та *дискретні*. Безперервні величини можуть набувати будь-яких значень із деякого інтервалу (наприклад, вага людини, температура тіла тощо). Дискретні величини можуть набувати лише визначених значень, які можна пронумерувати (наприклад, кількість хворих, що пройшли обстеження, протягом кожної години).

В результаті досліджень дослідник отримує числові значення (*варіанти*), які відрізняються між собою. Часто за такими первинними даними складно зробити однозначні висновки, тому вони потребують обробки, яка починається з їх групування.

Групування — це процес систематизації, упорядкування первинних даних з метою отримання інформації, що містяться в них. Ряд варіант, що розташовані у порядку зростання числових значень, називається *варіаційним рядом*. Якщо кіль-

кість варіант велика, то варіаційний ряд розбивають на рівні інтервали. Перше завдання при групуванні варіаційного ряду полягає в тому, щоб розбити весь діапазон зміни ознаки у виборці (між максимальними і мінімальними варіантами вибірки) на інтервали. Це потребує визначення кількості інтервалів групування і ширини кожного із них. Зазвичай обирають інтервали однакової ширини. Групування роблять для того, щоб побудувати емпіричний розподіл і сформулювати за його допомогою припущення про форму розподілу досліджуваної ознаки у генеральній сукупності, з якої взята вибірка.

10.2 Характеристики вибірки

Питання про вибір кількості та ширини інтервалів групування вирішують у кожному конкретному випадку, виходячи із цілей дослідження, обсягу вибірки і ступеня варіювання ознаки у вибірці. Однак приблизно кількість інтервалів k можна оцінити тільки з обсягу вибірки n. Кількість інтервалів можна визначити за формулою Стерджеса:

$$k = 1 + 3,32 \lg(1,37n),$$

або за допомогою таблиці

Обсяг вибірки, n	, **Кількість інтервалів,** k
$25 \ldots 40$	$5 \ldots 6$
$40 \ldots 60$	$6 \ldots 8$
$60 \ldots 100$	$7 \ldots 10$
$100 \ldots 200$	$6 \ldots 12$
>200	$10 \ldots 15$

Коли кількість інтервалів обрано, то ширина кожного з них визначається за наступною формулою:

$$h = \frac{x_{max} - x_{min}}{k},$$

де x_{max} та x_{min} — відповідно максимальна і мінімальна варіанти вибірки. Зазначену різницю називають *розмахом варіації*:

$$R = x_{max} - x_{min}.$$

Але інформативність цього показника невелика, оскільки можна навести багато прикладів розподілів, які значно відрізняються а формою, але мають однаковий розмах. Тому він здебільшого використовується при малих (не більше 10) обсягах вибірки.

Частоти інтервалів n_i — частоти того, наскільки часто у вибірці зустрічаються варіанти, які належать до кожного інтервалу групування. Загальна кількість частот завжди дорівнює обсягу вибірки n.

Вибіркове середнє значення ($\overline{X}$) — центр вибірки, біля якого групуються елементи вибірки:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

де n — кількість спостережень; x_i — значення величини (варіанти), що досліджується.

Знаючи середнє арифметичне значення даних експерименту, виникає питання: як обчислити середню величину, на яку вірізняються дані від середнього арифметичного? Різницю між будь-яким вимірюванням у вибірці та середнім арифметичним цієї ж вибірки називають *відхиленням варіанти* або ж *вибірковим середнім* x_i від $\overline{X}$: $x_i - \overline{X}$. Якщо обчислити відхилення для усіх варіант, то серед отриманих значень будуть від'ємні та

додатні, які у сумі даватимуть 0, тобто взаємно компенсуються. Для того, щоб уникнути компенсації додатніх і від'ємних значень, існує декілька способів. Найпоширеніший — піднесення кожної різниці $x_i - \overline{X}$ до квадрату (квадрати як від'ємних, так і додатніх величин є величинами додатними). Додаючи квадрати усіх різниць та ділячи на кількість цих різниць, отримаємо величину, яка називається ***дисперсією*** (позначається D). Фактично вона показує середнє арифметичне квадратів відхилень. Для того, щоб позбутися квадрату величини, обчислюємо корінь квадратний з дисперсії. Отримане значення називають ***середнім квадратичним відхиленням*** (позначається σ).

Вибіркове середнє квадратичне відхилення (σ) — це ступінь відхилення елементів вибірки щодо середнього значення. Чим більше значення середнього квадратичного відхилення, тим далі відхиляються значення елементів вибірки від середнього значення:

$$\sigma = \sqrt{D} = \sqrt{\frac{\sum\limits_{i=1}^{n} \left(x_i - \overline{X}\right)^2}{n-1}}.$$

Знайдемо ширину довірчого інтервалу, в який з довірчою ймовірністю α потраляє справжнє значення (тобто математичне очікування μ генеральної сукупності).

Стандартна похибка (m_X) показує, наскільки значення вибіркової середньої близьке до середнього значення генеральної сукупності.

Стандартне відхилення виражається у таких одиницях, у яких і вимірювана ознака. Якщо потрібно порівняти між собою ступінь варіювання ознак, виражених у різних одиницях виміру, використовують коефіцієнт варіації V, що є відношенням середнього квадратичного відхиленя до математичного очікування (виражається у відсотках):

$$V = \frac{\sigma}{\overline{X}} \cdot 100\%.$$

Надійний інтервал дозволяє визначити межі, в яких з тією чи іншою імовірністю можуть знаходитися істинні значення величини, яка досліджується. Як надійні використовуються такі значення ймовірностей: $\alpha_1 = 0,95$; $\alpha_2 = 0,99$; $\alpha_3 = 0,999$. У медицині, у разі особливо відповідальних експериментів, вибирають $\alpha_3 = 0,999$, в інших випадках — $\alpha_1 = 0,95$.

Точність (*надійність межі помилки*) прямого вимірювання визначається формулою

$$\delta = \pm \left| t \cdot m_X \right|,$$

де t — коефіцієнт нормованого відхилення (*критерій Стьюдента*), що залежить від кількості ступенів свободи і вибраної надійної імовірності α. Надійний інтервал визначається за формулою

$$\overline{X} - \delta \leq X \leq \overline{X} + \delta.$$

10.3 Виявлення вірогідності відмінності середніх значень двох вибірок

Випадкова подія — подія, яка може трапитися чи не трапитися без певної закономірності для цього.

Випадкова величина — величина, яка набуває різних значень без певної закономірності, тобто випадково.

Ймовірність (p) — це параметр, який характеризує частоту випадкової події. Ймовірність змінюється від 0 до 1. Випадок $p = 0$ означає, що випадкова подія ніколи не трапиться, випадок $p = 1$ означає, що випадкова подія трапляється завжди.

Незалежними називаються події, коли настання однієї з них не змінює імовірність настання іншої. У протилежному випадку події називаються *залежними*.

Задача виявлення вірогідності відмінностей середніх арифметичних значень двох незалежних вибірок нерідко трапляє-

ться в медичній практиці. Використовуючи цей метод, можна встановити, чи різниця двох незалежних вибірок спричинена випадковим фактором, чи вона зумовлена якимись зовнішніми чинниками. Так, наприклад, порівнюючи середні значення частоти серцевих скорочень контрольної групи хворих ($\overline{X} = 145,7$) та групи, що досліджується ($\overline{X} = 125,6$), можна бачити, що вони відрізняються. Мета цього методу полягає у вирішенні проблеми: чи можна за цими даними зробити висновок про більшу ефективність нового технічного засобу реєстрації чи обробки сигналу або впливу нового препарату?

Для розв'язання задач такого типу використовують *критерій відмінності*, або *t-критерій Стьюдента*. Критерій Стьюдента найбільш часто використовується для перевірки гіпотези: «Середні двох вибірок належать до однієї і тієї самої сукупності». Критерій дозволяє знайти імовірність того, що ці середні відносяться до однієї сукупності. Якщо ця імовірність p нижче рівня значущості ($p < 0,05$), тоді треба вважати, що вибірки відносять до двох різних сукупностей.

Рівень значущості — це максимальне значення імовірності виникнення події, при якому подія вважається практично неможливою. У медицині найбільш поширений рівень значущості $p = 0,05$. Тому, якщо імовірність, з якою подія може трапитися випадковим чином $p < 0,05$, то треба вважати, що ця подія малоймовірна, і якщо вона все таки трапилася, то це не було випадково.

При використанні t-критерію виділяють два випадки:

- для перевірки гіпотези про однаковість генеральних середніх двох незалежних, непов'язаних вибірок. Так, наприклад, є контрольна група пацієнтів та група, що досліджується. Ці групи складаються з різних пацієнтів, кількість яких може бути різною. У цьому разі використовується двовибірковий t-критерій;

- для перевірки гіпотези про однаковість генеральних середніх двох залежних, пов'язаних вибірок. Так, наприклад, вимірюється вязкість крові звичайним, лабораторним чином, та в'язкість крові у тих самих людей непрямим методом, тобто виходячи із математичної залежності іншої величини. Тобто, коли одна і та сама група об'єктів породжує чисельний матеріал. У цьому разі використовується парний t-критерій.

Для використання обох цих критеріїв, ознака, що досліджується в кожній із груп, повинна мати нормальний закон розподілу. У разі, коли розподіл спостережень має складний, невідомий закон розподілу, відмінний від нормального закону, використовують *непараметричні методи статистики.*

10.4 Виявлення взаємозв'язку двох випадкових величин

Важливим завданням статистичної обробки даних медичних досліджень є також виявлення взаємозв'язку між вибірками. Для оцінки ступеня взаємозв'язку використовують коефіцієнт кореляції.

Коефіцієнт кореляції (r) — це параметр, що характеризує степінь лінійного взаємозв'язку між двома вибірками. Коефіцієнт кореляції змінюється від -1 (сувора обернена лінійна залежність) до +1 (сувора пряма пропорційна залежність). При значенні 0 лінійної залежності між двома вибірками не існує. На практиці коефіцієнт кореляції набуває деякого проміжне значення. Оцінюють глибину кореляційного зв'язку між величинами, виходячи з таких критеріїв:

- $0,0 < r < 0,4$ — лінійного взаємозв'язку між параметрами виявити не вдалося;
- $0,3 < r < 0,6$ — зв'язок між параметрами помірний;

- $0,6 < r < 0,8$ — присутній лінійний зв'язок між параметрами;
- $0,8 < r < 0,95$ — зв'язок між параметрами сильний;
- $0,95 < r < 1,0$ — зв'язок між параметрами дуже сильний.

Кореляційний аналіз дозволяє отримати кореляційну матрицю, яка містить коефіцієнти кореляції між різними параметрами.

10.5 Регресійний та дисперсійний аналізи даних результатів досліджень

Змінна — будь-яка величина, що варіюється. *Незалежна змінна* — змінна, варіювання якої трапляється незалежно від інших величин. *Залежна змінна* — величина, що змінюється при зміні однієї чи більшого числа незалежних змінних.

Регресійний аналіз — це метод визначення функції $Y = f(X)$, крива якої найкраще апроксимує (характеризує напрямок розміщення) серію експериментальних точок. В основу цього методу покладено вимогу найбільшої відповідності шуканого рівняння взаємозв'язку ознак X та Y, тобто функції $Y = f(X)$, графік якої найкраще буде наближатися до точок емпіричної кривої, що побудована за даними досліду.

Регресія використовується для аналізу впливу на окрему незалежну змінну значень однієї чи більше незалежних змінних. Так, наприклад, на ступінь захворюваності людини впливає декілька факторів: вік, вага та імунний статус. Регресія пропорційно розподіляє міру захворюваності між цими факторами на підставі даних захворювання, що досліджується.

Експериментальні дані апроксимуються лінійним рівнянням до 16-го порядку:

$$Y = a_0 + a_1X_1 + a_2X_2 + \ldots a_{16}X_{16}$$

де Y — залежна змінна;
$X_1, \ldots, X_{16}$ — незалежні змінні;
$a_0, a_1, \ldots, a_{16}$ - шукані коефіцієнти регресії.

10.6 Сучасні програми медичної статистики для обробки даних досліджень

У медичній статистиці сучасні інформаційні технології застосовуються на кожному етапі розробки і проведення спостережень, а саме: при розробці форм, формуванні плану вибірки, збору і введенні даних, їх обробці та аналізі, а також наданні інформації користувачеві.

Програми комп'ютерної обробки статистичних даних поділяють на професійні, напівпрофесійні і спеціалізовані. Професійні пакети володіють значною кількістю методів аналізу даних, напівпрофесійні — мають універсальні функції, а спеціалізовані пакети орієнтуються лише на вузьку область аналізу.

Найбільш популярним додатком для роботи зі статистичними даними є **MS Excel**. Це табличний процесор з математичними можливостями та статистичними функціями. Цей додаток впорається із задачею накопичення даних, виконанням проміжних обчислень та побудовою нескладних діаграм. Однак він не має засобів для побудови якісних наукових графіків. Тому краще статистичний аналіз даних виконувати в програмах, що призначені саме для таких цілей. Наприклад, можна скористатися макрос-додатком XLSTAT-Pro для MS Excel який, у який вбудовано більше 50 статистичних функцій.

STADIA. Це вітчизняний додаток. Він включає в себе усі необхідні функції для роботи та аналізу статистичних даних. Проте функціональні можливості програми практично не змі-

нилися з 1996 року, а тому графіки та діаграми, побудовані з допомогою додатку, виглядають архаїчно. Колірне співвідношення (червоний шрифт на зеленому фоні) втомлює при тривалій роботі.

SPSS. Використовується найчастіше для статистичної обробки даних. Відрізняється гнучкістю та потужністю. Додаток може бути використаним для різних видів статистичних розрахунків у біомедицині. Є у наявності русифікована версія SPSS 12.0.2 для Windows. Також 2002 року Київським видавництвом «Діасофт» було видано підручник про SPSS під назвою «SPSS 10: Мистецтво обробки інформації. Аналіз статистичних даних і відновлення прихованих закономірностей».

STATA. Професійний статистичний програмний пакет, що може бути використай у біомедичних цілях. Є одним із найпопулярніших додатків серед освітніх та наукових установ США. Для користувачів системи видається спеціальний журнал. Недоліком додатку є те, що немає можливості використання демо-версії.

STATISTICA. Виробником програми є фірма StatSoft Inc. (США), котра працює на ринку статистичних додатків починаючи з 1985 року. STATISTICA вміщує у собі значну кількість методів статистичного аналізу (більш ніж 250 функцій), що об'єднані спеціалізованими статистичними модулями. Даний додаток не є складним в освоєнні, а тому може бути рекомендований для різних біомедичних досліджень. На сьогоднішній день випущена сьома версія. Також пропонується повністю русифікована 6-а версія програми. Сам пакет STATISTICA описаний в декількох книгах, одна з яких, для медичних працівників: О.Ю. Реброва «Статистичний аналіз медичних даних. Застосування пакета прикладних програм STATISTICA.»

JMR. Додаток лідирує на ринку обробки та аналізу статистичних даних. Реалізує цей додаток SAS Institute. Однак

особливих переваг для медико-біологічної статистики цей програмний продукт не має.

NCSS. Програма вийшла на ринок 1981 року та розрахована на непрофесіоналів в області статистичної обробки. Інтерфейс системи дещо незвичний у використанні, однак усі дії супроводжуються підказками. Доступна також демо-версія NCSS 11.

SYSTAT. Зазначений додаток призначений для персональних комп'ютерів. Компанія Systat Software має у доробку досить популярні пакети SigmaStat і SigmaPlot, які є відповідно, програмою побудови діаграм (SP) та програмою статистичної обробки (SS). Можна використовувати у комплексі, що дозволяє не лише статистичну обробку а і візуалізацію даних.

STATGRAPHICS PLUS. Ця статистична програма є досить потужною, адже містить у собі більше 250 статистичних функцій. Остання доступна версія - 5.1. Є можливість ознайомлення з допомогою демо-версії. Додаток є досить популярним у вітчизняних дослідників.

MINITAB 14. Є у наявності демо-версія програми, яка працює 30 днів. Даний програмний пакет досить зручний у роботі, має гарний інтерфейс, та реалізує можливості візуалізації результатів роботи.

PRISM. Додаток створений спеціально для біомедичних цілей. Має зрозумілий інтерфейс, що дозволяє швидко проаналізувати дані та побудувати якісні графіки. Додаток включає основні статистичні функції. Однак програма не може повністю замінити серйозні статистичні пакети.