 © ДУ ІГМЕ

УДК 004.421:504.054

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ЕКСПРЕС-ОБРОБКИ ЕКОЛОГІЧНИХ ДАНИХ

Пашинська С.Л.[1], Антомонов М.Ю.[1], Гущук І.В.[2]
[1] ДУ "Інститут гігієни та медичної екології ім. О.М. Марзєєва НАМН України", м. Київ
[2] Рівненська Обл.СЕС

Актуальність проблеми. Для визначення та аналізу існуючої екологічної ситуації необхідно зібрати значний об'єм інформації з метою виявлення біологічних, хімічних, фізичних, соціальних та інших факторів середовища, що впливають на процеси життєдіяльності населення. У процесі збору даних необхідно періодично відстежувати якість зібраних матеріалів, контролювати дотримання прийнятих на організаційному етапі правил і принципів. Все це дозволяє правильно зібрати доброякісний статистичний матеріал. Зібраний у процесі спостереження матеріал потребує певної обробки різноманітними чисельними та статистичними методами, що передбачає оснащеність відповідною технікою та прикладним програмним забезпеченням. Проте не в усіх районах, у яких здійснюється попередній збір інформації, є можливість придбання спеціалізованого програмного забезпечення та організації його супроводу. У той же час, без певної автоматизації процесу обробки значного об'єму даних на даному етапі не уникнути похибок або й часткової втрати інформації, а також невиправданих затрат часу й зусиль.

Метою нашої роботи є розробка технології експрес-обробки статистичних даних на базі найбільш доступних і простих у використанні програмних засобів.

Методи дослідження. В даний час статистичну обробку даних проводять переважно за допомогою спеціалізованих програмних пакетів (наприклад, SPSS, «Statistica for Windows»). Однак розуміння суті статистичних методик дозволяє при необхідності автоматизувати цей процес за допомогою більш простих і універсальних засобів. Найбільш зручним у розглядуваній ситуації видається такий інструмент, як табличний процесор Microsoft Excel, за допомогою якого можна вирішити достатньо широкий спектр задач обмеженої складності. Не дивлячись на те, що ця програма не призначена спеціально для статистичних операцій, гнучкість її інтерфейсу, наявність формул і надбудов, а також можливість створення макросів дозволяють здійснити експрес-аналіз статистичних даних. Указаний програмний додаток відносно простий у використанні та має високу ступінь сумісності з іншими додатками. При необхідності таблиці Excel легко експортуються у програмний пакет Statistica для подальшої ускладненої статистичної обробки.

Виклад основного матеріалу. При малій кількості спостережень результати подаються у вигляді таблиці первинних даних. Виявлення закономірностей при цьому, як правило, не потрібно. Однак у більшості випадків для аналізу отриманих даних необхідно попередньо описати отриману статистичну сукупність за допомогою ряду параметрів. Опис кількісної ознаки проводиться за такими етапами: визначення виду розподілу ознаки; оцінка центральної тенденції досліджуваної сукупності; оцінка її розсіювання (розкиду). Розрахунок таких статистик досить простий і в той же час дозволяє зробити про вибірку певні припущення. Тому описові статистики є важливим і часто використовуваним статистичним інструментом.

До найбільш поширених статистичних характеристик, одержаних за результатами аналізу, відносяться стандартне відхилення, розмах варіації, коефіцієнт варіації, середнє значення, медіана, мода, дисперсія, асиметрія, ексцес та ін. Вибір характеристик центральної тенденції і розсіювання насамперед залежить від виду розподілу. У разі нормального розподілу використовують показники параметричної статистики (середнє значення, стандартне відхилення, дисперсію), у разі розподілу, відмінного від нормального або невизначеному, застосовують показники непараметричної статистики (медіану, квантилі, інтерквартильний інтервал). Непараметричні методи не вимагають знання

закону розподілу досліджуваних сукупностей і обчислення їх основних параметрів. Крім того, методи непараметричної статистики прості в застосуванні і при розподілах, близьких до нормального, дають хороший результат. Ще одна перевага непараметричних методів - застосовність до якісних ознак, які виражаються порядковими номерами або індексами. Однак, зазначені методи менш потужні, ніж параметричні і мають істотні обмеження у застосуванні щодо кількості спостережень. Таким чином, перш ніж приступати до опису ознаки, слід визначити його тип і вид розподілу. Від цього залежить вибір статистичного методу його узагальнення.

Відповідність експериментального розподілу до нормального можна перевірити такими найпростішими способами.

1. За числами Вестергарда (0,3; 0,7; 1,1; 3). При нормальному розподілі в межах: x±0,3σ знаходиться 25% всіх одиниць спостереження; x±0,7σ знаходиться 50% всіх одиниць спостереження; x±l,lσ знаходиться 75% всіх одиниць спостереження; x±3,0σ знаходиться 99% всіх одиниць спостереження [5].

2. За правилом «двох третин» Юла (за співвідношенням середньої арифметичної і структурних середніх). При нормальному розподілі, якому притаманна симетричність:

$$\bar{x} = \frac{3\,Me - Mo}{2} \quad \text{або} \quad Mo = 3(Me - \frac{2}{3}\bar{x}),$$

де $\bar{x}$ – середнє значення, а Me – медіана, Mo – мода.

3. За коефіцієнтом асиметрії:

$$As = \frac{x - Mo}{\sigma}; \quad -3 \le As \le +3.$$

3. За медіаною. Якщо Me займає серединне положення між 25-м і 75-м процентилем, то розподіл близький до нормального.

4. За обчисленням спеціальних критеріїв узгодження та ін.

При n>100 непараметричні методи використовувати не доцільно, навіть якщо не виконується вимога нормальності.

Так як значна частина статистичних методів (параметрична статистика) заснована на припущенні, що розподіл близький до нормального, то, якщо експериментальні дані не лягають на криву нормального розподілу, їх намагаються перетворити таким чином, щоб отримана крива відповідала нормальному розподілу. Найбільш часто використовуються такі способи «нормалізації перетворення» даних: гармонійне перетворення; добування квадратного кореня; логарифмування (дає найбільш точне наближення). Успішність перетворення даних оцінюють за коефіцієнтом асиметрії: чим ближче він до 0, тим ближче експериментальний розподіл до нормального.

Використання формальних методів обробки даних без належного логічного аналізу зібраного матеріалу і отриманих результатів може призвести до неправильних висновків. Аналіз результатів проводиться з використанням сучасних математико-статистичних методів за умови їх відповідності природі досліджуваного явища. Логічний аналіз у першому наближенні дасть уявлення про надлишкову розмірність масиву. Дослідник при візуальному огляді може виявити стовпці з відсутніми даними, або ж з усіма однаковими значеннями. Таблиці можуть містити стовпці з проміжними змінними, що не будуть використовуватися в подальшій обробці, або ж допоміжними значеннями, потрібними лише для обчислення інших характеристик. Таким чином, значна частина показників може бути вилучена уже на цьому етапі. Розрахувавши максимальне та мінімальне значення можна на цьому етапі візуально виявити сумнівні варіанти з подальшою перевіркою їх на аномальність.

Результати спостережень, що викликають сумнів своєю відмінністю від інших у більшу чи в меншу сторону, необхідно перевірити на наявність грубої похибки. Як правило, такі значення вибраковують. Проте в окремих випадках аномальні значення можуть нести інформацію про процес, і їх не

слід автоматично відкидати. Такі сигнали слід зафіксувати та проаналізувати. Доцільно провести аналіз даних з викидами і без них та порівняти результати. Після вилучення виявлених викидів, характеристики повинні бути перераховані за даними скороченої вибірки. Процес ітераційний і повторюватиметься до вилучення всіх викидів.

Найпростіші непараметричні способи виявлення викидів базуються на визначенні квартилів. Викидами вважаються значення, що не потрапляють у діапазон:

$$[Q_1 - 1.5*IQR;\ Q_3 + 1.5*IQR]\ ,$$

де, Q_1 і Q_3 – 1-й і 3-й квартилі відповідно, IQR – інтерквартильний інтервал.

При n>100 непараметричні методи використовувати не доцільно, навіть якщо не виконується вимога нормальності.

Для результатів спостережень, розподілених за нормальним законом зазвичай використовують параметричний критерій, що базується на обчисленні нормованого відхилення. При кількості вимірів 20<n<50 застосовують правило «трьох сигм» (3σ). При великій кількості спостережень коефіцієнт може змінюватись в залежності від обсягу вибірки [2]: при 50≤n<100 користуються критерієм 4σ; при 100<n<1000 – 4,5σ; при 1000<n<10000 – 5σ. При наявності малої вибірки (n<20) може бути використаний метод, що базується на критерії В.І. Романовського. При об'ємі вибірки n<10 можна скористатись варіаційним критерієм Діксона. За вищевказаними методами розрахунки середнього та стандартного відхилення проводяться без урахування екстремальних значень. Деякі інші методи виявлення викидів (критерії Шовене, Граббса, Ірвіна) включають у розрахунки середнього та стандартного відхилення сумнівні значення.

Вибір критерію залежить від мети вимірювань, але рішення відкинути якісь дані залишається суб'єктивним. Для прийняття рішення пропонується така схема обробки сумнівних значень.

1. Виявити можливі викиди.
2. Визначити причини появи викидів.
3. Виключити викиди, які є грубими промахами і не представляють популяцію спостережень.
4. Прийняти до уваги викиди, які несуть інформацію про процес і проводити подальший аналіз з їх урахуванням.
5. Провести аналіз даних з викидами і без них та порівняти результати.
6. Проаналізувати ситуацію, що спричинила викид.

Експрес-аналіз масиву даних пропонується здійснити за таким алгоритмом (рис. 1).

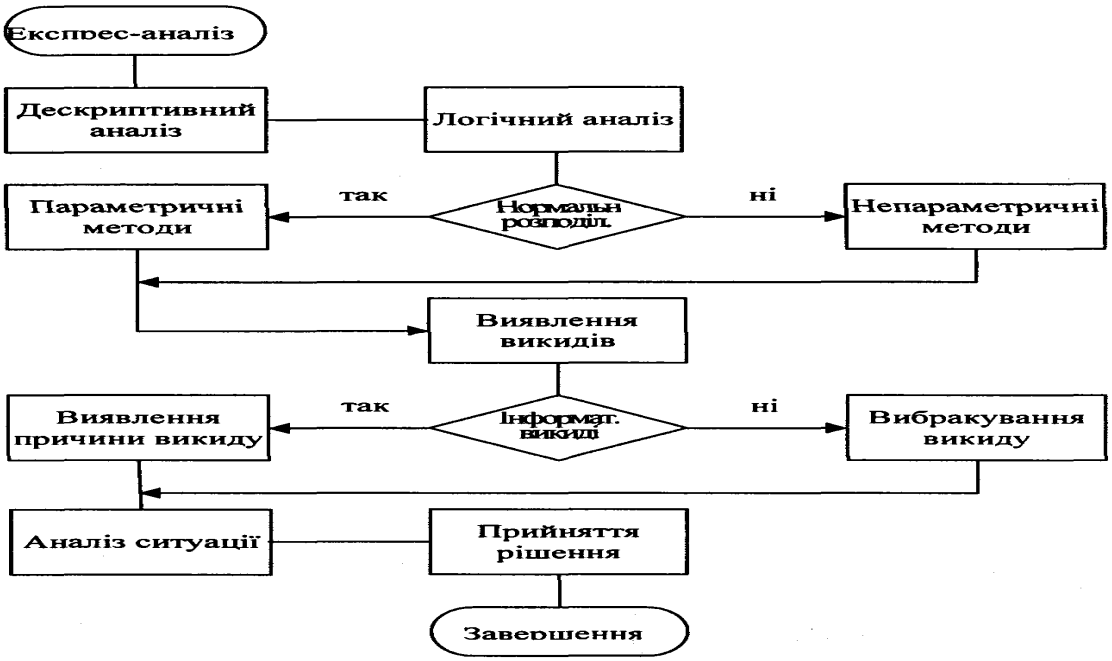

Рисунок 1. Блок-схема алгоритму первинного аналізу даних.

Обробка масиву починається з дескриптивного аналізу (розрахунок стандартного відхилення, розмаху варіації, коефіцієнта варіації, середнього значення, медіана, моди, дисперсії, асиметрії, ексцесу та ін.). Наступний крок – логічний аналіз, який дозволить зробити про даний масив певні припущення і, при необхідності, скорегувати. Перевірка на відповідність експериментального розподілу до нормального дозволить визначитися з методами обробки (параметричні або непараметричні). Надалі слід зробити перевірку сукупності на наявність викидів та їх аналіз. Якщо аномальні значення є грубими помилками, їх необхідно вибракувати. Якщо ці значення є, так званими, «м'якими викидами», вони можуть бути інформативними і сигналізувати про реальні відхилення в екологічній ситуації конкретних населених місць.

Для реалізації алгоритму зручними є, наприклад, офісні додатки, зокрема Microsoft Office Excel із вбудованим засобом VBA.

Висновки

Запропонована технологія дозволяє оперативно здійснити обробку масиву даних спочатку найбільш простими, не трудомісткими методами, а при необхідності підключити більш потужні критерії. Аналіз аномальних значень дає можливість у першому наближенні виявити загрозливі сигнали в екологічній ситуації певних районів для подальшого їх детального обстеження. Розробленні засобами VBA форми та макроси дозволяють реалізувати вищеописану технологію експрес-обробки статистичних даних.

ЛІТЕРАТУРА

1. Excel для экономистов и менеджеров / А.Г. Дубина, С.С. Орлова, И.Ю. Шубина, А.В. Хромов. – СПб : Питер, – 2004. – 295 с.
2. Гублер Е.В. Вычислительные методы анализа и распознавания патологических процессов / Е.В. Гублер. – Л., – 1978. – 294 с.
3. Кулаичев А.П. Методы и средства комплексного анализа данных / А.П. Кулаичев. – М. : Форум–Инфра. – М, – 2006. – 162 с.
4. Лапач С.Н. Статистика в науке и бизнесе / С.Н. Лапач, А.В. Чубенко, П.Н. Бабич. – К. : МОРИОН, – 2002. – 640 с.
5. Метрология, стандартизация и сертификация : рук-во к выполнению лаб. Работ / сост. И.Л. Дудинцева. – М. : РГОТУПС, – 2007. – 12 с.
6. Просветов Г.И. Эконометрика: Задачи и решения: учебно-методическое пособие. – 4-е изд., доп. / Г.И. Просветов – М. : Издательство РДЛ, – 2007. – 192 с.
7. Радкевич Я.М. Метрология, стандартизация и сертификация / Я.М. Радкевич. – М. : Высшая школа, 2004. – 767 с.
8. Сергеев А.Г. Метрология, стандартизация и сертификация: учеб. Пособие Мин-ство образования РФ / А.Г. Сергеев. – М. : ЛОГОС, – 2003. – 406 с.

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ ЭКСПРЕСС-ОБРАБОТКИ ЭКОЛОГИЧЕСКИХ ДАННЫХ

Пашинская С.Л., Антомонов М.Ю., Гущук И.В.

Цель. Разработка технологии экспресс-обработки статистических данных.

Методы. Математические, статистические.

Результаты. Предложена технология оперативной обработки массива данных. Рассмотрена схема обнаружения аномальных значений. Построен алгоритм первичного анализа данных. Разработан пакет макросов для реализации алгоритма.

INFORMATION TECHNOLOGY FOR EXPRESS-PROCESSING OF ECOLOGICAL DATA

S.L. Pashinskaia, M.Yu. Antomonov, I.V. Guschuk

Objective. Development of technology for routine processing of statistical data.

Methods. Mathematical, statistical.

Results. The technology for the routine processing of data massive is proposed. A scheme for the detection of abnormal values was considered. An algorithm of data primary analysis was constructed. A package of macros for the implementation of the algorithm was created.

УДК 613.955:373.51

РИЗИК РОЗВИТКУ ЗАХВОРЮВАНЬ У ДІТЕЙ В ЗАЛЕЖНОСТІ ВІД ОРГАНІЗАЦІЇ ФІЗИЧНОГО ВИХОВАННЯ В ЗАГАЛЬНООСВІТНІХ НАВЧАЛЬНИХ ЗАКЛАДАХ

Гозак С.В., Єлізарова О.Т.
ДУ «Інститут гігієни та медичної екології ім. О.М. Марзєєва НАМН України», м. Київ

Фізична культура є одним з вагомих засобів підтримання і зміцнення здоров'я та працездатності учнів на індивідуальному, колективному і популяційному рівнях [1-3]. Проте тенденція до погіршення здоров'я дитячого населення, в тому числі учнів загальноосвітніх навчальних закладів (ЗНЗ), вимагає з'ясовування ролі фізичного виховання у закладах освіти.

В теперішній час актуальність здоров'язберігаючої спрямованості процесу фізичного виховання в ЗНЗ не викликає сумніву і дослідження у цьому напрямку направлені на детекцію існуючої ситуації з метою визначення напрямків її оптимізації і розробки та впровадження управлінських рішень. Тому **метою** нашого дослідження було визначення впливу умов та організації фізичного виховання на формування порушень опорно-рухового апарату, а також розвитку хронічних захворювань у дітей, що навчаються в міських загальноосвітніх навчальних закладах.

Матеріали і методи. Проведено викопіювання відомостей з медичних карт 52 225 учнів 83 загальноосвітніх навчальних закладів міст Київ, Суми та Львів щодо хронічної захворюваності, порушень постави, плоскостопості та зниження зору. У цих навчальних закладах за єдиною схемою була проведена гігієнічна оцінка умов та організації фізичного виховання за участі лікарів з гігієни дітей і підлітків державної санітарно-епідеміологічної служби. Вибірка є репрезентативною для рівня значимості, що дорівнює 0,05.

Визначали частоту хронічних захворювань, порушень постави, плоскостопості та зниження зору серед дітей обстежених ЗНЗ. Кількісну характеристику впливу факторів «умови фізичного виховання» та «організація фізичного виховання» оцінювали за допомогою розрахунку відносного ризику (RR) та етіологічної фракції (EF) [4].

Систематизація матеріалу і первинна математична обробка були виконані за допомогою таблиць Microsoft EXCEL 2010. Статистична обробка проводилась з використанням пакету STATISTICA 8.0.

Результати. Результати медичних оглядів учнів свідчать про те, що хронічні захворювання у містах має 16,0%±0,5% дітей. Серед учнів 1-4 класів цей показник становить 13,2%±0,9%, 5-9 класів – 16,4%±0,7%, 10-11 класів – 21,8%±1,4%. Така тенденція є однаковою для всіх регіонів спостереження, її можна описати лінійною функцією $y=4,3x+8,53$ ($R^2=0,98$) (рис. 1).

З'ясовано, що вже сформовані порушення постави зустрічаються у 14,4%±0,2% учнів, плоскостопість – у 11,7%±0,2%, порушення зору – у 14,6%±0,2% (табл. 1). Вивчення динаміки показників, що характеризують виявлену патологію, також свідчить про достовірне їх зростання протягом навчання ($p<0,001$). Так, частка дітей з порушенням постави збільшується в 1,9 раз, з плоскостопістю – в 1,4 рази, зі зниженням зору – в 2,1 рази.