КОНЦЕПТ
научно-методический электронный журнал
ART 14178 УДК 378.147:658.512.2
Алексеев В. П., Степаньян В. В. Повышение адекватности и достоверности модели обработки ненормированных массивов данных в исследованиях образовательных систем методом планирования эксперимента // Концепт. – 2014. – № 07 (июль). – ART 14178. – 0,3 п. л. – URL: http://e-koncept.ru/2014/14178.htm. – Гос. рег. Эл № ФС 77-49965. – ISSN 2304-120X.

Алексеев Валерий Павлович,
кандидат технических наук, доцент НОУ ВПО «Международный инновационный университет», г. Сочи
alekseev.valeri@yandex.ru

Степаньян Владимир Владимирович,
студент экономико-технологического факультета НОУ ВПО «Международный инновационный университет», г. Сочи
step.wo@mail.ru

Повышение адекватности и достоверности модели обработки ненормированных массивов данных в исследованиях образовательных систем методом планирования эксперимента

Аннотация. *Статья посвящена вопросам повышения точности статистических математических моделей, применяемых для исследования эффективности образовательного процесса. Авторы предлагают изменить типовые подходы к обработке информационных ненормированных по величине массивов данных с использованием метода планирования эксперимента. При этом учитываются особенности информационных массивов, характерных для педагогических процессов. Используя эти предложения, можно повысить адекватность и достоверность модели обработки, а следовательно, точность обработки.*
Ключевые слова: *образовательная система, статистическая математическая модель, уравнение регрессии, адекватность и достоверность, нормальный закон распределения, медиана, мода.*
Раздел: *(01) педагогика; история педагогики и образования; теория и методика обучения и воспитания (по предметным областям).*

Введение

В [1] проведены исследования влияния индивидуальных личностных факторов на эффективность обучения студентов по дисциплине «Информатика». Исследования проводились с использованием оригинального метода обработки ненормируемых по величине массивов статистических данных на основе теории планирования эксперимента. Особенностью таких массивов является их относительная субъективность, обусловленная наличием вероятности правильно угаданного ответа при тестировании. Преодоление субъективности и составляет проблематику настоящей статьи. Далее предлагаются возможные пути повышения адекватности и достоверности обработки массивов, полученных в ходе обработки результатов тестирования.

Нами апробирована следующая методика исследования образовательных систем. Пусть мы имеем образовательную систему с главным выходным параметром $Y(x)$, где x – факторы, влияющие на величину $Y(x)$. Примерами главного выходного параметра могут быть успеваемость, качество знаний, количество дипломов с «отличием» и т. п. Факторами могут быть средний балл ЕГЭ, посещаемость, средние затраты на одного обучаемого, стоимость материальной базы и многие другие, реально влияющие на образовательный процесс. Как факторы, так и выходные параметры могут быть взяты из статистических данных, или определены методом тестирования, экспертных оценок, или измерены в ходе исследований. При этом как выходные параметры, так и факторы будут иметь закон распределения и его статистические величины, такие как среднее квадратичное отклонение, математическое ожидание, мода,

медиана, диапазон разброса и другие. Результат исследования однозначно представляется в виде адекватного уравнения регрессии, которое связывает выходной параметр *Y* с факторами x с помощью коэффициентов регрессии [2], рассчитанных компьютером в соответствии с теорией планирования эксперимента (ТПЭ) [3]. Однако при исследованиях образовательных систем невозможно задавать значения факторов на заранее назначенных уровнях значений, называемых нормированными, так как эти факторы отражают субъективные стороны обучающихся и индивидуальны. В [4] нами предложена теория и методика обработки результатов вычислительного эксперимента с использованием ненормированных значений факторов. При этом основные правила обработки информационных массивов, характерных для ТПЭ, остаются прежними, в частности подчинение нормальному закону распределения массива значений выходного параметра. Массивы значений факторов на соответствие нормальному закону не проверяются. В этом случае оказалось, что при проверке на адекватность уравнения регрессии очень часто расчётное значение *F*-критерия Фишера близко к табличному, т. е. условие адекватности находится на пределе допустимого или адекватность не обеспечивается.

Суть предложений

В нашем случае используется ненормированный массив экспериментальных данных – массив, в котором вектор столбец представляет собой неограниченный поток данных, полученных в ходе тестирования. В таком столбце, не исключены грубые ошибки, как и в векторе-столбце выходного параметра, так как зачастую такие данные снимаются автоматизированными системами, не имеющими возможность исключать такие ошибки, а обучаемые могут вводить заведомо некорректные или непроверенные данные.

Нами предложено обрабатывать исходный массив ненормированных данных правилом три сигма для каждого вектор-столбца, включая выходной параметр.

Линдебергом и Леви было доказано, что если независимые случайные величины ξ_1, ξ_2, … имеют все одно и то же нормальное распределение вероятностей и если каждое ξ_n имеет среднее значение m_1 и стандартное отклонение σ_1, то сумма $\xi = \xi_1 + \ldots + \xi_n$ асимптотически нормальна [5]. В данной теореме предполагается, что каждая из случайных величин подчиняется нормальному закону распределения. В нашем случае конечная линейная математическая модель, уравнение регрессии и есть не что иное, как сумма величин распределенных асимптотически нормально, т. е. каждая случайная величина распределена асимптотически нормально. Практика показала, что такие случаи встречаются нередко. Так, например, в обработанном нами ранее массиве данных о результатах исследования влияния индивидуальных личностных факторов студентов на результативность обучения дисциплине «Информатика» [6] каждый фактор и выходной параметр представляли собой случайные величины, распределенные асимптотически нормально.

При распределении асимптотически нормально случайных величин факторов также асимптотически нормально распределена и случайная величина выходного параметра. Таким образом, при обработке результатов если как можно ближе подвести распределение величин факторов к нормальному закону распределения, то распределение величины выходного параметра будет больше соответствовать нормальному закону распределения. Для этого необходимо обработать каждый столбец по правилу три сигма и тем самым исключить более грубые ошибки с большей вероятностью. На практике это означает, что при исследованиях образовательных систем необходимо набирать как можно большее количество значений факторов, чтобы выбрать из полученного информационного массива реплику значений каждого фактора,

КОНЦЕПТ
научно-методический электронный журнал
ART 14178 УДК 378.147:658.512.2
Алексеев В. П., Степаньян В. В. Повышение адекватности и достоверности модели обработки ненормированных массивов данных в исследованиях образовательных систем методом планирования эксперимента // Концепт. – 2014. – № 07 (июль). – ART 14178. – 0,3 п. л. – URL: http://e-koncept.ru/2014/14178.htm. – Гос. рег. Эл № ФС 77-49965. – ISSN 2304-120X.

наиболее соответствующую нормальному закону распределения. В этом случае повышается вероятность подчинения этому закону вектора-столбца выходного параметра, значит, и адекватность конечного уравнения регрессии.

Обработав два полученных массива с помощью программы обработки массивов ненормированных данных методом планирования эксперимента [7], мы получили два уравнения регрессии:

$$Y_1 = 3{,}9 + 0{,}13 \times q_1 + 0{,}55 \times q_2 + 0{,}25 \times q_3, \quad (1)$$

$$Y_2 = 3{,}9 + 0{,}14 \times q_1 + 0{,}54 \times q_2 + 0{,}26 \times q_3, \quad (2)$$

где Y_1 – уравнение регрессии, полученное после обработки массива данных по правилу три сигма только по выходному параметру; Y_2 – уравнение регрессии, полученное после обработки массива данных по правилу три сигма по всем столбцам векторам.

Из уравнений (1) и (2) видно, что качественная оценка влияния факторов на выходной параметр не изменилась, а количественная изменилась существенно.

При этом линейная модель второго уравнения регрессии более адекватна:

$$F_{расч}\,1 = 2{,}9, \quad (3)$$

$$F_{расч}\,2 = 1{,}7, \quad (4)$$

где $F_{расч}\,1$ – расчётное значение критерия Фишера для Y_1; $F_{расч}\,2$ – расчётное значение критерия Фишера для Y_2. При этом максимальная относительная погрешность первого уравнения составила 14,7%, а второго уравнения – 6,5%.

Таким образом, показано, что случайная величина выходного параметра распределена асимптотически нормально и тогда, когда случайные величины факторов также распределены асимптотически нормально. Кроме того, предварительная обработка массива данных по правилу три сигма не только выходного параметра, но и всех факторов дает нам более адекватную модель, что позволяет повысить точность при оптимизации системы. При компьютерном моделировании обработка больших массивов численных значений факторов не представляет больших затруднений с точки зрения временных и материальных затрат.

При обработке больших массивов возникает ещё одна проблема. В практике статистической обработки данных значительное внимание уделяется определению при заданной вероятности попаданию значения полученных величин в интервал $\alpha < x < \beta$. Это необходимо для исключения из ряда измерений грубых ошибок с заданной вероятностью. Для такого исключения существует известное правило трех сигм, в котором говорится, что с вероятностью, близкой к единице (0,993), случайная величина, подчиняющаяся нормальному закону распределения, не выходит за пределы интервала $[m_x - 3\sigma, m_x + 3\sigma]$ [8].

В нашем случае при обработке ненормированных массивов данных методом теории планирования эксперимента количество тестовых значений величин может достигать десятков тысяч, вероятностное распределение такого массива для исследования другой образовательной системы изображено на рис. 1.

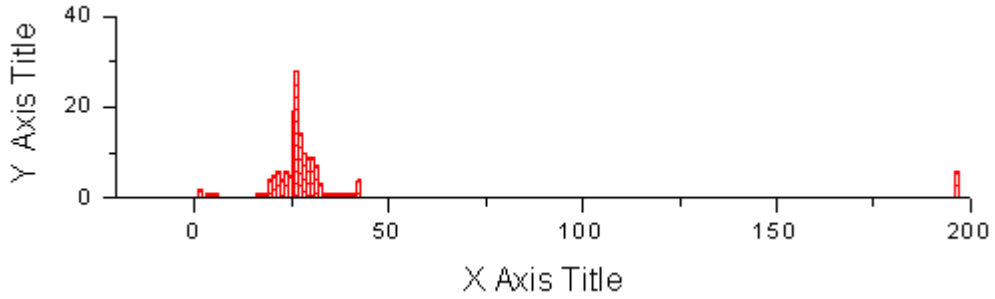

Рис. 1. Исходный массив данных

КОНЦЕПТ
научно-методический электронный журнал
ART 14178 УДК 378.147:658.512.2
Алексеев В. П., Степаньян В. В. Повышение адекватности и достоверности модели обработки ненормированных массивов данных в исследованиях образовательных систем методом планирования эксперимента // Концепт. – 2014. – № 07 (июль). – ART 14178. – 0,3 п. л. – URL: http://e-koncept.ru/2014/14178.htm. – Гос. рег. Эл № ФС 77-49965. – ISSN 2304-120X.

Математическое ожидание в этом случае принимает вид стандартного среднего арифметического: $\overline{X_n} \to \infty \to M(X)$.

При использовании автоматических систем для снятия измерений зачастую допускаются грубые ошибки – резкие выбросы за пределы нормального распределения (рис. 1). Такие ошибки задают значительное смещение математического ожидания относительно центральной точки нормального распределения. Следовательно, обработка массива по правилу трех сигм неприменима для таких случаев.

Для такого случая необходимо ввести другую величину, которая определяла бы центральную точку нормального распределения случайной величины, тем самым охарактеризовать расположение значений случайной величины в данном распределении с помощью некоторого типичного значения этой величины. Наиболее важными характеристиками такого рода являются математическое ожидание, мода и медиана. При широком диапазоне данных математическое ожидание применять для нашего случая нецелесообразно.

Мода – это точка максимума плотности вероятности $f(x)$. Единственная мода существует только в случае, когда кривая плотности $y = f(x)$ имеет один максимум, т. е. распределение является унимодальным. В нашем случае кривая плотности имеет три максимума, т. е. мультимодальна. Следовательно, мода, так же как и математическое ожидание, не может определить точно центральную точку асимптотически нормального распределения оптимизируемого параметра [9].

Медианой распределения μ называется такая точка x_0, которая разделяет распределенную массу на две равные части, каждая из которых содержит массу 1/2. Таким образом, медиана есть любой корень уравнения $F(x) = 1/2$. Медиана обладает следующим свойством: первый абсолютный момент $E(|\xi - c|)$ принимает минимальное значение, когда c есть медиана. Момент $E(|\xi - c|)$ имеет одинаковое значение при любом возможном значении медианы c [10]:

$$E(|\xi - c|) = \begin{cases} E(|\xi - \mu|) + 2\int_{\mu}^{c} (c - x)\partial F(x), \text{при} \quad c > \mu \\ E(|\xi - \mu|) + 2\int_{c}^{\mu} (x - c)\partial F(x), \text{при} \quad c < \mu \end{cases} \qquad (5)$$

Следовательно, целесообразней будет взять за типичный параметр, характеризующий асимптотически нормальное распределение, медиану.

Действительно, если взять асимптотически нормально распределенную случайную величину с резким выбросом (см. рис. 1), то медиана будет лежать наиболее близко к центральной точке нормального распределения (рис. 2).

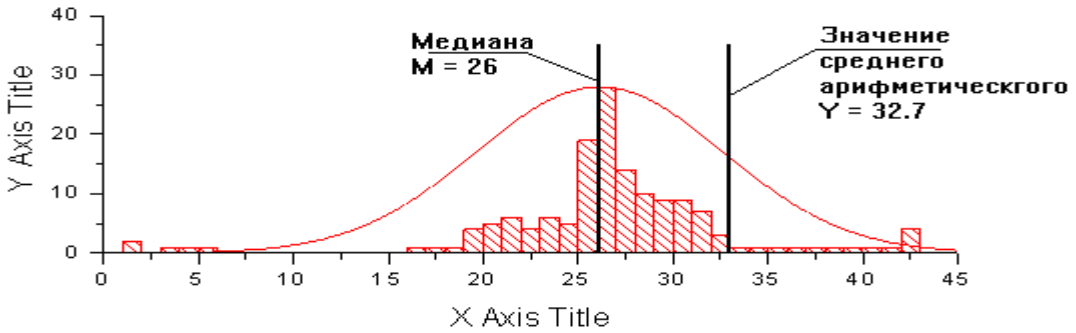

Рис. 2. Графическое представление различия математического ожидания от медианы

Таким образом, чем больше по количеству и по качеству грубых ошибок, тем дальше значение среднего арифметического будет лежать от истинного.

После обработки по правилу трех сигм с учетом медианы данный массив стал соответствовать нормальному закону распределения на 78% (рис. 3), а с учетом математического ожидания – на 71% (рис. 4).

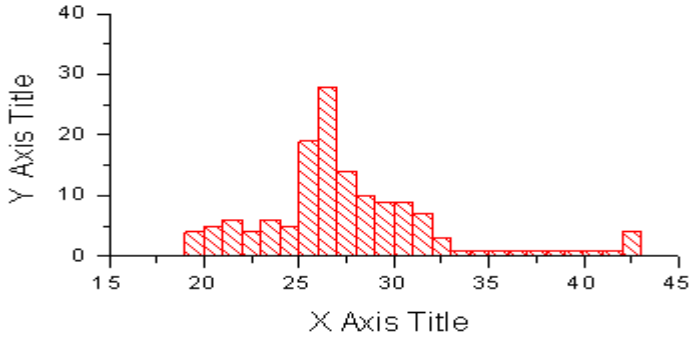

Рис. 3. Распределение случайной величины после обработки с учетом медианы

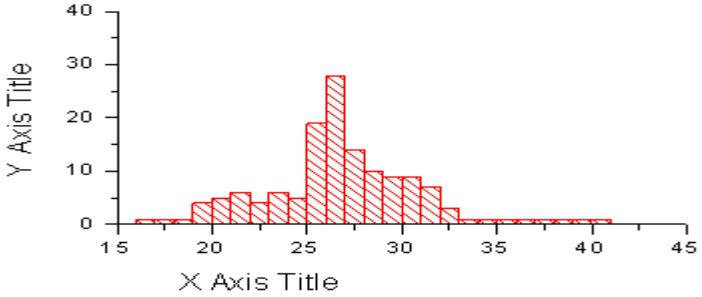

Рис. 4. Распределение случайной величины после обработки с учетом математического ожидания

Выводы

1. Предложен способ обработки ненормированных статистических массивов данных, полученных в ходе автоматизированного тестирования при исследованиях образовательных систем с использованием теории планирования эксперимента, отличающийся повышенной точностью.
2. Предложенный способ отличается обработкой по правилу трёх сигм не только массива выходных параметров, но и массивов всех факторов, а также расчётом не среднего арифметического, а медианы массивов измеренных величин.
3. Применение данного способа позволяет повысить точность обработки в 2 раза в зависимости от однородности, протяжённости массива данных и количества грубых ошибок и промахов.

Ссылки на источники

1. Степаньян В. В. Исследование влияния индивидуальных личностных факторов студентов на результативность обучения дисциплине «Информатика» // Концепт. – 2014. – № 03 (март). – ART 14070. – 0,3 п. л.
2. Там же.
3. Адлер Ю. П., Маркова Е. В., Грановский Ю. В. Планирование эксперимента при поиске оптимальных условий. – М.: Наука, 1976. – 139 с.
4. Алексеев В. П., Кузнецов А. В. Алгоритм и программа обработки массивов непрерывных данных методом планирования эксперимента. // Вестник ТГПУ. –2005. – № 51.
5. Крамер Г. Математические методы статистики / под ред. акад. А. Н. Колмогорова. – М.: «Мир», 1975. – 648 с.

КОНЦЕПТ
научно-методический электронный журнал
ART 14178 УДК 378.147:658.512.2
Алексеев В. П., Степаньян В. В. Повышение адекватности и достоверности модели обработки ненормированных массивов данных в исследованиях образовательных систем методом планирования эксперимента // Концепт. – 2014. – № 07 (июль). – ART 14178. – 0,3 п. л. – URL: http://e-koncept.ru/2014/14178.htm. – Гос. рег. Эл № ФС 77-49965. – ISSN 2304-120X.

6. Степаньян В. В. Указ. соч.
7. Там же.
8. Крамер Г. Указ. соч.
9. Там же.
10. Там же.

ISSN 2304-120X

Valery Alexeev,
Candidate of Engineering Sciences, Associate Professor, International Innovative University, Sochi
alekseev.valeri @ yandex.ru
Vladimir Stepanian,
Student of economics and technology department, International Innovative University, Sochi
step.wo @ mail.ru

Improve of the adequacy and reliability of no standardized processing model datasets in studies of educational systems method of experiment planning

Abstract. The article is devoted the questions of improvement of the accuracy of statistical mathematical models, used to investigate the effectiveness of the educational process. The authors suggest changing standard approaches to information processing large datasets, using the method of experiment planning. This takes into account features of data sets characteristic of pedagogical processes. Using these proposals one can improve the accuracy of the model and the value processing, and hence the accuracy of machining.

Key words: educational system, statistical mathematical model, regression equation, the adequacy and accuracy, normal distribution, median, mode.

References

1. Stepan'jan, V.V. (2014) “Issledovanie vlijanija individual'nyh lichnostnyh faktorov studentov na rezul'tativnost' obuchenija discipline ‘Informatika’”, *Koncept*, № 03 (mart). ART 14070. 0,3 p. l. (in Russian).
2. Ibid.
3. Adler, Ju.P., Markova, E.V. & Granovskij, Ju.V. (1976*) Planirovanie jeksperimenta pri poiske optimal'nyh uslovij*, Nauka, Moscow, 39 p. (in Russian).
4. Alekseev, V.P., Kuznecov, A.V. (2005) “Algoritm i programma obrabotki massivov ne-preryvnyh dannyh metodom planirovanija jeksperimenta”, *Vestnik TGPU*, № 51 (in Russian).
5. Kramer, G. & Kolmogorov, A.N. (ed.) (1975) *Matematicheskie metody statistiki,* “Mir”, Moscow, 648 p. (in Russian).
6. Stepan'jan, V.V. (2014) Op. cit.
7. Ibid.
8. Kramer, G. & Kolmogorov, A.N. (ed.) (1975) Op. cit.
9. Ibid.
10. Ibid.

Рекомендовано к публикации:
Горевым П. М., кандидатом педагогических наук, главным редактором журнала «Концепт»