

Способы определения непрерывности или дискретности исходных данных

С.С. Акимов, аспирант, Оренбургский ГУ

Прерывность (дискретность) — пространственно-временная отграниченность элементов, состояний объекта; непрерывность — взаимосвязь (взаимообусловленность) элементов и состояний объекта [1].

Первым, кто дал строгое определение непрерывности как понятию математического анализа, был Огюстен Луи Коши. Его определение непрерывности опиралось на понятие бесконечно малого, которому он придал новый смысл: у Коши бесконечно малое — переменная величина, стремящаяся к нулю [2].

Таким образом, непрерывность и дискретность представляют собой значимые характеристики, которые отражают противоположные, но взаимосвязанные свойства материальных объектов.

В современном понимании все те объекты, переменные которых могут принимать несчётное множество сколь угодно близких друг к другу значений, называются непрерывными. Подавляющее большинство реальных физических и теоретических объектов, состояние которых характеризуется только макроскопическими физическими величинами (температура, давление, скорость, ускорение, сила тока, напряжённость электрического или магнитного полей и т.д.) обладают свойством непрерывности [3]. Математические структуры, адекватно описывающие такие объекты, тоже должны быть непрерывными. Поэтому при модельном описании таких объектов используется главным образом аппарат дифференциальных и интегро-дифференциальных уравнений. Объекты, переменные которых могут принимать некоторое, практически всегда конечное число наперёд известных значений, называются дискретными. Основой формализованного описания дискретных объектов является аппарат математической логики (логические функции, аппарат булевой алгебры, алгоритмические языки). В связи с развитием ЭВМ дискретные методы анализа получили широкое распространение также и для описания и исследования непрерывных объектов [4].

В связи со сказанным выше возникает вопрос: имеется ли какая-либо возможность отличить непрерывные данные от дискретных?

Вопрос бессмыслен, если исследователь точно знает объект исследования. поголовье скота, производство каких-либо штучных изделий или подсчёт количества пассажиров в транспорте по своему характеру являются дискретными величинами, тогда как измерение температуры, длины, стоимости является процессом измерения непрерывных величин.

Но если исследователь имеет на руках только данные, о характере которых у него нет ни малейшего представления, то поставленный вопрос является весьма актуальным, поскольку в условиях неопределённости свойства распределения могут быть неизвестны, поэтому необходим метод для определения непрерывности и дискретности [5].

Сформулируем задачу: имеется некоторая совокупность неизвестных чисел, при этом не имеется никакой информации об их источнике, свойствах, никакой характеристики, которая бы указывала на принадлежность данной совокупности к чему-либо. Необходимо определить один параметр данной совокупности, а именно принадлежность её к непрерывному ряду или к дискретному.

В литературе решений подобной задачи встречается не много, поскольку единого, «безотказного» механизма не существует. Приведём здесь некоторые приближённые способы решения данной задачи [6].

Рассмотрим первый способ. Он основан на предположении, что основная масса дискретных величин являются целыми, тогда как непрерывные величины по самой своей природе целыми быть не могут. Под целым числом, согласно определению, будем понимать все положительные или отрицательные числа, не являющиеся дробями, и нуль, например ... -3, -2, -1, 0, 1, 2, 3 ... [1].

Однако целочисленность — это частный случай дискретного распределения, потому этот способ является наиболее простым и наименее приближённым. Дискретность ещё не гарантирует целочисленность, хотя не станем отрицать, что подавляющее большинство дискретных совокупностей данных являются целочисленными. Поскольку целочисленность необычайно просто определить, можно использовать этот метод в качестве дополнительного доказательства в пользу того или иного решения.

Таким же частным случаем решения можно считать наличие в выборке отрицательных значений. Все известные дискретные распределения — распределение Пуассона, биномиальное, Бернулли и т.д. задаются исключительно на положительном промежутке, потому наличие отрицательных значений показывает, что исследуемое распределение относится к непрерывным. Разумеется, в общем случае дискретное распределение может задаваться и на отрицательном промежутке. Кроме того, существует достаточно большое количество непрерывных распределений, которые задаются исключительно на положительном промежутке (распределение бета, гамма, экспоненциальное), и к тому же на практике зачастую исследуются именно положительные величины. Наглядным

примером здесь служит группировка людей исходя из их роста — классический пример нормального распределения, однако он задаётся исключительно на положительном промежутке, ибо не бывает людей с отрицательным ростом. Потому данное свойство не может быть полноправно причислено к определяющим факторам дискретности.

Второй способ заключается в поиске частоты повторений $f(x)$ отдельных вариантов значений исследуемой совокупности.

В преддверии описания этого способа введём понятие «совпадение». В толковом словаре совпадение — это [7]:

- 1) одновременность явлений, событий и т.п.;
- 2) сходство, общность, одинаковость;
- 3) соответствие, согласованность;
- 4) соединение, сочетание, совмещение;
- 5) совмещение при наложении (о геометрических фигурах, линиях).

В данном случае под совпадением мы будем понимать сумму повторения каждого числа, встречающегося в совокупности. Иными словами, если некое число x_i встречается в совокупности всего один раз, количество его совпадений равно нулю. Если другое число x_j встречается, например, три раза, то количество совпадений будет равно двум. Сумма же всех подобных повторений и даст общее количество совпадений в совокупности.

Математически эту процедуру можно описать следующим образом. Допустим I — частота повторения числа x_i в совокупности исследуемых значений $Z(x_n)$. Тогда число совпадений $C(x_i) = I(x_i)^{-1}$. Общее же число совпадений $C = \sum C(x_n)$.

Если максимальное значение $\max C(x)$ не является слишком большим, то справедливо утверждать, что распределение является непрерывным. Алгоритм для определения дискретности (непрерывности) будет сведён к следующим действиям: нахождению наибольшего числа повторений признака $\max C(x)$; оценка этого значения. На основании этих данных можно делать предположение о непрерывности или дискретности исследуемых значений. Однако этот способ имеет очень большие трудности при интерпретации результатов, поскольку необходимо рассчитать пороговый уровень количества повторений, по которому можно будет оценить, к какой категории относить данную совокупность.

Рассмотрим третий способ решения поставленной задачи. Будем исходить из того, что дискретное значение всегда изменяется на одну и ту же величину (так называемый «шаг»), т.е. разница между двумя соседними результатами $x_2 - x_1 = \text{const}$. Обозначим результат этой операции $x_2 - x_1 = a$.

В соответствии с этим предположением разница между любыми двумя переменными $x_m - x_n = a(m - n)$ [8].

Приведём алгоритмы решений для каждого из перечисленных способов. Для первого способа

задача сводится к поиску дробных значений, если таковые находятся (притом в достаточно значительном количестве), совокупность будет считаться непрерывной. Вопрос о том, какое количество можно считать достаточным, пока оставим открытым.

Для второго способа необходимо произвести некоторые вычисления. Во-первых, необходимо отсортировать имеющуюся совокупность данных по возрастанию. Далее необходимо найти разность между двумя ближайшими значениями. При выполнении этого пункта все совпадающие величины обратятся в нуль. После чего достаточно просто подсчитать количество нулей в полученном ряду данных (назовем его «интервальный ряд»), которое и будет равно количеству совпадений разнородных значений в данной совокупности. Опять же вопрос о том, какое количество совпадений необходимо для окончательного решения в пользу того или иного варианта, оставим без внимания.

Наконец, опишем алгоритм определения непрерывности или дискретности для третьего способа. Начало этого алгоритма аналогично предыдущему. Также необходимо отсортировать данные по возрастанию и найти интервальный ряд. После чего для интервального ряда найдём наибольший общий делитель (НОД). Поскольку, как уже говорилось выше, дискретное значение всегда изменяется на одну и ту же величину, эта величина должна выступать в роли общего делителя, иными словами, она должна совпадать с НОД. Далее переходим к оценке самого НОД. Тут можно воспользоваться уже приведённым выше алгоритмом, который используется для второго способа определения дискретности: отсортируем полученный интервальный ряд по возрастанию (НОД должен оказаться первым значением в данном ряду или последующим, если первое и следующие за ним значения нуль), и вновь найдём интервальный ряд. Далее вновь подсчитаем количество нулей — при достаточно большом количестве нулей можно сделать вывод, что первоначальная совокупность является дискретной.

Из двух последних способов можно вывести некий общий принцип. Они оба представляют собой набор одних и тех же процедур, который повторяется один раз во втором способе и два раза в третьем. Если же продолжать обработку данных подобным способом, то с каждым разом количество совпадений (нулей) будет всё больше и больше.

Однако при большом числе итераций непрерывные значения также начинают показывать совпадения (уже на 4–6 итерации совпадения варьируются в интервале 95–100%). Потому наиболее разумно остановиться всего на трёх итерациях.

Важно учитывать, что при определении непрерывности или дискретности очень большую роль играет объём выборки. При количестве значений менее 10 результаты будут очень неточными, тем более что при каждой итерации количество зна-

чений уменьшается на единицу. От 10 значений и выше начинает проявляться зависимость. Наиболее чувствительная здесь третья итерация, потому ей надо отдать приоритет. Ближе к сотне значений третья итерация (как наиболее чувствительная) начинает определять совпадения (причём достаточно значительные) в заведомо непрерывных распределениях. Потому для значений от 100 более точной является вторая итерация. Свыше одной тысячи значений вторая итерация перестаёт давать точный результат и на первый план выходит первая [9].

Зная значимость каждой итерации при определённых объёмах выборки, мы имеем принципиальную возможность вывода ряда формул для стохастического определения дискретности или непрерывности выборки. Для этого необходимо присвоить ранги каждой итерации при определённых объёмах выборки. Наиболее частыми в задаче с тремя переменными выступают коэффициенты 0,1; 0,2 и 0,7. Обозначим первую, вторую и третью итерации соответственно А, В и С. Эмпирическим путём было установлено, что наиболее оптимальным соотношением для 10 значений является соотношение $0,1A + 0,2B + 0,7C$; для 100 значений соотношение $0,1A + 0,7B + 0,2C$ и для 1000 и выше — соотношение $0,7A + 0,2B + 0,1C$.

Для повышения точности необходим плавный переход от формулы к формуле. Этого можно добиться, используя скользящие значения, которые изменялись бы в соответствии с изменением объёма выборок. Для получения второй формулы из первой необходимо постепенно снижать ранг третьей итерации в пользу второй. Математически это будет выглядеть так:

$$0,1A + (0,2 + (0,5N/100))B + (0,7 - (0,5N/100))C,$$

где N — количество данных в выборке.

Аналогичным образом выполним преобразование и для других формул:

$$(0,1 + (0,6N/1000))A + (0,7 - (0,5N/1000))B + (0,2 - (0,1N/1000))C.$$

Вывод. Изученная в статье проблема определения непрерывности или дискретности решается методом ранжирования и последовательного нахождения совпадений. Выявленные эмпирическим путём формулы, приведённые в работе, дают возможность с большой точностью определить непрерывность или дискретность изучаемого массива данных.

Литература

1. Большой энциклопедический словарь. URL: <http://dic.academic.ru/dic.nsf/enc3p/243162>.
2. Огюстен Луи Коши. Биография. URL: <http://arifmetica.com/koshi.php>.
3. Непрерывные и дискретные модели. URL: <http://www.automationlab.ru/index.php/2011-06-26-21-10-36/397-6>—
4. Понятие о моделях и моделировании. URL: <http://do.gendocs.ru/docs/index-338446.html>
5. Кван О.В., Фомина М.В., Суханова О.Н., Акимов С.С. Изучение влияния алиментарных факторов питания на показатели крови лабораторных животных с использованием непараметрической статистики // Вестник Оренбургского государственного университета. 2011. № 15. С. 150—152.
6. Акимов С.С., Шепель В.Н. Эвристическая процедура определения подходящего распределения вероятности // Компьютерная интеграция производства и ИПИ-технологии: сб. матер. V Всеросс. науч.-практич. конф. Оренбург: Изд. ИП Осиночкин Я.В., 2011. С. 137—139.
7. Толковый словарь. URL: <http://slovaizbukv.ru/совпадение.html>
8. Акимов С.С. Комплексный подход к определению закона распределения вероятности по выборке из генеральной совокупности // Актуальные проблемы торгово-экономической деятельности и образования в современных условиях: электрон. сб. науч. тр. VII Междунар. науч.-практич. конф., посвящ. 50-летию Оренбургского филиала РГТЭУ. Оренбург: Оренбургский филиал РГТЭУ, 2012. С. 468—476.
9. Акимов С.С. Расчёт вероятности дискретности для массива данных // Научное обозрение. 2013. № 6. С. 78—82.