Маломерная статистическая выборка

В. А. Каладзе [1], E-mail: wakaladze@yandex.ru,
В. А. Работкин [2]

[1] Международный институт компьютерных технологий
[2] Воронежский государственный университет

Аннотация. *Рассмотрены применяемые до настоящего времени подходы к исследованию характеристик маломерных статистических выборок. Разработан подход к установлению её репрезентативности. Приведены примеры, иллюстрирующие предложенный метод.*

Ключевые слова: *Маломерная выборка, метод статистических испытаний, репрезентативность, проверка статистической гипотезы.*

Введение

Во многих инженерных и физических статистических задачах приходится сталкиваться с ситуацией, когда имеющиеся исходные данные для расчётов или исследований представлены небольшим, весьма скромным по объёму множеством значений. Более того, эти значения зачастую не сопровождаются соответствующим множеством их вероятностного выбора, т.е. невозможно сформировать вероятностный ряд, описывающий случайную величину.

Другими словами, перед нами обычная прикладная статистическая задача, когда на небольшом объёме статистически неполной информации требуется получить качественные оценки параметров некоторой системы или описать законы процессов, протекающие в ней.

1. Маломерная статистическая выборка

Будем рассматривать массив исходных данных такой задачи, как дискретное множество числовых значений, полученных из генеральной совокупности (ГС) на основе простого случайного выбора, т.е. статистическую выборку s, которая, при этом, по своему смыслу должна содержать необходимую информацию о системе.

Статистическую выборку s принято упрощённо рассматривать как дискретное множество над полем действительных чисел, численной характеристикой которого является её мера (мощность) m, определяемая через количество наблюдений в выборке. Вопрос о применимости s в конкретной задаче многие пытаются решить с помощью нахождения некой функциональной зависимости её меры от точности искомых оценок

Выборку можно использовать для исследования системы, если она содержит достаточно данных для получения искомых статистических

оценок, точность которых напрямую связана с объёмом s, однако величина мощности выборки ещё не гарантирует наличия у неё свойства представлять ГС. Такое несоответствие связано с тем, что повышение точности статистической оценки приводит к снижению её достоверности. Этот парадокс наглядно усматривается в общеизвестном «правиле трёх сигм».

При этом, возникает важная самостоятельная задача определения условий, при которых s в достаточной мере сможет адекватно представлять ГС, поскольку эта адекватность обеспечивается не только мощностью множества наблюдений, отобранных в s из ГС. Вообще-то говоря, эту выборку следует воспринимать не только как множество возможных значений случайной величины, поскольку она должна наследовать важнейшую вероятностную характеристику ГС, такую как конкретный закон распределения вероятностей над этим множеством, который необходимо учитывать. Однако это требование в большинстве прикладных задач невыполнимо.

В этой связи, в прикладных задачах закон распределения, рассматриваемой генеральной совокупности, определяется в «широком смысле»: только через его числовые моменты (как начальные, так и центральные). При этом его функциональная зависимость (в интегральной или дифференциальной форме) не оценивается и не выводится из условий исследуемого процесса (что зачастую невозможно), но считается известной на некоторых законных основаниях, например, из проведённых ранее в данной предметной области теоретических исследований, и в предположении, что она известна исследователю.

Тем не менее, определять репрезентативность полученной выборки можно только с учётом её вероятностных свойств. Нельзя исследовать характеристики s, как множества действительных чисел детерминированными методами, чем обычно грешат исследователи.

Данная задача определения условий адекватности маломерной статистической выборки сводится к двум направлениям исследования: 1) определение допустимого минимального объёма выборки; 2) оценке достаточности статистической близости s к ГС. Т.е. требуется провести формирование показателей, определяющих оба эти направления.

Возникает резонный вопрос: при каком минимальном количестве элементов выборки может рассматривать вопрос о её репрезентативности, т.е. при каких ограничениях можно провести оценку адекватности для маломерных s, имеющих небольшую меру m.

Так, легко видеть, что на ограничении снизу $|S| = m < 3$ выборка S не содержит статистического смысла и однозначно не может представлять генеральную совокупность. Если же статистическая выборка по мощности сопоставима с ГС, то вопрос о репрезентативности такой выборки считается не актуальным.

Поскольку в большинстве прикладных задач расчёты проводятся на основе ограниченных выборок мощности $m < 300$, а нижний предел меры репрезентативной выборки определяется как 200, 300, то будем считать, что мера достаточной выборки начинается с 300.

В литературных источниках [1,2] можно усмотреть различные оценки величины m , определяющие уровень малости выборки: 25, 30, 35, на которой можно получить удовлетворительное решение той или иной прикладной задачи. Однако общего подхода нет и единственным обоснованием правильности принятого решения в большинстве случаев, является опыт исследователя в конкретной предметной области.

Чтобы разобраться в этой проблеме, проведём по материалам нескольких исследований [3,4,5,6] условную классификацию меры статистических выборок в соответствии с характером решаемых задач. Так для решения задач статистического оценивания неизвестных характеристик системы требуется привлечения случайной выборки мощности $30 < m < 300$. В задачах статистического контроля уже известных характеристик достаточно будет маломерной выборки $3 < m < 30$. При этом условность такой классификации объясняется субъективностью её источников, на которых сформированы эти ограничения.

Кроме того, в ряде работ были попытки проводить исследования характеристик S информационными методами через оценивание информации, содержащейся в ней [3,6]. Для этого использовались приёмы группировки данных: построение гистограмм, проверка через ХИ-квадрат критерий. Однако малость меры затрудняет использовать информационный подход. Поэтому исследования сводились к непосредственным методам построения над всей выборкой статистической функции распределения, использования критерия Уилкоксона и т.п.

Решать вопрос об объёме адекватной статистической выборки только с точки зрения её меры, как это делается в большинстве случаев [7,8,9,10], не следует, поскольку такой подход не обеспечивает получение достоверных результатов. Необходимую оценку m можно получить только лишь оценив статистическую значимость исследуемой выборки и только в условиях конкретной задачи. Что указывает на

локальность решения и невозможность получения единой оценки минимальной меры, обеспечивающей адекватность выборки.

Решение о минимальном размере выборки и её адекватности следует принимать на статистической основе с использованием статистических критериев в ходе проверки статистической гипотезы H_0, оценив доверительную вероятность, с которой эта выборка S может быть принята к решению статистической задачи.

Поскольку мера m статистической выборки – это число степеней свободы по Фишеру, то процедура определения нижней грани меры репрезентативной выборки основана на подборе к имеющейся величине ч.с.с. эффективного значения уровня значимости, обеспечивающего получение оптимального значения, используемого статистического критерия.

Если в инженерной задаче требуется установить адекватность (генеральной совокупности) используемого небольшого массива данных [7], т.е., по существу, репрезентативность маломерной статистической выборки, то на основе классического подхода [5] проверяются статистические гипотезы равенства центра маломерной выборки с эталонной и равенство их дисперсий, поскольку среднеквадратическое отклонение определяет меру выборки.

2. Определение закона распределения на маломерной выборке

В задачах радиометрии и спектрометрии излучений различной физической природы применяются методики многократной регистрации числа частиц потока излучения $K(\Delta t) = K(\cdot)$ за фиксированный интервал времени Δt. Последовательность из таких значений $K(\cdot)$ можно рассматривать как реализацию случайного процесса с дискретным временем и дискретными значениями, т.е. временного ряда. Для анализа и классификации дискретных статистических распределений по их типам или формам используется случайный вектор, представляющий эмпирическое распределение ЭР$(K(\cdot))$ $(n_0, \ldots, n_i, \ldots, n_l)$, $\sum_{i=1}^{l} n_i = n$, где $n_i(K = i)$ – число одинаковых K_i в выборке объёма n. Сложность ситуации состоит в том, что на маломерной выборке, при малых n, известные стандартные статистические методы [11] не позволяют выявить нестационарность и периодичность процесса. В частности, критерий Хи-квадрат, указывая на соответствие их распределению Пуассона, не позволяет разделить такие ЭР$(K(\cdot))$ по их формам, различие которых связано различными

значениями параметров распределения. В таком случае, при проверке статистической гипотезы об адекватности закона распределения, выявление соответствующей доверительной вероятности (или уровня значимости) будет определять уровень надёжности принятого решения.

При анализе и обработке таких распределений большую информативность обеспечивает наблюдение не интегрального счёта за интервал времени, а исследование интервалов между отдельными событиями (т.е. наблюдение накопления событий во времени). В этом случае необходимо регистрировать время появления каждого событий [12]. Такой подход позволяет не терять информацию о временных характеристиках потоков излучения и произвольно выбирать интервал квантования времени Δt. При проверке эмпирических данных на соответствие их распределению Пуассона и определения параметров этих распределений дополнительно можно исследовать распределение интервалов времени между отдельными событиями на соответствие их экспоненциальному распределению.

Проверка гипотез согласия ЭР(K) основана обычно на критерии Хи-квадрат, в котором применяется неоднозначная процедура группирования выборочных значений, а асимптотические значения его квантилей могут существенно отличаться от значений при небольших объёмах выборок. В работе [13] рассмотрена задача модификации критерия согласия на основе метода эмпирических производящих функций вероятности (ЭПФВ).

При статистическом анализе данных, если вид закона распределения дискретной случайной величины известен, то параметры функции распределения можно оценить, используя несколько первых целых эмпирических моментов распределения на основе метода моментов. Если же вид функции распределения не задан, то совокупность всех целочисленных центральных моментов, как теоретических, так и тем более эмпирических, не всегда позволяет однозначно установить функцию распределения. Такая задача существенно усложняется в условия случайных выборок малого объема. С увеличением порядка возрастает разброс значений моментов. Поэтому при анализе распределений ограничиваются моментами не более четвертого порядка. При сравнении нескольких ЭР $(K(\cdot))$ можно воспользоваться анализом проекций фазовых траекторий функций дробного порядка $1 < S < 5$ комплексных центральных моментов $\mu(S_l V_q(\cdot))$ СВР [14].

Заключение

Разработан адекватный подход к установлению репрезентативности маломерной статистической выборки её генеральной совокупности.

Список литература

1. Статистический анализ малого числа наблюдений / И. П. Демаков [и др.]; Л., 1973. – 27 с.
2. Методы статистического анализа и обработка малого числа наблюдений при контроле качества и надёжности приборов и машин / Л., 1974. – 92 с.
3. Демаков, И. П. Статистические методы определения законов распределения при анализе точности и надёжности промышленных изделий по результатам эксперимента / И. П. Демаков, В. Е. Потепун; Л.. – 39 с.
4. Еременко, И. В. Об одном методе построения законов распределения величин при малом числе испытаний / И. В. Еременко, А. Н. Свердлик; В кн.: Некоторые вопросы специального применения вычислительной техники. Л., ЛВИКА им. А. Ф. Можайского, 1963. – 18-29 с.
5. Кендалл, М. Статистические выводы и связи / М. Кендалл, А. Стьюарт; М., Наука, 1973. – 900 с.
6. Шор, Я. Б. Статистические методы анализа и контроля качества и надёжности / Я. Б. Шор // Советское радио, М., 1962. – 552 с.
7. Хан, Г. Статистические модели в инженерных задачах / Г. Хан, С. Шапиро; М., Мир, 1969. – 396 с.
8. Хахубия, П. Г. Об эффективности различения по малым выборкам нормального и равномерного типов распределений / П. Г. Хабубия // Теория вероятностей и её применения. – М., 1966. – Т. 11. – № 1. – С. 120.
9. Brinbaum, Z. W Small sample distribution for multi-sample statistics of Smirnov type / Z. W. Brinbaum, R. A. Hall //Annals of Math Statist., 1960. – V. 31. – N. 3.
10. Burr, Ë. I. Small sample distribution of two-sample Cramer-von Mise's W and Watson's V. / Ë. I. Burr // Annals of Math. Statist., 1964. – V. 35. – N. 3.
11. Чавчанидзе, В. В. Об определении законов распределения на основе малого числа наблюдений / В. В. Чавчанидзе, В. А. Кумсишвили //В кн.: Применение вычислительной техники для автоматизации производства (Труды совещания 1959 г.). М., Машгиз, 1961. – С. 71-75.
12. Акиндинова, Е. В. Multichannel Spectrometer of Time Distribution / Е. В. Акиндинова, А. Г. Бабенко, В. М. Вахтель, Н. А.

Евсеев, В. А. Работкин, Д. Д. Харитонова // Exotic Nuclei : EXON 2014. Proceeding of the International Symposium, Kaliningrad, Russia, 8-13 Sept. 2014 . – Singapore, 2014 . – P. 651-658 .

13. Акиндинова, Е. В. Критерий согласия на основе производящей функции вероятности / Е. В. Акиндинова, А. Г. Бабенко, В. М. Вахтель, В. А. Работкин, К. С. Рыбак // Воронежская зимняя математическая школа С.Г. Крейна – 2018: Материалы международной конференции, Воронеж, Научная книга, 2018. – ISBN 978-5-4446-1094-7. – С. 117.

14. Нигматуллин, Р. Р. Статистика дробных моментов: новый метод количественного «прочтения» произвольной случайной последовательности / Р. Р. Нигматулин // Учён. зап. Казан. гос. ун-та. Сер. Физ.-матем. науки, 2005. – Т. 147. – книга 2. – С.129-161