

УДК 519.25

doi:10.15217/issn1684-8853.2016.6.96

## ИССЛЕДОВАНИЕ ЛИНЕЙНОЙ КОРРЕЛЯЦИОННОЙ СВЯЗИ В МНОГОМЕРНОМ МЕТОДЕ ТОЧЕЧНЫХ РАСПРЕДЕЛЕНИЙ

В. С. Попукайло<sup>а, 1</sup>, преподаватель

<sup>а</sup>Приднестровский государственный университет им. Т. Г. Шевченко, Инженерно-технический институт, Тирасполь, Молдова

**Постановка проблемы:** современное производство имеет тенденцию к уменьшению объемов выборок при проведении контрольных испытаний. Существующие методы нахождения линейного коэффициента корреляции не дают необходимой точности результатов в случаях выборок малого объема. **Цель:** анализ возможностей применения коэффициента корреляции Пирсона после виртуального увеличения исходной таблицы данных многомерным методом точечных распределений. **Результаты:** исследования показали, что применение виртуального увеличения таблицы исходных данных, основанного на методе точечных распределений, позволяет избежать ошибочного определения слабой отрицательной вместо слабой положительной линейной корреляционной связи при анализе многомерных выборок малого объема. Кроме того, предложенная методика дает возможность сузить разброс оценки значений коэффициентов корреляции Пирсона, что может быть применено для дополнительной оценки величины линейной корреляционной связи при построении математических моделей по пассивным данным на ранних этапах исследования. Новизна подхода состоит в использовании информации о каждой отдельной реализации малой выборки с учетом знаний о видах законов распределения одномерных случайных величин. Формирование виртуально увеличенной таблицы (состыковка значений) производится по максимальному уровню плотности вероятности. **Практическая значимость:** многомерный метод точечных распределений может применяться для оперативного управления технологическими процессами на этапе накопления информации об исследуемом объекте.

**Ключевые слова** — корреляция, математическое моделирование, метод точечных распределений, малая выборка.

### Введение

При решении задач контроля качества в промышленных производствах большой интерес представляет задача определения значимости корреляционной связи между исследуемыми величинами, так как отбор сильносвязанных факторов позволяет существенно сократить количество браковочных контрольных операций при сохранении того же качества изделия. Насколько связь между случайными величинами близка к строго линейной, показывает коэффициент корреляции. Для его применения исходная информация должна отражать наиболее характерные черты изучаемой совокупности, а число наблюдений должно быть достаточно большим. Однако в задачах технического контроля не всегда можно получить соответствующий набор данных для их последующего анализа.

В предыдущей работе автора [1, с. 27] было показано, что процедура «бутстрэппинга» [2, с. 135] и применение метода точечных распределений [3, с. 75] не увеличивают точности при нахождении наивероятнейшего коэффициента корреляции Пирсона. Более того, «бутстрэппинг» увели-

чивает разброс получаемых значений, а нахождение коэффициента корреляции по методу Чебышева после применения метода точечных распределений и вовсе не позволяет с точностью определить наличие высокого уровня связи. По этим причинам встал вопрос о правомерности нахождения коэффициентов корреляции после построения виртуальной увеличенной таблицы данных в многомерном методе точечных распределений (ММТР) [4, с. 207].

### Методы исследования

Для поиска ответа на этот вопрос было решено из массива данных большого объема, полученных при производстве кристаллов интегральных микросхем, распределенных по нормальному закону распределения, с помощью таблицы равномерно распределенных случайных величин получить выборки по 10 элементам. Данный объем определен не случайно, так как в промышленности существуют производства, по технологическим особенностям которых невозможно обеспечить большой объем контрольной выборки. Например, при производстве кристаллов интегральных микросхем из-за специфики топологии пластин на них имеются от пяти до 10 тестовых ячеек, измерения в которых должны с некоторой вероятностью отражать поведение одноименных параметров 400–5000 рабочих кристаллов.

На следующем шаге к полученной выборке был применен алгоритм многомерного метода

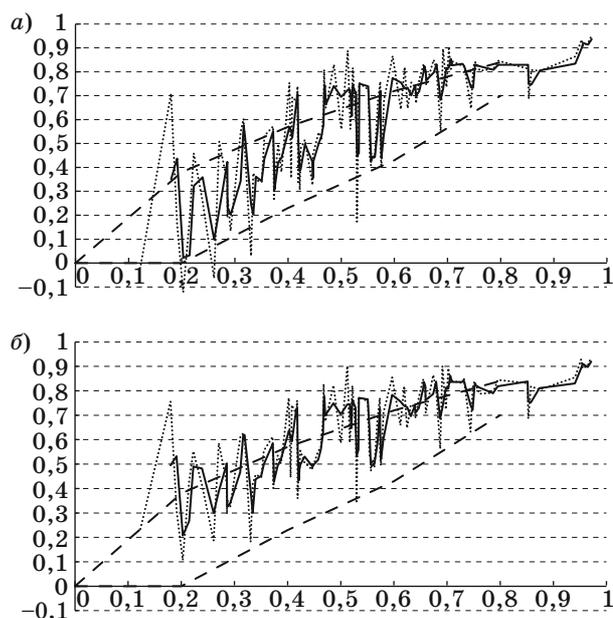
<sup>1</sup> Научный руководитель — доктор технических наук, профессор кафедры информационных технологий и автоматизированного управления производственными процессами Инженерно-технического института Приднестровского государственного университета им. Т. Г. Шевченко Ю. А. Долгов.

точечных распределений [4, с. 207]. Таким образом, был получен новый, виртуально увеличенный массив данных, содержащий 280 строк. После чего были рассчитаны коэффициенты парной корреляции Пирсона для исходного массива данных, для массива извлеченных данных малого объема и для массива данных, полученных после применения ММТР.

### Анализ результатов корреляционного анализа таблиц исходных и виртуально увеличенных данных

Для сравнения результатов приведем графики, на которых по оси  $x$  располагаются значения коэффициентов корреляции Пирсона, полученные по первоначальной выборке большого объема, а на оси  $y$  — коэффициенты корреляции по малой выборке (рис. 1, а) и коэффициенты, полученные после применения процедуры ММТР (рис. 1, б). На графиках точечно-пунктирная линия показывает результаты произведенных расчетов; сплошной линией обозначается линия тренда, аппроксимируемая по двум точкам, которая хорошо отражает тенденцию искомой зависимости; штрихпунктирными линиями на графиках обозначены доверительные интервалы для коэффициента корреляции, полученного по первоначальной выборке [5, с. 416].

Анализ полученных результатов дает основания утверждать, что применение многомерного метода точечных распределений снижает разброс между рассчитанными коэффициентами корреляции, особенно в случае слабых связей, при



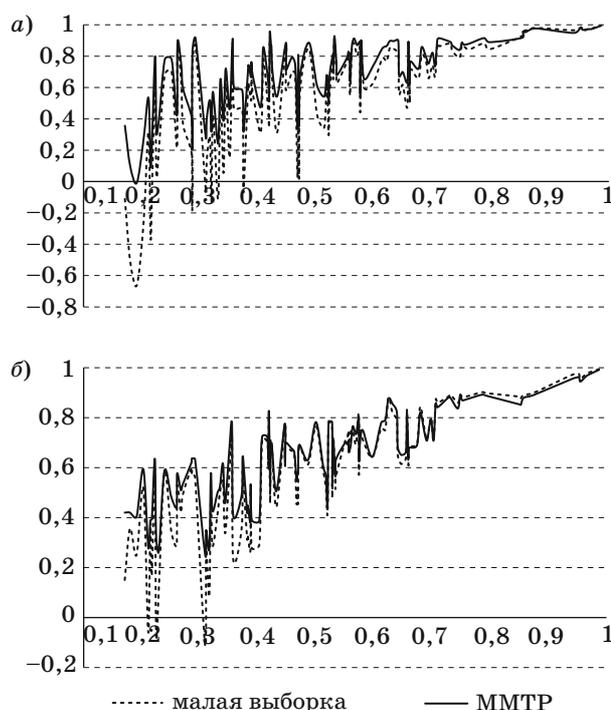
■ Рис. 1. Коэффициент корреляции Пирсона по малой выборке (а) и после применения ММТР (б)

общем сохранении вида линии тренда. При этом отмечается тенденция к завышению рассчитанного значения по сравнению с коэффициентом корреляции, рассчитанным по малой выборке, из-за чего достоверно распознать сильную линейную связь возможно только при получении величины  $r > 0,75$ . Необходимо обратить внимание на тот факт, что расчет коэффициента корреляции Пирсона по малой выборке в ряде случаев показывал слабую отрицательную связь в случаях слабой положительной. Этот недостаток усугубляется после применения виртуального увеличения выборки «бутстрэппингом» [1, с. 30], но он исчезает, если увеличивать объем данных многомерным методом точечных распределений.

Проведенные исследования показали, что разработанная методика дает более точные показатели при оценивании линейного коэффициента корреляции, вне зависимости от исследуемого объема выборки.

В качестве примера приведем графики отношения оцениваемого коэффициента корреляции по выборкам малого объема и оцениваемого коэффициента корреляции после виртуального увеличения выборки многомерным методом точечных распределений (значения по оси  $y$ ) к коэффициенту корреляции генеральной совокупности (значения по оси  $x$ ) для объема выборки  $n = 6$  (рис. 2, а) и  $n = 16$  (рис. 2, б).

Приведенные графики подтверждают, что при уменьшении объема исходных данных оценива-



■ Рис. 2. Оценка коэффициентов корреляции при  $n = 6$  (а) и  $n = 16$  (б)

ние выборочного коэффициента корреляции приводит к большим ошибкам, и его значение не может с достаточной достоверностью подтверждать наличие значимой наблюдаемой линейной связи между исследуемыми случайными величинами. Применение разработанной методики позволяет сузить доверительный интервал при оценивании коэффициента корреляции Пирсона по данным малого объема, а также избежать ошибочного определения отрицательной корреляционной связи.

## Заключение

Таким образом, можно рекомендовать рассчитывать коэффициент корреляции после виртуального увеличения объема выборки ММТР. В случае обнаружения в выборках  $n > 10$  линей-

ных корреляционных связей,  $r > 0,75$ , также следует признать данную наблюдаемую связь сильной и значимой. В данных объемом  $10 \geq n \geq 7$  наблюдаемую связь следует признать значимой при  $r \geq 0,8$ . При объеме исследуемой выборки  $n \leq 6$  значение порогового коэффициента корреляции следует принять на уровне  $r = 0,9$ .

Таким образом, применение виртуального увеличения таблицы исходных данных предложенным методом, в отличие от методов «бутстрэппинга», позволяет сохранить знание о виде закона распределения случайной величины и о величине линейной корреляционной связи между исследуемыми факторами, что дает возможность принимать более обоснованные решения по малому количеству исходных данных, а также улучшить качество статистического оценивания параметрическими методами.

## Литература

1. Попукайло В. С. Исследование линейной корреляционной связи в парных выборках малого объема // *Технология и конструирование в электронной аппаратуре*. 2016. № 1. С. 27–32. doi:10.15222/ТКЕА2016.1.27
2. Efron B. *The Jackknife, the Bootstrap and Other Resampling Plans*. — Philadelphia, Pa.: SIAM, 1982. — 135 p.
3. Столяренко Ю. А. Метод точечных распределений // *Радиоэлектронные и компьютерные системы*. 2012. № 6(58). С. 75–77.
4. Popukaylo V. Analysis of Statistical Modeling Methods for Small-Volume Samples // *Telecommunications, Electronics and Informatics: proc. of the 5th Intern. Conf., Chişinău, Moldova, May 20–23, 2015*. P. 207–211.
5. Болшев Л. Н., Смирнов Н. В. *Таблицы математической статистики*. — М.: Наука, 1983. — 416 с.

UDC 519.25

doi:10.15217/issn1684-8853.2016.6.96

### Linear Correlation Communication in Multidimensional Method of Point Distributions

Popukaylo V. S.<sup>a</sup>, Lecturer, vsp.science@gmail.com

<sup>a</sup>Pridnestrovian State University, Engineering School, 2a, Vosstania St., 3300, Tiraspol, Moldova

**Introduction:** Modern production tends to decrease the volume of samples during routine tests. The existing methods of finding the linear correlation coefficient do not provide the necessary precision of the results when the sample volume is small. **Purpose:** We analyze the possibilities to apply the Pearson correlation coefficient after the virtual growth of the initial data table using the multidimensional method of point distributions. **Results:** The study showed that the usage of virtual growth of the original data table based on the point distribution method allows you to avoid a wrong determination of the linear correlation communication as weak negative instead of weak positive when analyzing small-volume multidimensional samples. Besides, the proposed method allows you to narrow down the rating spread of Pearson correlation coefficient values which can be used to additionally assess the size of the linear correlation communication when creating mathematical models according to passive data at the early steps of the research. The novelty of this approach is using the information on each single small-sample implementation, relying on the knowledge about the types of one-dimensional random variable distribution laws. The formation of a virtually increased table (joining of values) is performed by the maximum level of the probability density. **Practical relevance:** The multidimensional method of pointed distributions can be applied to control technological processes in real time at the stage of accumulating the information about the studied object.

**Keywords** — Correlation, Mathematical Modeling, Method of Point Distributions, Small-Volume Sample.

## References

1. Popukaylo V. S. The Linear Correlation Research in Paired Small Volume Samples. *Tekhnologiya i konstruirovaniye v elektronnoy apparature*, 2016, no. 1, pp. 27–32 (In Russian). doi:10.15222/ТКЕА2016.1.27
2. Efron B. *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia, SIAM, 1982. 135 p.
3. Stolyarenko Y. A. Method of Point Allocations. *Radioelektronnye i komp'yuternyye sistemy*, 2012 no. 6(58), pp. 75–77 (In Russian).
4. Popukaylo V. S. Analysis of Statistical Modeling Methods for Small-Volume Samples. *Proc. of the 5th Intern. Conf. "Telecommunications, Electronics and Informatics"*, Chişinău, Moldova, 20–23 May, 2015, pp. 207–211.
5. Bol'shev L. N., Smirnov N. V. *Tablitsy matematicheskoy statistiki* [Tables of Mathematical Statistics]. Moscow, Nauka Publ., 1983. 416 p. (In Russian).