Мартышенко Н.С., Мартышенко С.Н., Кустов Д.А. Моделирование многомерных данных и компьютерный эксперимент // Техника и технология. – 2007. – №2. С. 47–52.

МОДЕЛИРОВАНИЕ МНОГОМЕРНЫХ ДАННЫХ И КОМПЬЮТЕРНЫЙ ЭКСПЕРИМЕНТ

к.т.н., профессор С.Н. Мартышенко, к.э.н. , Н.С. Мартышенко,

Д.А. Кустов

Владивостокский государственный университет экономики и сервиса

С развитием экономических процессов в России возрастает роль маркетинговых исследований. Основу маркетинговых исследований составляют данные анкетных опросов. Для обработки этих данных используются методы математической статистики. С накоплением опыта использования статистических методов естественно возникает потребность в решении все более сложных задач, требующих более совершенных методов обработки.

В настоящее время исследователи далеко не всегда уже удовлетворяются только методами обработки одномерных данных. Поэтому в последние годы стали получать все большее распространение задачи требующие использования методов многомерного анализа данных.

Для исследования возможностей применения того или иного статистического метода или разработки новой компьютерной технологии решения задач незаменимыми инструментом являются средства моделирования статистических данных.

Однако и в EXCEL и даже в таких известных пакетах по обработке статистических данных, как STATISTICA или SPSS представлены программные модули, позволяющие моделировать только одномерные статистические распределения.

Чтобы расширить возможности исследователей в области анализа многомерных данных нами были разработаны программные средства моделирования и анализа многомерных данных. Программные модули были реализованы в виде дополнительной надстройки к EXCEL, как самому распространенному среди практиков пакету по обработке данных. Тем более, что данные из пакета EX-

CEL без труда могут быть экспортированы в любой специализированный пакет по обработке статистических данных.

Для более адекватного отражения реальной ситуации при моделировании многомерных данных необходимо иметь возможность воспроизведения зависимости признаков. Среди многомерных законов распределения, учитывающих зависимость признаков наиболее известен многомерный нормальный закон. Поэтому основным модулем разработанного моделирующего комплекса является программа моделирования многомерной нормальной выборки. В основу программы был положен алгоритм моделирования нормального распределения, изложенный в работе [1]. Рассмотрим формальное описание алгоритма.

Многомерное нормальное распределение $\xi = (\xi_1, \xi_2, ..., \xi_m)$ описывается вектором математических ожиданий $\mu = (\mu_1, \mu_2, \mu_3, \ldots, \mu_m)$ и ковариационной матрицей K:

$$K = \begin{Vmatrix} K_{11} & K_{12} & \ldots & K_{1m} \\ K_{21} & K_{22} & \ldots & K_{1m} \\ \ldots & \ldots & \ldots & \ldots \\ K_{n1} & K_{n2} & \ldots & K_{mm} \end{Vmatrix}, \qquad (1)$$

где $K_{ij} = \mathbf{M}[(\xi_i - \mu_i)(\xi_j - \mu_j)]$; (2)

i=1,2,3,…m, j=1,2,3,…, m;

m – количество признаков многомерной нормальной выборки.

Вектор ξ с таким распределением можно получить специальным линейным преобразованием вектора $\eta = (\eta_1, \eta_2, ..., \eta_n)$, компоненты которого есть нормально распределенные случайные величины с параметрами $\mu = 0, \sigma = 1$. Для моделирования одномерной нормальной случайной величины существует множество способов. Самый простой способ моделирования состоит в преобразовании двух случайных чисел α_1 и α_2:

$$\eta_1 = \sqrt{-2\ln\alpha_1}\sin 2\pi\alpha_2, \ \eta_2 = \sqrt{-2\ln\alpha_1}\cos 2\pi\alpha_2 \qquad (3)$$

Преобразование η в ξ производится по формуле:

$$\xi = \mathrm{A}\eta + \mu \qquad (4)$$

В преобразовании участвует некоторая треугольная матрица А:

$$\mathrm{A} = \begin{Vmatrix} a_{11} & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n1} & \dots & a_{mm} \end{Vmatrix} \tag{5}$$

Коэффициенты a_{ij} могут быть определены с помощью рекуррентной процедуры. Общая рекуррентная формула имеет вид:

$$a_{ij} = \frac{K_{ij} - \sum_{k=1}^{j-1} a_{ik} a_{jk}}{\sqrt{K_{jj}} - \sum_{k=1}^{j-1} a_{jk}^2}, \tag{10}$$

где индексы изменяются в диапазоне $1 \le j \le i \le m$, а суммы с верхним нулевым приделом равны нулю ($\sum_{k=1}^{0} a_{ik} a_{jk} = 0$, $\sum_{k=1}^{0} a_{jk}^2 = 0$).

При задании параметров многомерного нормального распределения необходимо помнить, что ковариационная матрица должна быть положительно определенной. Во избежание ошибок программа осуществляет такую проверку и если матрица не удовлетворяет требованию, то выводится сообщение о том, что данные с такими параметрами не могут быть сгенерированы.

При штатном завершении работы программного модуля будет получена таблица данных, содержащая *N* наблюдений многомерной нормальной выборки из *m* признаков, обладающей заданными характеристиками.

Рассмотрим пример моделирования трех многомерных нормальных выборок с тремя зависимыми признаками. Совокупность объектов выборок составляют некоторые классы. При запуске программы для трех классов были установлены следующие объемы выборок $N_1 = 200, N_2 = 100, N_3 = 300$. Остальные параметры многомерной выборки представлены в таблице 1. Различия в степени зависимости признаков задаются коэффициентами корреляционной матрицы. Как видно из таблицы 1, степень зависимости между признаками первого класса гораздо ниже, чем между признаками во втором и третьем классах. В силу того, что при моделировании используется датчик случайных чисел, пара-

метры выборок, рассчитанные по модельным данным (таблица 2), будут несколько отличаться от заданных параметров (таблица 1).

Таблица 1

Параметры, заданные при моделировании многомерной нормальной выборки

№ класса	Признак	Среднее	Дисперсия	Корреляционная матрица		
Класс 1	Признак 1	5	1	1,0	0,3	0,2
	Признак 2	5	1	0,3	1,0	0,1
	Признак 3	4	1	0,2	0,1	1,0
Класс 2	Признак 1	10	2	1,0	0,5	0,5
	Признак 2	10	2	0,5	1,0	0,5
	Признак 3	10	2	0,5	0,5	1,0
Класс 3	Признак 1	8	3	1,0	0,9	0,8
	Признак 2	8	3	0,9	1,0	0,9
	Признак 3	2	3	0,8	0,9	1,0

Таблица 2

Параметры, рассчитанные по данным многомерной модельной выборки

№ класса	Признак	Среднее	Дисперсия	Корреляционная матрица		
Класс 1	Признак 1	5,02	1,16	1,00	0,25	0,27
	Признак 2	5,00	1,12	0,25	1,00	0,14
	Признак 3	3,95	0,83	0,27	0,14	1,00
Класс 2	Признак 1	9,89	1,53	1,00	0,37	0,47
	Признак 2	9,85	1,76	0,37	1,00	0,53
	Признак 3	10,08	1,83	0,47	0,53	1,00
Класс 3	Признак 1	8,02	3,08	1,00	0,92	0,80
	Признак 2	8,05	2,99	0,92	1,00	0,90
	Признак 3	2,10	3,06	0,80	0,90	1,00

Приведем расчетные данные по матрице преобразований A для 1-го,2-го и 3-го классов.

$$A_1 = \begin{Vmatrix} 1,0 & 0 & 0 \\ 0,3 & 0,953939 & 0 \\ 0,2 & 0,041931 & 0,978898 \end{Vmatrix}$$

$$A_2 = \begin{Vmatrix} 1,414214 & 0 & 0 \\ 0,707107 & 1,224745 & 0 \\ 0,707107 & 0,408248 & 1,154701 \end{Vmatrix}$$

$$A_3 = \begin{Vmatrix} 1{,}732051 & 0 & 0 \\ 1{,}558846 & 0{,}754983 & 0 \\ 1{,}385641 & 0{,}715247 & 0{,}753937 \end{Vmatrix}$$

Графический образ, сгенерированных данных можно проанализировать на трехмерном графике на рисунке 1 и на трех двухмерных графиках, каких как на рисунке 2.

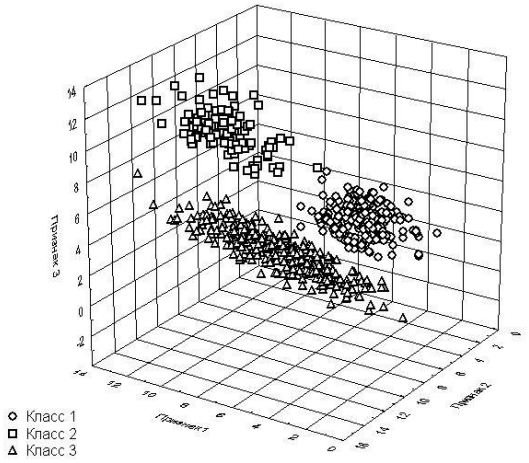

Рис. 1. Трехмерная диаграмма рассеивания смоделированной многомерной выборки.

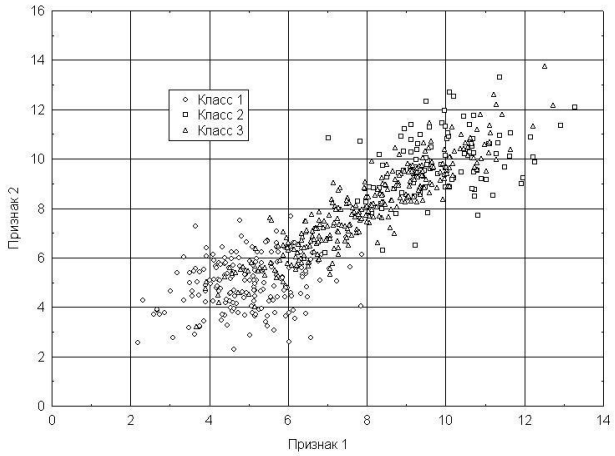

Рис. 2. Проекция диаграмма рассеивания многомерной выборки на плоскость с осями “Признак 1” – “Признак 2”

Из диаграмм рассеивания многомерной выборки следует, что классы, визуально хорошо различимые в пространстве трех признаков. Различия классов проявляются и в проекции выборок на плоскость с осями “Признак 1” – “Признак 3” и в проекции на плоскость с осями “Признак 2” – “Признак 3”. Однако

классы достаточно плохо визуально различимы на плоскости в координатах “Признак 1” – “Признак 2” (рис. 2).

Основное диалоговое окно программного модуля моделирования многомерных нормальных выборок представлено на рисунке 3.

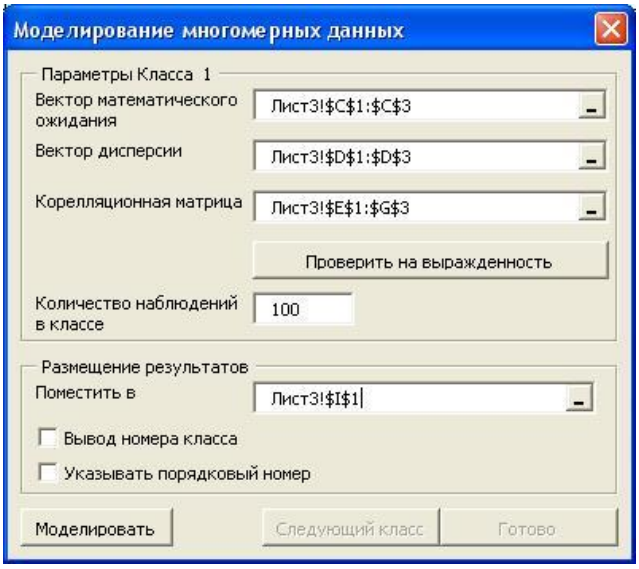

Рис. 3 Основное диалоговое окно программного модуля моделирования многомерных нормальных выборок

Рассмотренный алгоритм позволяет генерировать многомерные выборки непрерывных признаков. Признаки, получаемые в результате сбора анкетных данных, далеко не всегда носят числовой характер. Разнообразие типов признаков обусловлено разнообразием измерительных шкал, используемых в анкетах.

Использование различных измерительных шкал обусловлено не прихотью исследователя, а его стремлением предоставить вопросы в форме наиболее удобной для респондентов и в конечном итоге стремлением получить более достоверную информацию об изучаемом объекте или явлении.

Поэтому именно в анкетных данных получили наибольшее распространение признаки нечисловой природы [5]. Как показали многочисленные опыты, человек более правильно (и с меньшими затруднениями) отвечает на вопросы качественного, например, сравнительного, характера, чем количественного [4].

Теория обработки нечисловых данных находится только на этапе своего становления [4]. В связи с этим представляет большой интерес возможность компьютерного эксперимента с модельными нечисловыми признаками. Моде-

лирование статистической зависимости признаков нечисловой природы потребует гораздо большего количества параметров, чем для числовых признаков. Даже если решить такую задачу, графический образ выборки потеряет свою наглядность. Поэтому для моделирования многомерных выборок признаков нечисловой природы мы разработали специальный модуль, преобразующий многомерную выборку непрерывных значений признаков в многомерную выборку дискретных значений. В результате преобразования свойство зависимости признаков сохраняется. В программе задается количество уровней по всем признакам или по каждому в отдельности. Основное диалоговое окно программного модуля дискретизации представлено на рисунке 4.

Такой способ моделирования был использован нами для постановки экспериментов при исследовании возможностей методов повышения качества данных анкетных опросов [2]. Рассмотренные программные модули входят в состав разработанного нами программного комплекса обработки первичных данных анкетных опросов [3].

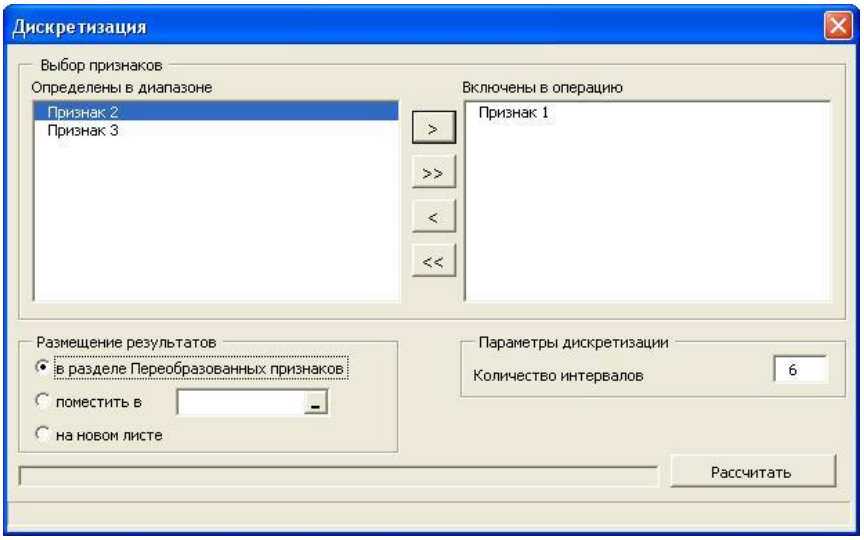

Рис. 4. Диалоговое окно программного модуля дискретизации

Разработанные программные средства моделирования данных могут быть полезны широкому кругу исследователей, производящих эксперименты с новыми методами обработки многомерных данных. Такие эксперименты позволяют выявить возможности разрабатываемых информационных технологий анализа многомерных данных. Кроме того, разработанные модули могут быть

использованы при постановке практических занятий для студентов, изучающих многомерные статистические методы

Список литературы

1. Ермаков С.М., Михайлов Г.А. Курс статистического моделирования. М.: Наука, 1976. 320 с.
2. Кустов Д.А., Методика повышения достоверности анкетных данных / Д.А. Кустов, Н.С. Мартышенко, // Интеллектуальный потенциал вузов – на развитие Дальневосточного региона России: Материалы VIII Международной конференции аспирантов и молодых ученых. 24–26 мая 2006 г. В 6 кн. Кн.2. Владивосток: Изд-во ВГУЭС, 2006. С. 24–39.
3. Мартышенко С.Н. Совершенствование математического и программного обеспечения обработки первичных данных в экономических и социологических исследованиях / С.Н. Мартышенко, Н.С. Мартышенко, Д.А. Кустов // Вестник ТГЭУ, 2006. № 2. С. 91–103
4. Орлов А.И. Нечисловая статистика. М.: МЗ-Пресс, 2004. 513 с
5. Толстова Ю.Н. Анализ социологических данных. Методология, дескриптивная статистика, изучение связей между номинальными признаками. М.: Научный мир, 2000. 352 с.