Секция «Социология»

О дефинициях понятия "методы многомерного анализа данных"

Кученкова Анна Владимировна
Аспирант
Российский Государственный Гуманитарный Университет, Социологический факультет, Москва, Россия
E-mail: anja.dernovaya@mail.ru

В сложившейся исследовательской и научной практике используются несколько понятий для обозначения математических методов анализа данных, применяющихся в социологии и других социальных науках. Например: методы математической статистики, методы анализа данных, методы многомерного анализа, интеллектуальный анализ данных (data mining). Одно из таких понятий – *методы многомерного анализа данных.* Важно понимать, что за ним стоит, каковы его границы, как оно соотносится с другими терминами, чем обусловлено его появление и распространение.

Наиболее близким к рассматриваемому понятию является термин «многомерный статистический анализ». Это одна из областей математической статистики, объектом изучения которой являются вектора: когда «каждое наблюдение представляется не одним-единственным числом, а некоторым конечным набором чисел, в котором в заданном порядке записаны все измеренные характеристики объекта» [1. С. 315]. Методы многомерного статистического анализа направлены на изучение многомерных данных, то есть на анализ распределений нескольких переменных одновременно. Они предназначены для решения таких задач [6. С. 5-6], как: исследование зависимостей между объектами и признаками; классификация объектов или признаков; снижение размерности пространства признаков. К этой области обычно относят регрессионный, дисперсионный, дискриминантный, факторный, кластерный анализ. Однако эти же виды анализа относят и к методам многомерного анализа данных, и к области data minig. В чм же тогда заключается различия этих терминов, каковы границы их употребления?

Ответ на этот вопрос кроется в том, что методы многомерной статистики неоднородны, делятся на две большие подгруппы. Первая включает методы, предназначенные для ситуаций, в которых исследуемый многомерный признак интерпретируется как многомерная случайная величина, а совокупность многомерных наблюдений – как выборка из генеральной совокупности. Выполнения ряда условий (допущений) «относительно природы многомерного (совместного) закона распределения вероятностей изучаемого многомерного признака» [1. С. 315-316] позволяет распространять полученные выводы на генеральную совокупность, строить доверительные интервалы для оценивания отдельных параметров. Это касается методов регрессионного, дисперсионного, дискриминантного анализа. Другие же (факторный, кластерный анализ) не основаны на указанных предположениях, они позволяют изучать структуру данных, однако, выводы, полученные на выборке объектов с помощью этих методов, нельзя распространить на генеральную совокупность. С этим связано то, что первая группа методов часто встречается в учебной литературе под названием «многомерный статистический анализ» [2, 3], вторая же обозначается просто как «многомерный анализ данных» [4, 5, 8]. Последнее понятие включает в себя не только методы, направленные на изучение случайных величин, значения которых распределены по закону нормального распределения, но

и на анализ жесткоструктурированных данных, не зависимо от формы распределения значений переменных.

Причинами подмены понятий, как показано в [7. С. 82-94], стала объективная специфика практики социологических исследований, выражающаяся в нарушении условий вероятностного порождения данных, отсутствии возможности проверить адекватность математической модели, необходимости анализа «чисел», полученных по шкалам низких типов.

«Классические» статистические методы рассчитаны в большей степени на так называемые «количественные» данные (переменные, измеренные на уровне выше порядкового), а также предполагающие подчинение распределений переменных закону нормального распределения. Эти условия далеко не всегда выполнимы в социологии, что привело к двум последствиям. Во-первых, стали разрабатываться альтернативные методы (например, статистика нечисловой природы), появляться новые направления (data mining). Во-вторых, получило распространение понятие «многомерного анализа данных», использующееся для обозначения методов анализа многомерных данных не зависимо от природы их происхождения и условий применения.

Литература

1. Айвазян С.А. Анализ многомерный статистический // Энциклопедический социологический словарь / РАН ИСПИ; под общ.ред. Г.В. Осипова. - М.: ИСПИ РАН, 1995. - 940 с.

2. Дубина И.Н. Математико-статистические методы в эмпирических социально-экономически исследованиях: учеб.пособие. – М.: Финансы и статистика, - 2010. – 416 с.

3. Толстова Ю.Н. Математико-статистические модели в социологии (математическая статистика для социологов): учеб.пособие. – М.: ГУ ВШЭ, 2008. – 243 с.

4. Крамер Д. Математическая обработка данных в социальных науках. Современные методы: учеб.пособие. – М.: Издательский центр «Академия», 2007. – 288 с.

5. Крыштановский А.О. Анализ социологических данных с помощью пакета SPSS: учеб.пособие. – М.: ГУ ВШЭ, 2006. – 281 с.

6. Многомерный статистический анализ в экономических задачах: компьютерное моделирование в SPSS: учеб. пособие. – М.: Вузовский учебник, 2009. – 310 с.

7. Толстова Ю.Н. Анализ социологических данных: методология, дескриптивная статистика, изучение связей между номинальными признаками. - М.: Научный мир, 2000. – 252 с.

8. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере: учеб.пособие. – М.: ИД «ФОРУМ», 2008. – 368 с.