Колонка редколлегии Editorial column

DOI: https://doi.org/10.26896/1028-6861-2020-86-7-5-6

ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКИЕ МОДЕЛИ ДАННЫХ — ОСНОВА МЕТОДОВ ПРИКЛАДНОЙ СТАТИСТИКИ

© **Александр Иванович Орлов**

PROBABILISTIC AND STATISTICAL DATA MODELS: A BASIS OF APPLIED STATISTICS METHODS

© **Aleksandr I. Orlov**

При обсуждении процедур анализа статистических данных обычно сосредотачивают внимание на расчетных формулах — не зная их, нельзя провести расчеты. Однако начинать надо с вероятностно-статистических моделей порождения изучаемых данных.

Например, в прикладной статистике наиболее распространенная модель выборки — это конечная последовательность независимых одинаково распределенных случайных величин[1], моделирующих результаты измерений (наблюдений, испытаний, опытов, анализов, обследований). Если общая функция распределения этих случайных величин произвольна, то обращаемся к методам непараметрической статистики. Для корректности математических рассуждений обычно принимают, что функция распределения результатов измерений непрерывна, следовательно, вероятность совпадения каких-либо двух результатов наблюдений (элементов выборки) равна нулю. Как известно, для реальных данных совпадения результатов встречаются достаточно часто. Следовательно, в таких случаях наблюдаются отклонения от непараметрической модели. Модель анализа совпадений при расчете непараметрических ранговых статистик представлена[2] в нашем журнале. Статистика интервальных данных создана для обработки округленных данных и данных с совпадениями.

До сих пор распространены реликтовые представления о том, что функция распределения результатов измерений относится к одному из популярных семейств распределений — нормальных, экспоненциальных, Вейбулла – Гнеденко, гамма-распределений и др. Для выборок из таких семейств в прошлом тысячелетии разработаны и изучены методы оценивания параметров и проверки статистических гипотез. Эта совокупность методов прочно заняла место в учебниках по теории вероятностей и математической статистике.

Отметим устойчивость предрассудков. Например, до сих пор пропагандируется использование метода максимального правдоподобия, хотя одношаговые оценки имеют столь же хорошие свойства, что и оценки максимального правдоподобия. Однако во многих случаях система уравнений максимального правдоподобия не имеет явного решения и соответствующие оценки рекомендуется находить итерационными методами, сходимость которых не изучают, хотя есть примеры, в которых отсутствие сходимости продемонстрировано. Между тем одношаговые оценки вычисляют по конечным формулам, без всяких итераций.

Особенно заметна любовь теоретиков к многомерным нормальным распределениям. Именно для таких распределений найдены явные формулы для различных характеристик в многомерном статистическом анализе, прежде всего в регрессионном. Причина в том, что удается использовать хорошо развитую в линейной алгебре теорию квадратичных форм.

Распределения почти всех реальных данных ненормальны. Это утверждение хорошо обосновано экспериментально, путем анализа результатов измерений[3]. Теоретические аргументы в пользу нормального распределения также не выдерживают критики. Например, говорят, что зависимость значения случайной величины от многих факторов влечет нормальность. Иногда добавляют, что факторы являются независимыми и сравнимыми по величине. Однако нормальность распределения можно ожидать лишь в случае аддитивной модели, когда факторы складываются (в силу Центральной предельной теоремы). Если же случайная величина формируется путем перемножения

[1] **Орлов А. И.** Прикладная статистика. — М.: Экзамен, 2006. — 671 с.

[2] **Орлов А. И.** Модель анализа совпадений при расчете непараметрических ранговых статистик / Заводская лаборатория. Диагностика материалов. 2017. Т. 83. № 11. С. 66 – 72.

[3] **Орлов А. И.** Распределения реальных статистических данных не являются нормальными / Научный журнал КубГАУ. 2016. № 117. С. 71 – 90.

(мультипликативная модель), то ее распределение является (в асимптотике) логарифмически нормальным. Если справедлива модель «самого слабого» звена (или «самого сильного»), т.е. значение случайной величины равно крайнему члену вариационного ряда значений факторов (соответственно минимуму или максимуму), то имеем в пределе распределение Вейбулла – Гнеденко.

Модель на основе семейства нормальных распределений или распределений из иного параметрического семейства можно сравнить с моделью поиска под фонарем потерянных в темных кустах ключей. Очевидно, под фонарем искать легче. Можно продемонстрировать активность. Однако надеяться на благоприятный исход поисков нельзя.

Из проведенного анализа следует необходимость использования непараметрических моделей распределений результатов измерений. Отметим, что интервалы их возможных значений, как правило, ограничены, т.е. распределения являются финитными. Следовательно, все моменты рассматриваемых случайных величин существуют и их выборочные аналоги могут использоваться в вычислениях.

Рассмотрим роль вероятностно-статистических моделей в многомерном статистическом анализе. Используют четыре основные класса регрессионных моделей.

Начнем с моделей первого типа — метода наименьших квадратов с детерминированной независимой переменной и параметрической зависимостью (линейной, квадратической и т.п.). Распределение отклонений произвольно (т.е. модель является непараметрической), для получения предельных распределений оценок параметров и регрессионной зависимости предполагаем выполнение условий Центральной предельной теоремы.

Второй тип моделей основан на выборке случайных векторов. Зависимость является параметрической, распределение двумерного вектора — произвольным. Об оценке дисперсии независимой переменной можно говорить только в модели на основе выборки случайных векторов, равно как и о коэффициенте детерминации как критерии качества модели[4].

Третий тип моделей регрессионного анализа, основанный на выборке случайных векторов, — непараметрическая регрессия, в которой как зависимость, так и отклонения от нее являются непараметрическими. Зависимость (как условное среднее) оценивается с помощью непараметрических оценок плотности.

Четвертый тип — промежуточный вариант — модель, в которой тренд линеен, а периодическая и случайная составляющие являются непараметрическими.

В моделях четвертого типа малые погрешности имеются в значениях как зависимой, так и независимой переменных. В прошлом этот раздел прикладной статистики назывался конфлюэнтным анализом, сейчас он входит в статистику интервальных данных.

К регрессионному анализу примыкают задачи сглаживания временных рядов и статистики случайных процессов, в которых отклонения от функции времени зависимы.

Анализ многообразия моделей регрессионного анализа приводит к выводу, что не существует единой «стандартной модели»[5]. Другими словами, при решении задачи восстановления зависимости необходимо начинать с выбора и обоснования вероятностно-статистической модели.

Необходимо исходить из теории измерений, согласно которой первый шаг при анализе данных — выявление шкал, в которых они измерены. Известно, что для данных, измеренных в порядковой шкале, в качестве средних величин можно использовать только члены вариационного ряда, прежде всего медиану, а применение среднего арифметического или среднего геометрического недопустимо. Как следствие, поскольку ранги или баллы, как правило, измерены в порядковой шкале, складывать их нельзя. В частности, нельзя оценивать успеваемость учащихся по среднему баллу экзаменационных оценок.

Статистические выводы должны быть инвариантны относительно допустимых преобразований шкал измерения данных. Значит, следует выяснить, какими алгоритмами анализа данных из рассматриваемого семейства можно пользоваться в данной шкале. Обратная задача — для определенного алгоритма анализа данных выяснить, в какой шкале можно им пользоваться. Коэффициент линейной парной корреляции Пирсона соответствует шкале интервалов, а непараметрические ранговые коэффициенты корреляции Спирмена и Кендалла позволяют изучать взаимосвязи порядковых переменных.

С позиций теории измерений обсудим метод анализа иерархий. Исходные данные — результаты парных сравнений, измеренные в порядковых шкалах. Результаты же расчетов выражены в шкале интервалов. С точки зрения теории измерений такое недопустимо. Следовательно, методом анализа иерархий пользоваться не следует. Рекомендуем применять адекватные методы анализа экспертных оценок, в частности, методы средних арифметических рангов, медиан рангов, согласования кластеризованных ранжировок[6].

---

[4] **Орлов А. И.** Ошибки при использовании коэффициентов корреляции и детерминации / Заводская лаборатория. Диагностика материалов. 2018. Т. 84. № 3. С. 68 – 72.

[5] **Орлов А. И.** Многообразие моделей регрессионного анализа (обобщающая статья) / Заводская лаборатория. Диагностика материалов. 2018. Т. 84. № 5. С. 63 – 73.

[6] **Орлов А. И.** Организационно-экономическое моделирование: учебник. В 3-х ч. Ч. 2. Экспертные оценки. — М.: Изд-во МГТУ им. Н. Э. Баумана, 2011. — 486 с.