Колонка редколлегии Editorial column

ПАРАМЕТРИЧЕСКИЕ И НЕПАРАМЕТРИЧЕСКИЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

© А. И. Орлов

Статья поступила 16 марта 2018 г.

PARAMETRIC AND NONPARAMETRIC STATISTICAL METHODS

© A. I. Orlov

Submitted March 16, 2018.

Статистические методы анализа данных (результатов изменений, наблюдений, испытаний, анализов, опытов, обследований) опираются на математическую статистику как теоретическую базу. Эта наука прошла в своем развитии ряд этапов, которые отразились в методической и справочной литературе. Каждый следующий этап в определенном смысле отрицает предыдущий, и противоречия между этапами иногда создают трудности у лиц, начинающих заниматься анализом данных. Обсудим одну из трудностей — часто обсуждаемую проблему соотношения параметрических и непараметрических статистических методов.

Математическая статистика как наука создана в начале XX в. Одна из основных задач того времени — описание статистических данных. Для ее решения К. Пирсон предложил использовать четырехпараметрическое семейство распределений. В настоящее время более популярны его подсемейства — нормальных распределений, экспоненциальных, логарифмически-нормальных, гамма-распределений, распределений Вейбулла – Гнеденко. Все они зависят от одного, двух или трех параметров, поэтому для полного описания распределения достаточно знать или оценить одно, два или три числа.

Следующий этап в развитии математической статистики — создание теории и алгоритмов оценивания параметров и проверки гипотез в предположении, что исходные данные описываются случайными величинами, распределения которых входят в то или иное подмножество четырехпараметрического семейства распределений Пирсона. На этом этапе математическая статистика пополнилась замечательными математическими теоремами, например, описывающими асимптотическое поведение оценок метода максимального правдоподобия и одношаговых оценок, задающих нижнюю границу дисперсии несмещенной оценки параметра (неравенство Рао – Крамера). Накопленные научные результаты позволили составить учебники математической статистики, по которым и сейчас проводится обучение.

Многие специалисты, связанные с анализом данных, и в настоящее время думают, что распределения рассматриваемых ими случайных величин являются нормальными. Сотрудничая с подобными специалистами, математики без сопротивления принимают указанный постулат нормальности и развивают математический аппарат статистики.

Есть ли основания априори предполагать нормальность результатов измерений?

Иногда утверждают, что если погрешность измерения (или иная случайная величина) определяется в результате совокупного действия многих малых факторов, то в силу Центральной Предельной Теоремы (ЦПТ) теории вероятностей эта величина хорошо приближается (по распределению) нормальной случайной величиной. Это утверждение, вообще говоря, неверно.

Точнее, такое утверждение справедливо, если малые факторы действуют аддитивно и независимо друг от друга. Если же они действуют мультипликативно (и независимо друг от друга), то в силу той же ЦПТ аппроксимировать распределение рассматриваемой величины надо логарифмически нормальным распределением. В прикладных задачах обосновать аддитивность, а не мультипликативность действия малых факторов обычно не удается.

Если же зависимость имеет общий характер, не приводится к аддитивному или мультипликативному виду, а также нет оснований принимать известные модели, дающие экспоненциальное, Вейбулла – Гнеденко, гамма или иные распределения, то о распределении итоговой случайной величины практически ничего не известно, кроме внутриматематических свойств типа регулярности в том или ином смысле.

При обработке конкретных данных иногда по традиции считают, что погрешности измерений имеют нормальное распределение. На предположении нормальности построены классические модели регрессионного, дисперсионного, факторного анализов, метрологические модели, которые еще продолжают встречаться как в отечественной нормативно-технической документации, так и в международных стандартах. На то же предположение опираются модели расчетов максимально достигаемых уровней тех или иных характеристик, применяемые при проектировании систем обеспечения безопасности функционирования экономических структур, технических устройств и объектов. Однако теоретических оснований для такого предположения нет. Необходимо экспериментально изучать распределения погрешностей.

Это было сделано. Оказалось, что практически все распределения реальных данных являются ненормальными. Такой вывод был сделан по результатам изучения многих тысяч выборок. Сводки экспериментальных данных приведены, например, в статьях А. И. Орлова[1]. Следовательно, параметрическая статистика не является адекватной при анализе реальных статистических данных. Необходимы другие инструменты статистического анализа, не опирающиеся на конкретный вид функций распределения.

Совокупность таких инструментов называют непараметрической статистикой. Эта область математической статистики развивается с довоенных времен. К третьему этапу развития математической статистики относятся, в частности, критерии Колмогорова, Смирнова, коэффициенты ранговой корреляции Спирмена и Кендалла. К настоящему времени с помощью непараметрической статистики можно решать тот же набор задач, что и с помощью параметрической статистики.

Несмотря на приведенные выше факты, у отдельных исследователей возникает желание применить тот или иной метод параметрической статистики. Они начинают с проверки нормальности. Если гипотеза нормальности не отклоняется, применяют алгоритмы, основанные на нормальности. К сожалению, такой подход некорректен. Дело в том, что отклонение от нормальности должно быть весьма выражено, чтобы обычно используемые критерии привели к отклонению гипотезы нормальности. Весьма полезен вывод о том, что по выборкам объема 6 – 50, как правило, не удается отличить нормальное распределение от других видов распределений[2].

Для определения функции распределения с точностью 0,01 с помощью критерия согласия Колмогорова необходимо несколько тысяч наблюдений[3], что для большинства задач прикладной статистики нереально. Поскольку функция $\Phi(x)$ стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1 и функция

$$\Psi(x) = e^x(1 + e^x)^{-1}$$

стандартного логистического распределения удовлетворяют соотношению

$$\sup_{x \in R^1} |\Phi(x) - \Psi(1{,}7x)| < 0{,}01,$$

то из сказанного следует, что различить по реальным данным нормальное и логистическое распределения почти всегда невозможно.

Учебники по математической статистике были составлены в первой половине XX в. и содержали результаты параметрической статистики. С небольшими модификациями эти учебники используются и в настоящее время, обычно в них упоминаются лишь отдельные непараметрические методы. Как правило, обсуждают критерий согласия Колмогорова, критерии однородности двух независимых выборок Смирнова и Вилкоксона, ранговые коэффициенты корреляции Спирмена и Кендалла. Отсутствие общего взгляда на непараметрическую статистику приводит к тому, что в среде пользователей статистических методов распространены различные заблуждения. Например, считают, что непараметрические методы — это методы ранговой статистики, в которой статистические критерии являются функциями от рангов наблюдений. Это заблуждение резко сужает сферу применения непараметрической статистики. Необходимо составить адекватное представление о непараметрической статистике, ее структуре[4] и отразить это в научной, учебной и методической литературе.

[1] **Орлов А. И.** Часто ли распределение результатов наблюдений является нормальным? / Заводская лаборатория. Диагностика материалов. 1991. Т. 57. № 7. С. 64 – 66; **Орлов А. И.** Распределения реальных статистических данных не являются нормальными / Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2016. № 117. С. 71 – 90.

[2] **Селезнев В. Д., Денисов К. С.** Исследование свойств критериев согласия функции распределения данных с гауссовой методом Монте-Карло для малых выборок / Заводская лаборатория. Диагностика материалов. 2005. Т. 71. № 1. С. 68 – 73.

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

[4] **Орлов А. И.** Структура непараметрической статистики (обобщающая статья) / Заводская лаборатория. Диагностика материалов. 2015. Т. 81. № 7. С. 62 – 72.