Колонка редколлегии Editorial column

DOI: https://doi.org/10.26896/1028-6861-2021-87-7-6-7

СМЕНА ПАРАДИГМ В ПРИКЛАДНОЙ СТАТИСТИКЕ

© **Александр Иванович Орлов**

CHANGE OF PARADIGMS IN APPLIED STATISTICS

© **Alexandr I. Orlov**

Статистические методы анализа данных широко применяются исследователями в различных областях науки. Обсудим смену парадигм прикладной статистики — изменения основ общепринятой модели действий в этой области математических методов исследования. Рассмотрим три парадигмы — примитивную, устаревшую, современную.

Поясним на примере. Исходя из примитивной парадигмы, применяют расчетные формулы критерия Стьюдента для проверки равенства нулю математического ожидания без какого-либо обоснования. Согласно устаревшей парадигме констатируют (без строгого обоснования), что результаты измерений имеют нормальное распределение, затем применяют критерий Стьюдента. В современной парадигме используют непараметрические методы, в рассматриваемой постановке — основанные на центральной предельной теореме[1].

Очевидно, обоснованность статистических выводов возрастает при переходе от примитивной парадигмы к устаревшей и далее — к современной. В настоящее время в практике научных работ используются все три парадигмы. Обсудим, как это влияет на качество результатов исследовательской деятельности.

Примитивная парадигма — следование кем-то составленным рецептам «поваренной книги». Программные продукты часто провоцируют такие расчеты. Приходится констатировать, что довольно часто итоговые выводы оказываются полезными с позиций прикладной области. Но иногда они могут быть и грубо ошибочными. Об опасности бездумного применения программных продуктов предупреждал[2] проф. В. В. Налимов, член раздела «Математические методы исследования» редколлегии нашего журнала в 1961 – 1997 гг.

В устаревшей парадигме (середина XX в.) элементы выборки рассматриваются как независимые случайные величины, распределения которых входят в то или иное параметрическое семейство распределений — нормальных, логистических, экспоненциальных, Вейбулла – Гнеденко, Коши, Лапласа, гамма-распределений и др. Все эти семейства выделены из четырехпараметрического семейства распределений, введенного основателем математической статистики К. Пирсоном в начале XX в. Он принял гипотезу, что распределения реальных данных всегда совпадают с каким-то элементом его четырехпараметрического семейства. Затем началось развитие теории параметрической математической статистики, в которой задачи оценивания и проверки гипотез решались для выборок из тех или иных параметрических семейств. Был получен ряд замечательных математических моделей и результатов, связанных, например, с методом максимального правдоподобия, критериями Пирсона (хи-квадрат), Пирсона, неравенством Рао – Крамера и др. Многомерное нормальное распределение оказалось весьма полезным для развития регрессионного и дискриминантного анализа.

Параметрической математической статистике посвящено основное содержание распространенных вузовских учебников по математической статистике. В отличие от примитивной парадигмы, имеется строгая математическая теория, позволяющая получать расчетные алгоритмы и на их основе — полезные практические рекомендации. Стоит отметить лишь один недостаток — распределения реальных данных, как правило, не являются нормальными и вообще не входят в четырехпараметрическое семейство Пирсона. Делают попытки проверить нормальность или, например, экспоненциальность реальных данных. Зачастую отклонить гипотезу нормальности не удается. Но это нельзя рассматривать как подтверждение нормальности распределения рассматриваемых данных, поскольку для тех же данных не удается отклонить гипотезу о том, что распределение данных соответствует другому популярному распределению. Причина очевидна — малый объем выборки. Например, для того чтобы выяснить, какому распределению соответствуют анализируемые данные — нормальному или логистическому, необходимо не менее 2500 наблюдений. Реальные объемы выборок обычно значительно меньше.

Развитие теории параметрической математической статистики продолжается и в настоящее

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

[2] **Налимов В. В.** Теория эксперимента. — М.: Наука, 1971. — 208 с.

время. В частности, сравнительно недавно выяснено, что вместо оценок максимального правдоподобия целесообразно использовать одношаговые оценки, разработаны методы доверительного оценивания для гамма-распределения и др. С помощью параметрической математической статистики решено много прикладных задач в конкретных областях исследования. Но в ряде случаев получены ошибочные выводы, хотя доля их заметно меньше, чем при опоре на примитивную парадигму.

Современная парадигма[3] основана на непараметрической и нечисловой статистике. В отличие от параметрической статистики предполагается, что элементы выборки с числовыми значениями имеют произвольную непрерывную функцию распределения. Центральной областью прикладной статистики стала статистика нечисловых данных[4], позволяющая единообразно подходить к анализу статистических данных произвольной природы.

Современную парадигму называем новой, хотя ее основы сформировались еще в 1980-х годах, когда во время подготовки к созданию Всесоюзной статистической ассоциации (учредительный съезд прошел в 1990 г.) понадобилось проанализировать состояние и перспективы прикладной статистики.

К настоящему времени непараметрическими методами можно решать практически тот же круг задач анализа данных, что и параметрическими. Преимущество непараметрики в том, что нет необходимости принимать необоснованные предположения о виде функции распределения. Недостатком является то, что реальные данные часто содержат совпадения. Если функция распределения элементов выборки непрерывна, то вероятность их совпадения равна нулю. Противоречие возникает из-за того, что свойства прагматических чисел, используемых для записи результатов измерений (наблюдений, испытаний, опытов, анализов, обследований), отличаются от свойств математических чисел (например, прагматические числа записываются с помощью конечного числа цифр, а почти все действительные числа теоретически требуют бесконечного ряда цифр). Разработаны подходы[5] к анализу совпадений при применении непараметрических статистик, позволяющие снять рассматриваемое противоречие.

В некоторых случаях параметрические методы позволяют обнаружить и предварительно изучить важные эффекты непараметрической статистики. Так, хорошо известно, что распределения реальных данных, как правило, не являются нормальными. Однако математический аппарат в случае нормальности зачастую является более простым. Согласно устаревшей парадигме в математической статистике широко используются многомерные нормальные распределения. Именно для таких распределений найдены явные формулы для различных характеристик в многомерном статистическом анализе, прежде всего в регрессионных постановках. Это связано с тем, что глубоко развита теория квадратичных форм в евклидовом пространстве (квадратичные формы стоят в степени экспоненты, описывающей плотность многомерного нормального распределения). Используя развитый математический аппарат, основанный на многомерной нормальности, удается разработать и изучить методы оценивания размерности вероятностно-статистической модели[6] в целях переноса полученных результатов на непараметрические постановки.

К настоящему времени теоретические исследования по прикладной статистике проводятся в основном в соответствии с современной парадигмой. Так, статистике нечисловых данных посвящено 63 % работ по прикладной статистике, опубликованных[7] в разделе «Математические методы исследования» журнала «Заводская лаборатория. Диагностика материалов» в 2006 – 2015 гг. Однако значительная доля прикладных работ осуществляется в традициях устаревшей или даже примитивной парадигм. Такие работы нецелесообразно огульно отрицать. Они могут приносить пользу в конкретных областях. Однако бесспорно, что переход на современную парадигму прикладной статистики повысит научный уровень исследований, а также позволит получить важные результаты в конкретных областях. Приходится констатировать, что исследователи, связанные с анализом данных, недостаточно знакомы с непараметрической и нечисловой статистикой. Необходимо шире распространять информацию о современной парадигме прикладной статистики.

---

[3] **Орлов А. И.** Новая парадигма прикладной статистики / Заводская лаборатория. Диагностика материалов. 2012. Т. 78. № 1. С. 87 – 93.

[4] **Орлов А. И.** Статистика нечисловых данных за сорок лет (обзор) / Заводская лаборатория. Диагностика материалов. 2019. Т. 85. № 7. С. 69 – 84.

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

[6] **Орлов А. И.** Оценивание размерности вероятностно-статистической модели / Научный журнал КубГАУ. 2020. № 162. С. 1 – 36.

[7] **Орлов А. И.** Развитие математических методов исследования (2006 – 2015 гг.) / Заводская лаборатория. Диагностика материалов. 2017. Т. 83. № 1. Ч. 1. С. 78 – 86.