ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 1 **МАТЕМАТИКИ** **Выпуск 2**
1994

НАГАЕВ А. В.

НЕКОТОРЫЕ ЭКСТРЕМАЛЬНЫЕ ЗАДАЧИ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Рассматривается ряд экстремальных задач, имеющих отношение к проблеме устойчивости статистических решений. Найдены явные выражения для возможного размаха квантиля данного уровня в классе симметричных унимодальных распределений, первые четыре момента которых совпадают с соответствующими моментами нормального распределения. На примере линейных оценок квантиля рассмотрена новая характеристика устойчивости оценивания. Приведены результаты вычисления расстояния между различными масштабно-сдвиговыми семействами в связи с проблемой выбора модели, объясняющей имеющиеся статистические данные.

Ключевые слова: устойчивость статистического решения, масштабно-сдвиговое семейство, расстояние по вариации, расстояние Хеллингера, линейная оценка, квантиль распределения.

Введение

Не только обоснование разумности, но и само применение многих статистических решающих процедур связано с решением тех или иных экстремальных задач. Простейшим примером задач подобного рода является вычисление значения оценки максимального правдоподобия в тех случаях, когда невозможно получить явное аналитическое выражение для этой оценки. К числу экстремальных задач математической статистики следует также отнести: вычисление максимального риска, связанного с использованием данной решающей процедуры, нахождение байесовской оценки, установление границы Рао–Крамера и т. п. Классическая лемма Неймана–Пирсона также представляет собой пример экстремальной задачи, которая решается замечательно просто, несмотря на глубину заключенного в ней смысла. Естественно, что поиск оптимального решающего правила всегда есть экстремальная задача той или иной сложности.

Все это традиционные задачи математической статистики. Однако, имеется целый ряд экстремальных задач, которые формально не являются задачами математической статистики. Например, отыскание расстояния между данными масштабно-сдвиговыми семействами распределений. Ясно, что, по сути, это — вычислительная проблема. Тем не менее, решение подобных проблем исключительно важно с точки зрения статистики малых и умеренных объемов наблюдений.

Близость масштабно-сдвиговых семейств говорит сразу о многом. Во-первых, зачастую становится бессмысленной задача выбора любого из этих семейств в качестве базовой модели, поскольку нет надежды различить их по выборке умеренного объема. Во-вторых, если одно из семейств допускает оптимальную решающую процедуру, то она окажется достаточно хорошей и в условиях второго.

Другим примером экстремальной задачи, непосредственно не являющейся задачей математической статистики, является задача об отыскании экстремальных значений данного функционала в некотором классе распределений. Это также чисто вычислительная проблема. Однако ее решение очень важно для прикладной статистики. Действительно, если нам известно распределение с точностью до принадлежности некоторому классу, то прежде всего хочется знать, как влияет подобная точность на интересующую нас характеристику этого распределения. Понятно, что можно сформулировать великое разнообразие таких задач. Ниже мы рассмотрим задачу об отыскании экстремальных значений квантиля заданного уровня.

Интуитивно понятно, что подобные постановки каким-то образом связаны с устойчивостью статистических решений.

Последнее время термин «устойчивость» постепенно вытесняется термином «робастность», хотя последний имеет вполне определенный смысл. Правильное употребление его связано с понятием засоренности выборки, т. е. попадания в нее наблюдений, значительно отличающихся от основной массы наблюдений. Робастность, строго говоря, есть малая чувствительность (устойчивость) статистического решения к засорению. В противовес ей устойчивость, как она понимается здесь, есть малая чувствительность к неточности определения истинного распределения. Решающее правило тем устойчивее, чем меньше изменяется риск, связанный с ее применением, при замене одного исходного распределения на другое.

Сама проблема устойчивости связана с тем, что обычно нет достаточных оснований для выбора конкретной модели. Гораздо чаще имеются веские основания в пользу некоторого класса моделей. Поэтому предпочтительнее заботиться не о выборе конкретной модели, а о подборе устойчивого решения.

Упомянем еще об одном важном обстоятельстве. Наиболее трудными задачами математической статистики являются задачи так называемого многомерного статистического анализа [1]–[3]. В большинстве из них выбор базовой модели на основе имеющегося эмпирического материала является безнадежным делом. Поэтому обоснованность применения статистических решающих процедур становится весьма сомнительной. С другой стороны, возможности современных компьютеров позволяют использовать целый набор эвристических алгоритмов — по сути дела, тех же решающих процедур. Отличие от традиционной статистики состоит здесь в том, что не делается никаких предположений о статистической природе наблюдений и, тем самым, нет возможности судить о риске, связанном с использованием подобного алгоритма. О качестве алгоритма обычно судят по результатам моделирования. Однако смоделировать все мыслимые ситуации невозможно. Поэтому естественно попытаться выделить некоторое, желательно небольшое, число «крайних» — наиболее плохих — ситуаций, моделируя которые, можно сделать заключение о возможном риске. Понятие «крайней» ситуации есть не что иное как нечетко сформулированная экстремальная задача, а возможный размах риска — естественная характеристика устойчивости алгоритма.

Совокупность эвристических алгоритмов и рекомендации по их использованию составляют существо так называемого анализа данных [4], [5]. Создается впечатление, что некоторое взаимное неприятие традиционных статистиков и представителей этого нового направления следовало бы несколько сгладить.

Поскольку в условиях умеренного объема наблюдений традиционная статистика и анализ данных, по существу, равнообоснованны, то полезно выработать некоторую методологию описания качества решающих процедур.

Именно с этой точки зрения следует рассматривать содержание данной работы. Полученные здесь результаты суть не что иное как попытка движения в указанном направлении.

Работа состоит из четырех разделов. Первый посвящен проблеме различимости масштабно-сдвиговых семейств, возникающей при выборе вероятностной модели, объясняющей имеющиеся данные. Наибольший интерес здесь представляют результаты вычисления расстояния Хеллингера и расстояния по вариации между различными масштабно-сдвиговыми семействами.

Во втором разделе сформулированы две теоремы, связанные с размахом квантиля заданного уровня в классе симметричных унимодальных распределений, первые четыре момента которых совпадают с моментами нормального распределения. Одна из этих теорем доказывается в последнем (четвертом) разделе. Доказательство другой будет приведено в следующих публикациях.

Третий раздел посвящен общим принципам оценки устойчивости линейных оценок квантиля с точки зрения размаха возможного смещения в заданном классе масштабно-сдвиговых семейств.