УДК 51: 101.8

В. М. Резников

Институт философии и права СО РАН
ул. Николаева, 8, Новосибирск, 630090, Россия

Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия
E-mail: rvm@philosophy.nsc.ru

МЕТОДОЛОГИЧЕСКИЕ ПРОБЛЕМЫ ПРИМЕНЕНИЯ СТАТИСТИЧЕСКИХ КРИТЕРИЕВ [*]

Вводится базовое понятие математической дисциплины. Понятие называется базовым, если фундаментальные результаты в данной дисциплине получены с использованием этого понятия. Такими понятиями в математической статистике являются распределение вероятностей и независимость данных. Показано, что самые сложные задачи статистики – проверка согласия данных гипотетическим распределениям и проверка независимости данных. Поэтому в курсе математической статистики для нематематиков особое внимание должно быть уделено овладению знаниями и умениями для практического анализа базовых понятий.

Ключевые слова: базовое понятие, фундаментальный результат, независимость, вероятностное распределение данных.

Проблема корректного применения статистических методов является значимой, так как среди математических наук наиболее популярной в приложениях является стохастическая математика. Прикладная теория вероятностей и математическая статистика популярны в биологии, медицине, психологии, а также в экономике и технических науках. Проблемы корректного применения математики возникли вместе с ее широким применением примерно в середине прошлого столетия. До этого времени математический аппарат применялся в основном физиками и астрономами, традиционно имевшими глубокую подготовку в математике. Кроме того, раньше математика использовалась в академическом режиме, так, например, астроном мог полжизни потратить на определение траектории движения исследуемой планеты. Осознание значимости математики для решения широкого круга проблем в различных областях знания связано с развитием прикладной математики, появлением компьютеров и прикладных компьютерных программ. Математика действительно стала популярной, во многих науках невозможно опубликовать статью, в которой нет статистических расчетов.

Практика реальных приложений статистического анализа показывает актуальность проблемы корректных применений статистики. Так, на примере анализа 250 диссертаций в биологии и медицине было показано, что в большинстве диссертаций применялись неадекватные или неоптимальные статистические методы, не применялись современные методы, которые являются менее зависимыми от точного соответствия данных выбранной статистической модели [Леонов, Ижевский, 1997]. Во многих диссертационных работах не проверялись условия применимости статистических методов. Авторы исследования [Там же] объясняют неудовлетворительное положение дел с применением вероятностной математики двумя связанными причинами. Во-первых, недостаточным вниманием к преподаванию вероятностной математики и недостаточно высокой подготовленностью преподавателей теории вероятностей и

[*] Работа выполнена при финансовой поддержке РФФИ (проект № 08-07-00272а) и Междисциплинарного интеграционного проекта СО РАН № 47.

математической статистики. Во-вторых, негативными социальными факторами. Известно, что когда генетика объявлялась лженаукой, то прекращалось преподавание биостатистики на биологических факультетах университетов. Мы полагаем, что каждая из названных причин влияет на неудовлетворительную ситуацию с применением стохастической математики. Однако существует и третья причина, связанная с неадекватностью математической статистики для приложений, это неадекватность статистического аппарата для корректных приложений [Резников, 2005]. Для улучшения качества преподавания и соответственно успешного овладения прикладным анализом некоторым базовым темам этой дисциплины необходимо уделить особое внимание. По нашему мнению, проблема выбора статистических критериев и методов, а также проверка условий их применимости должны быть определены как основные проблемы прикладного статистического анализа. Почему эти проблемы являются наиболее значимыми для приложений?

Во-первых, особенностью статистики является то, что проверка условий применимости статистических методов является неизмеримо более сложной проблемой, чем непосредственное применение самого статистического метода для решения исследуемой проблемы [Там же]. При всем многообразии статистических методов и моделей практически все они основаны на двух предположениях: данные имеют известное распределение вероятностей и(или) данные являются независимыми. Поэтому проверка соответствия данных гипотетическому распределению и проверка независимости данных являются базовыми проблемами статистического анализа. Если мы знаем вид теоретического распределения вероятностей, но не знаем некоторые или все параметры распределения, то методы статистики обеспечивают практически точное решение этой проблемы.

Однако знание теоретического распределения априори, но без знания параметров этого распределения не относится к типичной ситуации. Обычно не известно ни распределение данных, ни параметры распределения. В ситуации отсутствия гипотезы о распределении по данным строят гистограмму, по виду гистограммы формулируют гипотезу о виде распределения. Однако существуют известные сложности, связанные с формулированием гипотез. Дело в том, что по данным гипотеза не может быть однозначно сформулирована. Однозначность не достигается в силу недоопределенности теоретической величины эмпирическими данными. Поэтому иногда решается задача выбора наиболее подходящей гипотезы из нескольких гипотез. Если гипотеза определена правильно, но не полностью, как говорят, с точностью до неизвестных параметров, то современные методы оценивания параметров обеспечивают их безупречное определение, однако если данные даже в небольшой степени не соответствуют выбранному теоретическому распределению, то возникает проблема неробастности.

Современные методы статистики являются неробастными. Неробастность означает, что при небольшом несоответствии данных выбранной модели ошибки статистических расчетов являются непрогнозируемыми и ничем не ограниченными. Отметим, что некоторые классики статистики, в частности Р. Фишер, недооценивали трудности корректного описания распределения, адекватного данным. Фишер полагал, что схематичное изображение данных позволяет легко сформулировать гипотетическое распределение. А затем распределение уточняется путем оценивания параметров распределения. Однако он недооценил сложности проблемы недоопределенности теоретической величины данными исследований и проблему неробастности статистических методов [Фишер, 1958].

Во-вторых, все результаты о поведении статистических критериев имеют асимптотический характер. Они утверждают о сходимости полуэмпирического распределения к предельному теоретическому распределению, однако теоремы о сходимости доказывают только существование сходимости, но они не являются конструктивными. Поэтому теоремы не определяют объем данных, при котором полуэмпирическое распределение с достаточной точностью аппроксимирует теоретическое распределение.

В-третьих, для разного вида распределений оптимальными являются определенные критерии, поэтому выбор нужного критерия

является сложной проблемой. Так, критерий К. Пирсона хи-квадрат является универсальным, он адекватен для всех видов распределений, однако мощность критерия невелика. Критерий А. Н. Колмогорова – Е. Н. Смирнова имеет большую мощность, но он адекватен исключительно для непрерывных распределений. Не существует детальных описаний, какой критерий, в каком случае наиболее подходит, поэтому пользователь прикладного статистического анализа должен иметь достаточную квалификацию в области статистики, чтобы самостоятельно и корректно выбрать статистический критерий, так как он сам несет ответственность за его выбор. Имеет смысл решение проблемы разными методами, и если они приводят примерно к одинаковым решениям, то это является подтверждением правильности решения.

При анализе рекомендаций по применению статических методов и критериев возникают сложности, так как рекомендации являются вариабельными. В наибольшей степени вариабельность связана с группированием данных. Группирование означает разбиение данных на группы. Современные процедуры группирования не являются однозначно определенными. В настоящий момент известны три различных подхода к группированию данных. Это распределение данных в интервалы одинаковой длины, разбиение данных на интервалы, с одинаковой вероятностью попадания в каждый интервал, и разбиение на интервалы, при котором мощность критерия для проверки близких гипотез будет максимальной. Группирование выполняется весьма часто, например, оно является составной частью критерия Пирсона хи-квадрат. Данному критерию более ста лет, однако до сих пор рекомендации по его применению являются вариабельными. Алгоритм, реализующий критерий Пирсона, определяется следующими параметрами: нижняя граница объема данных, минимальное число интервалов, нижняя граница теоретически ожидаемого количества данных в группе, минимальное число данных в интервале, метод оценивания неизвестных параметров. Приведем некоторые рекомендованные оценки выделенных параметров. Нижняя граница объема данных от 50 до 400 [Мельник, 1983]. Для нижней границы числа групп предлагаются следующие оценки: 1) $r \geq \max(8, s + 1)$. Здесь r – число групп, s – число оцениваемых параметров. 2) Для выборки объемом 100 и более данных рекомендуют 8÷12 интервалов [Айвазян и др., 1983]. Вариабельными являются рекомендации для нижней границы теоретически ожидаемого количества данных в группе. Этот параметр обозначим символом e. Б. А. Севастьянов рекомендует $e \geq 10$, Н. Кендалл и А. Стюарт $e \geq 5$, Б. Ван-дер-Варден принимает $e = 1$ [Резников, 2005]. Рекомендации для минимального объема данных в группе варьируют от двух до десяти, иногда допускается, что в некоторых группах данных нет. В качестве методов оценивания предлагались следующие методы: метод минимума хи-квадрат, метод моментов по группированным данным, метод максимального правдоподобия по негруппированным данным. Вариабельность рекомендаций объясняется тем фактом, что для малых объемов данных имеет место неодинаковая сходимость полуэмпирической функции распределения критерия к теоретическому распределению хи-квадрат. Для правильной алгоритмизации критерия Пирсона предполагается оптимальное определение значений выделенных параметров. В работах Б. Ю. Лемешко путем моделирования определены оптимальные значения параметров алгоритма, реализующего критерий Пирсона [Лемешко, Постовалов, 2003].

От анализа распределений перейдем к анализу независимости.

В науке и в обыденной жизни роль независимости трудно переоценить. Подтверждение одного и того же факта независимо проведенными исследованиями с помощью разных методов повышает надежность полученного результата. А. Н. Колмогоров полагал, что после корректного определения вероятности вопрос о независимости является самым главным. Аргументы А. Н. Колмогорова касаются выделения теории вероятностей в самостоятельную науку и развития теории вероятностей как математической науки. По его мнению, благодаря особому интересу к проблеме независимости теория вероятностей отделилась от абстрактной теории меры. Он считал, что в классических доказательствах фундамен-

тальных теорем (в них определяются условия, при которых исследуемое событие почти обязательно происходит, т. е. происходит с единичной вероятностью) обязательно использовалась идея независимых экспериментов.

Значимость независимости для корректного применения прикладной статистики показана Ю. И. Алимовым [Алимов, Кравцов, 1992]. Корректная эмпирическая проверка независимости факторов *A* и *B* предполагает, что экспериментально получены частоты *m*(*A*) и *m*(*A*/*B*), и *m*(*A*) ≈ *m*(*A*/*B*). Здесь *m*(*A*) это частота события *A*, *m*(*A*/*B*) это условная частота события *A* при условии события *B*. Символ ≈ обозначает примерное равенство. Для анализа независимости в многофакторных (в частности, в трехфакторных) экспериментах необходимо определить частоты *m*(*A*) и *m*(*A*/*BCD*) и проверить их равенство. Учитывая трудоемкость полуэкспериментальной проверки независимости, в статистике разработаны частные формальные схемы, учитывающие известную информацию об исследуемых данных. Однако в современной статистике нет общего формального определения независимости. Так, для проверки независимости линейных связей используют линейный коэффициент корреляции. Если он равен нулю, то данные независимы, а если он равен ±1, то данные связаны прямо пропорциональной или обратно пропорциональной зависимостями. Коэффициент корреляции адекватен и для нормально распределенных данных. Если данные не связны линейным образом и не соответствуют нормальному распределению, то для определения независимости используют другие идеи. Так, для независимых событий верно, что их совместная вероятность мала. Так как из курса математической статистики известно [Айвазян и др., 1983], что для описания маловероятных событий адекватно распределение С. Пуассона, то соответствие данных этому распределению является доказательством их независимости.

На практике модель независимых экспериментов обычно принимают не на основе эмпирического анализа или формальных рассуждений, а с помощью интуитивно содержательных соображений. Приведем типичные аргументы, используемые для обоснования независимости данных. Во-первых, полагают, что если эксперименты проводились в полностью контролируемых условиях, то естественно принять модель независимых экспериментов. Однако за пределами физики говорить о полном контроле проводимых экспериментов нет оснований. Даже в физике бывает проблема с контролем условий проводимых экспериментов. Из истории физики известно, что когда в лаборатории Э. Резерфорда был открыт родон, то все приборы зашкаливали. Дело в том, что заранее не было известно, что новое вещество является газообразным. Когда догадались, что новое вещество является газом, то после рафинирования условий проведения экспериментов добились правильных показаний приборов. Во-вторых, к модели независимости приходят на основе содержательного анализа, при этом иногда ссылаются на А. Н. Колмогорова, который говорил, что в рамках математики невозможно полностью определить условия, при которых адекватна идея независимых экспериментов, и эта проблема относится к философскому анализу.

Известно, что начиная с Р. Декарта философский анализ апеллирует к интуиции. В то же время особенностью стохастической математики является то, что многие результаты в этой области являются неинтуитивными [Секей, 2003]. Для стохастической математики неизвестны собственные логические парадоксы, однако многие результаты являются неинтуитивными, или неудобными для приложений. Проблеме неинтуитивности в вероятностной математике посвящен ряд монографий [Секей, 2003; Стоянов, 1999]. Естественно, что интуиция изменяется в результате приобретения знаний в специальной области. Приведем неинтуитивный на первый взгляд результат, который может служить в качестве индикатора знаний студентов в теории вероятностей. Пусть события *A* и *B* определяются как попадание в одноименные области на плоскости, причем эти области не пересекаются. Встает вопрос, являются ли эти события независимыми? Для человека, не изучавшего теорию вероятностей, события *A* и *B* представляются независимыми.

Проверим, так ли это на самом деле. Для независимых событий верна следующая формула:

$$P(A \wedge B) = P(A) \times P(B). \qquad (1)$$

Здесь $P(A)$, $P(B)$ и $P(A \wedge B)$ – соответственно вероятности событий: A, B и $A \wedge B$.

Имеем $P(A \wedge B) = P(\varnothing) = 0$, однако в общем случае $P(A) \neq 0$ и $P(B) \neq 0$, поэтому формула (1) не выполняется, и независимости нет.

По нашему мнению, принятие независимости на основе интуитивно-содержательных результатов не является обоснованным, скорее верно обратное осторожное суждение. Если экспериментально установлено, что частота $m(A)$ события A равна условной частоте $m(A/B)$ этого события при условии B, то есть $m(A) \approx m(A/B)$, или на основе формального анализа установлена независимость событий A и B, то имеются определенные основания для гипотезы о содержательной независимости явлений.

Проблема определения условий, при которых может быть принята модель независимых наблюдений, относится к реальной практике применения статистического анализа. Во многих науках, например в генетике, пренебрегают коэффициентами корреляции, меньшими, чем 0,01. На первый взгляд это вполне интуитивно и естественно. Однако пренебрежение малокоррелированными связями не всегда является оправданным. В работе П. С. Эльясберга приведен пример расчета дисперсии для двух групп случайных величин [Эльясберг, 1986]. В первом случае рассматривалась 1000 независимых случайных величин, а во втором 1000 случайных величин с коэффициентом корреляции 0,01. Оказалось, что при таком достаточно большом объеме данных дисперсии в двух группах отличались более чем на порядок, причем в группе, где отбрасывали малокоррелированные связи, дисперсия получилась неоправданно оптимистичной.

В заключение отметим, что в современном курсе статистического анализа для нематематиков нет смысла в детальном доказательстве теорем, относящихся к теоретической статистике. Для прикладников прагматически значимым является умение применить статистический метод на практике. Корректное применение статистических методов предполагает проверку их условий применимости. Наибольшую сложность представляет проверка адекватности данных базовым понятиям статистики. Понятие называется базовым в конкретной математической дисциплине, если оно удовлетворяет двум условиям. Во-первых, с его помощью доказываются наиболее значимые результаты в данной дисциплине. Во-вторых, оно логически невыводимо на основе других понятий. Таким понятием в математической статистике является понятие распределения. Другим важнейшим понятием является понятие независимости. Так как проверка независимости данных и их соответствие гипотетическим распределениям являются самыми сложными проблемами статистики, то в курсе современного прикладного анализа этой тематике должно быть уделено максимальное внимание.

Список литературы

Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983. 471 с.

Алимов Ю. И., Кравцов Ю. А. Является ли вероятность «нормальной» физической величиной? // Успехи физических наук. 1992. Т. 162, № 7. С. 149–182.

Лемешко Б. Ю., Постовалов С. Н. Компьютерное моделирование как способ познания статистических закономерностей в технике // Вероятностные идеи в науке и философии: Материалы регион. науч. конф. (с участием иностр. ученых) 23–35 сентября 2003 г. Новосибирск: ИФПР СО РАН, 2003. С. 102–105.

Леонов В. П., Ижевский П. В. Об использовании прикладной статистики при подготовке диссертационных работ по медицинским и биологическим специальностям // Бюллетень Государственного Высшего Аттестационного комитета Российской Федерации. 1997. № 3. С. 56–61.

Мельник М. Основы прикладной статистики. М.: Энергоатомиздат, 1983. 414 с.

Резников В. М. Вероятностные концепции: анализ оснований и приложений. Новосибирск, 2005. 158 с.

Секей Г. Парадоксы в теории вероятностей и математической статистике. М.: Мир, 2003. 273 с.

Стоянов Й. Контрпримеры в теории вероятностей. М.: Факториал, 1999. 288 с.

Фишер Р. А. Статистические методы для исследователей. М.: Гос. стат. изд-во, 1958. 268 с.

Эльясберг П. С. Вычислительная информация: Сколько ее нужно? Как ее обрабатывать? М.: Наука, 1986. 208 с.

Материал поступил в редколлегию 01.09.2009

V. M. Reznikov

METHODOLOGICAL PROBLEMS OF APPLICABILITY OF STATISTICAL CRITERIA

A base notion for a mathematical discipline is introduced. The notion is base if fundamental results in the field were acquired with the use of this notion. The base notions in mathematical statistics are distribution of probabilities and independence of data. It is demonstrated that the most difficult problems of statistics are goodness-of-fit of data to hypothetical distributions and tests of independence. So a course on mathematical statistics for non-mathematicians has to focus on mastering knowledge and skills for practical analysis of the base notions.

Keywords: basic notion, fundamental result, independency, data probability distribution.