Методы оценки близости допредельных и предельных распределений статистик

 Рассматривается проблема оценки близости предельных распределений статистик и распределений, соответствующих конечным объемам выборок. При каких объемах выборок уже можно пользоваться предельными распределениями? Каков точный смысл термина "можно" в предыдущей фразе? Основное внимание уделяется переходу от точных формул допредельных распределений к пределу и применению метода статистических испытаний (Монте-Карло). Обсуждаются "подводные камни" на пути исследователя в рассматриваемой области.

**1. Асимптотическая математическая статистика**

**и практика анализа статистических данных**

 Как мы обычно подходим к обработке реальных данных в конкретной прикладной задаче? Первым делом строим статистическую модель. Если мы хотим перенести выводы с совокупности результатов наблюдений на более широкую совокупность, например, предсказать что-либо, то рассматриваем, как правило, вероятностно-статистическую модель. Например, традиционную модель выборки, в которой результаты наблюдений - реализации независимых (в совокупности) одинаково распределенных случайных величин. Очевидно, ***любая модель лишь приближенно соответствует реальности.*** В частности, естественно ожидать, что распределения результатов наблюдений несколько отличаются друг от друга, а сами результаты связаны между собой, хотя и слабо. И эти ожидания во многих конкретных случаях оправдываются (в терминах конкретной прикладной ситуации см. об этом, например, в монографии [1]).

 Итак, первый этап - переход от реальной ситуации к математической модели. Далее - неожиданность: на настоящем этапе своего развития математическая теория статистики зачастую не позволяет провести необходимые исследования для имеющихся объемов выборок. Более того, отдельные математики пытаются оправдать свой отрыв от практики соображениями о структуре этой теории, на первый взгляд убедительными. Неосторожная давняя фраза Б. В. Гнеденко и А. Н. Колмогорова: "Познавательная ценность теории вероятностей раскрывается только предельными теоремами" [2] взята на вооружение и более близкими к нам по времени авторами. Так, И. А. Ибрагимов и Р. З. Хасьминский пишут: "Решение неасимптотических задач оценивания, хотя и весьма важное само по себе, как правило, не может являться объектом достаточно общей математической теории. Более того, соответствующее решение часто зависит от конкретного типа распределения, объема выборки и т. д. Так, теория малых выборок из нормального закона будет отличаться от теории малых выборок из закона Пуассона" [3, с.7].

 Согласно цитированным и подобным им авторам, основное содержание математической теории статистики - предельные теоремы, полученные в предположении, что объемы рассматриваемых выборок стремятся к бесконечности. Эти теоремы опираются на предельные соотношения теории вероятностей, типа Закона Больших Чисел и Центральной Предельной Теоремы. Ясно, что сами по себе подобные утверждения относятся к математике, т. е. к сфере чистой абстракции, и не могут быть непосредственно применены для анализа реальных данных. Их использование опирается на важное предположение: *"При данном объеме выборки достаточно точными являются асимптотические формулы. "*

 Конечно, в качестве первого приближения представляется естественным воспользоваться асимптотическими формулами, не тратя сил на анализ их точности. Но это - лишь начало долгой цепи исследований. Как же обычно преодолевают разрыв между результатами асимптотической математической статистики и потребностями практики статистического анализа данных? Какие "подводные камни" подстерегают на этом пути? Обсуждению этих вопросов и посвящена настоящая статья.

**2. Точные формулы и асимптотика**

 Начнем с наиболее продвинутой в математическом плане ситуации, когда для статистики известны как предельное распределение, так и распределения при конечных объемах выборки.

 Примером является двухвыборочная односторонняя статистика Н.В.Смирнова. Рассмотрим две независимые выборки объемов m и n из непрерывных функций распределения F(x) и G(x) соответственно. Для проверки гипотезы однородности двух выборок

H0 : F(x) = G(x) для всех действительных чисел x

в 1939 г. Н.В.Смирнов в статье [4] предложил использовать статистику

D+(m,n) = sup ( Fm(x) - Gn(x) ) ,

где супремум берется по всем действительным числам x. Для обсуждения проблемы соотношения точных и предельных результатов ограничимся случаем равных объемов выборок, т.е. m = n. Положим

H(n, t) = P ( D+(n,n) $ t n - 1/2) .

В цитированной статье [4] Н,В. Смирнов показал, что при безграничном возрастании объема выборки n вероятность H(n, t) стремится к exp ( - t 2 ).

 В работе [5] 1951 г. Б.В.Гнеденко и В.С.Королюк показали, что при целом с = t n1/2 (именно при таких t вероятность H(n, t) как функция t имеет скачки, поскольку статистика Смирнова D+(n,n) кратна 1/ n ) рассматриваемая вероятность H(n, t) выражается через биномиальные коэффициенты, а именно,

 (1).

 К сожалению, непосредственные расчеты по формуле (1) возможны лишь при сравнительно небольших объемах выборок, поскольку величина n!.уже при n=100 имеет более 200 цифр и не может быть без преобразований использована в вычислениях. Следовательно, наличие точной формулы для интересующей нас вероятности не снимает необходимости использования предельного распределения и изучения точности приближения с его помощью.

 Широко известная формула Стирлинга для гамма-функции и, в частности, для факториалов позволяет преобразовать последнее выражение в асимптотическиое разложение, т.е. построить бесконечный степенной ряд (по степеням n ) такой что каждая следующая частичная сумма дает все более точное приближение для интересующей нас вероятности H(x, t) . Это и было сделано в работе А.А.Боровкова [6], опубликованной в 1962 г. Большое количество подобных разложений для различных статистических задач приведено в работах [7-9] В.М.Калинина и О.В. Шалаевского в конце 60-х - начале 70-х годов. (Интересно отметить, что асимптотические разложения в ряде случаев расходятся, т.е. остаточные члены имеют нетривиальную природу.)

 В наших работах конца семидесятых годов была сделана попытка теоретически оценить остаточный член второго порядка. Итоги подведены в статье [10] и монографии [11, § 2.2, с.37-45]. Справедливо равенство

H(n, t) = exp ( - t 2 ).(1 + f(t)/n + g(n,t)/ n2 ),

где

f(t) = t2 (1/2 - t2/ / 6 ).

Целью указанных работ было получение равномерных по n, t оценок остаточного члена второго порядка g(n,t) сверху и снизу в области, задаваемой условиями

0 < t n - 1/2 < А, 0 < t < t max , n $ n0 . (2)

С помощью длинных цепочек оценок остаточных членов в формулах, получаемых при преобразовании формулы (1) к предельному виду, сформулированная выше цель была достигнута, и для различных наборов параметров А, t max , n0 получены равномерные по n, t оценки остаточного члена второго порядка g(n,t) сверху и снизу в области (2). Так, например, при А = 0,5, t max = 1,73, n0 = 8 нижняя граница равна (- 0,71), а верхняя есть 2,65.

 Основными недостатками такого подхода являются являются , во первых, зависимость оценок от параметров А, t max , n0 , задающих границы областей, во-вторых, завышение оценок, иногда в сотни раз, обусловленное желанием получить равномерные оценки по области (оценкой реальной погрешности в точке является значение следующего члена асимптотического разложения).

 Поэтому при составлении рассчитанной на практическое использование методики [12] проверки однородности двух выборок с помощью статистики Смирнова мы перешли на другую методологию (назовем ее "методологией заданной точности"), которую кратко можно описать следующим образом.

а) Выбирается достаточно малое число р, например р = 0,05 или р = 0,20.

б) Приводятся точные значения H(n, t) для всех значений n таких, что

| H(n, t) - exp ( - t 2 ) | > p exp ( - t 2 ) .

в) Если же последнее неравенство не выполнено, то предлагается пользоваться вместо H(n, t) предельным значением exp ( - t 2 ).

 Таким образом, принятая нами в методике [12] методология предполагает интенсивное использование вычислительной техники. Результат расчетов - граничные значения объемов выборок n(p,t) такие, что при меньших значениях выброк рекомендуется пользоваться точными значениями, а при больших - предельными, - описывается таблицей, а не формулой. Отметим, что при построении реальных таблиц не обойтись без выбора того или иного конкретного значения р, задающего объемы таблиц.

**3. Оценки скорости сходимости**

 Теоретические оценки скорости сходимости в различных задачах математической статистики иногда формулируются в весьма абстрактном виде. Так, в 60-70-х годах была популярна задача оценки скорости сходимости распределения классической статистики омега-квадрат (Крамера-Мизеса-Смирнова). Для максимума модуля допредельной и предельной функций распределения этой статистики различные авторы доказывали, что для любого e>0 существует константа С(e) такая, что упомянутый максимум не превосходит С(e) n - w + e . Прогресс состоял в увеличении константы w. Сформулированный выше результат был доказал последовательно для w = 1/10, 1/6, 1/5, 1/4, 1/3, 1/2 и 1 (подробнее история этих исследований рассказана в § 2.3 монографии [11]).

 Конечно, все эти исследования не могли дать конкретных практических рекомендаций. Однако необходимой исходной точкой является само существование предельного распределения. Представим себе, что некто, не зная, что у распределения Коши нет математического ожидания, моделирует выборочные средние арифметические наблюдений из этого распределения. Ясно, что его попытки оценить скорость сходимости выборочных средних к пределу обречены на провал.

 Последовательное улучшение теоретических оценок скорости сходимости дает надежду на быструю реальную сходимость. Действительно, как показано в статье [13], предельным распределением для указанной статистики можно пользоваться уже при объеме выборки, равном 4.

**4. Использование датчиков псевдослучайных чисел**

 Если же предельное распределение известно то возникает возможность изучить скорость сходимости численно методом статистических испытаний (Монте-Карло). Именно так поступила Г.В.Рыданова в своей диссертации [14], реализуя описанную выше "методологию заданной точности". ПРи этом возникли две проблемы.

 Во-первых, откуда известно, что скорость сходимости монотонна? Если при данном объеме выборки различие мало, то будет ли оно мало и при дальнейших? Иногда отклонения допредельного распределения от предельного объясняются довольно сложными причинами. Так, для распределения хи-квадрат они связаны с до сих пор не решенными теоретико-числовыми проблемами о числе целых точек в эллипсоиде растущего диаметра. "Подводные камни", связанные с распределением хи-квадрат. разобраны в статье М. Мирвалиева М.С.Никулина [15].

 Во-вторых, с помощью датчиков псевдослучайных чисел получаем допредельные распределения с погрешностью, которая может преуменьшать различие. Поясним мысль аналогией. Растущий сисгнал измеряется с погрешностями. Когда можно гарантировать, что его величина наверняка превзошла заданную границу? Не будем здесь обсуждать известные подходы к решению этой задачи.

 Проблема качества датчиков псевдослучайных чисел продолжает оставаться открытой. В журнале "Заводская лаборатория" с 1985 г. по 1993 г. продолжалась активная дискуссия по этой проблеме, завершившаяся статьей С.М.Ермакова [16] и нашим комментарием [17] к нему. Для моделирования в пространствах фиксированной размерности датчики псевдослучайных чисел решают поставленные задачи. Но для рассматриваемых нами задач размерность не фиксирована - мы не знаем, при каком конкретно объеме выборки можно переходить к предельному распределению согласно "методологии заданной точности".

 Нужны дальнейшие работы по изучению качества датчиков псевдослучайных чисел в задачах неопределенной размерности. Поскольку критиков датчиков обычно обвиняют в том, что они сами их не используют, отмечу, что мы применяли этот инструментарий при изучении помех, создаваемых электровозами [11], при изучении статистических критериев проверки однородности двух выборок [18].

**5. А нужна ли вообще асимптотика?**

 В журнале "Заводская лаборатрия" в последние годы опубликован ряд работ Б.Ю.Лемешко. Они посвящены актуальному направлению прикладной статистики, связанному с интенсивным использованием вычислительной техники для изучения свойств статистических процедур. В диссертации Б.Ю.Лемешко [19] подводятся итоги более чем двадцатилетней (с 1973 г.) работы автора (в составе группы исследователей под руководством проф.В.И.Денисова).

 Как уже отмечалось, математические методы в статистике обычно позволяют получать лишь асимптотические результаты, и для переноса выводов на конечные объемы выборок приходится применять вычислительные методы. Диссертантом разработан и успешно применяется оригинальный подход, основанный на интенсивном использовании современной вычислительной техники. Основная идея такова: в качестве альтернативы асимптотическим методам математической статистики используется анализ результатов статистического моделирования (порядка 2000 испытаний) выборок конкретных объемов (200, 500, 1000). При этом анализ предельных распределений заменяется на анализ распределений соответствующих статистик при указанных объемах выборок.

 К достоинствам подхода диссертанта относится возможность замены теоретических исследований расчетами. Разработанная в исследовательском коллективе программная система дает в принципе возможность численно изучить свойства любого статистического алгоритма для любого конкретного распределения результатов наблюдений и любого конкретного объема выборки. К недостаткам подхода Б.Ю.Лемешко относится зависимость от свойств датчиков псевдослучайных чисел (проблемам качества таких датчиков посвящена упомянутая выше дискуссия в журнале "Заводская лаборатория" в 1985-1993 гг.), а также - что более важно - неизвестность предельного распределения (и даже самого факта его существования), а потому невозможность обоснованного переноса полученных выводов на объемы выборок, отличные от исследованных. Поэтому с точки зрения теории математической статистикии полученные диссертантом результаты следует пока рассматривать как правдоподобные (а не доказательные, как в классической математической статистике).

 Кроме того, они принципиально неточные. Даже в наиболее благоприятных условиях отклонения смоделированного распределения от теоретического предельного, по нашей оценке, может иметь порядок (1/2000 + 1/1000)1/2 = 0,038. Это означает, в частности, что процентные точки, сответствующие уровням значимости 0,05 и особенно 0,01, рассчитанные Б.Ю.Лемешко, могут сильно отличаться от соответствующих процентных точек предельных распределений. Очевидно, следующий этап работ - изучение точности полученных в диссертации выводов, прежде всего приближений и процентных точек.

 Однако сразу все не сделаешь. Поэтому Б.Ю.Лемешко совершенно прав, развивая новые компьютерные подходы к давним задачам прикладгной математической статистики. В частности, весьма полезными и интересными являются результаты, касающиеся непараметрических критериев согласия. Весьма интересным и полезным представляется также метод построения оптимального группирования, в частности, при использовании критериев типа хи-квадрат. Важен результат о неробастности (неустойчивости) оценок максимального правдоподобия по негруппированным данным. Надо поддержать идею использования одновременно двух оценок по группированным данным с использованием как оптимального, так и раввновероятного группирования. Этот подход диссертанта соответствует современным идеям в области устойчивости (робастности) статистических выводов, в частности, подходу монографии [11].

 На автора данной работы большое впечатление произвела статья Б.Р.Левина и Н.О.Демидовича [20], в которой сравниваются два плана контроля надежности. Оказывается, чтопри объемах выборки, меньших 150, лучше первй план, а при больших 150 - второй. Значит, если бы по методу Б.Ю.Лемешко сравнивались эти планы при n=100, то лучшим был бы признан первый план, что неверно.

 Другая относящаяся к делу ассоциация - из весьма содержательной монографии [21]. Будем суммировать бесконечный ряд с членами zn= 1/ n . Поскольку члены его убывают, то обычно используемые алгоритмы остановят вычисления на каком-то шагу. А сумма-то - бесконечна!

 Итак, Б.Ю.Лемешко предложил интересный инструментарий и проделал полезную работу, но его подход никоим образом не является панацеей.

**6. Необходимость скоординированных исследований**

 В прикладной математической статистике давно назрела необходимость координации новых исследований и критического анализа накопленных результатов. В статье [22] была сформулирована и обоснована программа превращения этой сферы научно-практических исследований в организованную отрасль науки наподобие метрологии. В статье [23] рассказано о первых шагах такой работы, предпринятой в рамках Центра статистических методов и информатики и Российской ассоциации статистических методов. К сожалению, развитие общей экономической ситуации в России ставит под сомнение не только возможность подобных нововведений, но и само существование современной науки.

Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект 97-06-80033).

**Литература**

1. Эльясберг П.Е. Измерительная информация. Сколько ее нужно, как ее обрабатывать? - М.: Наука, 1983. 208 с.

2. Гнеденко Б.В., Колмогоров А.Н. Предельные распределения для сумм независимых случайных величин. - М.-Л.: ГИТТЛ, 1949. 264 с.

3. Ибрагимов И.А., Хасьминский Р.З. Асимптотическая теория оценивания. -М.: Наука, 1979. 528 с.

4. Смирнов Н.В. / Бюлл. МГУ, Сер.А, 1939, т.2, № 2, с.3-14.

5. Гнеденко Б.В., Королюк В.С./ Докл. АН СССР, 1951, т.80, № 4, с.525-528.

6. Боровков А.А. / Изв. АН СССР, Сер. матем., 1962, т.26, с.605-624.

7. Калинин В.М. / Труды Матем. ин-та им. В.А.Стеклова АН СССР, 1968, т.104, с.88-134.

8. Калинин В.М. / Труды Матем. ин-та им. В.А.Стеклова АН СССР, 1970, т.111, с.163-194.

9. Калинин В.М., Шалаевский О.В. / Записки научн. семинаров Ленингр. отд-ния Матем. ин-та им. В.А.Стеклова АН СССР, 1972, т.26, с. 3-152.

10. Орлов А.И., Орловский И.В. / Статистические методы. Межвузовский сборник научн. трудов. - Пермь: Пермский гос. ун-т, 1978, с.100-109.

11. Орлов А.И. Устойчивость в социально-экономических моделях. - М.: Наука, 1979. - 296 с.

12. Методика. Проверка однородности двух выборок параметров продукции при оценке ее технического уровня и качества. Первая редакция. - М.: Всесоюзный научно-исследовательский институт стандартизации Госстандарта СССР, 1987. - 116 с.

13. Залесский Б.А., Ольшевская О.В. / Заводская лаборатория. 1989. Т. 55. №. 7. С. 103-105.

14. Рыданова Г.В. Некоторые вопросы статистического анализа случайных бинарных векторов. Автореф. дис. канд. физ.-мат. наук. - М.: 1988. - 16 с.

15. Мирвалиев М., Никулин М.С. / Заводская лаборатория. 1992. Т.58. №. 3. С. 52- 58.

16. Ермаков С.М. / Заводская лаборатория. 1993. Т. 59. №. 7. С. 48-50.

17. Орлов А.И. / Заводская лаборатория. 1993. Т. 59. №. 7. С. 51-51.

18. Камень Ю.Э., Камень Я.Э., Орлов А.И. / Заводская лаборатория. 1986. Т. 52. №. 12. С. 55-57.

19. Лемешко Б.Ю. Статистический анализ группированных, частично группированных и негруппированных наблюдений одномерных непрерывных случайных величин. Автореф. дис. докт. техн. наук. - Новосибирск: 1997. - 46 с.

20. Левин Б.Р., Демидович Н.О. / Надежность средств связи. - Киев: Технiка, 1976, с.59-72.

21. Блехман И.И., Мышкис А.Д., Пановко Я.Г. Механика и прикладная математика: Логика и особенности приложений математики. - М.: Наука, 1983. - 328 с.

22. Орлов А.И. / Заводская лаборатория. 1992. Т.58. №. 1. С. 67-74.

23. Орлов А.И. / Заводская лаборатория. 1997. Т. 63. №. 3. С. 55-62.