МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

В. П. Шуленин

РОБАСТНЫЕ МЕТОДЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

ИЗДАТЕЛЬСТВО НАУЧНО-ТЕХНИЧЕСКОЙ ЛИТЕРАТУРЫ

ТОМСК
«Издательство НТЛ»
2016

УДК 519.2
Ш955

Ш955 **Шуленин В. П.** Робастные методы математической статистики. – Томск: Изд-во НТЛ, 2016. – 260 с.

ISBN 978-5-89503-575-7

Классические методы статистики в большинстве своем обладают повышенной чувствительностью к исходным предпосылкам статистической модели, принятой при обработке данных эксперимента. При решении прикладных задач неизбежно возникают отклонения от исходных предпосылок модели, и применение стандартных методов в этих условиях может оказаться мало эффективным и часто приводит к существенным искажениям статистических выводов. В связи с этим возникает необходимость построения новых, нетрадиционных методов обработки информации, устойчивых (или робастных) к возможным отклонениям характеристик реальных данных от предполагаемых. Предлагаемая книга посвящена построению и исследованию робастных оценок параметров, представляемых в виде функционалов от эмпирической функции распределения вероятностей. Дается краткий обзор основных понятий и подходов к построению робастных процедур, характеристики качества которых «устойчивы» к возможным отклонениям от принятой модели. Приводятся примеры свойств робастности предложенных оценок параметров положения и масштаба, проводится их сравнение в рамках различных супермоделей. Предложены различные модификации и обобщения оценок Ходжеса – Лемана, средней разности Джини, интерквартильного размаха и др. Обсуждаются подходы к построению адаптивных оценок с использованием оценок функционалов, описывающих степень «затянутости хвостов» распределений вероятностей наблюдений.

Книга предназначается студентам и аспирантам вузов, научным работникам, а также может быть полезна преподавателям при разработке курсов лекций для магистрантов и аспирантов на факультетах прикладной математики и кибернетики.

УДК 519.2

Рецензенты:

А.А. Назаров, профессор, доктор технических наук, заведующий кафедрой теории вероятности и математической статистики ТГУ;

А.М. Кориков, профессор, доктор технических наук, заведующий кафедрой автоматизированных систем управления ТУСУР, заслуженный деятель науки РФ

ISBN 978-5-89503-575-7

ПРЕДИСЛОВИЕ

В данном предисловии я хочу представить книгу В.П. Шуленина её потенциальным читателям, с разными целями решившим ознакомиться с её содержанием.

Каждый, кто взял эту книгу в руки, заинтересовался, чем она может быть полезной непосредственно в его работе, поскольку в своём деле ему так или иначе приходится иметь дело со статистической обработкой реальных данных. Но конкретный характер заинтересованности связан с тем, какой именно аспект статистических процедур для него важен.

У одних этот интерес чисто практический: в своей работе они применяют статистические процедуры для использования получаемых статистических выводов, и их волнует точность, надёжность, достоверность, короче говоря, – степень пригодности результатов статистической обработки данных для дела. Их интересуют только эти характеристики алгоритмов обработки его данных. Для таких лиц книга может служить справочником рабочих характеристик используемых алгоритмов. В ней они найдут описание критериев качества получаемых результатов, способы вычисления этих характеристик и даже таблицы числовых значений этих критериев для некоторых конкретных ситуаций.

Другой круг потенциальных пользователей этой книги составляют прикладные математики, специализирующиеся в синтезе и анализе процедур обработки экспериментальных данных. Такой читатель найдёт в ней описание процесса изобретения новых подходов к развитию технологий статистического анализа данных и разработки методов оценивания качества получаемых выводов. Эта информация не только расширит его эрудицию в своей профессии, но вполне может вдохновить на дальнейшее собственное творчество.

Думаю, что книга может представить интерес и для «чистых» математиков. В ней собран обширный материал по созданию и исследованию специфических абстрактных математических пространств и введению различных метрик в этих пространствах, приводятся изощрённые методы доказательства теорем о качественных и количественных свойствах объектов в этих пространствах.

Несомненно, наиболее активной категорией пользователей данной книги будут преподаватели и студенты курсов теоретической и прикладной математической статистики. Такие курсы преподаются будущим специалистам прикладных профессий разного профиля: и физико-математического, и естественно-научного, и гуманитарного, – всем, кому приходится иметь дело с извлечением нужной информации из обширного массива экспериментальных данных (результатов наблюдений и измерений). Книга представляет богатый материал о современном состоянии самого молодого и весьма актуального раздела математической статистики – робастной статистики.

Робастная статистика возникла в результате стремления заполнить разрыв между двумя крайностями в выдвижении априорных предположений о природе данных, подлежащих обработке с целью как можно более эффективного извлечения полезной информации из экспериментальных данных. Проблема состоит в том, что для обеспечения высокого качества результата обработки данных поразному следует обрабатывать данные, имеющие разную природу.

К середине прошлого века в математической статистике сформировались два подхода к конкретизации особенностей полученных в эксперименте данных. Общим априорным предположением для них является только одно: данные эксперимента имеют *случайную* природу. Исчерпывающая информация о случайном объекте содержится в функции распределения вероятностей по возможным состояниям этого объекта. Результаты эксперимента говорят нам о том, какие возможные состояния наблюдаемого объекта осуществлялись в реальности и как часто они были «близкими» друг к другу. Алгоритм эффективной обработки этих данных зависит от того, что нам известно о распределении вероятностей. Вот на этом пункте и разошлись две классических ветви статистики: одна исходит из предположения, что функция распределения *известна* с точностью до конечного числа параметров (параметрическая статистика); вторая предполагает, что события случайны (т.е. подчиняются какому-то закону распределения), но функциональный вид закона *неизвестен* (непараметрическая статистика). Соответственно по-разному следует обрабатывать один и тот же протокол наблюдений для извлечения из него нужной информации. В практике встречаются оба варианта априорной осведомлённости исследователя о природе данных, и статистики-прикладники очень тщательно относятся к выбору подходящей процедуры для анализа своих данных.

Однако в практике встречаются случаи, когда знания о функции распределения вероятностей носят некоторый промежуточный характер между параметрической и непараметрической крайностями: есть основания полагать, что реальный закон точно *не известен*, но довольно «близок» к некоторому *известному* параметрическому, лежит в некоторой окрестности его. Это порождает идею использовать информацию о свойствах параметрического распределения при обработке данных из непараметрического, но близкого к нему распределения. Разработка этой идеи и породила третью ветвь математической статистики – *робастную* статистику. (Из всех смысловых значений английского термина *robust* русскому термину *робастный* соответствуют значения «крепкий», «стойкий», т.е. устойчивый к отклонениям.) Разрабатываются такие статистические процедуры, которые «близки» к оптимальным параметрическим при совпадении реального распределения с известным и устойчиво сохраняют свои качества, пока истинное распределение находится в его окрестности.

В книге В.П. Шуленина даётся обстоятельный обзор полученных к настоящему времени результатов развития робастной статистики – от строгих определений «супермодели» (комбинации параметрического распределения и его «окрестности») до методов синтеза робастной процедуры и анализа характеристик качества статистического вывода, полученного статистической процедурой. Излагаются технические подробности получения результатов (желающие разобраться более детально отсылаются к первоисточникам), изложение сопровождается комментариями («Замечаниями») автора, приводятся наглядные примеры. Всё это и делает книгу пригодной для разных пользователей математической статистики – практиков, теоретиков, преподавателей и студентов.

Робастная статистика – молодая, находящаяся в процессе роста и развития ветвь математической статистики. Видимо, поэтому автор ограничился описанием лишь достаточно оформившейся её части – проблем оценивания параметров. Будем следить за дальнейшими событиями в этой области: на очереди робастное оценивание закономерностей (как зашумлённых функциональных зависимостей, так и статистических связей типа корреляции и регрессии); робастная проверка гипотез; не лишено смысла и обсуждение робастного оценивания различных представлений функций распределения (с перспективой функциональной интерпретации робастных статистик).

Профессор Томского госуниверситета,
доктор технических наук (кибернетика и теория информации),
заслуженный деятель науки и техники РСФСР

Ф.П. Тарасенко

ОТ АВТОРА

Развитие научных исследований в настоящее время характеризуется всё более возрастающей ролью статистических процедур, их широким применением в различных областях практики и требованиями качественной обработки экспериментальных данных. Однако их успешное применение сдерживается тем фактом, что классические процедуры математической статистики обеспечивают заданное качество решений лишь при справедливости некоторых предположений о статистической модели, которая принята в рамках проводимого эксперимента. На практике часто возникают отклонения характеристик реальных данных эксперимента от предполагаемых в модели, и применение стандартных процедур в этих условиях может оказаться малоэффективным и даже может привести к существенным искажениям статистических выводов. Многие стандартные процедуры нормальной теории гарантируют качество решений лишь в рамках гауссовской модели наблюдений. В связи с этим возникает необходимость построения новых, нетрадиционных методов обработки информации, «свободных от распределения» и устойчивых (или робастных) к возможным отклонениям характеристик реальных данных от предполагаемых в принятой модели.

Современное состояние и развитие математической статистики характеризуется тремя направлениями: параметрическая, непараметрическая и робастная статистика. Эти направления различают по типам рассматриваемых статистических моделей, которые отражают уровень и полноту априорной неопределенности в статистическом описании изучаемого случайного явления. Классические методы параметрической статистики используют в явном виде известный функциональный характер распределения вероятностей наблюдений. Многие из этих методов разработаны для гауссовской модели наблюдений, которая не всегда является адекватной моделью при решении практических задач. В таких случаях целесообразнее использовать методы непараметрической и робастной статистики, которые бурно развиваются и появились как альтернатива классическим методам, в большинстве своем обладающим повышенной чувствительностью к исходным предпосылкам в статистических моделях. В данной книге приводиться описание статистических процедур, которые обладают устойчивостью (робастны) к отклонениям от исходных предпосылок в принятой статистической модели.

В разделах 1–6 дается краткий обзор основных понятий и подходов к построению робастных процедур, характеристики качества которых «устойчивы» к возможным отклонениям от принятой модели. Вводится понятие «функция влияния Хампеля», и на её основе определяются числовые характеристики робастности оценок параметров. Обсуждаются качественный и количественный подходы к определению робастности статистических процедур. Для изучения асимптотических распределений оценок функционалов от распределения вероятностей наблюдений приводится метод Мизеса.

В разделах 7–18 обсуждаются свойства общих классов *M*-, *L*-, *R*- и *MD*-оценок функционалов от распределения вероятностей наблюдений и устанавливаются связи между этими классами оценок. Предложены обобщения класса *L*-оценок, *R*-оценок, *U*-статистик и *MD*-оценок, которые вычисляются на основе порядковых статистик, из которых удалена определенная часть наименьших и наибольших значений. Приводятся примеры свойств робастности предложенных оценок параметров положения и масштаба, проводится их сравнение в рамках различных супермоделей. Предложены различные модификации и обобщения оценок Ходжеса – Лемана, средней разности Джини, интерквартильного размаха и др. Приводятся примеры построения адаптивных оценок, с использованием оценок функционалов, описывающих степень «тяжести хвостов» распределений вероятностей наблюдений.

В книге принята *сплошная* нумерация формул, определений, теорем, примеров и замечаний в пределах данного раздела. Например, ссылка (4.7) соответствует седьмой *формуле* четвёртого раздела, а следующее за ней *замечание* отмечено ссылкой 4.8. Ссылка (П.1.2) соответствует второй формуле в первом приложении. Для рисунков и таблиц используется самостоятельная нумерация в пределах раздела. При изложении общих результатов и понятий использовались работы различных авторов. Все заимствованные результаты снабжены ссылками в тексте.

Благодарю рецензентов профессора А.А. Назарова, профессора А.М. Корикова, которые внимательно прочитали рукопись книги и сделали целый ряд полезных критических замечаний, которые были учтены при редакторской обработке рукописи.

Особая благодарность моему научному руководителю профессору Ф.П. Тарасенко, который прочитал рукопись и любезно предоставил своё предисловие к данной книге.

Автор заранее выражает свою признательность за любые замечания со стороны читателей книги. Их можно направить по адресу E-mail: shvp@fpmk.tsu.ru

В. Шуленин

1. ВВЕДЕНИЕ

Представление различных знаний в виде, удобном для практической деятельности, основано на процессе построения моделей и проверки их адекватности. В связи с тем, что модель определяет характер всей последующей деятельности, её адекватность является предметом особой заботы всех исследователей. Не является исключением и процесс построения статистических моделей, при котором используются явные или неявные, не всегда легко проверяемые допущения типа независимости, одинаковой распределенности, принадлежности распределения заданному параметрическому либо непараметрическому семейству функций и т.п. Задачи статистического анализа возникают в тех случаях, когда принятая статистическая модель хотя бы частично неизвестна, и они сводятся к уточнению модели на основе наблюдений. При этом основой любых статистических выводов, конечно же, являются наблюдения, но не менее важную роль играют предположения, которые использованы при построении статистической модели. В литературе обычно рассматриваются параметрические и непараметрические статистические модели. Эти модели отличаются друг от друга уровнями априорной неопределенности в статистическом описании наблюдений [1]. При построении параметрических моделей функциональный вид распределения наблюдений задается, то есть считается, что ф.р. наблюдений *известна с точностью до параметров*, а при построении непараметрических моделей функциональный вид распределения наблюдений не используется, то есть считается, что ф.р. наблюдений *неизвестна*. Различие в способах задания этих моделей имеет тенденцию к сглаживанию, достигаемому путем введения промежуточных моделей, обычно являющихся расширением параметрических моделей с использованием «непараметрических окрестностей». Это объясняется тем фактом, что статистическая модель, как и вообще любая модель, является неизбежной идеализацией и может оказаться в лучшем случае лишь удачной аппроксимацией реальных процессов (в таких случаях говорят, что модель адекватна). Подчеркивая этот факт, Бокс (см. [2, с. 165, и 3]) пишет: «*Все модели неправильные, но некоторые из них полезны*». Итак, при решении практических задач всегда возникает вопрос: являются ли верными статистические выводы, полученные в рамках принятой модели, в условиях неизбежных отклонений от модели на практике? Если отклонения от принятой модели легко обнаруживаемы, то мы считаем её грубой и, естественно, пересматриваем модель. Следовательно, основной заботой должны быть необнаруживаемые отклонения и трудно проверяемые допущения модели. Примеры конкретных отклонений от статистических моделей и последствия, к которым приводят эти отклонения при использовании стандартных процедур, обсуждаются [4].

Итак, при использовании статистических процедур желательно иметь информацию о том, какие отклонения и от каких предположений модели эти отклонения оказывают решающее влияние на конечный вывод при статистическом анализе, и нарушения каких допущений являются относительно безвредными. В частности, могут возникнуть вопросы о применимости стандартных процедур нормальной теории, когда истинное распределение не является нормальным. Ответам на сформулированные вопросы и построению статистических процедур, нечувствительных к отклонениям от предположений, лежащих в их основе, посвящено новое, быстро развивающее направление, названное робастной статистикой, которое было выделено американским математиком Дж. Тьюки в «*статистику третьего поколения*», после параметрической и непараметрической статистики. Поток публикаций по этому направлению постоянно увеличивается, уже имеется ряд монографий, среди них первая книга Хьюбера [5], книга Ф. Хампеля и др. [6], обе переведены на русский язык, также имеется и учебная литература (см., например, [7–9]).

Термин «робастность» соответствует английскому слову «robust», дословный перевод которого «грубый, сильный, крепкий, здравый, ясный», в статистическую литературу этот термин был введен Боксом [3] в 1953 году, и с середины шестидесятых годов этот термин стал общепризнанным для раздела статистики, в котором развиваются статистические процедуры, нечувствительные к отклонениям от предположений принятой модели. Отметим, что идеи робастности имеют давнишнюю историю, которая прослежена в работе Стиглера [10]. Они появляются в работах К. Гаусса, С. Ньюкомба, А.Эдингтона, Г. Джефриса, Джини и др. Однако систематическое развитие идей робастности начинается с работ Дж. Тьюки (1941–1972 гг., см., например, [11–14]) и, особенно, после выхода работы Хьюбера [15] в 1964 г.

Итак, робастная статистика появилась как средство преодоления отрицательных последствий возможной неадекватности статистических моделей. Здесь может возникнуть вопрос: почему мы вынуждены использовать неадекватную модель? Может быть, естественнее направить усилия на построение подходящей модели и затем воспользоваться соответствующей статистической процедурой? Однако все осложняется тем фактом, что не всегда просто установить адекватность модели. Для этого порой требуется такое количество наблюдений, которым не располагает экспериментатор. Например, при использовании критериев согласия как инструмента для проверки адекватности статистической модели следует осознавать, что критерии согласия хотя и позволяют отвергнуть неправильную модель, которая противоречит наблюденным данным, но они не позволяют доказать, что модель верна. Часто объема экспериментальных данных недостаточно, чтобы различить статистические модели. При этом даже если критерий согласия и отклонит нулевую гипотезу, например о нормальности распределения, то мы не знаем, в каком направлении двигаться дальше при выборе статистической модели. Отвечая на вопрос, что является основанием для выбора статистической модели, многие ссылаются либо на эксперимент, либо на некоторые теоретические обоснования. В основу построения параметрической модели часто кладется нормальное распределение. Это является следствием использования центральной предельной теоремы и того факта, что в начальной стадии развития статистики основным объектом её изучения были ошибки независимых наблюдений, для которых допущение нормальности распределения часто является обоснованным. Однако следует напомнить, что появление нормального закона вызвано, прежде всего, математическими удобствами, см. цитату Гаусса в [16], а также Тьюки [13], который отмечает, что «*предполагаемые достоинства средней арифметической были использованы для введения гауссовского распределения, а предполагаемая «истинность» гауссовского распределения была использована для доказательства оптимальности средней арифметической. Нет сомнений в том, что этот круговорот был очевиден для многих, работавших в этой области…*». Здесь также уместно привести высказывание Анскомба [17], который отмечает, что «*единственным основанием для повсеместного использования процедур нормальной теории является убежденность некоторых статистиков в том, что предположение нормальности слишком удобно, чтобы не быть правильным*». В последнее время во многих работах всё чаще привлекается внимание исследователей к эффекту возможной «негауссовости» при решении практических задач. Стало уже популярным приводить высказывание Гири: «*нормальность – это миф, нормального распределения никогда не было и никогда не будет*». Следуя традиции, приведем и известное замечание Липмана (см. [18]): «*каждый уверен в справедливости закона ошибок, экспериментаторы – потому, что они думают, что это математическая теорема, математики – потому, что они думают, что это экспериментальный факт*». К этим высказываниям приведем для полноты картины и результаты детального обследования тщательно выполненных реальных физических экспериментов. В работе Хампеля [19], со ссылками на многочисленных авторов, приводится следующий вывод. *Наличие 5–10 % «ложных» наблюдений в реальных статистических данных – обычно правило, а не исключение. Для описания таких данных более подходящим являются распределения «с более затянутыми хвостами», чем у нормального распределения.* Анализируя причины, приводящие к отклонениям от параметрических моделей в условиях реальных экспериментов, Хампель [19] выделяет три фактора.

1. Наличие грубых ошибок, возникающих при измерении данных, при их введении в ЭВМ, при их передаче по каналам связи и т.п.

2. Ограниченная точность измерительных приборов, приводящая к округлению и группировке наблюдений.

3. Возможность неадекватности принятой параметрической модели.

В первых двух случаях отклонения от модели вызваны искажением данных. Это, пожалуй, один из первых и частных типов отклонений от модели, который стимулировал развитие робастных процедур. Отметим, что в статистике разработаны методы обнаружения выбросов, или критерии отбраковки грубых ошибок (см. например, [8]). Эти методы применимы для небольших объемов наблюдений, описываемых, как правило, нормальным распределением. Для больших совокупностей данных, особенно при их автоматизированной обработке на ЭВМ, тщательный анализ данных с целью обнаружения «выпадающих» наблюдений и их корректировки трудно осуществим. В таких ситуациях используются процедуры, которые, с одной стороны, оказываются нечувствительными к наличию умеренного засорения данных грубыми ошибками, а с другой – ведут себя достаточно хорошо при идеальных условиях нормальности или какого-либо другого предположения о типе распределения данных. Про такие процедуры говорят, что они обладают «защищенностью от выпадающих наблюдений» или еще их называют робастными к наличию грубых ошибок или выбросов [20]. В третьем случае речь идёт о трудно обнаруживаемых отклонениях от формы распределения, то есть истинная функция распределения наблюдений может незначительно отличаться от распределения, принятого в модели. Это наиболее важный и интересный с практической точки зрения тип отклонения от модели и, в связи с этим, наиболее изученный. В большинстве работ рассматриваются отклонения от модели, связанные, как правило, с «утяжелением хвостов» по сравнению с нормальным распределением. Отметим, что для многих процедур нормальной теории характерна повышенная чувствительность к утяжелению хвостов распределений (примеры приводятся в [4, 9]). Отметим также, что «робастность по распределению» и «защищенность от выпадающих наблюдений» – это понятия, различные по сути, на практике же выступают как синонимы [5].

Итак, принимая за основу некоторую достаточно правдоподобную модель (ниже будем её называть идеальной), всегда следует помнить, *что в реальном эксперименте возможны различные от-*

клонения от идеальной модели. Для описания идеальной модели вместе с возможными отклонениями от нее в условиях реального эксперимента используют понятие «*супермодель*». С использованием этого понятия, под робастной процедурой понимают такую процедуру, которая «работает достаточно хорошо» как при идеальной модели, так и в рамках супермодели. Термин «работает достаточно хорошо» каждый раз требует уточнения, и его конкретное содержание зависит от типа рассматриваемой процедуры, от принятой супермодели и от требований к поведению характеристик качества процедуры в рамках супермодели. Таким образом, при определении робастной процедуры требуется конкретизировать ответы на следующие вопросы.

1. *Робастность чего*? При ответе на этот вопрос уточняется тип статистической процедуры (проверка гипотез с помощью некоторого критерия либо построение точечной или интервальной оценки параметра и т.п.).

2. *Робастность к чему*? Для ответа на этот вопрос необходимо охарактеризовать идеальную статистическую модель и принять некоторую супермодель, которая описывает возможные отклонения от идеальной модели.

3. *Робастность в каком смысле*? В данном случае следует конкретизировать критерий качества процедуры и сформулировать требования к его поведению в рамках супермодели или, короче, сформулировать цель, которую мы хотим достичь.

Ясно, что как статистические процедуры, так и супермодели, а также и наши цели могут быть самыми различными, однако их всегда следует оговаривать, чтобы исключить разногласия в толковании термина «робастная процедура». Отметим, что разные авторы по-разному трактуют этот термин. Исключение, пожалуй, составляет лишь качественный подход к определению робастных процедур, развитый Хампелем [21, 22] для статистик, представляемых в виде функционала от эмпирической функции распределения вероятностей. Этот подход основан на использовании понятия непрерывности функционала относительно метрики, порождающей слабую сходимость. Что касается количественного подхода к определению робастной процедуры, то здесь имеется большое разнообразие, которое объясняется рассмотрением различных вариантов супермоделей, а также использованием различных характеристик качества процедур. Подводя итог, отметим, что наряду с разнообразной трактовкой в литературе термина «робастный», общность проявляется в том, что робастность может рассматриваться как тип гарантии: *незначительная потеря в эффективности при идеальной модели достигается одновременно с отсутствием катастрофических последствий при отклонениях от модели, то есть в рамках супермодели.*

2. ОСНОВНЫЕ ПОНЯТИЯ И ОБОЗНАЧЕНИЯ

Следуя принятой традиции в научной литературе по статистике, всюду ниже будем обозначать случайные величины прописными латинскими буквами. Возможные состояния (или реализации) случайных величин будем обозначать соответствующими строчными буквами. Например, X – случайная величина, а ее возможное состояние (или реализацию) будем обозначать через x, или $\vec{X}=(X_1,...,X_n)$ – случайный вектор, а его реализация – $\vec{x}=(x_1,...,x_n)$. Для случайной величины X ее функцию распределения (ф.р.) вероятностей везде ниже будем обозначать в виде F_X. Функция распределения вероятностей является универсальным инструментом, пригодным для изучения любой случайной величины, одномерной или многомерной, и непрерывного, дискретного или смешанного типов. Если X одномерная случайная величина непрерывного типа с бесконечным числом возможных значений $\{x\}$ на действительной оси $R^1=\{x:-\infty<x<+\infty\}$, то она характеризуется одномерной ф.р. $F_X(x)=P(X\le x)$, $x\in R^1$. Иногда удобнее описывать случайную величину X одномерной плотностью распределения вероятностей (п.р.в.) $f_X(x)=F'_X(x)$, $x\in R^1$. Для описания случайного вектора $\vec{X}=(X_1,...,X_n)$ используют функцию n переменных $x_1,...,x_n$, которая в точке $(x_1,...,x_n)\in R^n$, где R^n обозначает n-мерное евклидовое пространство, определяется с помощью вероятности совместного осуществления событий в фигурных скобках, то есть функция распределения случайного вектора $\vec{X}=(X_1,...,X_n)$ определяется в виде

$$F_{X_1,...,X_n}(x_1,...,x_n)=F_{\vec{X}}(x_1,...,x_n)=P\{X_1\le x_1,...,X_n\le x_n\},$$

$$(x_1,...,x_n)\in R^n.$$

Итак, для того чтобы охарактеризовать случайную величину X необходимо задать ее функцию распределения вероятностей F_X. Однако при решении реальных задач на практике функция распределения вероятностей F_X обычно полностью или частично неизвестна. В связи с этим и возникают статистические задачи, которые по своей сути направлены на устранение этой «неизвестности» на основе информации, содержащейся в результатах опыта, полученных путем наблюдений над изучаемой случайной величиной X. Отметим, что в реальных задачах степень «неизвестности» относительно ф.р. F_X может быть самой различной, она уточняется при формулировке статистической модели, которая играет важную роль при решении любой статистической задачи. Статистическую модель задают в виде пары множеств $(\mathrm{X}, \Im)$, где X – выборочное пространство, а $\Im$ – множество допустимых функций распределения вероятностей в условиях данного опыта, результатом которого является *наблюденная* реализация $\vec{x} = (x_1, ..., x_n)$ *наблюдаемого* случайного вектора $\vec{X} = (X_1, ..., X_n)$. Отметим, что множество $\Im$ задается на основе тех априорных сведений о распределении вероятностей, которыми располагает экспериментатор до проведения опыта. Выборочное пространство X определяется условиями проведения опыта. Выборочным пространством $\mathrm{X} = \{\vec{x}\}$ называют множество всех возможных реализаций $\vec{x} = (x_1, ..., x_n)$ векторной случайной величины $\vec{X} = (X_1, ..., X_n)$, на котором задано ее распределение вероятностей, однозначно определяемое функцией распределения $F_{\vec{X}}(x_1, ..., x_n) \in \Im$, $(x_1, ..., x_n) \in \mathrm{X}$. Обычно выборочное пространство $\mathrm{X} = \{\vec{x}\}$ является либо всем n-мерным евклидовым пространством R^n, либо его подмножеством, либо оно может состоять из отдельных точек R^n, если случайная величина X дискретная.

Замечание 2.1. Отметим, что при решении многих статистических задач предполагается, что компоненты случайного вектора $\vec{X} = (X_1, ..., X_n)$ являются независимыми и одинаково распределенными (н.о.р.) случайными величинами с такой же функцией распределения вероятностей, какой обладает изучаемая случайная величина *X*. Эти предположения соответствуют условиям, при которых статистические данные получены в схеме повторных и независимых наблюдений над изучаемой с.в. *X*. В частности, предположение одинаковой распределенности формально означает выполнение равенства $F_{X_i}(x) = F_X(x)$, $\forall i \in (1,...,n)$, а предположение независимости означает, что результат последующего опыта не зависит от результата предыдущего опыта. Математически эти предположения проявляются в том, что совместная функция распределения случайных величин $X_1,...,X_n$ факторизуется, то есть записывается в виде

$$F_{X_1,...,X_n}(x_1,...,x_n) = F_{X_1}(x_1)\cdots F_{X_n}(x_n) = \prod_{i=1}^{n} F_X(x_i). \qquad \textbf{(2.2)}$$

При этом последовательность $X_1,...,X_n$ н.о.р. случайных величин коротко называют *выборкой объема n* из распределения F_X (или, точнее, *выборкой*, которая порождается распределением F_X).

Итак, при решении любой статистической задачи одинаково важную роль играют два понятия:

1) *статистические данные* $(X_1,...,X_n)$ *– выборка*;

2) *статистическая модель* $(\mathrm{X}, \Im)$.

Суть и цель любой статистической процедуры состоит в том, чтобы на основе информации, содержащейся в выборке $X_1,...,X_n$, *доопределить статистическую модель* $(\mathrm{X}, \Im)$. Эту цель кратко можно выразить следующим образом:

$$\left\{\begin{matrix} X_1,...,X_n \\ \textit{стат. данные} \end{matrix}\right\} \Rightarrow \left\{\begin{matrix} (\mathrm{X}, \Im) \\ \textit{стат. модель} \end{matrix}\right\}. \qquad \textbf{(2.3)}$$

В математической статистике различают три общих типа моделей: параметрические непараметрические и модели, используемые при изучении и построении робастных статистических процедур, которые коротко называют супермоделями. Для этих типов моделей множество допустимых функций распределения $\Im$ будем соответственно обозначать в виде $\Im_\theta, \Im_N$, $\Im_R$. Наглядно это можно представить следующим образом:

$$\Im = \begin{cases} \to \ \Im_\theta = \{F_X : F_X(x;\theta), \theta \in \Theta\} - \text{параметрическая модель,} \\ \to \ \Im_N = \{\text{мн-во аб.непр.р.}\} - \text{непараметрическая модель,} \\ \to \ \Im_R = \{F_X : d(F_0, F_X) < \varepsilon\} - \text{супермодель роб.статистики.} \end{cases} \quad \textbf{(2.4.)}$$

Для статистических моделей параметрической (классической) статистики характерно, что множество допустимых функций распределений $\Im_\theta$ задается в виде параметрического семейства функций с конечным числом неизвестных параметров θ, которые принадлежат заданному параметрическому множеству Θ, то есть множество $\Im_\theta$ задается в виде

$$\Im_\theta = \{F_X : F_X(x\,;\theta)\,,\, \theta \in \Theta\,\}\,. \qquad \textbf{(2.5)}$$

Отметим, что в общем случае параметр θ может быть векторным. При задании параметрических моделей функциональный характер распределения считается известным, то есть вся неопределенность модели связана с неопределенностью параметров $\theta \in \Theta$. В связи с этим, статистические процедуры, реализуемые в рамках параметрических моделей, направлены на построение точечных и интервальных оценок этих параметров, а также на построение критериев проверки статистических гипотез относительно параметров θ на основе информации, содержащейся в выборке $X_1,...,X_n$.

Для моделей непараметрической статистики характерно, что множество допустимых функций распределений $\Im_N$ очень «широкое», оно задается из некоторых общих соображений типа непрерывности, симметрии, равенства и т.п. и не может быть представлено в виде параметрического семейства функций (этим и объясняется название этих моделей). Подчеркнём, что при конкретизации множества $\Im_N$ функциональный характер распределения F_X не используется, то есть считается, что ф.р. F_X неизвестна. Типичным примером является задача сравнения некоторого нового типа обработки по сравнению со стандартным. Например, пусть имеются статистические данные $X_1,...,X_n$, полученные в результате наблюдений над с.в. X с ф.р. F_X в стандартных условиях, и статистические данные $Y_1,...,Y_m$, полученные в результате наблюдений над с.в. Y с ф.р. F_Y в условиях, отличных от стандартных. По этим статистическим данным проверяется гипотеза о том, что изменение условий эксперимента не сказывается на его результатах. Эта гипотеза сводится к проверке факта равенства распределений F_X и F_Y. Если при этой проверке предполагается лишь равенство распределений, а их функциональный вид не конкретизируется и предполагается лишь, что они непрерывны, то эта задача однородности решается в рамках непараметрической статистической модели.

Особенности моделей робастной статистики состоят в том, что некоторые предположения принятой за основу идеальной модели в виде заданной ф.р. F_0 подвергаются сомнению. По этой причине принятая статистическая модель «расширяется» до «супермодели», которую обычно задают в виде некоторой окрестности идеальной модели F_0, конкретизируемой выбранной метрикой $d(F_0, F_X)$ на множестве допустимых ф.р. $\Im$, то есть, супермодель задается в виде множества $\Im_R = \{F_X : d(F_0, F_X) < \varepsilon\}$, $\varepsilon > 0$. В литературе используют различные метрики $d(F_0, F_X)$ и описаны также другие подходы к заданию супермоделей, они будут подробно обсуждаться в примерах ниже (см. также [4, 9]). Наиболее подходящей метрикой в пространстве функций распределения вероятностей $\Im$, позволяющей одновременно описывать отклонения, вызванные как наличием грубых ошибок, так и округлением и группировкой, является метрика Леви (см., например, [5]), которая определяется для двух функций распределения вероятностей $F, G \in \Im$. Её геометрический смысл состоит в том, что она равна диагонали максимально возможного квадрата, который можно вписать между функциями $F, G \in \Im$. Обобщением этой метрики является метрика Прохорова, которая определяется для произвольного полного, сепарабельного, метрического пространства. Пусть $(\Omega, \mathrm{B}, \Im)$ – вероятностное пространство с метрикой $\rho(\cdot\,,\cdot)$, где Ω – множество доступных наблюдению величин, B – некоторая σ-алгебра, $\Im$ – множество распределений вероятностей на (Ω, B). Пусть F и G – вероятностные меры на (Ω, B). Определим для $\forall B \in \Omega$ δ-окрестность B в виде

$B^{\delta} = \{x \in \Omega : \inf_{y \in B} \rho(x, y) \le \delta\}$. Окрестность Прохорова $\Pi_{\delta,\varepsilon}(G)$ для идеальной модели G определяется в виде

$$\Pi_{\delta,\varepsilon}(G) = \{F \in \Im : F(B) \le G(B^{\delta}) + \varepsilon, \forall B \in \mathrm{B}\},$$

При этом $\delta > 0$ соответствует локальным изменениям вероятности (например, описывает ошибки округления и группировку), а $\varepsilon > 0$ соответствует глобальным изменениям вероятности на (Ω, B), то есть описывает грубые ошибки, небольшая доля которых ε располагается произвольно на Ω. Обычно масштаб измерений выбирают таким образом, чтобы заданием одного $\varepsilon > 0$ можно было бы описать оба типа ошибок. При этом условии метрика Прохорова для вероятностных мер F и G определяется в виде

$$d_{\Pi}(F, G) = \inf\{\varepsilon : F(B) \le G(B^{\varepsilon}) + \varepsilon, \forall B \in \mathrm{B}\}.$$

Таким образом, супермодель, описывающая одновременно наличие грубых ошибок, группировку и округление наблюдений, заданная с помощью метрики Прохорова $d_{\Pi}(F, G)$, имеет вид

$$\Im_{\Pi,\varepsilon}(G) = \{F \in \Im : d_{\Pi}(F, G) \le \varepsilon\}, \quad 0 \le \varepsilon \le 1.$$

Если пространством наблюдений Ω является действительная ось R^1, то метрика Прохорова $d_{\Pi}(F, G)$ эквивалентна метрике Леви, которая соответствует замене множества B на интервал $(-\infty, x]$ и для функций распределения $F(x)$ и $G(x)$ определяется в виде

$$d_{\mathrm{L}}(F, G) = \inf\{\varepsilon : G(x - \varepsilon) - \varepsilon \le F(x) \le G(x + \varepsilon) + \varepsilon, \ \forall x \in R^1\}. \quad \textbf{(2.6)}$$

Отметим, что $\sqrt{2}\, d_{\mathrm{L}}(F, G)$ – это максимальное расстояние между графиками функций распределения $F(x)$ и $G(x)$, где оно имеет наклон в 45°, то есть это диагональ максимального квадрата, который можно вписать между функциями $F(x)$ и $G(x)$ (см., например, [5, с. 35]). Отметим также, что сходимость по метрике Леви влечет слабую сходимость, то есть $F_n(x) \to F(x)$ при $n \to \infty$ в том и только в том случае, когда $d_{\mathrm{L}}(F, G) \to 0$ при $n \to \infty$. Этот факт делает удобной метрику Леви при решении статистических задач. Отметим также, что расстояние по метрике Прохорова всегда больше чем расстояние по метрике Леви, то есть выполнение неравенства $d_{\Pi}(F,G) \le \varepsilon$ влечет выполнение неравенств

$$G(x-\varepsilon)-\varepsilon \le F(x) \le G(x+\varepsilon)+\varepsilon,\ \forall x \in R^1$$

(см., например, [5]). Супермодель, заданная с помощью метрики Леви, имеет вид

$$\Im_{\mathrm{L},\varepsilon}(G) = \{F \in \Im : d_L(F,G) \le \varepsilon\}. \quad \mathbf{(2.7)}$$

Отметим также, что при построении супермоделей могут использоваться и другие метрики, в частности метрика Колмогорова $d_{\mathrm{K}}(F,G) = \sup |F(x)-G(x)|$. Супермодель, заданная с помощью метрики Колмогорова, записывается в виде

$$\Im_{\mathrm{K},\varepsilon}(G) = \{F \in \Im : d_K(F,G) \le \varepsilon\}.$$

Обсуждение связей между различными метриками и их применение к построению супермоделей, приводится в дискуссии к работе [23], (см. также [5]).

Итак, в теории робастной статистики вместе с понятием базовой (идеальной) статистической модели вводится дополнительное понятие, названное супермоделью. Понятие супермодели служит для того, чтобы описать некоторую базовую модель, в которую мы верим и выбираем ее в качестве основы, а также описать возможные в условиях реального эксперимента отклонения от предположений базовой модели. При этом основная цель состоит в разработке таких статистических процедур, которые должны «работать достаточно хорошо» как при идеальной модели F_0, так и в рамках принятой супермодели $\Im_R$. В математическом отношении построение таких процедур сводится к изучению свойств некоторых функционалов $T(F)$, заданных на множестве $\Im_R$, а также к изучению свойств оценок этих функционалов, как при идеальной модели F_0, так и в рамках принятой супермодели $\Im_R$.

Замечание 2.8. При математическом описании статистических данных используют понятия упорядоченной статистики

$X_{(\cdot)} = (X_{(1)},...,X_{(n)})$ и рангового вектора $\vec{R} = (R_1,...,R_n)$ выборки $\vec{X} = (X_1,...,X_n)$ (см., например, [8]). Отметим, что информация, содержащаяся в исходной выборке, условно может быть разделена на две части. Одна из них содержится в упорядоченной статистике $X_{(\cdot)} = (X_{(1)},...,X_{(n)})$, а другая – в ранговом векторе $\vec{R} = (R_1,...,R_n)$. Этот факт можно представить следующим выражением:

$$\{(X_1,...,X_n)\} \Leftrightarrow \left\{ \left(X_{(1)},...,X_{(n)} \right) ; (R_1,...,R_n) \right\}, \tag{2.9}$$

которое означает, что совместное распределение векторов $\vec{X}_{(\cdot)}$ и $\vec{R}$ определяется распределением вектора $\vec{X}$.

Таким образом, статистические процедуры, направленные на уточнение статистической модели $(\mathrm{X}, \Im)$, могут быть основаны либо на исходной выборке $\vec{X} = (X_1,...,X_n)$, либо на порядковых статистиках $X_{(\cdot)} = (X_{(1)},...,X_{(n)})$, или на рангах наблюдений $\vec{R} = (R_1,...,R_n)$.

3. ПОДХОДЫ К ОПРЕДЕЛЕНИЮ РОБАСТНЫХ ПРОЦЕДУР

Различают два подхода к определению робастных процедур. Качественный подход был представлен в диссертационной работе Хампеля [21] путем обобщения понятий непрерывности функций применительно к функционалам от распределения вероятностей. Количественный подход подробно описывается в обзорной работе Хьюбера [16] и монографиях [5], Хампель и др. [6], (см. также [4, 9]).

3.А. Качественный подход к определению робастных процедур

Пусть задана последовательность $X_1,...,X_n$ н.о.р. случайных величин с функцией распределения $F \in \Im$, где $\Im$ – множество непрерывных распределений, на котором для заданной пары элементов $(F,G) \in \Im$ определена метрика Леви $d_{\text{L}}(F,G)$ вида (2.6). Далее, пусть $F_n \in \Im_n$ – эмпирическая функция распределения, построенная по выборке $X_1,...,X_n$, которая записывается в виде

$$F_n(x) = \frac{\mu_n(x)}{n} = \frac{1}{n}\sum_{i=1}^{n} C(x - X_i), \qquad \textbf{(3.1)}$$

где $\forall x \in R^1$ случайная величина $\mu_n(x)$ определяется в виде $\mu_n(x) = [$числу $X_i \le x$, $i = 1,...,n]$, и $C(\cdot)$ – единичная функция Хевисайда, то есть

$$C(u) = \begin{cases} 1, & u \ge 0, \\ 0, & u < 0. \end{cases} \qquad \textbf{(3.2)}$$

Заметим, что для непрерывной с.в. X с возможными значениями на действительной оси R^1 и с дополнительно определёнными порядковыми статистиками $X_{(0)} = -\infty$ и $X_{(n+1)} = +\infty$ эмпирическая функция распределения $F_n(x)$ может быть компактно записана через порядковые статистики $X_{(1)},...,X_{(n)}$ исходной выборки $X_1,...,X_n$ в виде

$$F_n(x) = i/n \ \ \textit{при} \ \ X_{(i)} \le x < X_{(i+1)}, \ i = 0,1,...,n. \qquad \textbf{(3.3)}$$

Отметим, что многие оценки параметров и статистики критериев $T_n = T_n(X_1,...,X_n)$, используемые на практике, могут быть представлены в виде функционала от эмпирической функции распределения, то есть в виде $T_n = T(F_n)$, где $T(F)$ – некоторый функционал, являющийся отображением множества непрерывных распределений $\Im$ в действительную ось, то есть $T(F): \Im \to R^1$. При этом, отображение $T_n : \Im_n \to R^1$ порождает на действительной оси R^1 распределение вероятностей статистики T_n при исходной ф.р. F наблюдений, которое обозначим через $L_F(T_n)$.

В своей диссертационной работе Хампель обобщил понятия непрерывности функций на понятия устойчивости последовательности оценок $\{T_n\}$, основываясь на принципе, согласно которому *произвольно малые изменения в распределении наблюдений должны вызывать лишь достаточно малые изменения в распределении оценки* T_n. Приведем необходимые определения и утверждения, взятые из работ [21, 22].

Определение 3.4. Последовательность оценок $\{T_n\}$ называется устойчивой в $F \in \Im$, если $\forall \varepsilon > 0 \ \exists \delta(\varepsilon) > 0$, такое, что $\forall G \in \Im$ и $\forall n$ выполняется выражение

$$d_{\mathrm{L}}(F,G) < \delta \Rightarrow d_{\mathrm{L}}\{L_F(T_n), L_G(T_n)\} < \varepsilon. \qquad \textbf{(3.5)}$$

Последовательность оценок $\{T_n\}$ называется устойчивой в окрестности F, если существует $\eta > 0$, такое, что выполнение неравенства $d_{\mathrm{L}}(F,G) < \eta$ влечет устойчивость последовательности оценок $\{T_n\}$ в G. Последовательность оценок $\{T_n\}$ называется устойчивой в классе $E \in \Im$, если она устойчива для всех $F \in E$.

Отметим, что выполнение выражения

$$d_{\mathrm{L}}(F_n, G_n) < \delta \Rightarrow |T(F_n) - T(G_n)| < \varepsilon$$

влечет выполнение выражения (3.5), то есть устойчивость последовательности оценок $\{T_n\}$ в $F \in \Im$.

Определение 3.6. Последовательность оценок $\{T_n\}$ непрерывна в F, если $\forall \varepsilon > 0 \ \exists \delta(\varepsilon) > 0$, такое, что $\exists n_0$ и $\forall n, m > n_0$ выполняется выражение

$$\{d_{\mathrm{L}}(F, F_n) < \delta \cap d_{\mathrm{L}}(F, F_m) < \delta\} \Rightarrow |T(F_n) - T(F_m)| < \varepsilon .$$

Определение 3.7. Функционал $T(F): \Im \to R^1$ непрерывен в F, если $\forall \varepsilon > 0 \ \exists \delta(\varepsilon) > 0$, такое, что $\forall G \in \Im$ выполняется выражение

$$d_{\mathrm{L}}(F, G) < \delta \Rightarrow |T(F) - T(G)| < \varepsilon .$$

Лемма 3.8. Пусть $T_n = T(F_n)$, $F_n \in \Im_n$, тогда если $T(F)$ непрерывен в F, то $T(F_n)$ является состоятельной оценкой $T(F)$, то есть $T(F_n)$ сходится по вероятности к $T(F)$.

Теорема 3.9. Пусть для последовательности оценок $\{T_n\}$ выполнены следующие два условия:

1) $T_n = T(F_n) = T_n(X_1, ..., X_n)$ – непрерывная функция в R^n для каждого n;

2) функционал $T(F)$ непрерывен в F. Тогда последовательность оценок $\{T_n\}$ является устойчивой в $F \in \Im$.

Таким образом, для проверки качественной устойчивости последовательности оценок $\{T_n\}$ нужно убедиться в непрерывности n-мерной функции $T_n = T(F_n)$ в R^n и непрерывности функционала $T(F)$, используя определение (3.7). Подробности и дополнительные сведения могут быть найдены в работе [21] (см. также [4, 9]).

3.Б. Количественный подход к определению робастных процедур

Так же как и качественный, количественный подход к определению робастности процедур опирается на требование, согласно которому произвольно малые изменения в распределении наблюдений должны вызывать лишь достаточно малые изменения в характеристиках качества процедур. Для уточнения этого требования необходимо конкретизировать критерий качества процедуры и наложить определенные ограничения на его поведение в рамках принятой супермодели, описывающей возможные изменения в распределении наблюдений. Ниже мы будем рассматривать несмещенные оценки, которые имеют асимптотически нормальное распределение, поэтому естественной количественной характеристикой оценок является дисперсия асимптотического распределения, и различные определения робастности будут связаны с некоторыми требованиями на ее поведение в рамках принятой супермодели. Перейдем теперь к более точным формулировкам.

Пусть интересующий нас параметр θ задан с помощью некоторого функционала $T(F)$, заданного на множестве $\Im$, то есть $\theta = T(F)$, $F \in \Im$. Для данного параметра θ обозначим через $\aleph$ множество допустимых функционалов. Далее, пусть определена идеальная модель F_0 и задана некоторая супермодель $\Im_\varepsilon(F_0) = \{F \in \Im : d_L(F, F_0) < \varepsilon\}$, причем для любого $F \in \Im$ функционал $T(F) \in \aleph$. Рассмотрим последовательность оценок $\{T_n\}$ вида $T_n = T(F_n)$, $F_n \in \Im_n$. Предполагаем, что оценки T_n состоятельны, то есть $T(F_n)$ сходится по вероятности к $T(F)$ при $n \to \infty$, и асимптотически нормальны, то есть выполняется выражение $L\{\sqrt{n}[T_n - T(F)]/\sigma_F(T_n)\} = N(0,1)$ при $n \to \infty$, где $\sigma_F^2(T_n)$ – асимптотическая дисперсия $\sqrt{n}\,T_n$-оценки. При этих предположениях используют для определения робастности оценок T_n в рамках супермодели $\Im_\varepsilon(F_0)$ такие количественные характеристики, как асимптотическое смещение $T(F) - T(F_0)$ и асимптотическая дисперсия $D_F(T_n) = \sigma_F^2(T_n)/n$. Наиболее часто (см., например, [5]) используют следующие две характеристики. Максимальное смещение в супермодели $\Im_\varepsilon(F_0)$, то есть

$$\sup\{|T(F)-T(F_0)|: F\in\Im_\varepsilon(F_0)\}, \qquad \textbf{(3.10)}$$

и максимальную асимптотическую дисперсию $\sqrt{n}\,T_n$-оценки, то есть

$$\sup\{\sigma_F^2(T_n): F\in\Im_\varepsilon(F_0)\}. \qquad \textbf{(3.11)}$$

Замечание 3.12. При изучении робастности оценок $T_n=T(F_n)$ значения параметра $\theta=T(F)$, $F\in\Im_\varepsilon(F_0)$, должны оставаться постоянными в рамках принятой супермодели $\Im_\varepsilon(F_0)$. Иначе возникает вопрос: что оценивается? Отметим, что это требование выполняется для состоятельных по Фишеру функционалов [6]. В частности, оно выполняется для функционалов, описывающих параметр положения (точку симметрии) для симметричных распределений. В таких случаях робастные оценки $T_n=T(F_n)$ определяются с помощью некоторых требований, накладываемых на поведение лишь асимптотической дисперсии $D_F(T_n)=\sigma_F^2(T_n)/n$ в рамках принятой супермодели, то есть при $F\in\Im_\varepsilon(F_0)$. Обширный перечень таких ограничений приведен в обзорной статье Хьюбера [16] (см. также работы [6, 4, 23]).

Определение 3.13. Последовательность оценок $\{T_n^*\}$ функционала $T(F)$, $F\in\Im_\varepsilon(F_0)$, называют минимаксно-устойчивой относительно расширения идеальной модели до супермодели $\Im_\varepsilon(F_0)$ (или просто минимаксно-робастной в $\Im_\varepsilon(F_0)$), если выполняется неравенство

$$\sup\{\sigma_F^2(T_n^*): F\in\Im_\varepsilon(F_0)\}\le\sup\{\sigma_F^2(T_n): F\in\Im_\varepsilon(F_0)\},$$
$$\forall T(F)\in\aleph. \qquad \textbf{(3.14)}$$

Отметим, что минимаксно-робастная оценка имеет минимальную дисперсию в классе оценок, порождаемых функционалами $T(F)\in\aleph$ при «наихудшей ситуации» в супермодели $\Im_\varepsilon(F_0)$. Если супермодель $\Im_\varepsilon(F_0)$ содержит лишь один элемент (идеальную модель F_0), то минимаксно-робастная оценка является оптимальной в обычном смысле. Отметим также, что минимаксно-робастная оценка T_n^*, являясь оптимальной в смысле (3.14) для супермодели $\Im_\varepsilon(F_0)$ в классе оценок, порождаемых функционалами $T(F)\in\aleph$, может иметь низкую эффективность при идеальной модели $F_0\in\Im_\varepsilon(F_0)$ в сравнении с оптимальной в обычном смысле оценкой для идеальной модели F_0. Это обстоятельство указывает на противоречие, которое возможно между робастностью и эффективностью.

Замечание 3.15. В тех случаях, когда конкретный вид несмещенной оценки с асимптотической дисперсией $\sigma^2(\psi,F)/n$ определяется заданием некоторой функции $\psi\in C_M$, критерий (3.13) может быть записан в виде

$$\inf_{\psi}\sigma^2(\psi,\tilde{F}_0)=\sigma^2(\psi_0,\tilde{F}_0)=\sup_{F}\sigma^{2}(\psi_0,F). \quad \textbf{(3.16)}$$

При этом говорят, что пара $\{\psi_0,\tilde{F}_0\}$, $\psi\in C_M, \tilde{F}_0\in\Im_\varepsilon(F_0)$, решает минимаксную задачу и $\tilde{F}_0$ называют наименее благоприятным распределением в супермодели $\Im_\varepsilon(F_0)$, а функция ψ_0 определяет минимаксно-робастную оценку в рамках супермодели $\Im_\varepsilon(F_0)$. Отметим, что критерий (3.13) был использован Хьюбером [15] для построения *М*-оценок параметра положения минимаксно-робастных в супермодели с засорением.

Замечание 3.17. Сравнение оценок между собой можно проводить по разным числовым характеристикам (см. разд. 2.6). Наиболее часто сравнение двух несмещенных и асимптотически нормальных оценок $\hat{\theta}_1$ и $\hat{\theta}_2$ параметра θ проводится с помощью асимптотической относительной эффективности $АОЭ_F(\hat{\theta}_1:\hat{\theta}_2)$, которая определяется через обратное отношение асимптотических дисперсий $\sqrt{n}\,\hat{\theta}_i$-оценок, $i=1,2$, то есть

$$АОЭ_F(\hat{\theta}_1:\hat{\theta}_2)=\frac{D_F(\hat{\theta}_2)}{D_F(\hat{\theta}_1)}=\frac{\sigma^2(\hat{\theta}_2,F)}{\sigma^2(\hat{\theta}_1,F)}. \quad \textbf{(3.18)}$$

Конечно, эта характеристика может быть вычислена при каждом заданном распределении *F*, однако обычно нас интересуют свойства оценок при различных *F*, принадлежащих некоторому семейству. При этом часто бывает так, что при одном распределении *F* оценка $\hat{\theta}_1$ предпочтительнее оценки $\hat{\theta}_2$, а при другом распределении *F* – картина обратная. Поэтому при изучении свойств робастности оценок не всегда удается провести их сравнение с помощью асимптотической относительной эффективности при изменении распределения *F* в рамках принятой супермодели. Отметим, что некоторая общность достигается при установлении границ изменения относительной эффективности оценок для супермоделей с упорядоченными распределениями (см. [4, 9]).

При изучении качества конкретной оценки $\hat{\theta}$ параметра θ также используют понятие асимптотической эффективности оценки (или абсолютной эффективности оценки), которую определяют в виде

$$AЭ_F(\hat{\theta}) = AOЭ_F(\hat{\theta} : \hat{\theta}_{\text{эфф}}) = \frac{\sigma^2(\hat{\theta}_{\text{эфф}}, F)}{\sigma^2(\hat{\theta}, F)} = \{\sigma^2(\hat{\theta}, F) i_1(\theta)\}^{-1}, \quad \textbf{(3.19)}$$

где $\hat{\theta}_{\text{эфф}}$ – эффективная оценка (или МГД-оценка, оценка с минимальной граничной дисперсией) параметра θ, дисперсия которой равна обратной величине количества информации Фишера $i_1(\theta)$.

Отметим также, что для сравнения совокупности несмещенных и асимптотически нормальных оценок $\hat{\theta}_1, ..., \hat{\theta}_k$ параметра сдвига θ при заданном симметричном распределении *F* используют понятие дефекта оценки [24]. Дефект оценки $\hat{\theta}_i$, $i = 1, ..., k$, среди сравниваемых оценок $\hat{\theta}_1, ..., \hat{\theta}_k$ параметра θ при заданном распределении *F* определяется в виде

$$DE_F(\hat{\theta}_i) = 1 - \min\{\sigma_F^2(\hat{\theta}_1), ..., \sigma_F^2(\hat{\theta}_k)\} / \sigma_F^2(\hat{\theta}_i), \ i = 1, ..., k. \quad \textbf{(3.20)}$$

Заметим, что если среди сравниваемых оценок $\hat{\theta}_1, ..., \hat{\theta}_k$ есть эффективная оценка $\hat{\theta}^*$ параметра сдвига θ при заданном распределении *F*, для которой дисперсия равна обратной величине информации Фишера, то есть $\sigma_F^2(\hat{\theta}^*) = 1 / I(f)$, тогда $\min\{\sigma_F^2(\hat{\theta}_1), ..., \sigma_F^2(\hat{\theta}_k)\} =$

$=1/I(f)$ и, следовательно, в этом случае абсолютный дефект оценки $\hat{\theta}_i$ равен единице минус ее абсолютная эффективность, то есть

$$ADE_F(\hat{\theta}_i) = 1 - АЭ_F(\hat{\theta}_i), \; i = 1,..., k. \qquad \textbf{(3.21)}$$

При изучении свойств робастности сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$ параметра сдвига θ в рамках супермодели $\Im$, состоящей из конечного набора симметричных распределений $\Im = \{F_1,..., F_r\}$, рассматривают поведение дефективностей оценок на плоскости двух распределений (см., например, [24, 4, 9]). По оси абсцисс обычно откладывают дефективность для базовой (идеальной модели, обычно гауссовской), а по оси ординат – дефективность для альтернативной модели, входящей в супермодель $\Im = \{F_1,..., F_r\}$. При таком наглядном представлении дефективностей оценок на плоскости двух распределений предпочтение отдается той оценке, которая окажется ближе к началу координат. Если же мы хотим сделать вывод о предпочтительности оценки среди сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$ параметра θ в рамках всей рассматриваемой супермодели $\Im = \{F_1,..., F_r\}$, то можно использовать евклидову метрику, которая, с использованием введенных обозначений, запишется в виде

$$d(\hat{\theta}_i; \Im) = \left\{ \sum_{j=1}^{r} [DE_{F_j}(\hat{\theta}_i)]^2 \right\}^{1/2}, \; i = 1,..., k. \qquad \textbf{(3.22)}$$

Предпочтение в рамках всей рассматриваемой супермодели $\Im = \{F_1,..., F_r\}$ отдается той оценке $\hat{\theta}_i$ среди сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$, для которой вычисленное значение $d(\hat{\theta}_i; \Im)$ минимальное, то есть

$$d(\hat{\theta}_i; \Im) = \min\{d(\hat{\theta}_1; \Im),..., d(\hat{\theta}_k; \Im)\}. \qquad \textbf{(3.23)}$$

Примеры вычисления этих характеристик для оценок конкретных параметров будут приведены ниже в следующих разделах. Отметим, что к наиболее важной локальной характеристике устойчивости оценок относится понятие «функция влияния», которое обсуждается в следующем разделе.

4. ФУНКЦИЯ ВЛИЯНИЯ И ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ РОБАСТНОСТИ ОЦЕНОК

Рассмотрим оценки $T_n = T_n(X_1,...,X_n)$ в виде функционала от эмпирической функции распределения, то есть $T_n = T(F_n)$, где $T(F)$ – некоторый заданный функционал, являющийся отображением множества непрерывных распределений $\Im$ в действительную ось, то есть $T(F):\Im \to R^1$. Наиболее важной локальной характеристикой устойчивости таких оценок является функция влияния, введенная Хампелем [19 , 21] в виде

$$IF(x;F,T) = \lim_{\lambda \to 0+} \frac{T\{(1-\lambda)F + \lambda\Delta_x\} - T(F)}{\lambda},\ 0 \le \lambda \le 1,\ x \in R^1, \quad \textbf{(4.1)}$$

для тех $x \in R^1$, при которых предел существует. Здесь Δ_x обозначает вырожденную функцию распределения в точке $x \in R^1$. Функция влияния дополнительно будет обсуждаться в разделе 6. Её упоминание здесь вызвано лишь необходимостью охарактеризовать числовые характеристики устойчивости оценок, которые вычисляются на основе этой функции. Отметим, что функция влияния описывает асимптотическое влияние одного наблюдения равного x на величину оценки $T_n = T(F_n)$ и позволяет вычислить асимптотическую дисперсию $\sqrt{n}T(F_n)$-оценки, которую обозначают в литературе через $V(T,F)$, или $\sigma_F^2(T_n)$, и вычисляют (детали см. в разделе 6) по формуле

$$V(T,F) = \sigma_F^2(T_n) = \int_{-\infty}^{\infty} IF^2(x;F,T)dF(x). \quad \textbf{(4.2)}$$

Используя функцию влияния, Хампель [19] определил следующие числовые характеристики устойчивости оценок.

1. Чувствительность оценки $T_n = T(F_n)$ к грубым ошибкам определяется в виде

$$\gamma^*(F,T) = \sup_x |IF(x;F,T)|. \qquad \textbf{(4.3)}$$

Отметим, что величина $\gamma^*(F,T)$ связана с максимальным смещением (3.10) оценки $T_n = T(F_n)$ в рамках заданной супермодели. Она характеризует влияние фиксированной доли «засорения» выборки на значение оценки и может быть использована в качестве приближенной границы для модуля смещения оценки. В частности, для супермодели с засорением вида

$$\Im_\varepsilon(F_0) = \{F : F(x) = (1-\varepsilon)F_0(x) + \varepsilon H(x)\},\ 0 \le \varepsilon \le 1, \qquad \textbf{(4.4)}$$

где F_0 – заданная функция распределения (идеальная модель) и $H(x)$ – произвольная непрерывная функция распределения, максимальное смещение оценки $T_n = T(F_n)$ пропорционально величине $\gamma^*(F,T)$. В самом деле, используя функцию влияния $IF(x;F,T)$, исходный функционал $T(F)$ можно линеаризовать вблизи идеальной модели F_0. Если распределения F и F_0 «близки», то согласно (6.5) с учетом (6.19), можно записать приближенное равенство

$$T(F) \approx T(F_0) + \int IF(x;F_0,T)d[F(x) - F_0(x)] + \cdots . \qquad \textbf{(4.5)}$$

Учитывая, что $F \in \Im_\varepsilon(F_0)$, то есть $F = F_0 + \varepsilon(H - F_0)$, и равенство $\int IF(x;F_0,T)dF_0(x) = 0$ (см. формулу (6.14)), выражение (4.5) перепишем в виде

$$T(F) - T(F_0) \approx \varepsilon \int IF(x;F_0,T)\,dH(x).$$

С учетом этого выражения, приближенное значение для максимального смещения (3.10) в рамках супермодели $\Im_\varepsilon(F_0)$ вида (4.4) записывается как

$$\sup_{F \in \Im_\varepsilon(F_0)} |T(F) - T(F_0)| \approx \varepsilon \sup_H \left| \int IF(x; F_0, T) dH(x) \right| =$$

$$= \varepsilon \sup_x |IF(x; F_0, T)| = \varepsilon \gamma^*(F_0, T). \qquad \textbf{(4.6)}$$

2. *Чувствительность оценки* $T_n = T(F_n)$ *к локальным изменениям наблюдений* определяется в виде

$$\lambda^*(F,T) = \sup_{x \neq y} |IF(x; F, T) - IF(x; F, T)| / |x - y|. \qquad \textbf{(4.7)}$$

Величина $\lambda^*(F,T)$ позволяет учитывать влияние эффектов округления и группировки наблюдений и, по существу, для дифференцируемой функции влияния характеризует ее максимальный наклон.

Итак, желательными с точки зрения робастности свойствами оценок $T_n = T(F_n)$ являются ограниченность функции влияния и конечность характеристик $\gamma^*(F,T)$ и $\lambda^*(F,T)$.

Замечание 4.8. Практическая целесообразность введения функции влияния состоит не только в том, что мы получаем возможность наглядного представления о влиянии отдельного наблюдения на величину оценки, но и в том, что с ее помощью можно вычислить такие характеристики оценки, как асимптотическую дисперсию, чувствительность оценок к грубым ошибкам и округлению и, кроме того, эта функция, как отмечается в [16, 22], является основным инструментом для построения оценок с заданными свойствами робастности. Например, полагая функцию влияния равной нулю вне некоторого интервала, мы тем самым вводим жесткое правило отбраковки выделяющихся наблюдений и устанавливаем связь между свойствами робастных оценок и критериями обнаружения выбросов в выборке. В связи с этим Хампель [22] ввел понятие «точка отбраковки – rejection point ».

3. *Точка отбраковки для оценки* $T_n = T(F_n)$ определяется в виде

$$\rho^*(F,T) = \inf\{ r > 0 : IF(x; F, T) = 0, \ |x| > r \}. \qquad \textbf{(4.9)}$$

Если оценка имеет функцию влияния, равную нулю вне интервала $[-r, r]$, то это означает, что все наблюдения вне этого интервала при построении оценки полностью игнорируются. В связи с этим желательным свойством для робастной оценки к наличию выбросов является конечность величины $\rho^*(F,T)$.

Замечание 4.10. Отметим, что асимптотическая дисперсия $V(T,F)$, вычисляемая по формуле (4.2), также представляет некоторый функционал от распределения наблюдений F, зависящий и от исходного функционала $T(F)$. По-видимому, с учетом этого факта, в работе [6] введено по аналогии с функцией влияния понятие «функция изменения дисперсии – change-of-variance function».

4. *Функция изменения дисперсии* $V(T,F)$ *оценки* $T_n = T(F_n)$ определяется в виде

$$CVF(x;F,T) = \lim_{\lambda \to 0+} \frac{V\{T,[(1-\lambda)F + \lambda\Delta_x]\} - V\{T,F\}}{\lambda}, \quad 0 \le \lambda \le 1, \tag{4.11}$$

для тех $x \in R^1$, при которых предел существует. Отметим, что формула (4.2) для вычисления асимптотической дисперсии оценок $T_n = T(F_n)$ справедлива только для гладких функций распределения F. Если условия гладкости не выполняются, то функция (4.11) не может быть определена. Однако в работе [6] отмечается, что эту трудность можно преодолеть, определив функцию равенством $CVF(x;F,T) = \chi(x;F,T)$, где функция $\chi(x;F,T)$ задается уравнением для некоторой гладкой ф.р. H:

$$\frac{\partial}{\partial\lambda}[V(T,(1-\lambda)F + \lambda H)]\big|_{\lambda=0} = \int \chi(x F,T)dH(x). \tag{4.12}$$

Числовая характеристика $\gamma^*(F,T)$ вида (4.3) была введена на основе функции влияния $IF(x;F,T)$. Используя функцию изменения дисперсии $CVF(x;F,T)$ вида (4.11), в работе [6] вводится понятие «чувствительность к изменению дисперсии».

5. *Чувствительность к изменению дисперсии оценки* $T_n = T(F_n)$ определяется в виде

$$\omega^*(T,F) = \sup_x CVF(x;F,T)/V(T,F). \tag{4.13}$$

Отметим, что данная числовая характеристика связана с максимальной дисперсией оценки в рамках заданной супермодели (детали см. в [6, с. 209]). Используя (4.5), для супермодели (4.4) получаем приближенное равенство для максимального значения асимптотической дисперсии в виде

$$\sup_{F\in\Im_\varepsilon(F_0)} V(T,F)=\exp\{\sup_H[\ln V(T,(1-\varepsilon)F_0+\varepsilon H)]\}\approx$$

$$\approx\exp\{\sup_H[\ln V(T,F_0)+\varepsilon\int[CVF(x;T,F_0)/V(T,F_0)]dH(x)]\}=$$

$$=\exp\{\ln V(T,F_0)+\varepsilon\sup_x[CVF(x;T,F_0)/V(T,F_0)]\}=$$

$$=V(T,F_0)\exp\{\varepsilon\omega^*(T,F)\}. \qquad \textbf{(4.14)}$$

Таким образом, числовые характеристики $\gamma^*(F,T)$ и $\omega^*(T,F)$ тесно связаны с максимальным смешением (3.10) и максимальной дисперсией (3.11) для супермодели с засорением (4.4). Эти числовые характеристики использованы в [6] для введения следующего определения.

Определение 4.15. Оценка $T_n = T(F_n)$ функционала $T(F)$, $F\in\Im$, называется B-робастной в заданной супермодели, если $\gamma^*(F,T)<\infty$, и называется V-робастной, если $\omega^*(T,F)<\infty$. Оценку, имеющую минимальное значение числовой характеристики $\gamma^*(F,T)$ в заданной супермодели, называют *наиболее B-робастной*, а оценку, обладающую наименьшим значением $\omega^*(T,F)$ в заданной супермодели, называют *наиболее V-робастной*.

Замечание 4.16. Описанные выше числовые характеристики относятся к локальным характеристикам, которые основаны либо на функции влияния, либо на функции изменения дисперсии. К глобальным характеристикам устойчивости оценок относится понятие «предел устойчивости, или пороговая точка – breakdown point», введенное в [21, 22].

Определение 4.17. Пороговая точка $\varepsilon^*(T,F)$ оценки $T_n=T(F_n)$ функционала $T(F)$, $F\in\Im$, определяется в виде

$$\varepsilon^*(T,F)=\sup\{\,\varepsilon:\sup_{F:d_L(F,F_0)<\varepsilon}|T(F)-T(F_0)|<\infty\,\}, \qquad \textbf{(4.18)}$$

где F_0 – идеальная модель и $d_L(F,F_0)$ – метрика Леви, определенная в (2.6).

Отметим, что пороговая точка $\varepsilon^*(T,F)$ относится к глобальным числовым характеристикам робастности оценок $T_n = T(F_n)$, её значение не зависит от распределения наблюдений, и она характеризует максимально возможное отклонение распределения наблюдений F (в смысле метрики Леви) от идеальной модели F_0, при котором смещение оценки $T_n = T(F_n)$ еще остается ограниченным. Отметим также, что в определении пороговой точки могут быть использованы и другие метрики. Метрика Леви использована здесь лишь для определенности. В частности, для супермодели с засорением вида (4.4) пороговая точка $\varepsilon^*(T,F)$ указывает на максимально возможную пропорцию засорения выборки, при которой смещение оценки $T_n = T(F_n)$ еще остается ограниченным. Так как пороговая точка $\varepsilon^*(T,F)$ не зависит от распределения F, то для оценки $T_n = T(F_n)$ функционала $T(F)$ будем обозначать её в виде $\varepsilon^*(T_n)$.

Замечание 4.19. Аналогичное понятие толерантности оценки T_n к наличию экстремальных значений в выборке, применимое при конечном объеме выборки n, было введено В. Ходжесом в работе [27] в следующем виде. Пусть существуют целые числа $r > 0$ и $m > 0$, такие, что выполняются неравенства $x_{(n-m)} \le T_n \le x_{(r+1)}$, и при любых фиксированных $x_{(r+2)},\dots,x_{(n)}$ условие $x_{(r+1)} \to \infty$ влечет, что оценка $T_n \to \infty$, а при любых фиксированных $x_{(1)},\dots,x_{(n-m-1)}$ условие $x_{(n-m)} \to \infty$ влечет, что оценка $T_n \to -\infty$. Тогда говорят, что оценка T_n толерантна к r экстремальным значениям справа и к m экстремальным значениям слева. Обычно $r = m$ и их общее значение обозначается через $\varepsilon^*(n)$, которое при больших объемах выборки совпадает с асимптотическим значением $\varepsilon^*(T_n)$. Обобщение понятия толерантности оценок на статистики критериев проверки гипотез приводится в работах [6, 9], (см. также работы: Ylvisaker D. Test resistance // J. Am. Stat. Assoc. 1977. V. 72. P. 551–556; Rieder H. Qualitative robustness of rank tests // Ann. Statist. 1982. V. 10. P. 205–211).

Замечание 4.20. Аналогом функции влияния при конечных объемах выборки n является «кривая чувствительности – sensitivity curve», которая была введена Тьюки (см., например, [24]) в виде

$$SC_n(x) = n[T_n(x_1,...,x_{n-1},x) - T_{n-1}(x_1,...,x_{n-1})]. \qquad \mathbf{(4.21)}$$

Для оценок T_n в виде функционала $T(F)$ от эмпирической ф.р. F_n, то есть для оценок вида $T_n = T(F_n)$, кривая чувствительности записывается как

$$\begin{aligned} SC_n(x) &= n[T\{(1-(1/n))F_{n-1} + (1/n)\Delta_x\} - T(F_{n-1})] = \\ &= [T\{(1-(1/n))F_{n-1} + (1/n)\Delta_x\} - T(F_{n-1})]/(1/n), \end{aligned} \qquad (4.22)$$

где F_{n-1} – эмпирическая функция распределения, построенная по выборке $X_1,...,X_{n-1}$, и Δ_x – вырожденная в точке x функция распределения. Отметим, что правая часть выражения (4.22) по форме совпадает с выражением для функции влияния (4.1), при этом F_{n-1} используется в качестве аппроксимации ф.р. F и пропорция засорения равна $\lambda = 1/n$. Во многих случаях кривая чувствительности приближается к функции влияния, то есть $SC_n(x) \to IF(x;F,T)$ при $n \to \infty$. Обсуждение этого понятия и конкретные примеры могут быть найдены в [6]. Отметим, что существует тесная связь метода расщепления (складного ножа) с функцией влияния оценок функционалов и кривой чувствительности. Эти вопросы обсуждаются в [9]. Отметим также, что понятие «функция влияния» для статистик критериев проверки гипотез, которое было введено Рончетти в 1979 году, и её связь с эффективностью критериев по Питмену (см. приложение П.2 и работы [20, 28]), обсуждаются в следующем разделе 4.А, см. также работы [6, 9].

4.А. Функция влияния для статистик критериев

Отметим еще раз, что термин «робастность» был введен в статистическую литературу в 1953 году Боксом при изучении им свойств критериев Бартлетта и F-критерия Фишера в ситуациях возможных отклонений от нормальной модели (см., например, [3, 9]). Работы Тьюки, Хьюбера и Хампеля послужили мощным толчком к изучению свойств робастности оценок параметров, представляемых в виде некоторого функционала $T(F)$, $F \in \mathfrak{I}$. При изучении свойств робастности оценок параметров вида $T_n = T_n(X_1,...,X_n) = T(F_n)$ важную роль играет очень удобное и конструктивное понятие «функция влияния» $IF(x;F,T_n)$ оценки T_n, предложенное Хампелем [19, 21, 22]. Следуя работе [6], обсудим понятие «функция влияния» для статистик критериев проверки гипотез, которое было введено Рончетти в 1979 году. Рассмотрим параметрическую модель в одновыборочном варианте. Обобщение на двувыборочный вариант приводится в [6].

Итак, предположим, что задана параметрическая статистическая модель $(\mathrm{X}, \mathfrak{I}_\theta)$. Здесь X – выборочное пространство и $\mathfrak{I}_\theta$ – заданное параметрическое множество допустимых распределений вероятностей, определяемое в виде $\mathfrak{I}_\theta = \{F : F_X(x,\theta), \theta \in \Theta\}$. Имеется выборка $X_1,...,X_n$ из распределения $F_X(x,\theta)$ с плотностью $f_X(x,\theta)$, $x \in R^1$, $\theta \in \Theta$, причем, функциональный вид распределения задан с точностью до неизвестного параметра θ, который принадлежит заданному параметрическому множеству Θ. Требуется по выборке $X_1,...,X_n$ из распределения $F_X(x,\theta)$ проверить гипотезу $H_0 : \theta = \theta_0$ против односторонней альтернативы $H_1^+ : \theta > \theta_0$. Отметим сразу же, что обсуждение варианта двусторонней альтернативы $H_1 : \theta \neq \theta_0$ проводится аналогично. Для удобства записи обозначим $F_\theta = F(x,\theta)$ и $F_0 = F(x,\theta_0)$, где θ_0 – заданное значение неизвестного параметра θ. Пусть критерий проверки гипотез $H_0 : \theta = \theta_0$, $H_1^+ : \theta > \theta_0$ основан на статистике $S_n = S_n(X_1,...,X_n)$ и определяется критической областью размера α в виде $\{S_n > s_{1-\alpha}\}$, где $s_{1-\alpha}$ – квантиль уровня $(1-\alpha)$ распределения тестовой статистики S_n при справедливости основной гипотезы H_0. Рассмотрим ситуацию, когда тестовая статистика S_n записывается в виде $S_n = S_n(X_1,...,X_n) = S(G_n)$, где G_n – эмпирическая функция распределения, построенная по наблюдениям с реальной ф.р. G, не обязательно совпадающей с ф.р. F_θ, принятой в параметрической модели $(\mathrm{X}, \mathfrak{I}_\theta)$.

Напомним (см. формулу (4.1)), что понятие «функция влияния» $IF(x;F,T_n)$ оценки $T_n = T(F_n)$ функционала $T(F)$, характеризующего интересующий нас параметр $\theta = T(F)$, было определено выражением (4.1). Определенная таким образом функция влияния является удобным инструментом при изучении свойств робастности оценок $T_n = T(F_n)$, *состоятельных по Фишеру* функционалов $T(F)$, для которых выполняется условие

$$T(F_\theta) = \theta \ \ \forall\theta \in \Theta, \qquad \textbf{(4.23)}$$

а также выполняется полезная и удобная формула (4.2) для вычисления асимптотической дисперсии $\sqrt{n}\,T_n$-оценки через функцию влияния $IF(x;F,T_n)$ в виде

$$\sigma_F^2(T_n) = \int\limits_{-\infty}^{\infty} IF^2(x;F,T)dF(x). \qquad \textbf{(4.24)}$$

При изучении свойств критериев проверки статистических гипотез возникают следующие особенности. В работе [6] отмечается, что тестовые статистики $S_n = S_n(X_1,...,X_n) = S(F_n)$, как правило, являются оценками функционалов, которые не удовлетворяют условию (4.23), то есть не являются состоятельными по Фишеру. По этой причине, в указанной работе понятие «функция влияния для статистик критериев» модифицируется следующим образом.

Пусть отображение $\xi_n : \Theta \to R^1$ определяется формулой $\xi_n(\theta) = M_\theta(S_n)$. Положим $\xi(\theta) = S(F_\theta)$ и предполагаем, что выполнены условия:

1) $\xi_n(\theta)$ *сходится к* $\xi(\theta)$ при $n \to \infty$ для всех $\theta \in \Theta$;

2) $\xi(\theta)$ строго монотонная непрерывная функция, для которой существует ξ^{-1}.

Определим новый функционал $U(G) = \xi^{-1}[S(G)]$. Этот функционал дает значение параметра, которое истинное распределение G имело бы, если бы оно совпадало с распределением F_θ, принятым в модели. Отметим, что функционал $U(G)$ является состоятельным по

Фишеру, поскольку $U(F_\theta) = \xi^{-1}[T(F_\theta)] = \xi^{-1}[\xi(\theta)] = \theta$, $\forall \theta \in \Theta$. Теперь для этого нового функционала $U(F_\theta)$ рассмотрим введенную Хампелем функцию влияния по аналогии с формулой (4.1).

Определение 4.25. Функция влияния $IF_{\text{кр}}(x; F_\theta, S)$ статистики S_n критерия проверки гипотез H_0 и H_1^+ по наблюдениям $X_1, \ldots, X_n$ из распределения F_θ определяется в виде

$$IF_{\text{кр}}(x; F_\theta, S) = IF(x; F_\theta, U) = \lim_{\lambda \to 0} \frac{U[(1-\lambda)F_\theta + \lambda \Delta_x] - U(F_\theta)}{\lambda}$$
$$0 \le \lambda \le 1,$$

в тех точках *x*, в которых предел существует. Здесь, как и прежде, Δ_x обозначает вырожденную в точке x функцию распределения.

Замечание 4.26. Функция влияния $IF_{\text{кр}}(x; F_\theta, S)$ статистики S_n критерия определяет влияние, которое оказывает резко выделяющееся наблюдение на значение (нормированной) статистики S_n, следовательно, и на решение (принять или отвергнуть гипотезу H_0). Таким образом, ограниченность этой функции, как и прежде, отвечает конечной чувствительности теста к наличию выбросов в выборке. Для практического вычисления этой функции используют формулу вида

$$\begin{aligned} IF_{\text{кр}}(x; F_\theta, S) &= \frac{\partial}{\partial \lambda} S\left[(1-\lambda)F_\theta + \lambda \Delta_x\right]\big|_{\lambda=0} / \xi'(\theta) = \\ &= IF(x; F_\theta, S) / \xi'(\theta). \end{aligned} \qquad \textbf{(4.27)}$$

На практике функцию влияния $IF_{\text{кр}}(x; F_\theta, S)$ статистики S_n критерия обычно вычисляют при справедливости основной гипотезы $H_0 : \theta = \theta_0$ и при этом используют обозначение $F_0 = F(x, \theta_0)$.

Отметим важные свойства функции влияния $IF_{\text{кр}}(x; F_\theta, S)$. Напомним, что для функции влияния $IF(x; F, T)$ оценки $T_n = T(F_n)$ функционала $T(F)$, определенной в (4.1), выполняется равенство $\int IF(x; F, T) dF(x) = 0$. Из этого равенства и формулы (4.27) следует аналогичное равенство и для функции влияния $IF_{\text{кр}}(x; F_\theta, S)$, то есть

$$\int IF_{\text{кр}}(x;F_0,T)dF_0(x)=0\,.$$

Далее, для функции влияния $IF(x;F,T)$ оценки $T_n=T(F_n)$ функционала $T(F)$ выполняется выражение (4.24). Использование этого выражения и неравенства Рао – Крамера позволяет проиллюстрировать связь между функцией влияния $IF(x;F,T)$ и асимптотической эффективностью оценок $T_n=T(F_n)$ состоятельных по Фишеру функционалов $T(F)$. Эта связь проявляется в том, что в рамках параметрической модели $(\mathrm{X},\Im_\theta)$ оценка $T(F_n)$ неизвестного значения θ_0 параметра θ асимптотически эффективна, если выполняется равенство

$$IF(x;F_0,T)=[\partial\ln f(x,\theta)/\partial\theta]|_{\theta_0}\;/\,I(F_0)\,.$$

Отметим, что равенство типа (4.24) выполняется и для функционалов, которые не являются состоятельными по Фишеру, то есть для функции влияния $IF_{\text{кр}}(x;F_\theta,S)$ критерия также выполняется равенство $D(S_n\,|\,H_0)=\sigma^2(F_0,S_n)/n$, где

$$\sigma^2(F_0,S_n)=\int\limits_{-\infty}^{\infty}IF_{\text{кр}}^2(x;F_0,S)dF_0(x)\,. \tag{4.28}$$

Однако это равенство не может быть использовано для построения оптимальных тестов проверки статистических гипотез. Для этой цели используют связь между функцией влияния $IF_{\text{кр}}(x;F_\theta,S)$ и эффективностью Питмена критериев (см. П.2) в виде

$$\int IF_{\text{кр}}^2(x;F_0,S)dF_0(x)=1/Э^2(S_n,F_0)\,, \tag{4.29}$$

где, согласно (П.2.5), квадрат эффективности Питмена $Э(T_n,F_0)$ для T_n-критерия, основанного на статистике $T_n=T_n(X_1,...,X_n)$, вычисляется при справедливости гипотезы $H_0:\theta=\theta_0$ по формуле

$$Э^2(T_n,F_{\theta_0})=\lim_{n\to\infty}\frac{[\partial M(T_n,F_\theta)/\partial\theta\,|_{\theta=\theta_0}]^2}{nD(T_n,F_{\theta_0})}=\lim_{n\to\infty}\frac{[\xi_n'(\theta_0)]^2}{nD(T_n,F_{\theta_0})}. \tag{4.30}$$

Для проверки формулы (4.29) следует использовать (4.27) и (4.30). В самом деле,

$$\int IF_{\text{кр}}^2(x;F_{\theta_0},S)dF_{\theta_0}(x) = \frac{1}{[\xi_n'(\theta_0)]^2}\int IF^2(x;F_{\theta_0},S)dF_{\theta_0} = \frac{\sigma^2(F_{\theta_0},S_n)}{[\xi_n'(\theta_0)]^2} =$$

$$= \lim_{n\to\infty}\frac{nD(S_n,F_{\theta_0})}{[\xi_n'(\theta_0)]^2} = \frac{1}{Э^2(S_n,F_{\theta_0})}.$$

Использование выражения (4.29) позволяет вычислять, согласно теореме (П.2.4), асимптотическую относительную эффективность двух критериев ($T_{1,n}$-критерия и $T_{2,n}$-критерия) через функции влияния этих статистик по формуле

$$ARE_F(T_{1,n}:T_{2,n}) = Э^2(T_{1,n},F)/Э^2(T_{2,n}F) = \frac{\int IF_{\text{кр}}^2(x;F,T_{2n})dF(x)}{\int IF_{\text{кр}}^2(x;F,T_{1n})dF(x)}. \qquad \textbf{(4.31)}$$

Отметим также, что асимптотическая мощность T_n-критерия может быть выражена, согласно теореме П.2.7, через функцию влияния $IF_{\text{кр}}(x;F,T_n)$ тестовой статистики T_n этого критерия. Согласно (П.2.8), асимптотическая мощность T_n-критерия проверки гипотез $H_0:\theta=\theta_0$, $H_1^+:\theta>\theta_0$ для последовательности альтернатив $\theta_n=\theta_0+\theta n^{-1/2}$, $\theta>0$, может быть вычислена по формуле

$$\lim_{n\to\infty} P\{T_n \ge k_{n,1-\alpha} \mid H_{1,n}^+\} = 1-\Phi(\lambda_{1-\alpha}-\theta\cdot Э(F,T_n)) =$$
$$= 1-\Phi\left(\lambda_{1-\alpha}-\theta\cdot\left(\int IF_{\text{кр}}^2(x;F,T_n)dF(x)\right)^{-1/2}\right), \qquad \textbf{(4.32)}$$

где $\lambda_{1-\alpha}=\Phi^{-1}(1-\alpha)$ и α – уровень значимости T_n-критерия.

Замечание 4.33. Напомним, что всякий T_n-критерий проверки статистических гипотез H_0,H_1, основанный на статистике $T_n(\vec{X})$, вычисляемой по выборке $X_1,...,X_n$ из распределения $F(x)$, характеризуется двумя вероятностями ошибок первого и второго рода. Вероятность ошибки первого рода используется для определения критической области тестовой статистики T_n путем задания уровня значимости α-критерия. Вероятность ошибки второго рода связана с понятием функции мощности критерия, которое отражает поведение вероятности правильного решения в пользу верной альтернативы. При изучении свойств робастности критериев в рамках заданной супермодели изучают стабильность поведения уровня значимости критерия и его мощности при изменении распределения наблюдений в заданной супермодели. В связи с этим возникает вопрос, почему используется понятие «функция влияния статистики критерия», а не такие характеристики, как «функция влияния на уровень» и «функция влияния на мощность»? В работе [6] вводятся такие характеристики робастности критериев проверки гипотез и отмечается, что они пропорциональны функции влияния $IF_{\text{кр}}(x;F,T_n)$, определенной в (4.25). Также отмечается, что при проверке гипотез $H_0:\theta=\theta_0$, $H_1^+:\theta>\theta_0$ с помощью T_n-критерия используют нормированные статистики $U_n=\xi_n^{-1}(T_n)$, которые сходятся по вероятности к $U=\xi^{-1}(T)$, где $T(F)$ – функционал, соответствующий тестовой статистике T_n. Критические значения $u_n(\alpha)$ статистики $U_n=\xi_n^{-1}(T_n)$ определяются выражением $P\{U_n\ge u_n(\alpha)\}=\alpha$, где α – заданный уровень значимости U_n-критерия проверки гипотез H_0 и H_1^+. Для асимптотически нормальных статистик $U_n=\xi_n^{-1}(T_n)$, для которых выполняется выражение

$$L\left\{\sqrt{n}[U_n-M(U_n\,|\,H_0)]/\left[\int IF^2(x;F,T_n)dF(x)\right]^{1/2}\right\}=N(0,\ 1), \quad \textbf{(4.34)}$$

«функция влияния на уровень $IF_L(\cdot)$» и «функция влияния на мощность $IF_P(\cdot)$» вычисляются (см. [6]) по формулам

$$IF_L(x;F,U_n)=Э(U_n,F)\phi(\lambda_{1-\alpha})IF_{\text{кр}}(x;F,T_n); \quad \textbf{(4.35)}$$

$$IF_P(x;F,U_n)=Э(U_n,F)\phi[\lambda_{1-\alpha}-\theta Э(F,U_n)]IF_{\text{кр}}(x;F,T_n), \quad \textbf{(4.36)}$$

где $\theta>0$ – заданная константа, используемая при построении последовательности контигуальных альтернатив $\theta_n=\theta_0+\theta n^{-1/2}$, которые используются при вычислении мощности U_n-критерия, $\lambda_{1-\alpha} = \Phi^{-1}(1-\alpha)$ и $Э(U_n, F)$ – эффективность Питмена для U_n-критерия. Отметим, что индексы (L) и (P) в обозначениях функций влияния соответствуют словам Level (уровень) и Power (мощность). Таким образом, как IF_L, так и IF_P прямо пропорциональны функции влияния статистики критерия $IF_{\text{кр}}(\cdot)$. Интерпретируются функции влияния $IF_L(\cdot)$ и $IF_P(\cdot)$, как и прежде, следующим образом. Если выборка содержит выбросы, то это оказывает влияние на вероятность того, что критерий отвергнет нулевую гипотезу. Например, факт наличия выбросов в выборке может привести к тому, что истинное распределение за счет засорения выбросами может попасть в область распределений альтернативы, и нулевая гипотеза будет ошибочно отвергнута. Напомним, что одновыборочные ранговые критерии (см. [34, 45, 46]) обладают свойством непараметричности и асимптотической непараметричности, то есть сохраняют заданный уровень значимости при выполнении условия, что выборка $X_1,...,X_n$ характеризуется как последовательность *н.о.р. случайных величин*. Отметим, что при выполнении *этого условия* асимптотическая нормальность статистики для двухвыборочного критерия Манна – Уитни (и эквивалентного рангового критерия Уилкоксона) доказана методом проекций (см. П.1) и работы [45, 8]. Наличие выбросов среди наблюдений нарушает это предположение и, как следствие, реальный уровень значимости может существенно отличатся от заданного (см. [2, 3, 45] и пример 6.2.4 в [9]). Для двухвыборочных критериев однородности уровень значимости остается постоянным лишь в том случае, когда распределения обеих выборок в точности совпадают. Наличие выбросов в выборках нарушает это условие и приводит к неконтролируемому изменению уровня значимости критерия однородности. Особенности, которые возникают при введении функций влияния $IF_L(\cdot)$ и $IF_P(\cdot)$ для двувыборочного варианта, обсуждаются в [6, с. 233] (см. также работу [45] и раздел 5.4 в [8]).

5. СВЯЗИ МЕЖДУ РАЗЛИЧНЫМИ ПОНЯТИЯМИ РОБАСТНОСТИ ОЦЕНОК

Основы современного подхода в теории робастности оценок были заложены в работах Тьюки [11–14], Хьюбера [5,15,16] и Хампеля [6, 19, 21, 22]. Хьюбер ввел понятие минимаксно-робастной оценки и построил для параметра положения *M*-оценки, оптимальные в минимаксном смысле для супермодели с засорением вида (4.4). Суть подхода Хьюбера состоит в следующем. Он рассмотрел класс *M*-оценок, определяемых некоторой функцией $\psi \in C_M$, асимптотическая дисперсия которых $V(\psi, F)$ зависит от выбора функции ψ и исходной ф.р. F наблюдений. При этом в качестве «окрестности» базовой симметричной ф.р. F_0 он использовал супермодель с симметричным засорением вида

$$\Im_\varepsilon^s(F_0) = \{F : F(x) = (1-\varepsilon)F_0(x) + \varepsilon H_s(x)\},\ 0 \le \varepsilon \le 1, \qquad \textbf{(5.1)}$$

где F_0 – заданная симметричная функция распределения (идеальная модель) и $H_S(x)$ – произвольная непрерывная симметричная функция распределения. Далее задача сводилась к минимизации асимптотической дисперсии $V(\psi, F)$ на множестве $\Im_\varepsilon^s(F_0)$, точнее, требовалось отыскать такую функцию $\psi_0 \in C_M$, для которой бы выполнялось выражение

$$\sup_{F \in \Im_\varepsilon(F_0)} V(\psi_0, F) = \min_{\psi \in C_M} \sup_{F \in \Im_\varepsilon(F_0)} V(\psi, F). \qquad \textbf{(5.2)}$$

Решение данной задачи было получено путем нахождения наименее благоприятного распределения $\tilde{F}_0$ с плотностью $\tilde{f}_0$, при которой информация Фишера $I(f_0)$ минимальна в рамках супермодели

$\Im_\varepsilon^s(F_0)$. При этом минимаксно-робастная оценка, определяемая функцией $\psi_0 = -\tilde{F}_0'' / \tilde{F}_0'$, является оценкой максимального правдоподобия для наименее благоприятного распределения $\tilde{F}_0$. Отметим также, что Хьюбер [16] рассматривал и другие супермодели, для которых при построении окрестностей базовых распределений использовались различные метрики, в частности метрика Колмогорова. Развитие подхода Хьюбера, рассмотрение супермоделей с асимметричным засорением и перенесение полученных результатов для *M*-оценок на *L*- и *R*-оценки, было осуществлено Джекелем в работах [25, 26] (см. также [4, 9]). Наиболее полно подход Хьюбера и связанные с ним результаты исследований других авторов представлены в его монографии [5].

Наряду с подходом Хьюбера, независимо и чуть позднее, Хампель и др. [6] развили подход, основанный на функции влияния. В 1968 году в своей диссертационной работе Хампель [21] ввел понятия «качественная устойчивость», «функция влияния» и определил через эту функцию различные числовые характеристики устойчивости оценок. Позднее, в работе [6] было введено понятие «*B*-робастная оценка». По аналогии с функцией влияния было введено понятие «функция изменения дисперсии» и числовая характеристика, названная чувствительностью изменения дисперсии, а также понятие «*V*-робастная оценка» (см. определение 4.15).

Следуя работе [6], обсудим связи между подходом Хьюбера и подходом, основанным на функции влияния. Для этого воспользуемся аппроксимациями (4.6) и (4.14) соответственно выражений (3.10) и (3.11). Отметим, что качество этих аппроксимаций при малых значениях ε очень хорошее. Например, для выборочной медианы $\overline{X}_{1/2}$ при $\varepsilon = 0,05$ в рамках супермодели $\Im_\varepsilon(\Phi)$, то есть при $F_0 = \Phi$, получаем точное значение равное (1,741) (см. [4, с. 176]), а приближенное значение, вычисленное по формуле (4.14), с учетом того, что $\sigma_\Phi^2(\overline{X}_{1/2}) = 1,571$ и $\omega^* = 2$, равно

$$\sup_{F\in\Im_\varepsilon(\Phi)} \sigma_F^2(\overline{X}_{1/2}) \approx \sigma_\Phi^2(\overline{X}_{1/2})\exp\{\varepsilon\omega^*(\overline{X}_{1/2},\Phi)\} =$$
$$=1,571\cdot\exp\{0,05\cdot 2\} = 1,571\cdot 1,1052 = 1,736.$$

Для оценки Ходжеса – Лемана HL при $\varepsilon = 0,05$ в рамках супермодели $\Im_\varepsilon(\Phi)$ по формуле (3.11) получаем точное значение равное 1,286 (см. [4, с. 176]), а приближенное значение, вычисленное по формуле (4.14), с учетом того, что $\sigma_\Phi^2(HL) = 1,047$ и $\omega^* = 4$, равно

$$\sup_{F\in\Im_\varepsilon(\Phi)} \sigma_F^2(HL) \approx \sigma_\Phi^2(HL)\exp\{\varepsilon\omega^*(HL,\Phi)\} =$$
$$=1,047\cdot\exp\{0,05\cdot 2\} = 1,047\cdot 1,2214 = 1,279.$$

Дополнительные примеры, иллюстрирующие качество аппроксимаций (4.6) и (4.14) для других оценок, могут быть найдены в работе [6, см. табл.1, с. 210].

Итак, принимая во внимание, что качество аппроксимации выражения $\sup\{V(\psi,F): F\in\Im_\varepsilon(F_0)\}$ с помощью приближенного выражения (4.14) является достаточно хорошим, минимаксная задача Хьюбера вида (5.2), может быть заменена оптимизационной задачей вида

$$\min_{\psi\in C_M} V(\psi,F_0)\cdot\exp\{\varepsilon\omega^*(\psi,F_0)\}. \qquad \textbf{(5.3)}$$

Постановка задачи Хампелем в таком виде и составляет суть подхода, основанного на функции влияния. Решение этой задачи связано с нахождением функции $\tilde\psi\in C_M$ при ограничении $\omega^*(\psi,F_0)\le\omega := \omega^*(\tilde\psi,F_0)$. Отметим, что M-оценка, определяемая такой функцией $\tilde\psi\in C_M$, называется оптимальной V-робастной в F_0 оценкой. Отметим также, что для неубывающих функций $\psi\in C_M$ понятия V-робастность и B-робастность эквивалентны, так как справедливо равенство $\omega^*(\psi,F) = 1+\gamma^*(\psi,F)/V(\psi,F)$ (см. работу [6, с. 166]).

Обсудим теперь связь между различными подходами, когда в качестве числовой характеристики робастности используется максимальное смещение (3.10) в заданной супермодели. При построении M-оценок, определяемых функцией $\psi \in C_M$, Хьюбер [5] рассматривал задачу минимаксного смещения оценки в виде

$$\min_{\psi \in C_M} \sup_{F \in A_\varepsilon(F_0)} |T(F) - T(F_0)|, \tag{5.4}$$

где $T(F)$ – функционал, определяющий M-оценку, и $A_\varepsilon(F_0)$ – супермодель с асимметричным засорением вида

$$A_\varepsilon(F_0) = \{F : F(x) = (1-\varepsilon)F_0(x) + \varepsilon H(x)\},\ 0 \le \varepsilon \le 1, \tag{5.5}$$

где F_0 – заданная функция распределения (идеальная модель) и $H(x)$ – произвольная непрерывная функция распределения (несимметричное ε-засорение). Использование аппроксимации (4.6) позволяет заменить минимаксную задачу Хьюбера (5.4) на оптимизационную задачу Хампеля вида

$$\min_{\psi \in C_M} [\varepsilon \cdot \gamma^*(\psi, F_0)]. \tag{5.6}$$

Решение данной задачи приводит к нахождению функции $\tilde{\psi} \in C_M$, которая определяет наиболее B-робастную оценку для супермодели $\Im_\varepsilon(F_0)$. Отметим, что и решение задачи Хьюбера (5.4), и решение задачи Хампеля (5.6) при построении наиболее B-робастной оценки приводит к выборочной медиане $\overline{X}_{1/2}$.

Связи между подходом Хьюбера и подходом, основанным на функции влияния, очень наглядно представлены в работе [6, см. табл. 2, с. 212].

6. ДИФФЕРЕНЦИАЛЬНЫЙ ПОДХОД МИЗЕСА К АНАЛИЗУ АСИМПТОТИЧЕСКИХ СВОЙСТВ СТАТИСТИК

В работах [29–35] описываются методы, позволяющие исследовать асимптотические свойства оценок параметров и статистик критериев. Метод проекций и связь функции влияния Хампеля с функцией, определяющей проекцию Гаека, обсуждается в [9] (см. также приложение П.1). В данном разделе обсуждается дифференциальный подход Мизеса к анализу асимптотических распределений статистик. Примеры анализа асимптотических свойств конкретных статистик методом Мизеса приводятся в [4, 8 и 9].

Пусть, как обычно, задана последовательность (н.о.р.) случайных величин $X_1,...,X_n$ с функцией распределения $F \in \Im$ и F_n – эмпирическая ф.р., построенная по выборке $X_1,...,X_n$. Рассмотрим статистики $T_n = T_n(X_1,...,X_n)$, которые являются симметричными функциями аргументов и, следовательно, могут быть представлены в виде функционалов от эмпирической функции распределения, то есть в виде

$$T_n = T_n(X_1,...,X_n) = T(F_n), \qquad \textbf{(6.1)}$$

где $T(F)$, $F \in \Im$, – подходящим образом выбранный функционал. Учитывая, что эмпирическая функция распределения F_n является разумной и «хорошей» оценкой ф.р. F (см. разд. 2.7.А в работе [7]), можно ожидать что и функционал $T(F_n)$ связан с $T(F)$ подобным образом при условии, что функционал $T(F)$ имеет достаточно «хорошее» поведение в окрестности F. Эти соображения приводят к рассмотрению распределения F в виде элемента множества $\Im$

(в виде «точки» на множестве $\Im$) и понятий непрерывности, дифференцируемости и других свойств регулярности функционала $T(F)$, заданного на множестве $\Im$.

Исследования асимптотических свойств статистик вида (6.1) были начаты фон Мизесом в 1947 году [31] и получили дальнейшее развитие в работах Серфлинга [29], Ференхольц [32], Ридза [30].

Суть подхода Мизеса состоит в аппроксимации статистики $T(F_n)$ более простой статистикой линейной структуры, асимптотическое распределение которой может быть получено с помощью центральной предельной теоремы либо с привлечением теории U-статистик Хёфдинга (см. [33, 35]). Основой такой аппроксимации является аналог ряда Тейлора для функционала $T(F)$, $F \in \Im$, с помощью которого «приращение» $T(G)-T(F)$ может быть представлено в терминах «дифференциалов» Гато $d_i T(F;G-F)$, $i=1,...,m$, функционала $T(F)$ для подходящим образом выбранного числа m. Для дальнейшего изложения введем необходимые понятия. Пусть заданы непрерывные функции распределения $F, G \in \Im$ и пусть функционал $T(F)$, $F \in \Im$, определён на выпуклом («звездчатом» в F) множестве вида $\{F+\lambda(G-F)\} \in \Im$, $0 \le \lambda \le 1$.

Определение 6.2. Если предел $\lim_{\lambda\to 0}\{T(F+\lambda(G-F))-T(F)\}/\lambda$ существует, то его называют дифференциалом Гато первого порядка функционала $T(F)$ в «точке» F по направлению G и определяют в виде

$$d_1 T(F;G-F) = \lim_{\lambda\to 0}\{T(F+\lambda(G-F))-T(F)\}/\lambda. \qquad \textbf{(6.3)}$$

Заметим, что дифференциал $d_1 T(F;G-F)$ может быть интерпретирован как обычная производная функции $Q(\lambda)=T(F+\lambda(G-F))$ действительной переменной λ при $\lambda=0$ (см., например, [5, 29]). Дифференциал Гато k-го порядка функционала $T(F)$ в «точке» F по направлению функции распределения G определяется в виде

$$d_k T(F;G-F) = \frac{d^{(k)}}{d\lambda^k} T[F+\lambda(G-F)]\,|_{\lambda=0},\ k=1,...,m. \qquad \textbf{(6.4)}$$

С целью аппроксимации статистики $T(F_n)$ получим сначала выражение для разности $T(G)-T(F)$. Пусть функция $Q(\lambda)$ удовлетворяет обычным условиям разложимости в ряд Тейлора в интервале $0 \le \lambda \le 1$. Тогда имеем

$$Q(\lambda) = Q(0) + \frac{Q^{(1)}(0)}{1!}\lambda + \frac{Q^{(2)}(0)}{2!}\lambda^2 + \cdots + \frac{Q^{(m)}(0)}{m!}\lambda^m + \\ + \frac{1}{(m+1)!}\frac{d^{(m+1)}}{d\lambda^{m+1}} T(F + \lambda(G-F))\Big|_{\lambda^*} \lambda^{m+1},$$

где $0 \le \lambda^* \le 1$. Учитывая далее, что $Q(0) = T(F)$, $Q(1) = T(G)$ и $Q^{(i)}(0) = d_i T(F; G-F)$, $i = 1,...,m$, получаем выражение для разности $T(G)-T(F)$:

$$T(G) - T(F) = \sum_{k=1}^{m} \frac{1}{k!} d_k T(F; G-F) + \\ + \frac{1}{(m+1)!}\frac{d^{(m+1)}}{d\lambda^{m+1}} T(F + \lambda(G-F))\Big|_{\lambda^*}. \qquad \textbf{(6.5)}$$

Используя теперь вместо функции распределения G эмпирическую функцию распределения F_n, получаем выражение для разности между интересующей нас оценкой функционала $T(F_n)$ и значением исходного функционала $T(F)$ в виде

$$T(F_n) - T(F) = V_{mn} + R_{mn}, \qquad \textbf{(6.6)}$$

где аппроксимационная статистика V_{mn} имеет линейную структуру вида

$$V_{mn} = \sum_{k=1}^{m} \frac{1}{k!} d_k T(F; G-F) \qquad \textbf{(6.7)}$$

и R_{mn}– остаточный член, характеризующий точность аппроксимации разности $T(F_n)-T(F)$ аппроксимационной статистикой V_{mn}.

Итак, если при соответствующей нормировке остаточный член по вероятности стремится к нулю, то есть $n^{m/2} R_{mn} \to^p 0$ при $n \to \infty$, тогда, согласно теореме Слуцкого (см., например, [9]), случайная величина $n^{m/2}[T(F_n)-T(F)]$ сходится по распределению к случайной величине $n^{m/2}V_{mn}$.

Таким образом, подход Мизеса для анализа асимптотического распределения статистик вида $T_n = T_n(X_1,...,X_n) = T(F_n)$ сводится к реализации следующих этапов:

1. Представление статистики $T(F_n)$ для подходящим образом выбранного числа m в виде (6.6).

2. Изучение асимптотического распределения аппроксимационной статистики $n^{m/2}V_{mn}$.

3. Доказательство сходимости по вероятности к нулю остаточного члена, то есть

$$n^{m/2}R_{mn} \to^{p} 0 \text{ при } n \to \infty .$$

Замечание 6.8. Отметим, что для доказательства сходимости по вероятности к нулю остаточного члена, то есть выражения $n^{m/2}R_{mn} \to^{p} 0$, $n \to \infty$, достаточно показать, что выполняется выражение

$$n^{m/2} \sup_{0\le\lambda\le1} \left| \frac{d^{(m+1)}}{d\lambda^{m+1}} T(F+\lambda(F_n - F)) \right| \xrightarrow{p} 0 \text{ при } n \to \infty . \qquad \textbf{(6.9)}$$

Однако в приведенном выражении (6.9) фигурирует, вообще говоря, производная большего порядка, чем это нужно для доказательства. На практике часто наиболее эффективным является непосредственный анализ остаточного члена R_{mn}, записанного в виде

$$R_{mn} = T(F_n) - T(F) - \sum_{k=1}^{m} \frac{1}{k!} d_k T(F; F_n - F) . \qquad \textbf{(6.10)}$$

При изучении асимптотического распределения статистик V_{mn} вида (6.7) отдельно рассматривают случаи $m=1$, $m=2$ и т.д. Наиболее распространенным, и его следует считать основным, является случай $m=1$. Тогда первый член V_{1n} в разложении (6.6) отличен от нуля и, с учетом (6.7), это разложение записывается как

$$T(F_n) - T(F) = V_{1n} + R_{1n} = d_1 T(F; F_n - F) + R_{1n} .$$

Отметим, что для широкого класса функционалов, оценки которых представляют практический интерес в статистике, дифференциал Гато k-го порядка может быть представлен в виде

$$d_k T(F;G-F)=\int\cdots\int S_k(x_1,...,x_k;F,T)\prod_{i=1}^{k} d[G(x_i)-F(x_i)],$$
$$1\le k\le m, \tag{6.11}$$

где $S_k(x_1,...,x_k;F,T)$ – функция, определённая в k-мерном пространстве, $(x_1,...,x_k)\in R^k$, конкретный вид которой определяется исходным функционалом $T(F)$ и которая в общем случае может зависеть от функции распределения F. В частном случае, при $m=1$, дифференциал Гато первого порядка записывается как

$$\begin{aligned} d_1 T(F;G-F)&=\int S_1(x;F,T)d[G(x)-F(x)]=\\ &=\int S_1(x;F,T)\,dG(x)-\int S_1(x;F,T)dF(x)=\\ &=\int\Big(S_1(x;F,T)-\int S_1(x;F,T)dF(x)\Big)dG(x)\\ &=\int IF(x;F,T)dG(x), \end{aligned} \tag{6.12}$$

где $IF(x;F,T)$ – функция влияния Хампеля, которая выражается через функцию $S_1(x;F,T)$ следующим образом:

$$IF(x;F,T)=S_1(x;F,T)-\int S_1(x;F,T)\,dF(x). \tag{6.13}$$

Заметим, что из формулы (6.13) следует для функции влияния $IF(x;F,T)$ важное равенство в виде

$$M_F\{IF(X;F,T)\}=\int IF(x;F,T)dF(x)=0. \tag{6.14}$$

Кроме того, заменив в формуле (6.12) функцию распределения G на вырожденную в точке x функцию распределения Δ_x, получаем явное выражение для функции влияния

$$IF(x;F,T)=d_1 T(F;\Delta_x-F). \tag{6.15}$$

Итак, функция влияния $IF(x;F,T)$ определяется в терминах дифференциала Гато первого порядка функционала $T(F)$ в «точке» F

по направлению вырожденной в точке x функции распределения Δ_x, то есть вычисляется по формуле

$$IF(x;F,T)=\lim_{\lambda\to 0+}\frac{T[(1-\lambda)F(y)+\lambda C(y-x)]-T(F)}{\lambda}=$$
$$=\frac{\partial}{\partial\lambda}T[F_{\lambda,x}]\big|_{\lambda=0},\ x\in R^1, \quad \textbf{(6.16)}$$

где $F_{\lambda,x}(y)=F(y)+\lambda[C(y-x)-F(y)]$, $0\le\lambda\le 1$ и $C(y-x)$ – единичная функция, определенная в (3.2). Отметим также, что согласно (6.12), дифференциал Гато первого порядка функционала $T(F)$ в «точке» F по направлению эмпирической функции $F_n(x)$, выражается через функцию влияния $IF(x;F,T)$ следующим образом:

$$d_1T(F;F_n-F)=\int IF(x;F,T)dF_n(x)=\frac{1}{n}\sum_{i=1}^{n}IF(X_i;F,T). \quad \textbf{(6.17)}$$

Следовательно, аппроксимационная статистика V_{1n} записывается в виде

$$V_{1n}=d_1T(F;F_n-F)=\frac{1}{n}\sum_{i=1}^{n}IF(X_i;F,T), \quad \textbf{(6.18)}$$

и так как $X_1,...,X_n$ – н.о.р. случайные величины, согласно центральной предельной теореме, имеет асимптотически нормальное распределение. С использованием функции влияния $IF(x;F,T)$, разложение (6.6) при $m=1$ принимает вид

$$T(F_n)=T(F)+d_1T(F;F_n-F)+R_{1n}=$$
$$=T(F)+\frac{1}{n}\sum_{i=1}^{n}IF(X_i;F,T)+R_{1n} \quad \textbf{(6.19)}$$

и с применением теоремы (6.20) служит основой для доказательства асимптотической нормальности оценок $T(F_n)$ функционалов $T(F)$.

Теорема 6.20. Пусть $X_1,...,X_n$ – последовательность независимых и одинаково распределенных случайных величин с ф.р. F, и пусть функционал $T(F)$ допускает разложение (6.19) и при этом выполняется выражение $n^{1/2}R_{1n} \to^{p} 0$, $n \to \infty$. Далее, пусть

$$0 < \sigma^2(T,F) = D_F\{IF(X;F,T)\} < \infty. \tag{6.21}$$

Тогда случайная величина $T(F_n)$ имеет асимптотически нормальное распределение со средним $T(F)$ и дисперсией $\sigma^2(T,F)/n$, то есть выполняется выражение

$$L\left(\frac{\sqrt{n}[T(F_n)-T(F)]}{\sigma(T,F)}\right) = N(0,1) \text{ при } n \to \infty. \tag{6.22}$$

Замечание 6.23. В работе [5, с. 44–47] отмечается, что асимптотическая нормальность оценки $T_n = T(F_n)$ функционала $T(F)$ доказывается «в одну строчку» при условии, что функционал $T(F)$ является дифференцируемым по Фреше. Однако не для всех функционалов, представляющих практический интерес, это сильное условие дифференцируемости выполняется и, кроме того, его довольно сложно проверять. При использовании подхода Мизеса необходимо затратить усилия на доказательство факта сходимости по вероятности к нулю остаточного члена в разложении (6.19). При этом, по существу, и выясняются условия регулярности, накладываемые на функционал $T(F)$ и функцию распределения F, при которых оценка $T_n = T(F_n)$ асимптотически нормальная.

Замечание 6.24. Если первый член V_{1n} в правой части разложения (6.6) равен нулю, то рассматривают вариант, когда $m = 2$, при этом асимптотическое распределение статистики V_{2n} не является нормальным. Однако это асимптотическое распределение статистики V_{2n} существует и может быть найдено, например, с помощью теории U-статистик (см., например, [35]). При произвольном m асимптотическое распределение аппроксимационной статистики V_{mn} вида (6.7) находится путем представления $d_m T(F;F_n - F)$ в виде стохастического интеграла [29].

Пример 6.25. Рассмотрим в качестве исходного функционала $T(F)$ квантиль уровня p, $0 < p < 1$, функции распределения $F(x)$, то есть

$$T(F) = F^{-1}(p),\ 0 < p < 1. \quad \textbf{(6.26)}$$

Оценкой $T(F_n) = F_n^{-1}(p)$ данного функционала является выборочный квантиль уровня p, который определяется в виде

$$T(F_n) = F_n^{-1}(p) = \begin{cases} X_{([np]+1)}, & (np) - \text{дробное}, \\ X_{(np)}, & (np) - \text{целое}, \end{cases} \quad \textbf{(6.27)}$$

где $X_{(1)},...,X_{(n)}$ – порядковые статистики исходной выборки $X_1,...,X_n$. Вычислим дифференциал Гато первого порядка $d_1T(F;G-F)$ функционала $T(F) = F^{-1}(p)$. Согласно определению (6.2), имеем

$$d_1T(F;G-F) = \frac{\partial}{\partial\lambda} F_\lambda^{-1}(p)\big|_{\lambda=0}, \quad \textbf{(6.28)}$$

где $F_\lambda^{-1}(t)$, $0 < t < 1$, – квантильная функция для ф.р. $F_\lambda = F + \lambda(G-F)$, $0 \le \lambda \le 1$. Рассмотрим тождество $F_\lambda(F_\lambda^{-1}(p)) \equiv p$ и перепишем его в виде $F(F_\lambda^{-1}(p)) + \lambda[G(F_\lambda^{-1}(p)) - F(F_\lambda^{-1}(p))] \equiv p$. Дифференцируя это тождество по λ и полагая $\lambda = 0$, с учетом (6.28) и равенства $F_\lambda^{-1}(p)\big|_{\lambda=0} = F^{-1}(p)$, получим выражение $f(F^{-1}(p))\,d_1T(F;G-F) + G(F^{-1}(p)) - p = 0$, из которого следует, что дифференциал Гато первого порядка функционала $T(F) = F^{-1}(p)$ вычисляется по формуле

$$d_1T(F;G-F) = \frac{p - G(F^{-1}(p))}{f(F^{-1}(p))},\ 0 < p < 1. \quad \textbf{(6.29)}$$

Отсюда следует, что функция влияния $IF(x;F,F^{-1}(p))$ оценки $T(F_n) = F_n^{-1}(p)$ функционала $T(F) = F^{-1}(p)$, $0 < p < 1$, запишется в виде

$$IF(x;F,F^{-1}(p)) = d_1T(F;\Delta_x - F) = \frac{p - C(F^{-1}(p) - x)}{f(F^{-1}(p))}\ ,$$
$$0 < p < 1,\ \ x \in R^1. \qquad \textbf{(6.30)}$$

Далее, аппроксимационная статистика V_{1n} в разложении (6.19) определяется выражением

$$V_{1n} = d_1T(F;F_n - F) = \frac{p - F_n(F^{-1}(p))}{f(F^{-1}(p))} =$$
$$= \frac{1}{n\,f(F^{-1}(p))}\sum_{i=1}^{n}[p - C(F^{-1}(p) - X_i)], \qquad \textbf{(6.31)}$$

и остаточный член R_{1n} записывается в виде

$$R_{1n} = F_n^{-1}(p) - F^{-1}(p) - [p - F_n(F^{-1}(p))]/f(F^{-1}(p)). \qquad \textbf{(6.32)}$$

Таким образом, в данном случае разложение (6.19) приобретает вид

$$F_n^{-1}(p) = F^{-1}(p) + \frac{1}{n\,f(F^{-1}(p))}\left[\sum_{i=1}^{n}[p - C(F^{-1}(p) - X_i)]\right] + R_{1n}. \qquad \textbf{(6.33)}$$

Можно убедиться (см., например, [36]), что при выполнении условия $f(F^{-1}(p)) > 0$ нормированный остаточный член R_{1n} сходится по вероятности к нулю, то есть выполняется выражение $n^{1/2}R_{1n} \to^{p} 0$, $n \to \infty$ и, следовательно, согласно теореме (6.20), выполняется другое асимптотическое выражение вида

$$L\{\sqrt{n}[F_n^{-1}(p) - F^{-1}(p)]/\sigma_F(F_n^{-1}(p))\} = N(0,1)$$
$$\text{при } n \to \infty,\ 0 < p < 1, \qquad \textbf{(6.34)}$$

где

$$\sigma_F^2(F_n^{-1}(p)) = \int IF^2(x;F,F^{-1}(p))dF(x) = \frac{p(1-p)}{f^2(F^{-1}(p))},$$
$$0 < p < 1. \qquad \textbf{(6.35)}$$

Заметим, что при $p = 1/2$ функционал $T(F) = F^{-1}(1/2)$ определяет медиану с.в. X с ф.р. F, то есть $T(F) = MED(X)$. Оценкой данного функционала является выборочная медиана $T(F_n)=med(X_1,...,X_n)=$ $=\overline{X}_{1/2}$, которая определяется в виде

$$T(F_n)=med(X_1,...,X_n)=\overline{X}_{1/2}=\begin{cases} X_{(k+1)}, & n=2k+1, \\ (X_{(k)}+X_{(k+1)})/2, & n=2k. \end{cases} \quad \textbf{(6.36)}$$

Функция влияния $IF(x;F,F^{-1}(1/2))$ выборочной медианы $\overline{X}_{1/2}$ является *ограниченной и разрывной* функцией вида

$$IF(x;F,F^{-1}(1/2))=\frac{1-2C(F^{-1}(1/2)-x)}{2f(F^{-1}(1/2))}=\frac{\operatorname{sign}(x-F^{-1}(1/2))}{2f(F^{-1}(1/2))}.$$
$$x\in R^1. \quad \textbf{(6.37)}$$

Из выражения (6.34) следует, что выборочная медиана $\overline{X}_{1/2}$ асимптотически нормальна, то есть выполняется выражение

$$L\{\sqrt{n}[\overline{X}_{1/2}-F^{-1}(1/2)]/\sigma_F(\overline{X}_{1/2}\}=N(0,1) \text{ при } n\to\infty, \quad \textbf{(6.38)}$$

где асимптотическая дисперсия $\sqrt{n}\,\overline{X}_{1/2}$-оценки вычисляется по формуле

$$\sigma_F^2(\overline{X}_{1/2}))=\int IF^2(x;F,F^{-1}(1/2))dF(x)=\frac{1}{4f^2(F^{-1}(1/2))}. \quad \textbf{(6.39)}$$

Пример 6.40. Рассмотрим теперь вариант, когда исходный функционал $T(F)$ задается неявно и определяется выражением вида

$$\int_{-\infty}^{\infty}[F(2T(F)-x)-(1/2)]dF(x)=0. \quad \textbf{(6.41)}$$

Оценка $T(F_n)$ этого функционала, называемая в литературе оценкой Ходжеса – Лемана, представляет собой половину медианы свертки эмпирической функции распределения F_n с собой, то есть

$$T(F_n)=HL=\frac{med(F_n*F_n)}{2}=med\{(X_i+X_j)/2,\ 1\le i\le j\le n\}=$$
$$=\begin{cases} W_{(r+1)}, & N=2r+1, \\ \{W_{(r)}+W_{(r+1)}\}/2, & N=2r, \end{cases} \quad \textbf{(6.42)}$$

где $W_{(1)},\dots,W_{(N)}$, $N = n(n+1)/2$, – упорядоченные значения средних Уолша $(X_i + X_j)/2$, $1 \le i \le j \le n$. Вычислим дифференциал Гато первого порядка $d_1 T(F;G-F)$ функционала $T(F)$, заданного выражением (6.41). Для этого сначала перепишем это выражение в другом виде

$$\int_0^1 F[2T(F) - F^{-1}(t)]dt = \frac{1}{2}. \qquad \textbf{(6.43)}$$

Далее, заменим ф.р. F в (6.43) на ф.р. $F_\lambda = F + \lambda(G-F)$, $0 \le \lambda \le 1$, в результате получим выражение вида

$$\int_0^1 \{ F(2T(F_\lambda) - F_\lambda^{-1}(t)) + \lambda[G(2T(F_\lambda) - F_\lambda^{-1}(t)) -$$

$$-F(2T(F_\lambda) - F_\lambda^{-1}(t))]\} dt = \frac{1}{2}. \qquad \textbf{(6.44)}$$

Дифференцируя это выражение по параметру λ и полагая $\lambda = 0$, с учетом (6.28) и (6.29), получим

$$\int_0^1 \{f(2T(F) - F^{-1}(t))\left[2d_1T(F;G-F) - \frac{t - G(F^{-1}(t))}{f(F^{-1}(t))} \right] +$$

$$+G(2T(F) - F^{-1}(t)) - F(2T(F) - F^{-1}(t))\}dt = 0.$$

Для простоты дальнейших преобразований рассмотрим случай симметричных распределений, то есть

$$F \in \Im_S = \{F : F(x) = 1 - F(-x) \ \ \forall x \in R^1\}.$$

В этом случае без потери общности можно считать, что $T(F) = 0$ и $f(-x) = f(x)$ $\forall x \in R^1$, а $F^{-1}(t) = -F^{-1}(1-t)$, $0 \le t \le 1$. С учетом этих равенств предыдущее выражение перепишется в виде

$$\int_0^1 \{f(F^{-1}(t))\left[2d_1T(F;G-F) - \frac{t - G(F^{-1}(t))}{f(F^{-1}(t))} \right] +$$

$$+G(-F^{-1}(t)) - F(-F^{-1}(t))\}dt = 0$$

или

$$2d_1T(F;G-F)\int_0^1 f(F^{-1}(t))dt-\int_0^1 tdt+2\int_0^1 G(F^{-1}(t))dt-\int_0^1(1-t)dt=0.$$

Отсюда получаем выражение для дифференциала Гато первого порядка функционала $T(F)$, $F\in\Im_S$, заданного выражением (6.41), в виде

$$d_1T(F;G-F)=\frac{1-2\int_0^1 G(F^{-1}(t))dt}{2\int_0^1 f(F^{-1}(t))dt}=\frac{1-2\int_{-\infty}^{\infty}G(x)dF(x)}{2\int_{-\infty}^{\infty}f^2(x)dx}=$$

$$=\frac{1}{\int_{-\infty}^{\infty}f^2(x)dx}\int_{-\infty}^{\infty}\left[F(x)-\frac{1}{2}\right]dG(x). \qquad \textbf{(6.45)}$$

Из формулы (6.45) получаем выражение для функции влияния оценки Ходжеса – Лемана (6.42) в виде

$$IF(x;F,HL)=d_1T(F;\Delta_x-F)=\frac{F(x)-(1/2)}{\int_{-\infty}^{\infty}f^2(x)dx}=\frac{F(x)-(1/2)}{\int_0^1 f(F^{-1}(t))dt},$$

$$x\in R^1,\ F\in\Im_S. \qquad \textbf{(6.46)}$$

Заметим, что данная функция *ограничена*. Далее, аппроксимационная статистика V_{1n} в разложении вида (6.19) определяется выражением

$$V_{1n}=d_1T(F;F_n-F)=\frac{1}{n\int_{-\infty}^{\infty}f^2(x)dx}\sum_{i=1}^{n}\left[F(X_i)-\frac{1}{2}\right]=$$

$$=\frac{1}{n\int_0^1 f(F^{-1}(t))dt}\sum_{i=1}^{n}\left[U_i-\frac{1}{2}\right], \qquad \textbf{(6.47)}$$

где $U_1,...,U_n$ – н.о.р. случайные величины с равномерным в интервале $[0,1]$ распределением. Таким образом, разложение вида (6.19) с учетом того, что в данном случае $T(F)=0$ для $F \in \Im_S$, записывается в виде

$$T(F_n) = HL = \frac{1}{n\int_0^1 f(F^{-1}(t))dt} \sum_{i=1}^{n} \left[U_i - \frac{1}{2} \right] + R_{1n}, \qquad \textbf{(6.48)}$$

где остаточный член R_{1n} в данном случае имеет вид $R_{1n} = T(F_n) - V_{1n}$. Далее, можно убедиться, что при условии $\int_0^1 f(F^{-1}(t))dt > 0$, выполняется выражение $n^{1/2}R_{1n} \to^p 0$, $n \to \infty$, и, согласно теореме (6.20), $\sqrt{n}HL$-оценка асимптотически нормальна, для которой асимптотическая дисперсия вычисляется по формуле

$$\sigma_F^2(HL) = \int_{-\infty}^{\infty} IF^2(x;F,HL)dF(x) = 1/12 \left(\int_0^1 f(F^{-1}(t))dt \right)^2,$$

$$F \in \Im_S. \qquad \textbf{(6.49)}$$

Замечание 6.50. В последующих разделах обсуждаются свойства основных типов робастных оценок функционалов от распределения вероятностей. В теории робастных оценок функционалов традиционно выделяют *M*-, *L*-, *R*-, и *MD*-оценки. В класс *M*-оценок, предложенных Хьюбером [15], включают оценки максимального правдоподобия, *L*-оценки являются линейными комбинациями порядковых статистик, *R*-оценки основаны на использовании ранговых критериев проверки статистических гипотез, *MD*-оценки строятся методом минимума выбранного расстояния между эмпирическим и модельным распределениями. К настоящему времени имеется обширная литература по изучению как асимптотических свойств робастных оценок функционалов, так и их свойств при конечных объемах выборки. Наиболее полно эти вопросы отражены в работах [2, 5, 6, 24, 29, 33] (см. также [4, 9]).

7. ОЦЕНКИ ТИПА МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ (*М*-ОЦЕНКИ)

К классу *М*-оценок относятся оценки, получаемые в результате решения уравнений, которые обычно возникают в задачах на нахождение экстремума. В этот класс входят оценки максимального правдоподобия и наименьших квадратов. По-видимому, толчком к интенсивному развитию этого класса оценок послужила работа Хьюбера [15], опубликованная в 1964 году, в которой предложены *М*-оценки параметра сдвига (параметра положения в симметричном случае) и обсуждены вопросы робастности этой оценки в рамках супермодели с засорением вида (4.4).

В данном разделе излагаются общие свойства *М*-оценок, доказывается их асимптотическая нормальность, вычисляются их функции влияния, кратко обсуждаются вопросы качественной робастности и приводятся необходимые формулы для вычисления числовых характеристик устойчивости.

Пусть, как обычно, задана последовательность (н.о.р.) случайных величин $X_1,...,X_n$ с функцией распределения $F \in \Im$ и F_n – эмпирическая ф.р., построенная по выборке $X_1,...,X_n$.

М-оценка, обозначаемая через $T_n = T(X_1,...,X_n)$, определяется как решение экстремальной задачи на минимум в виде $\sum \rho(X_i, T_n) = \min$ или как решение уравнения

$$\sum_{i=1}^{n} \psi(X_i, T_n) = 0, \qquad \textbf{(7.1)}$$

где ρ – заданная функция, а $\psi(x,t) = \partial\rho(x,t)/\partial t$. *М*-оценка, обозначаемая через $T_n = T(F_n)$, является оценкой функционала $T(F)$,

$F \in \Im$, определяемого неявно выражением вида

$$\int \psi(x, T(F)) dF(x) = 0. \tag{7.2}$$

В частности, в параметрическом случае, когда семейство функций распределения определяется в виде $\Im_\theta = \{F : F(x,\theta), \theta \in \Theta\}$, функционал $T(F) = \theta(F)$, определяющий параметр θ и являющийся состоятельным по Фишеру, то есть $T(F(x,\theta)) = \theta$, $\forall \theta \in \Theta$, выражается как

$$\int \psi(x, \theta(F)) dF(x,\theta) = 0.$$

При этом, если функция $\psi(x,\theta)$ связана с плотностью $f(x,\theta)$ соотношением $\psi(x,\theta) = -\partial[\ln f(x,\theta)]/\partial\theta$, то *М*-оценка параметра θ в виде $\theta(F_n)$ является оценкой максимального правдоподобия.

Обсудим асимптотические свойства *М*-оценок. Прежде всего опишем условия, накладываемые на функцию ψ и ф.р. *F*, при которых имеет место состоятельность и асимптотическая нормальность *М*-оценок. В литературе эти вопросы подробно описаны (см., например, [5, 6, 29, 37]). При этом были найдены различные наборы условий, одни из которых накладывают сильные ограничения на функцию ψ, определяющую конкретный вид *М*-оценки, и более слабые ограничения на ф.р. *F* наблюдений, другие – наоборот. Прежде чем сформулировать строгие результаты, приведем предварительное обсуждение асимптотических свойств *М*-оценок, используя подход Мизеса. Для удобства записи обозначим

$$\lambda_F(t) = \int \psi(x,t) dF(x), \; -\infty < t < \infty, \tag{7.3}$$

и пусть t_0 – корень уравнения $\lambda_F(t) = 0$, а T_n – корень «эмпирического» уравнения $\lambda_{F_n}(t) = 0$. Вычислим дифференциал Гато первого порядка функционала $T(F)$, $F \in \Im$, определяемого неявно выражением (7.2). Пусть, как обычно, $F_\lambda = F + \lambda(G - F)$, $0 \le \lambda \le 1$. Подставим в (7.2) F_λ вместо ф.р. *F*, дифференцируя по λ и полагая $\lambda = 0$, получим

$$d_1T(F;G-F) = -\frac{\int \psi(x,T(F))dG(x)}{\frac{\partial}{\partial t}\int \psi(x,t)dF(x)} = -\frac{\int \psi(x,T(F))dG(x)}{\lambda'_F[T(F)]}, \quad \textbf{(7.4)}$$

при условии, что $\lambda'_F[T(F)] \neq 0$. Заменив в (7.4) ф.р. G на вырожденную в точке x ф.р. Δ_x, получим выражение функции влияния $IF(x;F,T) = IF(x;F,\psi)$ для M-оценки, определяемой функцией ψ, в виде

$$IF(x;F,\psi) = -\psi(x,T(F))/\lambda'_F[T(F)],\ x \in R^1. \quad \textbf{(7.5)}$$

Отметим важную особенность M-оценки, состоящую в том, что её функция влияния $IF(x;F,\psi)$ пропорциональна функции ψ, которая определяет конкретный вид M-оценки. Напомним, что многие числовые характеристики оценок определяются через функцию влияния. Поэтому отмеченная особенность позволяет строить M-оценки с заданными свойствами робастности путем целенаправленного выбора функции ψ. Аппроксимационная статистика V_{1n} в разложении вида (6.19) в данном случае запишется в виде

$$V_{1n} = d_1T(F;F_n - F) = -\frac{1}{n\lambda'_F[T(F)]}\sum_{i=1}^{n}\psi(X_i,T), \quad \textbf{(7.6)}$$

а остаточный член, как обычно, равен $R_{1n} = T_n - t_0 - d_1T(F;F_n - F)$.

Итак, оставляя пока в стороне условия регулярности (они будут приведены ниже при доказательстве выражения $\sqrt{n}R_{1n} \to^p 0$ при $n \to \infty$), можно заключить, что M-оценка T_n асимптотически нормальна со средним значением $T(F)$ из (7.2) и дисперсией $D_F(T_n) = \sigma^2(\psi,F)/n$, где

$$\sigma^2(\psi,F) = \int IF^2(x;F,\psi)dF(x) = \frac{\int \psi^2(x,T(F))dF(x)}{[\lambda'_F(T(F))]^2}. \quad \textbf{(7.7)}$$

Строгие результаты в виде лемм, теорем и полные доказательства асимптотической нормальности M-оценок могут быть найдены в работах [5, 6, 29, 33] (см. также [4, 9]).

Замечание 7.8. Отметим, что M-оценки были введены Хьюбером [15] для оценивания параметра сдвига по выборке $X_1,...,X_n$ н.о.р. случайных величин с ф.р. $F(x-\theta)$ и плотностью $f(x-\theta)$, где θ – неизвестный параметр сдвига. В данном случае функция $\psi(x,t)=\psi(x-t)$. Можно убедиться, что M-оценка параметра сдвига $t_0=T(F)$ для ф.р. $F(x-t_0)$ является асимптотически эффективной, если определяющая её функция $\psi(x,t)$ пропорциональна отношению $f'(x)/f(x)$. Для этого используем приведенные выражения для асимптотических дисперсий M-оценок. Предполагаем, что условия, позволяющие менять порядок дифференцирования и интегрирования в (7.3), выполнены. Далее, интегрируя (7.3) по частям, при $\psi(x,t)=\psi(x-t)$ получим

$$\lambda'_F(t_0)=\int\psi'(x-t_0)f(x)dx=-\int\psi(x-t_0)f'(x)dx.$$

Учитывая предыдущее выражение и используя формулу (7.7), при $\psi(x,t)=\psi(x-t)$ имеем

$$\sigma^2(\psi,F)=\frac{\int\psi^2(x-t_0)f(x)dx}{\left(\int\psi(x-t_0)f'(x)dx\right)^2}.$$

Используя теперь неравенство Шварца, запишем

$$\left(\int\psi(x-t_0)f'(x)dx\right)^2\le\int\psi^2(x-t_0)f(x)dx\cdot\int[f'(x)/f(x)]^2dF(x).$$

Отсюда следует неравенство Рао – Крамера в виде

$$\sigma^2(\psi,F)\ge\frac{1}{\int[f'(x)/f(x)]^2dF(x)}=\frac{1}{I(F)},$$

причем равенство достигается, если функция $\psi(x,t)$ пропорциональна отношению $f'(x)/f(x)$. Таким образом, M-оценка параметра сдвига $t_0=T(F)$ для ф.р. $F(x-t_0)$ является асимптотически эффективной и совпадает с МП-оценкой максимального правдоподобия параметра сдвига θ.

Пример 7.9. Пусть ф.р. $F(x)$ имеет плотность $f(x)$ и $f(F^{-1}(p))>0$, $0<p<1$. Убедимся, что оценка квантиля уровня p может быть построена как *M*-оценка, определяемая функцией $\psi(x,t)=\psi(x-t)$ вида

$$\psi(x,t)=\psi(x-t)=\begin{cases} -1, & x<t, \\ p/(1-p), & x\ge t. \end{cases}$$

Сначала вычислим $\lambda_F(t)$ из (7.3). Для заданной функции $\psi(x-t)$ имеем

$$\lambda_F(t)=\int\psi(x-t)dF(x)=\frac{p}{1-p}\int_t^{\infty}dF(x)-\int_{-\infty}^{t}dF(x)=\frac{p-F(t)}{1-p}.$$

Таким образом, решением уравнения $\lambda_F(t)=0$ является квантиль уровня p, то есть $t_0=F^{-1}(p)$, $0<p<1$, а решением «эмпирического» уравнения $\lambda_{F_n}(t)=0$, является выборочный квантиль $T_n=F_n^{-1}(p)$. Убедимся теперь в асимптотической нормальности этой *M*-оценки. Отметим, что заданная функция $\psi(x-t)$ монотонна по t и интеграл

$$\int_{-\infty}^{\infty}\psi^2(x-t_0)dF(x)=\int_{-\infty}^{F^{-1}(p)}(-1)^2dF(x)+\int_{F^{-1}(p)}^{\infty}[p/(1-p)]^2dF(x)=\frac{p}{1-p}$$

является конечным. Кроме того, $\lambda'_F(t_0)\ne 0$, так как

$$\lambda'_F(t_0)=[-f(F^{-1}(t_0))/(1-p)]<0.$$

Таким образом, выполняется асимптотическое выражение

$$L\{\sqrt{n}(T_n-t_0)/\sigma(\psi,F)\}=N(0,1) \text{ при } n\to\infty,$$

где
$$\sigma^2(\psi,F)=\int\psi^2(x,t_0)dF(x)\ /\ [\lambda'_F(t_0)]^2=$$

$$=\frac{p/(1-p)}{[-f(F^{-1}(p))/(1-p)]^2}=\frac{p(1-p)}{f^2(F^{-1}(p))}. \qquad \textbf{(7.10)}$$

Отметим, что аналогичный результат был приведен ранее в примере (6.25).

Пример 7.11. Пусть $X_1,...,X_n$ – н.о.р. случайные величины с ф.р. $F(x-\theta)$ и плотностью $f(x-\theta)$, где θ – неизвестный параметр сдвига. Логарифм функции правдоподобия имеет вид

$$\ln L(\theta; x_1,...,x_n) = \sum_{i=1}^{n} \ln f(x_i - \theta) = -\sum_{i=1}^{n} \rho(x_i - \theta).$$

Так как оценка максимального правдоподобия $\hat{\theta}$ параметра θ находится из максимума функции правдоподобия $L(\theta;\vec{x})$ или, что эквивалентно, из минимума $\sum \rho(x_i - \theta) = -\int \ln f(x-\theta) dF_n(x)$, где $\rho(x) = -\ln f(x)$, то она является *M*-оценкой, определяемой функцией $\psi(x) = -f'(x)/f(x)$, и следует из уравнения

$$\sum_{i=1}^{n} \psi(X_i - \theta) = 0, \qquad \textbf{(7.12)}$$

которое является частным случаем уравнения (7.1). Рассмотрим конкретные распределения вероятностей.

1. Если исходное распределение F является нормальным, то $\rho(x) = x^2/2 + c$, $\psi(x) = x$ и решение уравнения $\sum \psi(x_i - \theta)$ дает *M*-оценку в виде выборочного среднего, то есть $\hat{\theta} = \overline{X}$. Эта оценка совпадает с МП-оценкой и совпадает с оценкой, построенной методом наименьших квадратов, причем $\sigma^2(\psi, F) = 1/I(f) = 1$.

2. Для логистического распределения $F(x) = [1 + \exp(-x)]^{-1}$ имеем $\psi(x) = 2F(x) - 1$, *M*-оценка совпадает с МП-оценкой и $\sigma^2(\psi, F) = 1/I(f) = 3$.

3. Для распределения Лапласа с плотностью $f(x) = \exp(-|x|)/2$ имеем $\rho(x) = |x| + c$, $\psi(x) = -1$ для $x < 0$ и $\psi(x) = 1$ для $x \geq 0$. В данном случае *M*-оценкой является выборочная медиана, то есть $\hat{\theta} = \overline{X}_{1/2}$. Эта оценка совпадает с МП-оценкой и с оценкой, построенной методом наименьших модулей, причем $\sigma^2(\psi, F) = 1/I(f) = 1$.

4. Для распределения Коши с плотностью $f(x)=1/[\pi(1+x^2)]$ имеем $\rho(x)=\ln(1+x^2)+c$, $\psi(x)=2x/(1+x^2)$. В данном случае *M*-оценка находится путем численного решения уравнения $\sum\psi(x_i-\theta)=0$. Она совпадает с МП-оценкой, и её асимптотическая дисперсия равна $\sigma^2(\psi,F)/n$, где $\sigma^2(\psi,F)=1/I(f)=2$.

Заметим, что функции ψ в примерах 2, 3 и 4 являются *ограниченными*, а в примере 4 функция ψ даже не монотонная и асимптотически стремится к нулю.

Замечание 7.13. Напомним, что оценку $T_n=T_n(X_1,...,X_n)$ называют эквивариантной относительно линейных преобразований, если выполняется равенство $T_n(AX_1+B,...,AX_n+B)=AT_n(X_1,...,X_n)+B$. Рассмотренные *M*-оценки T_n эквивариантны относительно параметра сдвига *B*. В самом деле, по исходным наблюдениям $X_1,...,X_n$, *M*-оценка T_n находится из уравнения

$$\sum\psi(X_i-T_n(X_1,...,X_n))=0,$$

а по преобразованным наблюдениям $X_1+B,...,X_n+B$ она определяется уравнением

$$\sum\psi[(X_i+B)-T_n(X_1+B,...,X_n+B)]=0$$

или

$$\sum\psi[X_i-\{T_n(X_1+B,...,X_n+B)-B\}]=0.$$

Сравнивая первое и третье уравнения, заключаем, что

$$T_n(X_1+B,...,X_n+B)=T_n(X_1,...,X_n)+B.$$

К сожалению, *M*-оценки не являются эквивариантными относительно масштабного параметра, то есть если значения случайной величины *X* умножить на некоторую константу, то начальная оценка не обязательно изменится на ту же константу. На практике масштабный параметр, как правило, неизвестен. В связи с этим Хьюбер [15] предложил использовать совместные оценки параметра сдвига и масштабного параметра. Однако, как показали обширные экспериментальные исследования (см., например, [24]), *M*-оценки с предварительной оценкой масштабного параметра по исходной выборке $X_1,...,X_n$ обладают явным преимуществом перед совместным вариантом. По этой причине на практике обычно используют M-оценки T_n параметра сдвига θ с предварительным масштабированием, которые эквивариантны относительно линейных преобразований и находятся путем решения уравнения

$$\sum_{i=1}^{n}\psi\left(\frac{X_i - T_n}{S_n}\right) = 0, \qquad \textbf{(7.14)}$$

где S_n – оценка масштабного параметра $S(F)$, построенная по исходной выборке $X_1,...,X_n$. В качестве оценки S_n обычно используют выборочную медиану абсолютных отклонений от медианы (MAD – Median Absolute Deviation), которая записывается в виде

$$S_n = 1,483 \cdot med\{\,| X_i - med(X_1,...,X_n)| \ ,\ 1 \le i \le n\ \}. \qquad \textbf{(7.15)}$$

Отметим, что нормировочная константа $1,483 = 1/\Phi^{-1}(0,75)$ для этой оценки введена для того, чтобы обеспечить в стандартном нормальном случае выполнение равенства $M_{\Phi}(S_n) = 1$, что обеспечивает состоятельность по Фишеру функционала $S(F)$ при $F = \Phi$.

Численное решение уравнения (7.14) обычно осуществляется с помощью итерационных схем. При построении *одношаговых* M-оценок (см. [2, 5]) вида

$$T_n = T_n^{(0)} + \frac{S_n^{(0)} \sum \psi[(X_i - T_n^{(0)})/S_n^{(0)}]}{\sum \psi'[(X_i - T_n^{(0)})/S_n^{(0)}]} \qquad \textbf{(7.16)}$$

в качестве начальных оценок $T_n^{(0)}$ и $S_n^{(0)}$ параметров сдвига и масштаба рекомендуется выбирать соответственно выборочную медиану и MAD, то есть $T_n^{(0)} = med(X_1,...,X_n) = \bar{X}_{1/2}$ и $S_n^{(0)}$ определятся выражением (7.15). При практической реализации процедуры вычисления одношаговых M-оценок может оказаться при некоторых функциях ψ и некоторых реализациях $x_1,...,x_n$ выборки $X_1,...,X_n$, что знаменатель $\sum \psi'[(x_i - T_n^{(0)})/S_n^{(0)}]$ в правой части (7.16) равен нулю. Эту опасность устраняют путем замены этого знаменателя на $n\int\psi'(z)d\Phi(z)$. Отметим, что для решения уравнения (7.14) также используют средневзвешенный метод наименьших квадратов (см., например, [2]), и *M*-оценку $\hat{\theta}_2$ параметра сдвига θ вычисляют по формуле

$$\hat{\theta}_2 = \frac{\sum W_i X_i}{\sum W_i} = \hat{\theta}_1 + \frac{S_n \sum \psi[(X_i - \hat{\theta}_1)/S_n]}{\sum W_i},$$

$$W_i = \frac{\psi[(X_i - \hat{\theta}_1)/S_n]}{(X_i - \hat{\theta}_1)/S_n},\ i = 1,...,n, \qquad \textbf{(7.17)}$$

где $\hat{\theta}_1$ – выборочная медиана. Сравнение этих итерационных схем вычисления оценок приведено в работе *Holland P.W., Welsch R.E.* Robust regression using interactively reweighted least-squares // Commun Statist. 1976. V. 4. P. 68–75. Асимптотические свойства обсуждаются в работе *Klein R., Yohai V.J.* Asymptotic behavior of iterative M-estimators for location // Bol. Soc. Bras. Mat. 1979. V. 10. No. 1. P. 27–42.

Обсудим асимптотические свойства *M*-оценки T_n параметра сдвига θ с предварительной оценкой $S_n = S(F_n)$ масштабного параметра $S(F)$ по исходной выборке $X_1,...,X_n$. Эта оценка T_n является решением уравнения (7.14), а также оценкой функционала $T(F)$, заданного неявно выражением вида

$$\int\psi\left(\frac{x - T(F)}{S(F)}\right)dF(x) = 0. \qquad \textbf{(7.18)}$$

Вычислим дифференциал Гато первого порядка этого функционала $T(F)$. Пусть, как обычно, $F_\lambda = F + \lambda(G - F)$, $0 \le \lambda \le 1$. Подставив в (7.18) вместо функции распределения F ф.р. F_λ, получим выражение

$$\int\psi\left(\frac{x - T(F_\lambda)}{S(F_\lambda)}\right)dF_\lambda(x) = 0.$$

Дифференцируя это выражение по параметру λ и полагая $\lambda = 0$, получим

$$\int \psi'\left(\frac{x-T(F)}{S(F)}\right)\left[\frac{-d_1T(F;G-F)\cdot S(F)-(x-T(F))d_1S(F;G-F)}{S^2(F)}\right]dF(x)+$$
$$+\int \psi\left(\frac{x-T(F)}{S(F)}\right)d[G(x)-F(x)]=0.$$

В этом выражении через $d_1S(F;G-F)$ обозначен дифференциал Гато первого порядка функционала $S(F)$, описывающего масштабный параметр. Предыдущее выражение перепишем в виде

$$-\frac{d_1T(F;G-F)}{S(F)}\int \psi'\left(\frac{x-T(F)}{S(F)}\right)dF(x)-$$
$$-\frac{d_1S(F;G-F)}{S(F)}\int \left(\frac{x-T(F)}{S(F)}\right)\psi'\left(\frac{x-T(F)}{S(F)}\right)dF(x)+$$
$$+\int \psi\left(\frac{x-T(F)}{S(F)}\right)dG(x)=0.$$

Отсюда получаем

$$d_1T(F;G-F)=\left[S(F)\int \psi\left(\frac{x-T(F)}{S(F)}\right)dG(x)-d_1S(F;G-F)\times\right.$$
$$\left.\times\int \left(\frac{x-T(F)}{S(F)}\right)\psi'\left(\frac{x-T(F)}{S(F)}\right)dF(x)\right]\Big/\int \psi'\left(\frac{x-T(F)}{S(F)}\right)dF(x). \qquad \textbf{(7.19)}$$

Существенное упрощение данного выражения достигается при дополнительных предположениях о том, что ф.р. F симметрична, то есть $F\in\Im_{S|\theta}$, и функция $\psi(x)$ нечетная, то есть $\psi(-x)=-\psi(x)$, $\forall x\in R^1$. Для нечетной функции $\psi(x)$ её производная $\psi'(x)$ является четной функцией и, следовательно, подынтегральная функция второго интеграла с симметричными пределами в числителе (7.19) нечетная, поэтому он равен нулю. Отметим также, что принятые предположения обеспечивают состоятельность по Фишеру функционала $T(F)$, заданного выражением $\int \psi(x-T(F))dF(x)=0$. В самом деле, пусть $\lambda(T)=\int \psi(x-T(F))dF(x)=0$, тогда $T(F)=\lambda^{-1}(0)$. Далее, учитывая, что $F(x)=1-F(2\theta-x)$ и $\psi(-x)=-\psi(x)$, $\forall x\in R^1$, имеем

$$0=\int\limits_{-\infty}^{\infty}\psi(x-T)dF(x)=-\int\limits_{-\infty}^{\infty}\psi(x-T)dF(2\theta-x)=\|2\theta-x=y|_{\infty}^{-\infty}\|=$$

$$=\int\limits_{-\infty}^{\infty}\psi(2\theta-y-T)dF(y)=-\int\limits_{-\infty}^{\infty}\psi[y-(2\theta-T)]dF(y)=-\lambda(2\theta-T). \quad \mathbf{(7.20)}$$

Отсюда следует, что $2\theta-T(F)=\lambda^{-1}(0)$, но $\lambda^{-1}(0)=T(F)$, поэтому $2\theta=2T(F)$ и, окончательно, $T[F(x-\theta)]=\theta$, $\forall\theta\in R^1$.

Итак, для $F\in\Im_{S|\theta}$, с учетом эквивариантности M-оценок, можно без потери общности считать, что $\theta=T(F)=0$, и для нечетной функции $\psi(x)$ из (7.19) имеем выражение для функции влияния оценки $T_n=T(F_n)$ функционала $\theta=T(F)$ в виде

$$IF(x;F,T)=\frac{S(F)\psi[x/S(F)]}{\int\psi'[x/S(F)]dF(x)}\ ,\ x\in R^1. \quad \mathbf{(7.21)}$$

Асимптотическая дисперсия $\sqrt{n}T_n$-оценки функционала $\theta=T(F)$, $F\in\Im_S$, вычисляется по формуле

$$\sigma^2(F;T,S)=\int IF^2(x;F,T)dF(x)=\frac{S^2(F)\int\psi^2(x/S(F))dF(x)}{\{\int\psi'(x/S(F))dF(x)\ \}^2}. \quad \mathbf{(7.22)}$$

Замечание 7.23. Приведем используемые на практике функции $\psi(x)$, определяющие конкретные M-оценки, а также рекомендованные значения констант настройки этих функций, которые на практике обычно выбирают, исходя из пятипроцентной потери эффективности M-оценки в нормальном случае.

1. Функция Хьюбера

$$\psi(x)=\begin{cases}-a, & x<-a,\\ x, & |x|\le a,\\ a, & x>a,\end{cases} \quad \mathbf{(7.24)}$$

где константа $a=1{,}345$ обеспечивает достижение 95 %-й эффективности M-оценки в нормальном случае при известной дисперсии.

2. Функция Хампеля

$$\psi(x)=\begin{cases}|x|, & 0\le|x|<a,\\ a, & a\le|x|<b,\\ \dfrac{c-|x|}{c-b}a, & b\le|x|<c,\\ 0, & c\le|x|,\end{cases} \tag{7.25}$$

где рекомендуемые константы следующие: $a=1,7$, $b=3,4$, $c=8,5$.

3. Функция Андрюса

$$\psi(x)=\begin{cases}\sin(x/a), & |x|\le a\pi,\\ 0, & |x|>a,\end{cases} \tag{7.26}$$

где величина $a=1,339$ соответствует 5 %-й потере эффективности при известном масштабе.

4. Биквадратная функция Тьюки

$$\psi(x)=\begin{cases}x[1-(x/a)^2]^2, & |x|\le a,\\ 0, & |x|>a.\end{cases} \tag{7.27}$$

Если масштаб известен, то $a=4,685$ соответствует 5 %-й потере эффективности в нормальном случае.

Замечание 7.28. Вопросы качественной и количественной устойчивости *M*-оценок достаточно полно описаны в книге Хьюбера [5, с. 59]. Здесь мы лишь отметим, что пороговая точка *M*-оценок равна нулю, если функция $\psi(x)$ неограниченная и она достигает своего максимального значения, равного 1/2 для ограниченной функции $\psi(x)$ при выполнении равенства $\psi(-\infty)=-\psi(+\infty)$. Далее, *M*-оценки удовлетворяют требованиям качественной устойчивости при идеальной модели F_0, если функция $\psi(x)$ ограниченная и решение уравнения $\lambda_{F_0}(t)=0$ является единственным. Эти условия являются необходимыми и достаточными для слабой непрерывности функционала $T(F_0)$ в «точке» F_0 (см. [5, 47]). Подробное описание свойств функции изменения дисперсии *M*-оценок приводится в [6]. Изучение свойств робастности *M*-оценок, их сравнение с другими оценками и связи с *L*-, *R*- и *MD*-оценками приводятся в разделах 10 и 13.

8. *L*-ОЦЕНКИ В ВИДЕ ЛИНЕЙНОЙ КОМБИНАЦИИ ПОРЯДКОВЫХ СТАТИСТИК

Оценки в виде линейной комбинации порядковых статистик имеют давнишнюю историю, и вновь повышенный интерес к этому классу оценок был вызван работами Беннета (1952) и Юнга (1955), в которых найдены оптимальные весовые коэффициенты для оценок параметров сдвига и масштаба. В последнее время этот интерес поддерживается тем фактом, что L-оценки обладают «хорошими свойствами робастности» и в вычислительном отношении они более просты, чем другие классы робастных оценок, типа M- и R-оценок. В данном разделе излагаются общие свойства L-оценок, доказывается их асимптотическая нормальность, вычисляются их функции влияния, кратко обсуждаются вопросы качественной робастности и приводятся необходимые формулы для вычисления числовых характеристик устойчивости. Изучение свойств робастности L-оценок и их сравнение с другими оценками приводятся в разделах 10 и 13. В литературе L-оценки достаточно полно описаны. Заинтересованного читателя мы отсылаем к работам [4–6, 29, 39, 41–43].

Пусть, как обычно, $X_1,...,X_n$ – последовательность н.о.р. случайных величин с функцией распределения $F(x)$, плотностью $f(x)$, $x \in R^1$. Далее, $X_{(1)},...,X_{(n)}$ – порядковые статистики и $F_n(x)$ – эмпирическая функция распределения, построенная по выборке $X_1,...,X_n$. Отметим, что большой класс практически важных статистик, таких, как выборочная медиана; урезанные и винзоризованные средние; интерквартильный размах; оценки Гаствирта; средняя разность Джини и многие другие, можно представить с помощью порядковых статистик $X_{(1)},...,X_{(n)}$ в виде линейной комбинации

$$T_n = \sum_{i=1}^{n} a_{ni} X_{(i)}, \tag{8.1}$$

где $\{a_{ni}\}$ – весовые коэффициенты и $X_{(i)}$ – i-я порядковая статистика. Удобно выделить два варианта задания весовых коэффициентов. В первом варианте весовые коэффициенты $\{a_{ni}\}$ определяются с помощью порождающей веса функции $J(t)$, $0 \le t \le 1$. Во втором варианте веса a_{ni} представляют собой дискретный набор заданных констант $a_1, ..., a_n$. В случае, когда для определения весов $\{a_{ni}\}$ используется порождающая веса функция $J(t)$, L-оценка записывается в виде

$$T_{n1} = \frac{1}{n} \sum_{i=1}^{n} J(i/n) X_{(i)}. \tag{8.2}$$

Эту оценку можно рассматривать в виде функционала от эмпирической функции распределения $F_n(x)$, то есть в виде $T_n = T(F_n)$, где функционал $T(F)$ определяется выражением

$$T_1(F) = \int_0^1 J(t) F^{-1}(t) dt = \int_{-\infty}^{\infty} x J[F(x)] dF(x). \tag{8.3}$$

Отметим, что весовые коэффициенты $\{a_{ni}\}$ статистики (8.1), задаваемые функцией $J(t)$, могут выражаться в различной форме. Некоторые авторы рассматривают их представление в интегральном виде:

$$a_{ni} = \int_{(i-1)/n}^{i/n} J(t) dt. \tag{8.4}$$

При такой форме задания весовых коэффициентов L-оценка эквивариантна относительно линейных преобразований исходных данных $X_1, ..., X_n$, то есть для неё выполняется равенство

$$T_n(AX_1 + B, ..., AX_n + B) = AT_n(X_1, ..., X_n) + B, \tag{8.5}$$

если на функцию $J(t)$ наложить ограничение вида

$$\int_0^1 J(t) dt = 1. \tag{8.6}$$

В самом деле, используя преобразованные данные

$$AX_1 + B, ..., AX_n + B,$$

получаем

$$T_n(AX_1 + B, ..., AX_n + B) = A\sum_{i=1}^{n} a_{in} X_{(i)} + B\sum_{i=1}^{n} a_{ni}.$$

Отсюда следует, что равенство (8.5) выполняется, если $\sum a_{ni} = 1$, то есть

$$\sum_{i=1}^{n} a_{ni} = \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} J(t)dt = \int_{0}^{1/n} J(t)dt + \int_{1/n}^{2/n} J(t)dt + \cdots + \int_{(n-1)/n}^{1} J(t)dt = \int_{0}^{1} J(t)dt = 1.$$

Итак, при указанном способе задания весовых коэффициентов, L-оценка записывается в виде

$$T_n^* = \sum_{i=1}^{n} \left[\int_{(i-1)/n}^{i/n} J(t)dt \right] F_n^{-1}(i/n), \qquad \textbf{(8.7)}$$

асимптотически эквивалентном выражению (8.2).

При задании весовых коэффициентов с помощью набора констант $a_1, ..., a_n$, для которых $\sum a_i = 1$, L-оценка записывается в виде

$$T_{n2} = \sum_{j=1}^{m} a_j X_{([np_j])}, \qquad \textbf{(8.8)}$$

где p_j заданы и удовлетворяют неравенству $0 < p_j < 1$, $j = 1, ..., m$. Эти оценки являются оценкой функционала $T_2(F)$ вида

$$T_2(F) = \sum_{j=1}^{m} a_j F^{-1}(p_j). \qquad \textbf{(8.9)}$$

Отметим, что некоторые L-оценки, например винзоризованное среднее, удобно представлять в виде суммы $T_{n1} + T_{n2}$, которая является оценкой функционала

$$T(F) = \int_{0}^{1} J(t) F^{-1}(t)dt + \sum_{j=1}^{m} a_j F^{-1}(p_j). \qquad \textbf{(8.10)}$$

Обсудим теперь вопрос, при каких условиях, накладываемых на весовые коэффициенты $\{a_{ni}\}$ и ф.р. F, оценки T_n вида (8.1) являются асимптотически нормальными. Этот вопрос тщательно исследовался в литературе (см., например, [9, 29,38, 41, 43]). В этих работах были найдены довольно сложные наборы условий, одни из которых накладывают сильные ограничения на веса $\{a_{ni}\}$ и более слабые на ф.р. F, другие – наоборот. Мы ограничимся здесь доказательством асимптотической нормальности L-оценок, используя подход Мизеса.

Рассмотрим отдельно статистики T_{n1} и T_{n2}, определенные соответственно в (8.2) и (8.8). Начнем с рассмотрения статистики T_{n1}, которая записывается в виде $T_{n1} = T(F_n)$, где функционал $T(F)$ определяется выражением

$$T(F) = \int_0^1 J(t)F^{-1}(t)dt = \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} J(t)F^{-1}(t)dt.$$

Учитывая, что $F_n^{-1}(t) \xrightarrow{p} F^{-1}(t)$, $\forall t \in (0,1)$ при $n \to \infty$, и равенство $F_n^{-1}(t) = X_{(nt)}$, можно, при выполнении неравенств $(i-1)/n \le t \le i/n$, заменить эмпирическую квантильную функцию $F_n^{-1}(t)$ на i-ю порядковую статистику $X_{(i)}$. В результате получим

$$T_{n1} = T(F_n) = \int_0^1 J(t)F_n^{-1}(t)dt = \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} J(t)F_n^{-1}(t)dt = \sum_{i=1}^{n} a_{ni}X_{(i)}.$$

Для изучения асимптотических свойств представим статистику T_{n1} в виде разложения

$$T_{n1} = T_1(F) + V_{1n} + R_{1n} \qquad \textbf{(8.11)}$$

и конкретизируем аппроксимационную статистику V_{1n} и остаточный член R_{1n}. Для этого сначала вычислим дифференциал Гато первого порядка функционала $T_1(F)$. Действуя стандартным образом, с учетом (6.29) получим

$$d_1T_1(F;G-F)=\int_0^1 J(t)\frac{t-G(F^{-1}(t))}{f(F^{-1}(t))}dt=$$

$$=\int_{-\infty}^{\infty}[F(y)-G(y)]J(F(y))dy. \quad \textbf{(8.12)}$$

Преобразуем данное выражение. Для этого обозначим через $U(t)$ интеграл с переменным верхним пределом вида

$$U(t)=\int_{1/2}^{t}\{J(v)/f(F^{-1}(v))\}dv,\ 1/2<t<1. \quad \textbf{(8.13)}$$

Отметим, что

$$U(F(x))=\int_{1/2}^{F(x)}\{J(v)/f(F^{-1}(v))\}dv=\int_{F^{-1}(1/2)}^{x}J(F(y))dy. \quad \textbf{(8.14)}$$

Отсюда следует, что $dU(F(x))=J(F(x))dx$ и выражение (8.12) можно переписать в виде

$$d_1T_1(F;G-F)=\int_{-\infty}^{\infty}[F(y)-G(y)]J(F(y))dy=\int_{-\infty}^{\infty}[F(y)-G(y)]dU(F(y))=$$

$$=[F(y)-G(y)]U(F(y))\,|_{-\infty}^{\infty}-\int_{-\infty}^{\infty}U(F(y))d[F(y)-G(y)]=$$

$$=\int_{-\infty}^{\infty}\left[U(F(y))-\int_{-\infty}^{\infty}U(F(y))dF(y)\right]dG(y). \quad \textbf{(8.15)}$$

Из полученной формулы (8.15) следует, что аппроксимационная статистика V_{1n} в формуле (8.2.11) записывается в виде суммы н.о.р. случайных величин, то есть

$$V_{1n}=d_1T_1(F;F_n-F)=\frac{1}{n}\sum_{i=1}^{n}\left[U(F(X_i))-\int_0^1 U(t)dt\right]=$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left[U(U_i)-\int_0^1 U(t)dt\right], \quad \textbf{(8.16)}$$

где $U_1,...,U_n$ – н.о.р. случайные величины с равномерным распределением в интервале $[0,1]$. Отметим, что данная аппроксимационная статистика, согласно центральной предельной теореме, имеет асимптотически нормальное распределение. Функция влияния $IF(x;F,T)=IF(x;F,J)$ L-оценки T_{n1}, весовые коэффициенты которой $\{a_{ni}\}$ порождаются функцией $J(t)$ и которая является оценкой функционала $T_1(F)$ из (8.3), определяется выражением

$$IF(x;F,J)=\int_0^1 J(t)[t-C(F^{-1}(t)-x)]dF^{-1}(t)=U(F(x))-\int_0^1 U(t)dt\,,$$

$$x\in R^1. \qquad \textbf{(8.17)}$$

Рассмотрим теперь остаточный член R_{1n} в разложении (8.11), который запишем в виде

$$R_{1n}=T_{n1}-T_1(F)-d_1T_1(F;F_n-F). \qquad \textbf{(8.18)}$$

Нашей дальнейшей целью является доказательство сходимости по вероятности к нулю нормированного остаточного члена. Следуя работе [29], сначала рассмотрим разность $T_1(G)-T_1(F)$. Для удобства записи введем обозначение

$$V(t)=\int_0^t J(u)du\,,\ 0\le t\le 1. \qquad \textbf{(8.19)}$$

Отметим, что

$$dV(t)=J(t)dt \text{ и } dV[F(x)]=J(F(x))dF(x),\ x\in R^1. \qquad \textbf{(8.20)}$$

С учетом введенных обозначений, функционал $T_1(F)$ из (8.3) перепишем в виде

$$T_1(F)=\int_0^1 J(t)F^{-1}(t)dt=\int_{-\infty}^{\infty} x\,J[F(x)]dF(x)=\int_{-\infty}^{\infty} xdV[F(x)]. \qquad \textbf{(8.21)}$$

Используя (8.21), после интегрирования по частям получаем выражение для разности $T_1(G)-T_1(F)$ в виде

$$T_1(G)-T_1(F)=-\int_{-\infty}^{\infty}\{V[G(x)]-V[F(x)]\}dx. \qquad \textbf{(8.22)}$$

Отсюда, с учетом (8.12) получаем

$$T_1(G) - T_1(F) - d_1 T_1(F; G - F) =$$

$$= -\int_{-\infty}^{\infty} \{V[G(x)] - V[F(x)]\} dx - \int_{-\infty}^{\infty} [F(x) - G(x)] J(F(x)) dx =$$

$$= -\int_{-\infty}^{\infty} \left\{ \frac{V[G(x)] - V[F(x)]}{G(x) - F(x)} - J(F(x)) \right\} [G(x) - F(x)] dx =$$

$$= -\int_{-\infty}^{\infty} W_{G,F}(x) [G(x) - F(x)] dx, \tag{8.23}$$

где функция $W_{G,F}(x)$ определяется для $G(x) \neq F(x)$ в виде

$$W_{G,F}(x) = \frac{V[G(x)] - V[F(x)]}{G(x) - F(x)} - J(F(x)),$$

$$G(x) \neq F(x),\ \forall x \in R^1, \tag{8.24}$$

и $W_{G,F}(x) = 0$ для $G(x) = F(x)$. В выражении (8.23) заменим функцию распределения $G(x)$ на эмпирическую функцию распределения $F_n(x)$, в результате получим

$$R_{1n} = T_{n1} - T_1(F) - d_1 T_1(F; F_n - F) =$$

$$= -\int_{-\infty}^{\infty} W_{F_n,F}(x) [F_n(x) - F(x)] dx. \tag{8.25}$$

Проведем теперь анализ порядка остаточного члена R_{1n}, определенного выражением (8.25). Это выражение допускает два варианта представления остаточного члена R_{1n} с помощью неравенств, которые приводят к различным ограничениям на функцию распределения $F(x)$ и на функцию $J(t)$, порождающую весовые коэффициенты $\{a_{ni}\}$ L-оценки вида (8.7).

Вариант 1. Из (8.25) имеем

$$|R_{1n}| \leq \sup_x |F_n(x) - F(x)| \int_{-\infty}^{\infty} |W_{F_n,F}(x)| dx. \tag{8.26}$$

Отметим, что согласно неравенству Дворецкого – Кифера – Вольфовитца (см., например, [29]), для случая н.о.р. случайных величин $X_1,...,X_n$ с ф.р. $F(x)$ и эмпирической функции распределения $F_n(x)$, построенной по выборке $X_1,...,X_n$, выполняется выражение

$$\sup_x |F_n(x)-F(x)| = O_p(n^{-1/2}) \text{ при } n \to \infty. \qquad \textbf{(8.27)}$$

Таким образом, из неравенства (8.26) следует, что для доказательства выражения $n^{1/2}R_{1n} \to^p 0$ при $n \to \infty$ нужно выяснить, при каких ограничениях на функции $F(x)$ и $J(t)$ интеграл в правой части (8.26) сходится по вероятности к нулю.

Вариант 2. Из (8.25) имеем

$$|R_{1n}| \le \sup_x |W_{F_n,F}(x)| \int_{-\infty}^{\infty} |F_n(x)-F(x)|\,dx. \qquad \textbf{(8.28)}$$

В данном случае для проверки выражения $n^{1/2}R_{1n} \to^p 0$ при $n \to \infty$ нужно выяснить, при каких ограничениях на функции $F(x)$ и $J(t)$ выполняются выражения

$$\int_{-\infty}^{\infty} |F_n(x)-F(x)|\,dx = O_p(n^{-1/2})$$

$$\text{и } \sup_x |W_{F_n,F}(x)| \xrightarrow{p} 0 \text{ при } n \to \infty. \qquad \textbf{(8.29)}$$

При выполнении этих условий, из неравенства (8.26) следует, что $|R_{1n}| = o_p(n^{-1/2})$ и, следовательно, $n^{1/2}R_{1n} \to^p 0$ при $n \to \infty$. Приведем теперь результаты об асимптотической нормальности L-оценок в виде теорем.

Теорема 8.30. Пусть L-оценка определяется статистикой $T_{n1} = T_1(F_n)$ вида (8.2). Предположим, что исходная ф.р. $F(x)$ непрерывна, а функция $J(t)$, $0 \le t \le 1$, определяющая веса $\{a_{ni}\}$, также непрерывна, ограничена и $J(t)=0$ для $t \notin [\alpha,\beta]$, $0<\alpha<\beta<1$. Тогда случайная величина $\sqrt{n}[T_{n1}-T_1(F)]/\sigma(F,T_1)$ имеет асимптотически стандартное нормальное распределение, то есть выполняется выражение

$$L\{\sqrt{n}[T_{n1}-T_1(F)]/\sigma(F,T_1)\}=N(0,1) \text{ при } n\to\infty, \qquad \textbf{(8.31)}$$

где

$$\sigma^2(F,T_1)=\int_{-\infty}^{\infty} IF^2(x;F,J)dF(x)=\int_0^1 U^2(t)dt-\left(\int_0^1 U(t)dt\right)^2 \qquad \textbf{(8.32)}$$

и функции $IF(x;F,J)$, $U(t)$ определены соответственно в (8.17) и (8.13).

Доказательство. Согласно разложению (8.11), асимптотическое распределение $\sqrt{n}[T_{n1}-T_1(F)]$ определяется асимптотически нормальным распределением статистики $\sqrt{n}V_{1n}$ при выполнении выражения $n^{1/2}R_{1n}\to^p 0$ при $n\to\infty$. Чтобы убедиться в справедливости этого выражения, при условии, что выполняется (8.27), достаточно убедиться в выполнении другого выражения: $\int|W_{F_n,F}(x)|\,dx\xrightarrow{p} 0$ при $n\to\infty$. Это выражение действительно имеет место в соответствии с леммой (8.2.4А) из работы [29, с. 281] при выполнении следующих условий для функции $J(t)$. Функция $J(t)$ непрерывна, ограничена и $J(t)=0$ для $t\notin[\alpha,\beta]$, $0<\alpha<\beta<1$. Отметим, что при этих условиях на функцию $J(t)$ для функции $W_{G,F}(x)$ из (8.24) выполняется неравенство

$$|W_{G,F}(x)|\le[G(x)-F(x)]^{-1}\int_{F(x)}^{G(x)}|J(t)|\,dt+|J(F(x))|\le$$

$$\le 2\sup\nolimits_t|J(t)|<\infty. \qquad \textbf{(8.33)}$$

Отметим также, что при замене ф.р. G на F_n, функция $W_{G,F}(x)$ запишется в виде

$$W_{F_n,F}(x)=[F_n(x)-F(x)]^{-1}\int_{F(x)}^{F_n(x)}J(t)dt\ -J(F(x)). \qquad \textbf{(8.34)}$$

Итак, при выполнении (8.27) также выполняется, при сформулированных условиях на функцию $J(t)$, выражение $\int|W_{F_n,F}(x)|\,dx\to^p 0$

при $n \to \infty$ (детали см. в [29]). Таким образом, $|R_{1n}| = o_p(n^{-1/2})$, а также $n^{1/2}R_{1n} \to^p 0$ при $n \to \infty$. Учитывая теперь, что аппроксимационная статистика V_{1n} в разложении (8.11) является суммой независимых и одинаково распределенных случайных величин (см. формулу (8.16)), согласно центральной теореме, заключаем, что выполняется асимптотическое выражение

$$L\{\sqrt{n}V_{1n}\} = N\left(0, \int IF^2(x;F,J)dF(x)\right) \text{ при } n \to \infty, \qquad \textbf{(8.35)}$$

и, следовательно, согласно теореме Слуцкого, выполняются формулы (8.31) и (8.32). Доказательство завершено.

Замечание 8.36. В приведенной теореме (8.30) на функцию $J(t)$ кроме условий непрерывности и ограниченности накладывалось довольно ограничительное требование, чтобы эта функция равнялась нулю вне интервала $[\alpha,\beta]$. Этому ограничению, в частности, удовлетворяет α-урезанное среднее (см. (8.54)), для которого $J(t) = 1/(1-2\alpha)$ при $\alpha \le t \le 1-\alpha$, и $J(t) = 0$ вне интервала $[\alpha, 1-\alpha]$, $0 < \alpha < 1/2$. Однако не для всех практически важных статистик сформулированные условия на функцию $J(t)$ выполняются. Например, для средней разности Джини (см. пример 8.52) функция $J(t) = 4t - 2$, $0 \le t \le 1$, и, следовательно, для неё теорема 8.30 неприменима. В связи с этим приведем здесь другой вариант теоремы об асимптотической нормальности L-оценок, которая доказана в [29, 38]. В этой теореме снимаются условия ограниченности функции $J(t)$ и равенства нулю её вне интервала $[\alpha,\beta]$. Однако при этом на функцию распределения F кроме требования непрерывности накладывается дополнительное условие о конечности интеграла $\int [F(x)(1-F(x))]^{1/2} dx < \infty$.

Теорема 8.37. Пусть L-оценка определяется статистикой $T_{n1} = T_1(F_n)$ вида (8.2). Предположим, что функция $J(t)$, $0 \le t \le 1$, определяющая веса $\{a_{ni}\}$, непрерывна и для непрерывной функции распределения F выполнено условие $\int [F(x)(1-F(x))]^{1/2} dx < \infty$. Тогда случайная величина $\sqrt{n}[T_{n1} - T_1(F)]/\sigma(F,T_1)$ имеет асимптотически стандартное нормальное распределение, то есть выполняются выражения (8.31) и (8.32).

Доказательство этой теоремы проводится по аналогичной схеме доказательства теоремы 8.30, детали могут быть найдены в работах [9, 29, 38].

Замечание 8.38. Если L-оценка является суммой статистик T_{n1} и T_{n2}, то есть оценкой функционала $T(F)=T_1(F)+T_2(F)$, определенного в (8.10), то её функция влияния равна сумме функций влияния $IF(x;F,T_1)$ и $IF(x;F,T_2)$, то есть $IF(x;F,T)=IF(x;F,T_1)+$ $+IF(x;F,T_2)$. Кроме того, для оценки $T_n=T_{n1}+T_{n2}$ применимы теоремы 8.30 и 8.37 при сохранении их условий с добавлением требования, чтобы

$$f(F^{-1}(p_i))>0,\ i=1,...,m,$$

и $$0<\sigma^2(F,T)=\int[IF(x;F,T_1)+IF(x;F,T_2)]^2 dF(x)<\infty.$$

Пример 8.39. Оценка Гаствирта (см. [24]) является L-оценкой вида (8.8) и записывается в виде

$$T_{n2}=GAS=0,3F_n^{-1}(1/3)+0,4F_n^{-1}(1/2)+0,3F_n^{-1}(2/3)=$$
$$=0,3X_{([n/3])}+0,4X_{([n/2])}+0,3X_{([2n/3])}.$$

Функционал, соответствующий этой оценке, имеет вид (8.9) при $m=3$, $a_1=0,3$ и $p_1=1/3$, $a_2=0,4$ и $p_2=1/2$, $a_3=0,3$ и $p_3=2/3$. Функция влияния оценки Гаствирта (GAS)

$$IF(x;F,GAS)=0,3\frac{(1/3)-C[F^{-1}(1/3)-x]}{f(F^{-1}(1/3))}+0,4\frac{(1/2)-C[F^{-1}(1/2)-x]}{f(F^{-1}(1/2))}+$$
$$+0,3\frac{(2/3)-C[F^{-1}(2/3)-x]}{f(F^{-1}(2/3))}.$$

Заметим, что эта функция ограничена и имеет скачки в точках $F^{-1}(1/3)$, $F^{-1}(1/2)$ и $F^{-1}(2/3)$. При условии, что знаменатели в последнем выражении положительны, оценка Гаствирта GAS имеет асимптотически нормальное распределение. Отметим, что интерквартильный размах $\hat{S}_3(1/4)=[F_n^{-1}(3/4)-F_n^{-1}(1/4)]/2$ также является L-оценкой вида (8.8) при $m=2$, $a_1=1/2$ и $p_1=1/4$, $a_2=1/2$ и $p_2=3/4$.

Пример 8.40. Средняя разность Джини Δ_F может быть записана в виде функционала $T(F)$ от ф.р. F в разных вариантах:

$$\Delta_F=T(F)=\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}|x-y|\,dF(x)dF(y)=2\int\limits_{-\infty}^{\infty}F(x)(1-F(x))dx=$$

$$=\int\limits_{-\infty}^{\infty}x(4F(x)-2)dF(x)=\int\limits_{0}^{1}(4t-2)F^{-1}(t)dt.$$

Весовые коэффициенты a_{ni} L-оценки для функции $J(t)=4t-2$, $0\le t\le 1$,

$$a_{ni}=\int\limits_{(i-1)/n}^{i/n}J(t)dt=\int\limits_{(i-1)/2}^{i/n}(4t-2)dt=\frac{2}{n^2}(2i-n-1),\ i=1,...,n.$$

Для функционала $T(F)$ L-оценка средней разности Джини Δ_F записывается в виде

$$\Delta_n=T(F_n)=\frac{2}{n^2}\sum_{i=1}^{n}(2i-n-1)X_{(i)}.$$

Отметим, что данная оценка является V-оценкой Мизеса. Соответствующая U-статистика Хёфдинга записывается в виде

$$\Delta_n^*=\frac{2}{n(n-1)}\sum_{1\le i<j\le n}|X_i-X_j|=\frac{2}{n(n-1)}\sum_{1\le i<j\le n}|X_{(i)}-X_{(j)}|=$$

$$=\frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(X_{(j)}-X_{(i)})=\frac{2}{n(n-1)}\sum_{i=1}^{n}(2i-n-1)X_{(i)}.$$

Оценки Δ_n и Δ_n^* асимптотически эквивалентны (см. замечание 7.4.33 в [9]). Отметим, что приведенные общие результаты для L-оценок позволяют убедиться в асимптотической нормальности оценки средней разности Джини. В самом деле, для функции $J(t) = 4t - 2$, $0 \le t \le 1$, имеем

$$U(t) = \int_{1/2}^{t} \{J(v)/f(F^{-1}(v))\}dv = 4tF^{-1}(t) - 2F^{-1}(t) - 4\int_{-\infty}^{F^{-1}(t)} xdF(x) + C,$$

$$1/2 < t < 1.$$

Далее, выражение для функции влияния оценки средней разности Джини записывается в виде

$$IF(x; F, \Delta_F) = U(F(x)) - \int_0^1 U(t)dt = \varphi_F(x) - \int_{-\infty}^{\infty} \varphi_F(x)dF(x),\ x \in R^1,$$

где функция $\varphi_F(x)$ выражается через функцию $U(t)$ следующим образом:

$$\varphi_F(x) = 4xF(x) - 2x - 4\int_{-\infty}^{x} ydF(y) = U(F(x)),\ x \in R^1.$$

Учитывая далее, что функция $J(t) = 4t - 2$, $0 \le t \le 1$, непрерывна, заключаем, что случайная величина $\sqrt{n}(\Delta_n - \Delta_F)/\sigma_F(\Delta_n)$ асимптотически нормальна, где

$$\sigma_F^2(\Delta_n) = \int \varphi_F^2(x)dF(x) - \left(\int \varphi_F(x)dF(x)\right)^2.$$

Пример 8.41. В класс L-оценок входит семейство α-урезанных средних, которые широко используются в теории робастного оценивания и записываются в виде

$$\bar{X}_\alpha = T_\alpha(F_n) = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} X_{(i)},\ k = [\alpha n],\ 0 \le \alpha < 1/2, \qquad \textbf{(8.42)}$$

где функционал $T_\alpha(F)$ определяется выражением

$$T_\alpha(F) = \frac{1}{1-2\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} xdF(x) = \frac{1}{1-2\alpha}\int_{\alpha}^{1-\alpha} F^{-1}(t)dt,\ 0 \le \alpha < 1/2. \textbf{(8.43)}$$

Отметим, что функционал $T_\alpha(F)$ является состоятельным по Фишеру для семейства симметричных распределений, то есть выполняется выражение $T_\alpha(F_\theta)=\theta, \forall\theta\in R^1$, $F_\theta \in \Im_{S|\theta}$. В самом деле, для $F_\theta \in \Im_{S|\theta}$ выполняется равенство $F_\theta^{-1}(t)=2\theta - F_\theta^{-1}(1-t)$, $0\le t\le 1$, следовательно,

$$T_\alpha(F_\theta)=\frac{1}{1-2\alpha}\int_\alpha^{1-\alpha} F_\theta^{-1}(t)dt=\frac{1}{1-2\alpha}\int_\alpha^{1-\alpha}[2\theta-F_\theta^{-1}(1-t)]dt=2\theta-T_\alpha(F_\theta).$$

Отсюда следует, что $T_\alpha(F_\theta)=\theta$. Отметим также, что при $\alpha=0$ урезанное среднее превращается в выборочное среднее $\overline{X}$, а при $\alpha\to 1/2$ эта оценка совпадает с выборочной медианой $\overline{X}_{1/2}$ (см. (6.36)). Для L-оценки вида (8.42) функция $J(t)$, определяющая весовые коэффициенты a_{ni}, равна

$$J(t)=\begin{cases}1/(1-2\alpha), t\in[\alpha,1-\alpha],\\ \quad 0, \quad t\notin(\alpha,1-\alpha),\end{cases} \tag{8.44}$$

и весовые коэффициенты a_{ni}, согласно (8.4), вычисляются по формуле

$$a_{ni}=\int_{(i-1)/n}^{i/n} J(t)dt=\frac{1}{n(1-2\alpha)},\ [\alpha n]+1\le i\le n-[\alpha n]. \tag{8.45}$$

Используя формулу (8.17) и функцию $J(t)$ вида (8.44), запишем функцию влияния для α-урезанных средних в виде

$$IF(x;F,T_\alpha)=\int_0^1 J(t)[t-C(F^{-1}(t)-x)]dF^{-1}(t)=$$

$$=(1-2\alpha)^{-1}\{tF^{-1}(t)\,|_\alpha^{1-\alpha}-\int_\alpha^{1-\alpha}F^{-1}(t)dt-\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)}C(y-x)dy\}.$$

Введем обозначение

$$b(\alpha)=\alpha[F^{-1}(1-\alpha)+F^{-1}(\alpha)]+\int_\alpha^{1-\alpha}F^{-1}(t)dt. \tag{8.46}$$

Далее, учитывая, что

$$\int\limits_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} C(y-x)dy = \begin{cases} F^{-1}(1-\alpha)-F^{-1}(\alpha), & x<F^{-1}(\alpha), \\ F^{-1}(1-\alpha)-x, & F^{-1}(\alpha)\le x\le F^{-1}(1-\alpha),, \\ 0, & x>F^{-1}(1-\alpha), \end{cases}$$

окончательно получаем выражение функции влияния для α-урезанного среднего в виде

$$IF(x;F,T_\alpha)=\frac{1}{1-2\alpha}\begin{cases} F^{-1}(\alpha)-b(\alpha), & x<F^{-1}(\alpha), \\ x-b(\alpha), & F^{-1}(\alpha)\le x\le F^{-1}(1-\alpha), \\ F^{-1}(1-\alpha)-b(\alpha), & x>F^{-1}(1-\alpha), \end{cases} \quad x\in R^1. \quad \mathbf{(8.47)}$$

Заметим, что функция влияния α-урезанного среднего ограничена. Полученное выражение (8.47) приобретает более простой вид при дополнительном предположении о симметрии распределения. Учитывая эквивариантность L-оценок относительно линейных преобразований, можно без потери общности положить $T_\alpha(F)=0$, и учитывая, что $b(\alpha)=0$ для $F\in\Im_S$, получаем

$$IF(x;F,T_\alpha)=\frac{1}{1-2\alpha}\begin{cases} F^{-1}(1-\alpha)\,\mathrm{sign}(x), & |x|>F^{-1}(1-\alpha), \\ x, & |x|\le F^{-1}(1-\alpha), \end{cases} \quad x\in R^1. \mathbf{(8.48)}$$

Используя (8.32), получаем для $F\in\Im_S$ формулу вычисления асимптотической дисперсии $\sqrt{n}\bar{X}_\alpha$-оценки в виде

$$\sigma_F^2(\bar{X}_\alpha)=\int\limits_{-\infty}^{\infty} IF^2(x;F,T_\alpha)dF(x)=$$

$$=\frac{1}{(1-2\alpha)^2}\left\{\int\limits_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x^2dF(x)+2\alpha[F^{-1}(1-\alpha)]^2\right\}=$$

$$=\frac{1}{(1-2\alpha)^2}\left\{\int\limits_{\alpha}^{1-\alpha}[F^{-1}(t)]^2dt+2\alpha[F^{-1}(1-\alpha)]^2\right\}=$$

$$=\frac{2}{(1-2\alpha)^2}\left\{\int\limits_{1/2}^{1-\alpha}[F^{-1}(t)]^2dt+\alpha[F^{-1}(1-\alpha)]^2\right\}. \quad \mathbf{(8.49)}$$

Пример 8.50. В класс L-оценок также входит семейство α-винзоризованных средних, которые широко используются в теории робастного оценивания и записываются в виде

$$\tilde{X}_\alpha = n^{-1}\{kX_{(k)} + \sum_{i=k+1}^{n-k} X_{(i)} + kX_{(n-k)}\},\ k = [\alpha n],\ 0 \le \alpha < 1/2. \quad \mathbf{(8.51)}$$

Функционал $\tilde{T}_\alpha(F)$, соответствующий этим оценкам, определяется следующим выражением:

$$\tilde{T}_\alpha(F) = \int_\alpha^{1-\alpha} F^{-1}(t)dt + \alpha[F^{-1}(\alpha) + F^{-1}(1-\alpha)],\ 0 \le \alpha < 1/2. \quad \mathbf{(8.52)}$$

Отметим, что функционал $\tilde{T}_\alpha(F)$ является состоятельным по Фишеру для семейства симметричных распределений, то есть выполняется выражение $\tilde{T}_\alpha(F_\theta) = \theta, \forall\theta \in R^1,\ F_\theta \in \Im_{S|\theta}$. Убедимся в этом. Для симметричных относительно точки θ распределений, то есть для $F_\theta \in \Im_{S|\theta}$, выполняется равенство $F_\theta^{-1}(t) = 2\theta - F_\theta^{-1}(1-t)$, $0 \le t \le 1$, следовательно,

$$\tilde{T}_\alpha(F_\theta) = \int_\alpha^{1-\alpha} [2\theta - F_\theta^{-1}(1-t)]dt + \alpha[2\theta - F_\theta^{-1}(\alpha) + 2\theta - F_\theta^{-1}(1-\alpha)] =$$

$$= 2\theta - \int_\alpha^{1-\alpha} F^{-1}(t)dt + \alpha[F^{-1}(\alpha) + F^{-1}(1-\alpha)].$$

Отсюда получаем равенство $2\tilde{T}_\alpha(F) = 2\theta$, и, следовательно, $\tilde{T}_\alpha(F_\theta) = \theta,\ \forall\theta \in R^1,\ F_\theta \in \Im_{S|\theta}$. Можно также убедиться, что функция влияния для α-винзоризованных средних определяется выражением вида

$$IF(x;F,\tilde{T}_\alpha) = \begin{cases} F^{-1}(\alpha) - \alpha/f(F^{-1}(\alpha)) - b^*(\alpha), & x < F^{-1}(\alpha), \\ x - b^*(\alpha), & F^{-1}(\alpha) \le x \le F^{-1}(1-\alpha), \\ F^{-1}(1-\alpha) + \alpha/f(F^{-1}(1-\alpha)) - b^*(\alpha), & x > F^{-1}(1-\alpha), \end{cases} \quad \mathbf{(8.53)}$$

где $$b^*(\alpha) = b(\alpha) - \alpha^2/f(F^{-1}(\alpha)) - \alpha^2/f(F^{-1}(1-\alpha)). \quad \mathbf{(8.54)}$$

Для симметричных распределений функция влияния для α-винзоризованных средних записывается в виде

$$IF(x;F,\tilde{T}_\alpha)=\begin{cases}\{F^{-1}(1-\alpha)+\alpha/f(F^{-1}(\alpha))\}sign(x), & |x|>F^{-1}(1-\alpha),\\ x, & |x|\le F^{-1}(1-\alpha),\end{cases}$$
$$x\in R^1,\ F\in\mathfrak{I}_S. \qquad \textbf{(8.55)}$$

Асимптотическая дисперсия $\sqrt{n}\tilde{X}_\alpha$-оценки для $F\in\mathfrak{I}_S$ вычисляется по формуле

$$\sigma_F^2(\tilde{X}_\alpha)=\int_{-\infty}^{\infty}IF^2(x;F,\tilde{T}_\alpha)dF(x)=$$
$$=\int_\alpha^{1-\alpha}[F^{-1}(t)]^2dt+2\alpha\left[F^{-1}(1-\alpha)+\frac{\alpha}{f(F^{-1}(\alpha))}\right]^2. \qquad \textbf{(8.56)}$$

Отметим, что функция влияния α-винзоризованного среднего ограничена и имеет «скачки» (разрывы первого рода), равные величине $\alpha/f(F^{-1}(\alpha))$, в точках $F^{-1}(\alpha)$, $F^{-1}(1-\alpha)$, и, следовательно, эта оценка подвержена повышенному влиянию ошибок округления наблюдений и их группировки в окрестности этих точек.

Замечание 8.57. Пусть $X_1,\ldots,X_n$ – н.о.р. случайные величины с ф.р. $F(x-\theta)$ и плотностью $f(x-\theta)$, где θ – неизвестный параметр сдвига. Предполагаем, что ф.р. $F\in\mathfrak{I}_S$, то есть $F(x)=1-F(-x)$, $\forall x\in R^1$. Отметим, что в симметричном случае в классе L-оценок существует асимптотически эффективная оценка параметра сдвига θ, весовые коэффициенты которой определяются функцией $J(t)$ вида

$$J(t)=\psi'(F^{-1}(t))/I(f),\ 0\le t\le 1, \qquad \textbf{(8.58)}$$

где $\psi(x)=-f'(x)/f(x)$ и $I(f)=\int\psi^2(x)dF(x)=\int\psi'(x)dF(x)$ – информация Фишера относительно параметра сдвига θ. В самом деле, используя (8.56), из формулы (8.13), учитывая, что $F\in\mathfrak{I}_S$, получаем

$$U(t) = \int\limits_{1/2}^{t} J(v)dF^{-1}(v) = \frac{1}{I(f)} \int\limits_{1/2}^{t} \psi^{/}(F^{-1}(v))dF^{-1}(v) =$$

$$= \frac{1}{I(f)} \int\limits_{0}^{F^{-1}(t)} d\psi(x) = \frac{\psi(F^{-1}(t))}{I(f)}, \; 0 \le t \le 1.$$

Отметим, что для нечетной функции ψ и данной функции $U(t)$ второй интеграл в формуле (8.32) равен нулю, и поэтому выполняется выражение

$$\sigma^2(F,J) = \int\limits_0^1 U^2(t)dt = \frac{1}{I^2(f)} \int\limits_0^1 \psi^2(F^{-1}(t))dt =$$

$$= \frac{1}{I^2(f)} \int\limits_{-\infty}^{\infty} \psi^2(x)dF(x) = \frac{1}{I(f)}.$$

Отметим также, что функционал $T(F) = \int_0^1 J(t)F^{-1}(t)dt$, соответствующий L-оценке, весовые коэффициенты которой определяются функцией $J(t)$, $0 \le t \le 1$, является состоятельным по Фишеру. Это свойство проявляется в том, что выполняется равенство $T(F_\theta) = \theta$, $\forall \theta \in R^1$ и $F_\theta \in \Im_{S|\theta}$, для функций $J(t)$, $0 \le t \le 1$, удовлетворяющих следующим условиям:

$$J(t) = J(1-t), \; 0 \le t \le 1, \; \int_0^1 J(t)dt = 1.$$

В самом деле, для $F_\theta \in \Im_{S|\theta}$ получаем следующие равенства:

$$T(F_\theta) = \int_0^1 J(t)F_\theta^{-1}(t)dt = \int_0^1 J(t)[2\theta - F_\theta^{-1}(1-t)dt] =$$

$$= 2\theta \int_0^1 J(t)dt + \int_1^0 J(1-u)F_\theta^{-1}(u)\,du = 2\theta - \int_0^1 J(u)F_\theta^{-1}(u)du\,.$$

Отсюда следует, что $2T(F_\theta) = 2\theta$ или $T(F_\theta) = \theta$, $\forall \theta \in R^1$ и $F_\theta \in \Im_{S|\theta}$.

Замечание 8.59. Располагая приведенными формулами для функций влияния L-оценок, нетрудно вычислить их числовые характеристики, такие, как γ^* и λ^* (см. формулы (4.3) и (4.7)). Эти характеристики вычисляются при сравнении различных оценок ниже. Что касается вопросов качественной устойчивости и вычисления пороговых точек L-оценок, мы отсылаем читателя за подробностями к работе [5]. Здесь же лишь отметим, что пороговая точка L-оценок равна наибольшему вещественному числу α, $0 < \alpha \leq 1/2$, для которого интервал $[\alpha, 1-\alpha]$ служит областью определения функции $J(t)$, определяющей веса L-оценки. Например, для α-урезанного среднего вида (8.42), пороговая точка $\varepsilon^*(\bar{X}_\alpha) = \alpha$, и при выполнении неравенств $0 < \alpha \leq 1/2$ эта оценка удовлетворяет требованиям качественной устойчивости.

9. *R*-ОЦЕНКИ, ОСНОВАННЫЕ НА ИСПОЛЬЗОВАНИИ РАНГОВЫХ КРИТЕРИЕВ

Ходжес и Леман [44] предложили метод построения оценок параметров на основе критериев проверки гипотез. Они использовали статистики ранговых критериев и построили оценки параметра сдвига в двухвыборочных задачах и оценки параметра положения симметричных распределений в одновыборочных задачах. Позднее, Хеттманспергер [45], активно используя этот метод, построил различные точечные и интервальные оценки, основанные на критериях знаков, знаковых рангов Уилкоксона и ранговых сумм Манна – Уитни. Оценки, построенные этим методом, в литературе называют *R*-оценками. Отметим одну важную особенность этих оценок, состоящую в том, что их эффективность «наследует» эффективность критериев Питмена, с помощью которых они были построены процедурой Ходжеса – Лемана. В данном разделе обсуждаются *R*- и R_α-оценки, вычисляются их функции влияния и числовые характеристики устойчивости. В отличие от *R*-оценок, R_α-оценки, предложенные в [4], вычисляются по аналогии с α-урезанным средним не по полной исходной выборке $X_1,...,X_n$, а по выборке, в которой предварительно удалены $[\alpha n]$ наименьших и $[\alpha n]$ наибольших порядковых статистик, n – объем выборки и α – заданный параметр, $0 \le \alpha < 1/2$.

9.А. *R*-оценки

Рассмотрим сначала двухвыборочную задачу с альтернативой сдвига. Имеются две выборки $X_1,...,X_m$ и $Y_1,...,Y_n$ объемом m и n соответственно из распределений $F(x)$ и $G(x) = F(x-\Delta)$, где F –

произвольное непрерывное распределение. При данных предположениях двухвыборочная задача обнаружения эффекта «обработки» сводится (для случая $\Delta > 0$) к проверке гипотез

$$H_0 : \Delta = 0,\ H_1^+ : \Delta > 0 \tag{9.1}$$

и построению точечных и интервальных оценок параметра Δ по статистическим данным $X_1,...,X_m$ и $Y_1,...,Y_n$. Ранговые статистики с метками общего вида для построения критериев проверки гипотез (9.1) определяются в виде

$$V = \sum_{i=1}^{n} a(R_i), \tag{9.2}$$

где $R_1,...,R_n$ – ранги Y-ов в смешанной выборке объема $N = n + m$, состоящей из выборок $X_1,...,X_m$, $Y_1,...,Y_n$, и последовательность весов (меток) $a_i = a(i)$, $i = 1,...,n$, удовлетворяет неравенствам $0 = a(0) \le a(1) \le ... \le a(n)$, среди которых хотя бы одно строгое. Для конкретизации меток, как и в одновыборочном случае, используют φ-функцию, порождающую метки, которую для альтернативы сдвига $H_1^+ : \Delta > 0$ определяют в виде

$$\varphi_f(u) = -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))},\ 0 < u < 1. \tag{9.3}$$

Отметим, что в общем случае при построении двухвыборочных критериев проверки гипотез $H_0 : \Delta = 0$, $H_1^+ : \Delta > 0$ используют φ-функции, удовлетворяющие условиям:

$\varphi(u)$, $0 < u < 1$, – неубывающая функция, причем

$$0 < \int_0^1 (\varphi(u) - \overline{\varphi})^2 du < \infty, \text{ где } \overline{\varphi} = \int_0^1 \varphi(u) du. \tag{9.3a}$$

Приближенные метки ранговых статистик (9.2) определяются с помощью φ-функции в виде

$$a(i) = \varphi\left(\frac{i}{N+1}\right),\ i = 1,...,n. \tag{9.4}$$

С использованием приближенных меток (9.4) статистики (9.2) записываются в такой форме:

$$\bar{V} = \frac{1}{N}\sum_{j=1}^{n}\varphi\left(\frac{R_j}{N+1}\right), \qquad \textbf{(9.5)}$$

где $R_1,...,R_n$ – ранги Y-ов в смешанной выборке объема $N = n + m$. Отметим, что в литературе используют и другие способы задания весов при помощи φ-функции (см. [45, 46]), например в виде

$$a_i = (m+n)\int_{(i-1)/N}^{i/N}\varphi(u)du\,,\ i = 1,...,n. \qquad \textbf{(9.6)}$$

Для «хороших» функций φ и распределений F приведенные варианты получения весов приводят к асимптотически эквивалентным ранговым критериям. Везде ниже будем для простоты предполагать, что объемы выборок равны, то есть $m = n$ и $N = 2n$.

Для построения R-оценки параметра сдвига Δ, которую обозначим через $T_n = T(F_n)$, воспользуемся ранговой статистикой вида (9.5) и процедурой Ходжеса – Лемана (см. П.3). Согласно этой процедуре, оценка $T_n = T(F_n)$ определяется как решение уравнения $S(T_n) \approx 0$, где ранговая статистика $S(T_n)$ записывается в виде

$$S(T_n) = \sum_{i=1}^{n}\varphi\{R_i/(2n+1)\}. \qquad \textbf{(9.7)}$$

Здесь φ – заданная функция меток, а R_i – ранг $Y_i - T_n$, $i = 1,...,n$, в объединенной выборке объема $2n$, состоящей из $X_1,...,X_n$ и $Y_1 - T_n,...,Y_n - T_n$. Итак, вторая выборка Y-ов смещается до тех пор, пока критерий, основанный на ранговой статистике $S(T_n)$, не перестанет «чувствовать» различие в сдвиге. Заметим, что $S(T_n)$ является разрывной функцией, поэтому точное равенство нулю в уравнении $S(T_n) = 0$ может и не достигаться. По этой причине правило нахождения R-оценки $T_n(\varphi)$ параметра сдвига Δ, определяемой функцией меток φ, записывают в виде

$$T_n(\varphi) = \{T_n^*(\varphi) + T_n^{**}(\varphi)\}/2, \qquad \textbf{(9.8)}$$

где $$T_n^*(\varphi) = \sup\{r : S(r) > 0\},\ T_n^{**}(\varphi) = \inf\{r : S(r) < 0\}. \qquad \textbf{(9.9)}$$

Рассмотрим теперь одновыборочную задачу. Пусть имеется выборка $X_1,...,X_n$ с произвольным непрерывным распределением $F(x-\theta)$. При дополнительном предположении о симметрии распределения, то есть для $F \in \mathfrak{I}_S$, параметр сдвига θ в одновыборочном варианте также называют параметром положения наблюдаемой случайной величины X.

R-оценка $T_n(\varphi)$ параметра сдвига θ определяется как решение уравнения $S(T_n) \approx 0$, в котором ранговая статистика $S(T_n)$ вида (9.7) вычисляется по исходной выборке $X_1,...,X_n$ и в качестве отсутствующей второй выборки используется зеркальное отражение исходной выборки, то есть $2T_n - X_1,...,2T_n - X_n$. Отметим, что R-оценки могут быть представлены в виде функционала от эмпирической функции распределения F_n, то есть в виде $T_n = T(F_n)$, где функционал $T(F)$ задается неявно с помощью выражения

$$\int \varphi\{[F(x) + 1 - F(2T(F) - x)]/2\}\, dF(x) = 0, \qquad \textbf{(9.10)}$$

которое для непрерывных и строго монотонных ф.р. F, после замены $F(x) = t$, можно записать в виде

$$\int_0^1 \varphi\{[t + 1 - F(2T(F) - F^{-1}(t))]/2\}\, dt = 0. \qquad \textbf{(9.11)}$$

Отметим также, что при выполнении условий регулярности, накладываемых на ф.р. F и на функцию меток φ, которые вытекают из требования слабой непрерывности функционала $T(F)$, R-оценка, являющаяся решением уравнения $S(T_n) = 0$, сходится по вероятности к функционалу $T(F)$, заданному выражениями (9.10) либо (9.11) (см., например, [5, 47]).

Обсудим теперь асимптотические свойства R-оценок, используя подход Мизеса. Представим статистику $T_n(\varphi)$ в виде разложения

$$T_n(\varphi) = T(F) + V_{1n} + R_{1n}, \qquad \textbf{(9.12)}$$

и конкретизируем аппроксимационную статистику V_{1n} и остаточный член R_{1n}. Для этого сначала вычислим дифференциал Гато первого порядка функционала $T(F)$, заданного выражением (9.11). Пусть, как обычно, $F_\lambda = F + \lambda(G - F)$, $0 \le \lambda \le 1$, и F_λ^{-1} – обратная функция для ф.р. F_λ. Отметим, что

$$F_\lambda(2T(F_\lambda) - F_\lambda^{-1}(t)) =$$
$$= F(2T(F_\lambda) - F_\lambda^{-1}(t)) + \lambda[G(2T(F_\lambda) - F_\lambda^{-1}) - F(2T(F_\lambda) - F_\lambda^{-1})].$$

Далее, подставив в выражение (9.11) ф.р. F_λ вместо F, дифференцируя по λ и полагая $\lambda = 0$, получим

$$-\frac{1}{2}\int_0^1 \varphi'\left[\frac{t+1-F(2T(F)-F^{-1}(t))}{2}\right]\times$$
$$\times\left\{f(2T-F^{-1}(t))\left[2d_1T(F;G-F)-\frac{t-G(F^{-1}(t))}{f(F^{-1}(t))}\right]+\right.$$
$$\left.+G(2T-F^{-1}(t)-F(2T-F^{-1}(t))\right\}dt = 0. \qquad \textbf{(9.13)}$$

Для простоты дальнейших преобразований, примем следующие предположения на функцию распределения F и на функцию меток $\varphi(t)$. Предполагаем, что функция распределения F симметрична, то есть $F \in \Im_{S|0}$, а для функции меток $\varphi(t)$ предполагаем, что она неубывающая, дифференцируемая на интервале $[0,1]$, причем

$$\int_0^1 \varphi^2(t)dt < \infty, \ \varphi(t) = -\varphi(1-t), \ 0 \le t \le 1. \qquad \textbf{(9.14)}$$

Кроме того, функции $\varphi'(t)$ и $f(F^{-1}(t))$ имеют ограниченную вариацию на $[0,1]$ и интеграл

$$\int_\alpha^{1-\alpha} \varphi'(t)f(F^{-1}(t))dt \ne 0, \ 0 \le \alpha < 1/2. \qquad \textbf{(9.15)}$$

Класс таких функций меток обозначим через C_R. Отметим, что при этих предположениях выполняется условие

$$\int_0^1 \varphi(t)dt = 0, \qquad \textbf{(9.16)}$$

что соответствует равенству на метки a_i в виде

$$\sum_{i=1}^{n} a_i = 0. \qquad \textbf{(9.17)}$$

Кроме того, при этих предположениях также можно без потери общности положить $T(F) = 0$. Итак, для $F \in \Im_{S|0}$ и $\varphi(t) \in C_R$ выражение (9.13) преобразуется к следующему виду:

$$-\frac{1}{2}\int_0^1 \varphi'(t)\left[f(F^{-1}(t))\left\{ 2d_1T(F;G-F) - \frac{t - G(F^{-1}(t))}{f(F^{-1}(t))} \right\} + \right.$$
$$\left. + G(F^{-1}(1-t)) - 1 + t \right] dt = 0$$

или

$$-d_1T(F;G-F)\int_0^1 \varphi'(t) f(F^{-1}(t))dt - \frac{1}{2}\int_0^1 G(F^{-1}(t))d\varphi(t) -$$
$$-\frac{1}{2}\int_0^1 G(F^{-1}(1-t))d\varphi(t) + \frac{1}{2}\int_0^1 d\varphi(t) = 0.$$

Интегрирование по частям после преобразований дает окончательное выражение для дифференциала Гато первого порядка функционала $T(F)$ в виде

$$d_1T(F;G-F) = \left\{ \int_{-\infty}^{\infty} \varphi(F(x))dG(x) \right\} \Big/ \left\{ \int_0^1 \varphi'(t) f(F^{-1}(t))dt \right\}. \qquad \textbf{(9.18)}$$

Отсюда следует, что функция влияния $IF(x;F,T) = IF(x;F,\varphi)$ для R-оценки $T_n(\varphi)$ параметра сдвига θ определяется в виде

$$IF(x;F,\varphi) = d_1T(F;\Delta_x - F) = \frac{\varphi(F(x))}{\int_0^1 \varphi'(t) f(F^{-1}(t))dt},$$
$$x \in R^1,\ F \in \Im_S. \qquad \textbf{(9.19)}$$

Далее, аппроксимационная статистика V_{1n} в разложении (9.12) равна

$$V_{1n} = \frac{1}{n\int_0^1 \varphi'(t) f(F^{-1}(t))dt} \sum_{i=1}^{n} \varphi(F(X_i)) =$$

$$= \frac{1}{n\int_0^1 \varphi'(t) f(F^{-1}(t))dt} \sum_{i=1}^{n} \varphi(U_i), \qquad \textbf{(9.20)}$$

где $X_1,...,X_n$ – н.о.р. случайные величины с ф.р. F, а $U_1,...,U_n$ – н.о.р. случайные величины с равномерным распределением в интервале $[0,1]$. Остаточный член R_{1n} в разложении (9.12) равен $R_{1n} = T_n(\varphi) - T(F) - d_1T(F; F_n - F)$. Можно убедиться (см., например, [47]), что при сформулированных условиях на функцию меток φ, и для $F \in \Im_{S|0}$ выполняется выражение $\sqrt{n}R_{1n} \to^p 0$ при $n \to \infty$.

Теорема 9.21. Пусть $F \in \Im_{S|0}$ и функция меток $\varphi(t) \in C_R$. Тогда R-оценка $T_n(\varphi)$ параметра сдвига θ, определяемая как решение уравнения $S(T_n) \approx 0$, асимптотически нормальна, то есть выполняется выражение

$$L\{\sqrt{n}[T_n(\varphi) - T(F)]/\sigma(F,\varphi)\} = N(0,1) \text{ при } n \to \infty, \qquad \textbf{(9.22)}$$

где функционал $T(F)$ задан формулой (9.11) и асимптотическая дисперсия $\sqrt{n}T_n(\varphi)$-оценки вычисляется как

$$\sigma^2(F,\varphi) = \int_{-\infty}^{\infty} IF^2(x;F,\varphi)dF(x) =$$

$$= \frac{\int_{-\infty}^{\infty} \varphi^2(F(x))dF(x)}{\left(\int_{-\infty}^{\infty} \varphi'(F(x)) f^2(x)dx\right)^2} = \frac{\int_0^1 \varphi^2(t)dt}{\left(\int_0^1 \varphi'(t) f(F^{-1}(t))dt\right)^2}. \qquad \textbf{(9.23)}$$

Доказательство основано на использовании теоремы 6.20 и проверке справедливости выражения $\sqrt{n}R_{1n} \to^{p} 0$ при $n \to \infty$. Детали могут быть найдены в [5, 47].

Пример 9.24. Пусть функция меток равна $\varphi(t) = t - 1/2$, $0 \le t \le 1$. Можно убедиться, что данная функция меток принадлежит классу C_R. Задание такой функции меток соответствует использованию в двухвыборочном варианте линейной ранговой статистики Уилкоксона (см., например, [8]), а в одновыборочном варианте соответствует критерию знаковых рангов Уилкоксона (см. [8]). Соответствующие R-оценки параметра Δ в двухвыборочном варианте и параметра θ в одновыборочном варианте, построенные с помощью процедуры Ходжеса – Лемана (см. П.3), записываются в виде

$$\Delta_n = med\{\ (Y_i - X_j),\ i = 1,...,n,\ j = 1,...,n\ \},$$
$$\theta_n = HL = med\{(X_i + X_j)/2,\ 1 \le i \le j \le n\}. \qquad \textbf{(9.25)}$$

Отметим, что функционал $T(F)$, заданный в (9.10), для функции меток $\varphi(t) = t - 1/2, 0 \le t \le 1$, определяется выражениями

$$\int_{-\infty}^{\infty} \{F(2T(F) - x) - (1/2)\} dF(x) = 0$$

или

$$\int_{-\infty}^{\infty} F(2T(F) - x) dF(x) = \frac{1}{2}. \qquad \textbf{(9.26)}$$

Из последнего выражения следует, что функционал $T(F)$ для данной функции меток определяет половину медианы случайной величины $Y = X_1 + X_2$, где X_1 и X_2 – независимые случайные величины с функцией распределения F. В самом деле, функция распределения случайной величины $Y = X_1 + X_2$ равна

$$G(y) = P\{Y \le y\} = P\{(X_1 + X_2) \le y\} = \int F(y - x) dF(x). \qquad \textbf{(9.27)}$$

Сравнивая данное выражение со вторым равенством в (9.26), заключаем, что $2T(F) = G^{-1}(1/2)$. Отметим также, что свертка симметричных распределений симметрична, то есть, если $F \in \Im_S$, то и $G \in \Im_S$. Следствием этого факта является состоятельность по Фишеру функционала $T(F)$, то есть для него выполняется выражение

$T(F_\theta)=\theta$, $\forall\theta\in R^1$ и $F_\theta\in\Im_{S|\theta}$. Убедимся в этом. Введем обозначение $\lambda(T(F))=\int\limits_{-\infty}^{\infty}F(2T(F)-x)dF(x)$, тогда $T(F)=\lambda^{-1}(1/2)$. Далее, для $F_\theta\in\Im_{S|\theta}$ выполняется равенство $F(x)=1-F(2\theta-x)$, $\forall x\in R^1$. Используя это равенство, запишем

$$\frac{1}{2}=\int\limits_{-\infty}^{\infty}F(2T-x)dF(x)=-\int\limits_{-\infty}^{\infty}[1-F(2\theta-2T+x)]dF(2\theta-x)=\|2\theta-x=y\|=$$

$$=1-\int\limits_{-\infty}^{\infty}F[2(2\theta-T)-y]dF(y)=1-\lambda(2\theta-T(F)).$$

Отсюда получаем $\lambda(2\theta-T(F))=1/2$, или $2\theta-T(F)=\lambda^{-1}(1/2)$, или $2\theta=2T(F)$, $\forall\theta\in R^1$ и $F_\theta\in\Im_{S|\theta}$.

Выборочная оценка $T_n=T(F_n)=HL$ функционала $T(F)$ находится из уравнения

$$\int\limits_{-\infty}^{\infty}F_n(2T(F_n)-x)dF_n(x)=\frac{1}{2}\text{ или }\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}C\{T_n-(X_i+X_j)/2\}=\frac{1}{2}.$$

Отсюда получаем $T_n=T(F_n)=HL=med\{(X_i+X_j)/2,\ 1\le i,j\le n\}$. Заметим, что оценка Ходжеса – Лемана определяется в виде медианы средних Уолша $(X_i+X_j)/2$, причем в данном случае используются все n^2 пар средних Уолша (то есть рассматривается V_n-оценка Мизеса). В литературе также встречаются варианты записи этой оценки, когда используются только такие пары, для которых $i<j$ или $i\le j$. Все три варианта асимптотически эквивалентны. Функция влияния асимптотически нормальной оценки Ходжеса – Лемана ограничена при условии, что $\int f(F^{-1}(t))dt\ne 0$, и определяется, согласно (9.19), в виде

$$IF(x;F,HL)=\frac{F(x)-(1/2)}{\int\limits_0^1 f(F^{-1}(t))dt},\ x\in R^1,\ F(x)\in\Im_S. \quad \textbf{(9.28)}$$

Далее, из (9.23) следует, что асимптотическая дисперсия $\sqrt{n}HL$-оценки вычисляется по формуле

$$\sigma_F^2(HL)=\int_{-\infty}^{\infty} IF^2(x;F,HL)dF(x)=1/12\left(\int_0^1 f(F^{-1}(t))dt\right)^2, \qquad \textbf{(9.29)}$$

которая была приведена ранее в (6.49). Отметим, что для логистического распределения $F_{(2)}(x)=1/(1+e^{-x})$, $x\in R^1$, асимптотическая дисперсия $\sqrt{n}HL$-оценки равна трем, то есть $\sigma^2(F_{(2)},HL)=$ $=3=1/I(f_{(2)})$. Это означает, что HL-оценка Ходжеса – Лемана является асимптотически эффективной оценкой параметра сдвига θ логистического распределения $F_{(2)}(x-\theta)$. Можно убедиться, что чувствительность к грубым ошибкам $\gamma^*(F,HL)$ для HL-оценки вычисляется по формуле

$$\gamma^*(F,HL)=\frac{\sup_x|F(x)-1/2|}{\int_0^1 f(F^{-1}(t))dt}=\frac{1}{2\int_0^1 f(F^{-1}(t))dt}.$$

Для нормального распределения, то есть при $F=\Phi$, с учетом того, что интеграл

$$\int_0^1 \phi(\Phi^{-1}(t))dt=\int_{-\infty}^{\infty}\phi^2(x)dx=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-x^2}dx=\frac{1}{2\sqrt{\pi}},$$

получаем $\gamma^*(\Phi,HL)=\sqrt{\pi}\approx 1,77$. Далее, чувствительность к локальным изменениям наблюдений $\lambda^*(F,HL)$ для HL-оценки вычисляется по формуле

$$\lambda^*(F,HL)=\frac{\sup_x|f(x)|}{\int_0^1 f(F^{-1}(t))dt},$$

и при $F=\Phi$ $\qquad \lambda^*(\Phi,HL)=\dfrac{1/\sqrt{2\pi}}{1/2\sqrt{\pi}}=\sqrt{2}\approx 1,41.$

Пример 9.30. Пусть функция меток $\varphi(t)=\Phi^{-1}(t)$, $0\le t\le 1$. Использование такой функции меток соответствует линейной ранговой статистики Ван-дер-Вардена. Соответствующая *R*-оценка $T_n = T(F_n) = NS$ с нормальными метками в литературе обозначается NS (Normal Score) и определяется как решение уравнения $S(T_n)\approx 0$, где ранговая статистика $S(T_n)$ записывается в виде

$$S(T_n)=\sum_{i=1}^{n}\Phi^{-1}\left(\frac{R_i}{2n+1}\right). \qquad \textbf{(9.31)}$$

Здесь R_i – ранг $Y_i - T_n$, $i=1,...,n$, в объединенной выборке объема $2n$, состоящей из $X_1,...,X_n$ и $Y_1-T_n,...,Y_n-T_n$. Функция влияния $IF(x;F,NS)$ для *R*-оценки с нормальными метками не ограничена и определяется в виде

$$IF(x;F,NS)=\frac{\Phi^{-1}(F(x))}{\int\limits_0^1\{f(F^{-1}(t))/\phi(\Phi^{-1}(t))\}dt},\ x\in R^1,\ F(x)\in\Im_S, \qquad \textbf{(9.32)}$$

где $\phi(x)=\Phi'(x)$ – стандартная нормальная плотность. В частности, при $F(x)=\Phi(x)$ её функция влияния совпадает с функцией влияния выборочного среднего, то есть $IF(x;\Phi,NS)=x=IF(x;\Phi,\bar{X})$, $x\in R^1$.

Согласно (9.23), асимптотическая дисперсия $\sqrt{n}NS$-оценки для $F(x)\in\Im_S$ вычисляется по формуле

$$\sigma_F^2(NS)=\int\limits_{-\infty}^{\infty} IF^2(x;F,NS)dF(x)=\frac{\int\limits_0^1[\Phi^{-1}(t)]^2\,dt}{\left(\int\limits_0^1\{f(F^{-1}(t))/\phi(\Phi^{-1}(t))\}dt\right)^2}=$$

$$=\left\{\int_{-\infty}^{\infty}[f^2(x)/\phi(\Phi^{-1}(F(x)))]dx\right\}^{-2}. \qquad \textbf{(9.33)}$$

Отметим, что при $F(x)=\Phi(x)$ асимптотическая дисперсия $\sqrt{n}NS$-

оценки равна единице, то есть $\sigma^2_\Phi(NS) = 1 = 1/I(\phi)$. Это означает, что NS-оценка является асимптотически эффективной оценкой параметра сдвига θ нормального распределения $\Phi(x-\theta)$.

9.Б. R_α-оценки

В отличие от рассмотренных выше R-оценок, R_α-оценки, предложенные в [4], вычисляются по аналогии с α-урезанным средним не по полной исходной выборке $X_1,...,X_n$, а по выборке, в которой предварительно удалены $[\alpha n]$ наименьших и $[\alpha n]$ наибольших порядковых статистик, n – объем выборки и α – заданный параметр, $0 \le \alpha < 1/2$. Итак, R_α-оценки, построенные с помощью ранговой статистики с функцией меток φ, обозначим через $R_\alpha(\varphi) = T_{n,\alpha}$ и определим как решение уравнения $S_\alpha(T_{n,\alpha}) \approx 0$, где $S_\alpha(r)$ – ранговая статистика вида

$$S_\alpha(r) = \sum_{i=[\alpha n]+1}^{n-[\alpha n]} \varphi\left(\frac{r_\alpha(i)}{2(n-2[\alpha n])+1}\right),\ 0 \le \alpha < 1/2, \qquad \textbf{(9.34)}$$

$r_\alpha(i)$ – ранг $2r - X_{(i)}$ в смешанной выборке объема $2(n-2k)$, состоящей из $X_{(k+1)},...,X_{(n-k)}$ и $2r - X_{(k+1)},...,2r - X_{(n-k)}$, $k = [\alpha\, n]$. Функционал $T_\alpha(F)$, соответствующий R_α-оценке, определяется выражением вида

$$\int_\alpha^{1-\alpha} \varphi\{[t+1-F(2T_\alpha(F)-F^{-1}(t))]/2\}\,dt = 0,\ 0 \le \alpha < 1/2. \qquad \textbf{(9.35)}$$

Выбирая при заданном значении параметра α функцию меток $\varphi(t)$, $0 \le t \le 1$, мы тем самым конкретизируем $R_\alpha(\varphi)$-оценку. Ниже будем предполагать, что функция меток $\varphi(t)$ неубывающая, дифференцируемая на $[0,1]$, причем

$$\int_0^1 \varphi^2(t)dt < \infty,\ \varphi(t) = -\varphi(1-t),\ 0 \le t \le 1. \qquad \textbf{(9.36)}$$

Кроме того, функции $\varphi'(t)$ и $f(F^{-1}(t))$ имеют ограниченную вариацию на $[0,1]$ и интеграл

$$\int_{\alpha}^{1-\alpha} \varphi'(t) f(F^{-1}(t))dt \neq 0,\ 0 \leq \alpha < 1/2. \qquad \textbf{(9.37)}$$

Класс таких функций меток обозначим через C_R^*.

Замечание 9.38. Отметим, что при $\alpha = 0$ $R_\alpha(\varphi)$-оценка совпадает с рассмотренной ранее $R(\varphi)$-оценкой. Отметим также, что $R_\alpha(\varphi)$-оценку можно определить с помощью R-оценки, основанной на функции меток $\varphi_\alpha(t)$, заданной в виде

$$\varphi_\alpha(t) = \begin{cases} \varphi(\alpha), & t < \alpha, \\ \varphi(t), & \alpha \leq t \leq 1-\alpha, \\ \varphi(1-\alpha), & t > 1-\alpha. \end{cases} \qquad \textbf{(9.39)}$$

Это означает, что свойства $R_\alpha(\varphi)$-оценки совпадают со свойствами $R(\varphi_\alpha)$-оценки.

Обсудим теперь асимптотические свойства $R_\alpha(\varphi)$-оценок. Сначала выпишем выражение для дифференциала Гато первого порядка функционала $T_\alpha(F)$, заданного выражением (9.35). Для простоты считаем, что $F \in \Im_{S|0}$, при этом положим $T_\alpha(F) = 0$, и пусть функция меток $\varphi(t) \in C_R^*$. Действуя по такой же схеме, как и для R-оценок, получим

$$d_1 T_\alpha(F; G-F) =$$

$$= A_F^{-1}(\varphi,\alpha)\int \{[C(y - F^{-1}(1-\alpha)) - C(y - F^{-1}(\alpha))]\varphi(F(y)) +$$

$$+\varphi(\alpha)C(F^{-1}(\alpha) - y) + \varphi(1-\alpha)C(y - F^{-1}(1-\alpha))\}dG(y), \qquad \textbf{(9.40)}$$

где $A_F(\varphi,\alpha) = \int_{\alpha}^{1-\alpha} \varphi'(t) f(F^{-1}(t))dt$.

Отсюда следует, что для симметричных распределений (то есть $F \in \Im_{S|0}$), функция влияния $IF(x;F,T_\alpha) = IF(x;F,\varphi_\alpha)$ для R_α-оценки $T_{n,\alpha}(\varphi)$ параметра сдвига θ определяется в виде

$$IF(x;F,\varphi_\alpha)=A_F^{-1}(\varphi,\alpha)\begin{cases}\varphi(1-\alpha)\operatorname{sign}(x), |x|>F^{-1}(1-\alpha),\\ \varphi(F(x)), \quad |x|\le F^{-1}(1-\alpha),\end{cases}$$
$$F\in\Im_{S|0}. \tag{9.41}$$

Замечание 9.42. Отметим, что функция влияния вида (9.41) ограничена и, следовательно, R_α-оценки подвержены лишь ограниченному влиянию выбросов в выборке. Отметим также, что выражение (9.40) следует из (9.18) при использовании функции меток $\varphi_\alpha(t)$ из (9.39). В самом деле, подставив в формулу (9.18) функцию меток $\varphi_\alpha(t)$, получим

$$d_1T_\alpha(F;G-F)=\frac{\int\limits_{-\infty}^{\infty}\varphi_\alpha(F(x))dG(x)}{\int\limits_0^1\varphi_\alpha'(t)f(F^{-1}(t))dt}=$$

$$=\frac{\varphi(\alpha)\int_{-\infty}^{F^{-1}(\alpha)}dG(x)+\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)}\varphi(F(x))dG(x)+\varphi(1-\alpha)\int_{F^{-1}(1-\alpha)}^{\infty}dG(x)}{\int_\alpha^{1-\alpha}\varphi'(t)f(F^{-1}(t))dt}.$$

Теорема 9.43. Пусть $F\in\Im_{S|0}$ и функция меток $\varphi(t)\in C_R^*$. Тогда R_α-оценка $T_{n,\alpha}(\varphi)$ параметра сдвига θ, определяемая как решение уравнения $S(T_{n,\alpha})\approx 0$, асимптотически нормальна, то есть выполняется выражение

$$L\{\sqrt{n}[T_{n,\alpha}(\varphi)-T_\alpha(F)]/\sigma(F,\varphi,\alpha)\}=N(0,1) \text{ при } n\to\infty, \tag{9.44}$$

где функционал $T_\alpha(F)$ задан выражением (9.35) и асимптотическая дисперсия $\sqrt{n}T_{n,\alpha}(\varphi)$-оценки вычисляется по формуле

$$\sigma^2(F,\varphi,\alpha)=\int\limits_{-\infty}^{\infty}IF^2(x;F,\varphi)dF(x)=\frac{B(\varphi,\alpha)}{A_F^2(\varphi,\alpha)}, \tag{9.45}$$

где
$$B(\varphi,\alpha) = 2\alpha\varphi^2(\alpha) + \int_{\alpha}^{1-\alpha} \varphi^2(t)dt,$$

$$A_F(\varphi,\alpha) = \int_{\alpha}^{1-\alpha} \varphi'(t) f(F^{-1}(t))dt \,. \qquad \textbf{(9.46)}$$

Доказательство этой теоремы основано на использовании теоремы (6.20) и проводится по аналогии с доказательством асимптотической нормальности R-оценок, приведенном в [47]. Сравнение R_α- и R-оценок в рамках различных супермоделей приводится в работах [4, 9] (см. также раздел 17).

Пример 9.47. Пусть функция меток равна $\varphi(t) = t - 1/2$, $0 \le t \le 1$. Для такой функции меток R_α-оценка, предложенная в [4] и названная α-урезанной оценкой Ходжеса – Лемана, определяется в виде

$$HL_\alpha = med\,\{(X_{(i)} + X_{(j)})/2,\ k+1 \le i \le j \le n-k\},$$
$$k = [\alpha n],\ 0 \le \alpha < 1/2. \qquad \textbf{(9.48)}$$

Отметим, что в отличие от HL-оценки вида (9.25), которая вычисляется по исходной выборке $X_1,...,X_n$ путем определения медианы средних значений Уолша $(X_i + X_j)/2$, $1 \le i \le j \le n$, α-урезанная оценка Ходжеса – Лемана вычисляется на основе упорядоченной статистики $X_{(1)},...,X_{(n)}$, из которой предварительно удалены $k = [\alpha n]$ наименьших и наибольших порядковых статистик. В данном случае функционал $T_\alpha(F)$ задан неявно выражением

$$\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} F(2T_\alpha(F) - x)dF(x) = \frac{1-2\alpha}{2}. \qquad \textbf{(9.49)}$$

Пусть $F \in \Im_{S|0}$ и функция меток $\varphi(t) \in C_R^*$. Функционал $T_\alpha(F)$ является состоятельным по Фишеру, то есть для него выполняется выражение $T(F_\theta) = \theta$, $\forall\theta \in R^1$ и $F_\theta \in \Im_{S|\theta}$. Из (9.41) следует выражение для функции влияния в виде

$$IF(x;F,HL_\alpha)=\frac{1}{2\int_\alpha^{1-\alpha} f(F^{-1}(t))dt}\begin{cases}(1-2\alpha)\operatorname{sign}(x), |x|>F^{-1}(1-\alpha),\\ (2F(x)-1), \quad |x|\le F^{-1}(1-\alpha),\end{cases}$$

$$x\in R^1. \qquad \textbf{(9.50)}$$

Согласно (9.45), асимптотическая дисперсия $\sqrt{n}HL_\alpha$-оценки для $F\in\Im_{S|0}$ вычисляется по формуле

$$\sigma_F^2(HL_\alpha)=\frac{(1+4\alpha)(1-2\alpha)^2}{12\left(\int_\alpha^{1-\alpha} f(F^{-1}(t))dt\right)^2}. \qquad \textbf{(9.51)}$$

Отметим, что из (9.51) при $\alpha\to 0$ следует формула для асимптотической дисперсии $\sqrt{n}HL$-оценки, а при $\alpha\to 1/2$ получаем асимптотическую дисперсию выборочной медианы $\sqrt{n}\overline{X}_{1/2}$. В самом деле,

$$\lim_{\alpha\to 0}\sigma_F^2(HL_\alpha)=\lim_{\alpha\to 0}\frac{(1+4\alpha)(1-2\alpha)^2}{12(\int\limits_\alpha^{1-\alpha} f(F^{-1}(t))dt)^2}=$$

$$=\frac{1}{12(\int\limits_0^1 f(F^{-1}(t))dt)^2}=\sigma_F^2(HL).$$

Далее, используя формулу дифференцирования интеграла по параметру и применяя правило Лопиталя, с учетом того, что для $F\in\Im_S$ выполняется равенство $f(F^{-1}(1-\alpha))=f(F^{-1}(\alpha))$, получаем

$$\lim_{\alpha\to 1/2}\sigma_F^2(HL_\alpha)=\lim_{\alpha\to 1/2}\frac{(1+4\alpha)(1-2\alpha)^2}{12(\int\limits_\alpha^{1-\alpha} f(F^{-1}(t))dt)^2}=$$

$$=\lim_{\alpha\to 1/2}\left(\frac{1+4\alpha}{12}\right)\left\{\lim_{\alpha\to 1/2}\frac{1-2\alpha}{\int\limits_\alpha^{1-\alpha} f(F^{-1}(t))dt}\right\}^2=$$

$$= \frac{1}{4}\left\{ \lim_{\alpha\to 1/2} \frac{-2}{-f(F^{-1}(1-\alpha)) - f(F^{-1}(\alpha))} \right\}^2 = \frac{1}{4f^2(0)} = \sigma_F^2(\bar{X}_{1/2}) \; .$$

Приведем результаты вычислений асимптотических дисперсий $\sqrt{n}HL_\alpha$-оценок по формуле (9.51) для супермоделей, описывающих различные отклонения от нормального распределения.

1) Рассмотрим гауссовскую модель с масштабным засорением, которая определяется в виде

$$\Im_{\varepsilon,\tau}(\Phi) = \{F : F_{\varepsilon,\tau}(x) = (1-\varepsilon)\Phi(x) + \varepsilon\Phi(x/\tau)\},\; 0 \le \varepsilon \le 1,\; \tau \ge 1.$$

Можно убедиться (см., например, [9]), что асимптотическая дисперсия $\sqrt{n}HL_\alpha$-оценки для $F(x) \in \Im_{\varepsilon,\tau}(\Phi)$ вычисляется по формуле

$$\sigma^2(\Phi_{\varepsilon,\tau}, HL_\alpha) =$$

$$= \frac{\pi(1+4\alpha)(1-2\alpha)^2}{3\left\{(1-\varepsilon)^2 A + [2\sqrt{2}\,\varepsilon(1-\varepsilon)B/\sqrt{1+\tau^2}\,] + (\varepsilon^2/\tau)C\right\}^2}, \qquad \textbf{(9.52)}$$

где $A = 2\Phi(x^*\sqrt{2}) - 1$, $B = 2\Phi(x^*\sqrt{1+\tau^2}/\tau) - 1$, $C = 2\Phi(x^*\sqrt{2}/\tau) - 1$, $x^* = \Phi_{\varepsilon,\tau}^{-1}(1-\alpha)$ – квантиль уровня $(1-\alpha)$ для ф.р. $\Phi_{\varepsilon,\tau}(x) = (1-\varepsilon)\Phi(x) + \varepsilon\Phi(x/\tau)$. В частности, при $\alpha = 0$ получаем выражение для асимптотической дисперсии оценки Ходжеса – Лемана в виде

$$\sigma^2(\Phi_{\varepsilon,\tau}, HL) = \frac{\pi}{3\left\{(1-\varepsilon)^2 + [2\sqrt{2}\,\varepsilon(1-\varepsilon)/\sqrt{1+\tau^2}\,] + (\varepsilon^2/\tau)\right\}^2}. \qquad \textbf{(9.53)}$$

При $\alpha \to 1/2$, асимптотическая дисперсия α-урезанной оценки Ходжеса – Лемана совпадает с асимптотической дисперсией выборочной медианы и вычисляется как

$$\sigma^2(\Phi_{\varepsilon,\tau}, HL_{1/2}) = \sigma^2(F_{\varepsilon,\tau}, \bar{X}_{1/2}) = \pi/2(1-\varepsilon+\varepsilon/\tau)^2. \qquad \textbf{(9.54)}$$

Численные расчеты по формулам (9.52) – (9.54) приведены в табл. 9.1.

Таблица 9.1

Асимптотическая дисперсия $\sqrt{n}HL_\alpha$-оценок для $F(x)\in\Im_{\varepsilon,\tau}(\Phi)$ при $\tau=3$

$\varepsilon \setminus \alpha$	0,00	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,4	0,45	0,50
0.00	1,047	1,06	1,085	1,117	1,156	1,203	1,256	1,318	1,390	1,474	1,571
0.01	1,071	1,081	1,104	1,136	1,175	1,221	1,274	1,337	1,409	1,494	1,592
0.05	1,171	1,172	1,189	1,216	1,252	1,297	1,351	1,415	1,490	1,578	1,681
0.10	1,311	1,302	1,308	1,328	1,360	1,403	1,457	1,522	1,600	1,693	1,803
0.20	1,651	1,628	1,605	1,604	1,622	1,658	1,709	1,776	1,861	1,965	2,091
0.30	2,090	2,062	2,004	1,971	1,966	1,987	2,032	2,100	2,191	2,308	2,454
0.40	2,655	2,627	2,543	2,465	2,425	2,423	2,455	2,519	2,616	2,749	2,921
0.50	3,376	3,351	3,257	3,136	3,049	3,010	3,020	3,077	3,179	3,329	3,534

2) Рассмотрим супермодель в виде конечного набора стандартных распределений, упорядоченных по степени затянутости их хвостов (см. раздел 17). Пусть $F\in\Im_S^*=\{F_{(1)},F_{(2)},F_{(3)},F_{(4)}\}$, где $F_{(1)}=\Phi$ – стандартное нормальное распределение с плотностью ϕ, $F_{(2)}$ – логистическая ф.р., $F_{(3)}$ – распределение Лапласа, $F_{(4)}$ – распределение Коши.

А) *Нормальное распределение*. Вычислим интеграл

$$\int\limits_{\alpha}^{1-\alpha} f(F^{-1}(t))dt=\int\limits_{\Phi^{-1}(\alpha)}^{\Phi^{-1}(1-\alpha)}\phi^2(x)dx=$$

$$=\frac{1}{2\pi}\int\limits_{\Phi^{-1}(\alpha)}^{\Phi^{-1}(1-\alpha)}\exp^{-x^2}dx=\|x=y/\sqrt{2}\|=\frac{1}{2\sqrt{\pi}}\cdot\frac{1}{\sqrt{2\pi}}\int\limits_{\sqrt{2}\Phi^{-1}(\alpha)}^{\sqrt{2}\Phi^{-1}(1-\alpha)}\exp^{-y^2/2}dy=$$

$$=\frac{1}{2\sqrt{\pi}}\left[\Phi(\sqrt{2}\Phi^{-1}(1-\alpha))-\Phi(\sqrt{2}\Phi^{-1}(\alpha))\right]=\frac{1}{2\sqrt{\pi}}\left[2\Phi(\sqrt{2}\Phi^{-1}(1-\alpha))-1\right].$$

С учетом данного выражения, из формулы (9.51) получаем

$$\sigma_\Phi^2(HL_\alpha)=\frac{\pi(1+4\alpha)(1-2\alpha)^2}{3(2\Phi(\sqrt{2}\Phi^{-1}(1-\alpha))-1)^2}. \qquad \textbf{(9.55)}$$

Б) *Логистическое распределение*. Учитывая, что в данном случае $f(F^{-1}(t))=t(1-t)$, $0\le t\le 1$, получаем

$$\int_{\alpha}^{1-\alpha} f(F^{-1}(t))dt = \int_{\alpha}^{1-\alpha} t(1-t)dt = (1-6\alpha^2+4\alpha^3)/6.$$

С учетом данного выражения, из (9.51) получаем

$$\sigma^2_{F_{)2)}}(HL_\alpha) = \frac{3(1+4\alpha)(1-2\alpha)^2}{(1-6\alpha^2+4\alpha^3)^2}. \qquad \mathbf{(9.56)}$$

В) *Распределение Лапласа*. Учитывая, что в данном случае

$$f(F^{-1}(t)) = \begin{cases} t, & 0 \le t < 1/2, \\ 1-t, & 1/2 \le t \le 1, \end{cases}$$

получаем $\int_{\alpha}^{1-\alpha} f(F^{-1}(t))dt = \int_{\alpha}^{1/2} tdt + \int_{1/2}^{1-\alpha} (1-t)dt = \frac{1-4\alpha^2}{4}.$

С учетом данного выражения, из (9.51) получаем

$$\sigma^2_{F_{(3)}}(HL_\alpha) = \frac{4(1+4\alpha)(1-2\alpha)^2}{3(1-4\alpha^2)^2} = \frac{4(1+4\alpha)}{3(1+2\alpha)^2}. \qquad \mathbf{(9.57)}$$

Г) *Распределение Коши*. Учитывая, что в данном случае

$$f(F^{-1}(t)) = 1/\pi[1+tg^2(\pi(1-(1/2)))],$$

получаем

$$\int_{\alpha}^{1-\alpha} f(F^{-1}(t))dt = \frac{1}{\pi}\int_{\alpha}^{1-\alpha} \frac{dt}{1+tg^2(\pi(t-1/2))} = \|u = \pi(t-1/2)\| =$$

$$= \frac{1}{\pi^2}\int_{-\pi(1-2\alpha)/2}^{\pi(1-2\alpha)/2} \frac{du}{1+tg^2 u} = \frac{1}{\pi^2}\int_{-\pi(1-2\alpha)/2}^{\pi(1-2\alpha)/2} \cos^2(u)du =$$

$$= \frac{1}{2\pi^2}\left[\pi(1-2\alpha)+\sin(\pi(1-2\alpha))\right] = \frac{1}{2\pi}\left[(1-2\alpha)+\sin(\pi(1-2\alpha))/\pi\right].$$

С учетом данного выражения, из (9.51) получаем

$$\sigma^2_{F_{(4)}}(HL_\alpha) = \frac{\pi^2(1+4\alpha)(1-2\alpha)^2}{3\left[(1-2\alpha)+\sin(\pi(1-2\alpha))/\pi\right]^2}. \qquad \mathbf{(9.58)}$$

Численные расчеты по формулам (9.55) – (9.58) приведены в табл. 9.2.

Таблица 9.2

Асимптотическая дисперсия $\sqrt{n}HL_\alpha$-оценок для $\mathfrak{I}_S^* = \{F_{(1)}, F_{(2)}, F_{(3)}, F_{(4)}\}$

Ф.р. \ α	0,00	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
F1-Гаусса	1,047	1,06	1,085	1,117	1,156	1,203	1,256	1,318	1,39	1,474	1,571
F2-логист.	3,000	3,002	3,016	3,048	3,099	3,174	3,273	3,401	3,561	3,758	4,000
F3-Лапласа	1,333	1,322	1,296	1,262	1,224	1,185	1,146	1,107	1,07	1,034	1,000
F4-Коши	3,290	3,208	3,025	2,813	2,616	2,456	2,345	2,286	2,283	2,341	2,467

Таблица 9.3

Абсолютная эффективность $\sqrt{n}HL_\alpha$-оценок для $\mathfrak{I}_S^* = \{F_{(1)}, F_{(2)}, F_{(3)}, F_{(4)}\}$

Ф.р. \ α	0,00	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
F1-Гаусса	0,955	0,943	0,922	0,895	0,865	0,831	0,796	0,759	0,719	0,678	0,637
F2-логист.	1,000	0,999	0,995	0,984	0,968	0,945	0,917	0,882	0,842	0,798	0,750
F3-Лапласа	0,750	0,756	0,772	0,792	0,817	0,844	0,873	0,903	0,935	0,967	1,000
F4-Коши	0,608	0,623	0,661	0,711	0,765	0,814	0,853	0,875	0,876	0,854	0,811

Таблица 9.4

Асимптотические дисперсии и абсолютные эффективности оценок для $F \in \mathfrak{I}_S^*$

Характеристики оценок	Нормальное $F_{(1)}$, $I(F_{(1)}) = 1$	Логистическое $F_{(2)}$, $I(F_{(2)}) = 1/3$	Лапласа $F_{(3)}$, $I(F_{(3)}) = 1$	Коши $F_{(4)}$, $I(F_{(4)}) = 1/2$
$\sigma^2(F, \bar{X})$ $AЭ(F, \bar{X})$	1,00 1,00	$\pi^2/3 = 3,29$ 0,91	2,00 0,50	∞ 0
$\sigma^2(F, \bar{X}_{1/2})$ $AЭ(F, \bar{X}_{1/2})$	$\pi/2 = 1,571$ 0,637	4,00 0,75	1,00 1,00	$\pi^2/4 = 2,467$ $8/\pi^2 = 0,811$
$\sigma^2(F, HL)$ $AЭ(F, HL)$	$\pi/3 = 1,047$ 0,955	3,00 1,00	$4/3 = 1,333$ 0,75	$\pi^2/3 = 3,290$ 0,608

Для удобства сравнения характеристик оценок в табл. 9.4 приведены асимптотические дисперсии $D_F(\hat{\theta}) = \sigma^2(F, \hat{\theta})/n$ и абсолютные эффективности $AЭ_F(\hat{\theta}) = \{I(F) \cdot \sigma^2(F, \hat{\theta})\}^{-1}$ выборочного среднего $\bar{X}$, выборочной медианы $\bar{X}_{1/2}$ и оценки Ходжеса – Лемана HL для $F \in \Im_S^*$.

3) Рассмотрим в качестве супермодели семейство распределений Стьюдента $\Im_r \in \Im_S$ с r степенями свободы, которое характеризуется плотностью

$$f_r(x) = A(r)(1 + (x^2/r))^{-(r+1)/2},\ x \in R^1,$$

и функцией распределения

$$F_r(x) = A(r) \int\limits_{-\infty}^{x} (1 + (y^2/r))^{-(r+1)/2} dy,$$

где $A(r) = \Gamma((r+1)/2) / \sqrt{r\pi}\, \Gamma(r/2)$, $\Gamma(\cdot)$ – гамма-функция. Можно убедиться (см., например, [9]), что для $F_r(x) \in \Im_r$ асимптотическая дисперсия $\sqrt{n} HL_\alpha$-оценок вычисляется по формуле

$$\sigma^2_{F_r}(HL_\alpha) = \frac{1}{12}\left(\frac{2r+1}{r}\right)\frac{A^2(2r+1)}{A^4(r)} \times$$

$$\times \frac{(1+4\alpha)(1-2\alpha)^2}{\{2F_{2r+1}[\sqrt{(2r+1)/r}\, F_r^{-1}(1-\alpha)] - 1\}^2} = \frac{r\pi}{12}\left[\frac{\Gamma(r+1)}{\Gamma(r+(1/2))}\right]^2 \times$$

$$\times \left[\frac{\Gamma(r/2)}{\Gamma((r+1)/2)}\right]^4 \frac{(1+4\alpha)(1-2\alpha)^2}{\{2F_{2r+1}[\sqrt{(2r+1)/r}\, F_r^{-1}(1-\alpha)] - 1\}^2}. \qquad \textbf{(9.59)}$$

В частности, из формулы (9.59) при $\alpha = 0$ получаем формулу для вычисления асимптотической дисперсии оценки Ходжеса – Лемана в виде

$$\sigma^2_{F_r}(HL)/n = \frac{1}{12n}\left(\frac{2r+1}{r}\right) \cdot \frac{A^2(2r+1)}{A^4(r)} =$$

$$= \frac{r\pi}{12n} \cdot \left[\frac{\Gamma(r+1)}{\Gamma(r+(1/2))}\right]^2 \left[\frac{\Gamma(r/2)}{\Gamma((r+1)/2)}\right]^4. \qquad \textbf{(9.60)}$$

Далее, из (9.59) при $\alpha \to 1/2$ следует формула для вычисления асимптотической дисперсии выборочной медианы $\bar{X}_{1/2}$ в виде

$$\sigma^2_{F_r}(\bar{X}_{1/2})/n = \frac{1}{4n f^2(0)} = \frac{\pi r}{4n}\left[\frac{\Gamma(r/2)}{\Gamma((r+1)/2)}\right]^2. \qquad \textbf{(9.61)}$$

Отметим, что для выборочного среднего $\bar{X}$ асимптотическая дисперсия $\sqrt{n}\bar{X}$-оценки для $F_r(x) \in \Im_r$ вычисляется по формуле

$$\sigma^2_{F_r}(\bar{X}) = A(r)\int_{-\infty}^{\infty} x^2(1+\ x^2/r)^{-\frac{r+1}{2}}\,dx = r/(r-2),\ r > 2. \qquad \textbf{(9.62)}$$

Численные значения асимптотических значений дисперсий оценок, рассчитанные по приведенным формулам (9.60) – (9.62) для $F_r(x) \in \Im_r$ при различных степенях свободы r, приведены в табл. 9.5. Абсолютная эффективность оценок $\hat{\theta}$, вычисляемая по формуле $AЭ(F_r,\hat{\theta}) = \{I(f_r)\cdot\sigma^2(F_r,\hat{\theta})\}^{-1}$, где $I(f_r)$ – информация Фишера, приведена в круглых скобках табл. 9.5. Из табл. 9.5 следует, что оценка Ходжеса – Лемана обеспечивает высокую абсолютную эффективность в рамках всего семейства распределений Стьюдента для числа степеней свободы $r \ge 3$.

Таблица 9.5

Асимптотические дисперсии и абсолютные эффективности оценок для $F_r(x) \in \Im_r$

Ст. св. r	1	2	3	5	7	9	$r \to \infty$	$Ad(\hat{\theta},\Im_r)$
$\sigma^2(F_r,HL)$	3,29 (0,61)	1,92 (0,87)	1,58 (0,95)	1,34 (0,99)	1,25 (1,00)	1,20 (1,00)	1,047 (0,96)	**0,41**
$\sigma^2(F_r,\bar{X}_{1/2})$	2,47 (0,81)	2 ,00 (0,81)	1,85 (0,80)	1,73 (0,77)	1,67 (0,75)	1,66 (0,72)	1,571 (0,64)	0,66
$\sigma^2_{F_r}(\bar{X})$	∞ (0,00)	∞ (0,00)	3,00 (0,50)	1,67 (0,80)	1,40 (0,89)	1,28 (0,94)	1, 000 (1,00)	1,52

Абсолютные дефекты оценок $Ad(\hat{\theta},\Im_r)$ в семействе распределений Стьюдента, с приведенными значениями числа степеней свободы r, вычислены по формуле

$$Ad(\hat{\theta}, \Im_r) = \left(\sum_r [1 - AЭ(F_r, \hat{\theta})]^2 \right)^{1/2}.$$

Согласно этому критерию, среди сравниваемых оценок HL, $\bar{X}_{1/2}$ и $\bar{X}$ предпочтение следует отдать оценке Ходжеса – Лемана, для которой значение абсолютного дефекта в рамках рассматриваемой супермодели $\Im_r$ равно $Ad(HL, \Im_r) = 0,41$ и является минимальным в правом столбце табл. 9.5.

Асимптотическая относительная эффективность оценки HL_α относительно оценки Ходжеса – Лемана HL, вычисляемая по формуле

$$AOЭ_F(HL_\alpha : HL) = \sigma_F^2(HL) / \sigma_F^2(HL_\alpha),$$

для $F_r(x) \in \Im_r$, запишется в виде

$$AOЭ_{F_r}(HL_\alpha : HL) = \frac{\{2F_{2r+1}[\sqrt{(2r+1)/r}\, F_r^{-1}(1-\alpha)] - 1\}^2}{(1+4\alpha)(1-2\alpha)^2}, \qquad \textbf{(9.63)}$$

где $F_r(x)$ – табулированная функция распределения Стьюдента с r степенями свободы и F_r^{-1} также табулированная квантильная функция. Из (9.63) при $r \to \infty$ получаем асимптотическую относительную эффективность $AOЭ_F(HL_\alpha : HL)$ для нормального распределения, то есть для $F = \Phi$, в виде

$$AOЭ_\Phi(HL_\alpha : HL) = \frac{\{2\Phi[\sqrt{2}\, \Phi^{-1}(1-\alpha)] - 1\}^2}{(1+4\alpha)(1-2\alpha)^2}. \qquad \textbf{(9.64)}$$

Из (9.63) при $r = 1$ получаем асимптотическую относительную эффективность $AOЭ_F(HL_\alpha : HL)$ для распределения Коши, то есть для $F = F_1$, в виде

$$AOЭ_{F_1}(HL_\alpha : HL) = \frac{\{2F_3[\sqrt{3}\, F_1^{-1}(1-\alpha)] - 1\}^2}{(1+4\alpha)(1-2\alpha)^2} =$$

$$= \frac{\{1 - 2\alpha + \pi^{-1}\sin(\pi(1-2\alpha))\}^2}{(1+4\alpha)(1-2\alpha)^2}. \qquad \textbf{(9.65)}$$

Пример 9.66. Пусть функция меток $\varphi(t) = \Phi^{-1}(t)$, $0 \le t \le 1$. Для такой функции меток R_α-оценка параметра сдвига θ в одновыборочном случае, предложенная в [4] и названная α-урезанной оценкой с нормальными метками, обозначается через NS_α и определяется как решение уравнения $S(T_{n,\alpha}) \approx 0$, где ранговая статистика $S(T_{n,\alpha})$ записывается как

$$S(r) = \sum_{i=[\alpha n]+1}^{n-[\alpha n]} \Phi^{-1}\left(\frac{r_\alpha(i)}{2(n-2k)+1}\right),\ k = [\alpha n],\ 0 \le \alpha < 1/2. \qquad \textbf{(9.67)}$$

Здесь $r_\alpha(i)$ обозначает ранг $2r - X_{(i)}$ в смешанной выборке объема $2(n-2k)$, состоящей из $X_{(k+1)},\dots,X_{(n-k)}$ и $2r - X_{(k+1)},\dots,2r - X_{(n-k)}$, $k = [\alpha n]$. Для симметричных распределений, то есть для $F \in \Im_{S|0}$ и функций меток $\varphi(t) \in C_R^*$, из выражения (9.42) получаем

$$IF(x; F, NS_\alpha) = \frac{1}{\int_\alpha^{1-\alpha} [f(F^{-1}(t)) / \phi(\Phi^{-1}(t))]dt} \times$$

$$\times \begin{cases} \Phi^{-1}(1-\alpha)\operatorname{sign}(x), |x| > F^{-1}(1-\alpha), \\ \Phi^{-1}(F(x)), \qquad |x| \le F^{-1}(1-\alpha), \end{cases} \quad x \in R^1, \qquad \textbf{(9.68)}$$

где ϕ обозначает стандартную нормальную плотность. Отметим, что при нормальном распределении функция влияния NS_α-оценки совпадает с функцией влияния α-урезанного среднего. Кроме того, в отличие от R-оценки с нормальными метками, NS_α-оценка имеет ограниченную функцию влияния для $0 < \alpha < 1/2$, и при $\alpha \to 1/2$ её функция влияния совпадает с функцией влияния выборочной медианы. Далее, согласно (9.45), асимптотическая дисперсия $\sqrt{n}NS_\alpha$-оценки для $F \in \Im_{S|0}$ вычисляется по формуле

$$\sigma_F^2(NS_\alpha) = \frac{\{1 - 2\alpha + 2\Phi^{-1}(\alpha)\phi(\Phi^{-1}(\alpha)) + 2\alpha[\Phi^{-1}(\alpha)]^2\}}{\left(\int_\alpha^{1-\alpha} [f(F^{-1}(t)) / \phi(\Phi^{-1}(t))]dt\right)^2}. \qquad \textbf{(9.69)}$$

9.В. Предел устойчивости R_α-оценок

Отметим, что R_α-оценки обладают качественной робастностью при условии, что функционал $T_\alpha(F)$ однозначно определяется выражением (9.35) при заданной идеальной модели F_0. При выполнении этого условия функционал $T_\alpha(F)$ слабо непрерывен в F_0 (см. теорему 4.1 в [5]). Получим формулу для вычисления пороговой точки R_α-оценок. Пусть, как обычно, F_0 обозначает идеальную модель и рассмотрим супермодель $\Im_{\mathrm{L},\varepsilon}(F_0)$, заданную с помощью метрики Леви $d_{\mathrm{L}}(F_0,F)$, в виде

$$\Im_{\mathrm{L},\varepsilon}(F_0)=\{F : d_{\mathrm{L}}(F_0,F)\le\varepsilon\},\ \varepsilon>0 .$$

Положим для простоты $T_\alpha(F)=0$ и определим максимальное смещение R_α-оценки при отклонении F от F_0 в окрестности $\Im_{\mathrm{L},\varepsilon}(F_0)$ в виде

$$b(\varepsilon)=\sup\{T_\alpha(F) : d_{\mathrm{L}}(F_0,F)\le\varepsilon\}.$$

Согласно определению 4.17, пороговая точка $\varepsilon^*(\varphi,\alpha)$ для R_α-оценки определяется выражением

$$\varepsilon^*(\varphi,\alpha)=\sup\{\varepsilon : b(\varepsilon)<\infty\}. \qquad \textbf{(9.70)}$$

Следуя [5], введем несобственное распределение

$$F_\varepsilon(x)=\begin{cases}F_0(x-\varepsilon)-\varepsilon, & x>x_0+\varepsilon,\\ 0, & x\le x_0+\varepsilon,\end{cases}\ \text{где } x_0=F_0^{-1}(\varepsilon). \qquad \textbf{(9.71)}$$

Отметим, что несобственное распределение $F_\varepsilon(x)$ является стохастически наибольшим в супермодели $\Im_{\mathrm{L},\varepsilon}(F_0)$, то есть выполняется выражение $d_{\mathrm{L}}(F_0,F_\varepsilon)=\varepsilon$ и $F_0 <_{St} F_\varepsilon$. Напомним, что символ $(<_{St})$ обозначает стохастическую упорядоченность распределений (см. определение 17.4). Введем обозначение

$$\mu_F(t;\varphi,\alpha)=\int\limits_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)}\varphi\{[F(x)+1-F(2t-x)]/2\}dF(x),\ 0\le\alpha<1/2 .$$

Учитывая, что функция φ, определяющая R_α-оценку, монотонная, и тот факт, что $F_0 <_{St} F_\varepsilon$, получим неравенство

$$\mu_F(t;\varphi,\alpha) \le \mu_{F_\varepsilon}(t;\varphi,\alpha),\ \forall F \in \Im_{L,\varepsilon}(F_0).$$

С учетом данного неравенства, пороговая точка $\varepsilon^*(\varphi,\alpha)$ является решением уравнения

$$\lim_{t\to\infty} \mu_{F_\varepsilon}(t;\varphi,\alpha) = 0,\ 0 \le \alpha < 1/2. \tag{9.72}$$

Напомним, что $\varphi(t) \in C_R^*$ и, следовательно, $\varphi(t) = -\varphi(1-t)$. С учетом данного равенства, из выражения (9.72) получаем (см. [4]) формулы для определения пороговой точки $\varepsilon^*(\varphi,\alpha)$ для R_α-оценки в виде

$$\varepsilon^*(\varphi,\alpha) = \frac{1}{\varphi(1-\alpha)} \int_{1/2}^{1-\alpha} \varphi(t)dt + \alpha,\ \alpha_0 \le \alpha < 1/2, \tag{9.73}$$

где α_0 является решением относительно α уравнения вида

$$\int_{1/2}^{1-\alpha} \varphi(t)dt + \alpha\varphi(\alpha) = 0. \tag{9.74}$$

Для неравенств $0 \le \alpha < \alpha_0$ пороговая точка $\varepsilon^*(\varphi,\alpha)$ является решением относительно ε уравнения

$$\int_{1/2}^{1-(\varepsilon/2)} \varphi(t)dt - \int_{1-(\varepsilon/2)}^{1-\alpha} \varphi(t)dt + \alpha\varphi(\alpha) = 0,\ 0 \le \alpha < \alpha_0. \tag{9.75}$$

Отметим, что при $\alpha = 0$ из (9.75) следует известная формула для вычисления пороговой точки $R(\varphi)$-оценки (см., например, [5]) в виде

$$\int_{1/2}^{1-(\varepsilon/2)} \varphi(t)dt = \int_{1-(\varepsilon/2)}^{1} \varphi(t)dt. \tag{9.76}$$

Пример 9.77. Пусть функция меток $\varphi(t)=\Phi^{-1}(t)$, $0\le t\le 1$. Для такой функции меток R_α-оценка параметра сдвига θ в одновыборочном случае называется α-урезанной оценкой с нормальными метками и обозначается через NS_α. Используя (9.72) – (9.75), получаем формулы вычисления пороговой точки для NS_α-оценки в виде

$$\varepsilon^*(NS_\alpha)=2\Phi\{-\sqrt{\ln 4-2\ln[1+\sqrt{2\pi}\,\alpha\Phi^{-1}(\alpha)+\exp(-[\Phi^{-1}(\alpha)]^2/2)]}\;\},$$
$$0\le\alpha\le\alpha_0,$$

$$\varepsilon^*(NS_\alpha)=\alpha+(1-\exp(-(1/2)[\Phi^{-1}(\alpha)]^2)/\sqrt{2\pi}\,\alpha\Phi^{-1}(1-\alpha),$$
$$\alpha_0<\alpha<1/2\,, \qquad \textbf{(9.78)}$$

где $\alpha_0\approx 0{,}16$ является решением уравнения

$$1+\sqrt{2\pi}\,\alpha\Phi^{-1}(\alpha)-\exp\{-(1/2)[\Phi^{-1}(\alpha)]^2\}=0.$$

Пример 9.79. Пусть функция меток $\varphi(t)=t-(1/2)$, $0\le t\le 1$. Для HL_α-оценок получаем формулы для вычисления пороговой точки в виде

$$\varepsilon^*(HL_\alpha)=1-(1/2)\sqrt{2(1-4\alpha^2)}\ \text{для}\ 0\le\alpha\le(1/6),$$

$$\varepsilon^*(HL_\alpha)=(1+2\alpha)/4\ \text{для}\ (1/6)<\alpha<(1/2). \qquad \textbf{(9.80)}$$

Из приведенных формул (9.78) и (9.80) следуют при $\alpha=0$ известные результаты. Для оценки Ходжеса – Лемана получаем $\varepsilon^*(HL)=1-\ (1/\sqrt{2})\approx 0{,}29$. Для R-оценки с нормальными метками получаем $\varepsilon^*(NS)=2\Phi[-\sqrt{\ln 4}]\approx 0{,}24$. Отметим, что при малых значениях параметра α ($\alpha<0{,}2$) пороговые точки HL_α- и NS_α-оценок значительно выше пороговой точки α-урезанного среднего, для которого $\varepsilon^*(\bar{X}_\alpha)=\alpha$, $0\le\alpha\le 0{,}5$. Асимптотические дисперсии и числовые характеристики робастности HL_α- и NS_α-оценок параметра сдвига θ нормального распределения $\Phi(x-\theta)$ приведены в табл. 9.6.

Таблица 9.6

Асимптотические дисперсии и числовые характеристики робастности оценок для $F=\Phi$

$\hat{\theta}$	α	$\sigma^2(\Phi,\hat{\theta})$	$\gamma^*(\Phi,\hat{\theta})$	$\lambda^*(\Phi,\hat{\theta})$	$\varepsilon^*(\hat{\theta})$
HL	0,0	1,047	1,77	1,41	0,29
HL_α	0,1	1,085	1,52	1,52	0,31
	0,2	1,156	1,40	1,86	0,35
	0,3	1,256	1,31	2,62	0,40
	0,4	1,390	1,27	5,08	0,45
$\bar{X}_{1/2}$	0,5	1,571	1,25	∞	0,50
$\bar{X}$	0,0	1,000	∞	1,00	0.24
NS_α	0,1	1,061	1,60	1,25	0,28
	0,2	1,145	1,40	1,67	0,34
	0,3	1,252	1,31	2,50	0,39
	0,4	1.390	1,27	5,00	0,45

Пример 9.81. Рассмотрим два варианта обобщенных оценок Ходжеса – Лемана. Обобщенная оценка Ходжеса – Лемана $HL(m)$ записывается в виде

$$HL(m) = med\{(X_{i_1} + \ldots + X_{i_m})/m,\ \ 1 \le i_1 < \ldots < i_m \le n\},\ 2 \le m < n.$$

Другой вариант, названный обобщенной α-урезанной оценкой Ходжеса – Лемана $HL_\alpha(m)$ (см. [4, 9, 62]), записывается в виде

$$HL_\alpha(m) = med\{(X_{(i_1)} + \ldots + X_{(i_m)})/m,\ \ [\alpha n] \le i_1 < \ldots < i_m \le n - [\alpha n]\},$$
$$0 \le \alpha < 1/2.$$

Можно убедиться (см. [75]), что пороговая точка $\varepsilon^*(HL(m))$ для $HL(m)$-оценок вычисляется по формуле

$$\varepsilon^*(HL(m)) = 1 - 2^{-1/m},\ 2 \le m < n. \quad \textbf{(9.82)}$$

Толерантность $\varepsilon_n^*(HL_\alpha(m))$ к наличию выбросов в выборке для $HL_\alpha(m)$-оценок определяется как наибольшее целое число j, при котором выполняется неравенство

$$2\binom{n-2[\alpha n]-j}{m} - 1 \ge \binom{n-2[\alpha n]}{m},\ j > [\alpha n], 0 \le \alpha < 1/2. \quad \textbf{(9.83)}$$

В табл. 9.7 приведены значения отношений толерантности $\varepsilon_n^*(HL_\alpha(m))$ к объему выборки n, вычисленные по формуле (9.83) при $\alpha = 0$. Для сравнения, в правом столбце табл. 9.7 приведены значения пороговых точек $\varepsilon^*(HL(m)) = \lim_{n\to\infty}(\varepsilon_n^*/n)$, вычисленные по формуле (9.82). Из приведенной таблицы видно, что сходимость отношения (ε_n^*/n) к пороговой точке $\varepsilon^*(HL(m))$ «очень быстрая».

В табл. 9.8 приведены значения отношений толерантности $\varepsilon_n^*(HL_\alpha(m))$ к объему выборки n, вычисленные по формуле (9.83) при $\alpha = 0{,}3$.

Таблица 9.7

Отношения ε_n^*/n и пороговые точки $\varepsilon^*(HL(m))$ для $HL(m)$-оценок

m	$n=5$	$n=10$	$n=20$	$n=40$	$n=50$	$\varepsilon^*(HL(m))$
2	0,20	0,20	0,25	0,27	0,28	0,29
3	0,00	0,10	0,15	0,20	0,20	0,21
4	0,00	0,00	0,10	0,15	0,15	0,16
5	0,00	0,00	0,10	0,10	0,12	0,13
10	0,00	0,00	0,05	0,05	0,06	0,07

Таблица 9.8

Отношения ε_n^*/n для $HL_\alpha(m)$-оценок при $\alpha = 0{,}3$

m	$n=5$	$n=10$	$n=20$	$n=40$
2	0,20	0,30	0,40	0,40
3	0,20	0,30	0,35	0,37
4	0,00	0,30	0,30	0,35

Отметим, что при увеличении параметра m пороговая точка $\varepsilon^*(HL_\alpha(m))$ для $HL_\alpha(m)$-оценок убывает, однако, увеличивая параметр α, $0 \le \alpha < 1/2$, её можно приблизить к максимально возможному значению пороговой точки выборочной медианы, равному $\varepsilon^*(\bar{X}_{1/2}) = 1/2$.

10. СВЯЗИ МЕЖДУ *M*-, *L*- И *R*-ОЦЕНКАМИ ПАРАМЕТРА ПОЛОЖЕНИЯ

Пусть имеется выборка в виде последовательности $X_1,...,X_n$ н.о.р. случайных величин с ф.р. $F(x-\theta)$. Предполагаем, что ф.р. F непрерывна и симметрична, то есть $F \in \Im_{S|0}$. При этих предположениях параметр сдвига θ называют параметром положения наблюдаемой с.в. *X*. Следуя работе Джекеля [25], обсудим связи между классами *M*, *L* и *R*-оценками параметра положения θ. Используя общие асимптотические свойства этих оценок, приведем условия, при которых их асимптотические дисперсии совпадают. Введем необходимые обозначения и определим свойства классов функций C_M, C_L и C_R, которые содержат функции ψ, J и φ, соответственно определяющие *M*-, *L*- и *R*-оценки.

M-оценка $T_n = T_n(X_1,...,X_n)$ параметра положения θ определяется как решение уравнения

$$\sum_{i=1}^{n} \psi(X_i - T_n) = 0, \tag{10.1}$$

где заданная функция $\psi \in C_M$ определяет конкретный вид *M*-оценки и класс функций C_M содержит дифференцируемые нечетные возрастающие функции $\psi(x)$, для которых ψ' имеет компактный носитель, ограниченную вариацию и

$$\int \psi'(x) dF(x) \neq 0. \tag{10.2}$$

Для $F \in \Im_{S|0}$ и $\psi \in C_M$ *M*-оценка, являющаяся решением уравнения (10.1), имеет асимптотически нормальное распределение и, согласно (7.7), асимптотическая дисперсия $\sqrt{n}T_n$-оценки, определяемой функцией ψ, вычисляется по формуле

$$\sigma^2(F,\psi)=\frac{\int\psi^2(x)dF(x)}{\left(\int\psi'(x)dF(x)\right)^2}. \qquad \textbf{(10.3)}$$

Отметим, что для функции ψ, согласованной с плотностью f в виде $\psi_f(x)=-f'(x)/f(x)$, M-оценка является асимптотически эффективной оценкой параметра θ, так как для такой функции $\psi_f(x)=-f'(x)/f(x)$ выполняется равенство $\sigma^2(F,\psi_f)=1/I(f)$.

L-оценка $T_n=T_n(X_1,...,X_n)$ параметра положения θ определяется в виде линейных комбинаций порядковых статистик $X_{(1)},...,X_{(n)}$, то есть в виде

$$T_n=\sum_{i=1}^{n}a_{ni}X_{(i)}, \qquad \textbf{(10.4)}$$

где веса a_{ni} определяются функцией $J(t)\in C_L$ и вычисляются по формуле

$$a_{ni}=\int\limits_{(i-1)/n}^{i/n}J(t)dt. \qquad \textbf{(10.5)}$$

Класс функций C_L содержит функции $J(t)$, $0\le t\le 1$, которые непрерывны, имеют ограниченную вариацию, причем

$$\int\limits_0^1 J(t)dt=1,\ J(t)=J(1-t),\ 0\le t\le 1. \qquad \textbf{(10.6)}$$

Для $F\in\Im_{S|0}$ и $J\in C_L$, L-оценка вида (10.4), веса которой определяются функцией $J(t)$ по формуле (10.5), согласно теореме (8.30), имеет асимптотически нормальное распределение. Отметим, что для $F\in\Im_{S|0}$ и $J\in C_M$, выполняются равенства $U(t)=-U(1-t)$, $\int_0^1 U(t)dt=0$ и асимптотическая дисперсия $\sqrt{n}T_n$-оценки вычисляется по формуле

$$\sigma^2(F,J)=\int_0^1 U^2(t)dt, \tag{10.7}$$

где
$$U(t)=\int_{1/2}^{t}\frac{J(v)dv}{f(F^{-1}(v))}. \tag{10.8}$$

R-оценка $T_n=T_n(X_1,...,X_n)$ параметра положения θ определяется как решение уравнения

$$\sum_{i=1}^{n}\varphi\left(\frac{r_i}{2n+1}\right)=0, \tag{10.9}$$

где r_i – ранг $2T_n-X_i$ в смешанной выборке, состоящей из $X_1,...,X_n$ и $2T_n-X_1,...,2T_n-X_n$.

Предполагаем, что функция меток $\varphi(t)$, $0\le t\le 1$, принадлежит классу функций C_R, который содержит дифференцируемые, возрастающие функции, причем φ' имеет компактный носитель, $\varphi'(f(F^{-1}(t)))$ имеет ограниченную вариацию и

$$\int_0^1\varphi'(t)f(F^{-1}(t))dt\ne 0,\ \varphi(t)=-\varphi(1-t),\ 0\le t\le 1. \tag{10.10}$$

Для $F\in\Im_{S|0}$ и $\varphi\in C_R$, R-оценка, являющаяся решением уравнения (10.9) относительно T_n, согласно теореме 9.21, имеет асимптотически нормальное распределение. Асимптотическая дисперсия $\sqrt{n}T_n$-оценки, определяемой функцией меток $\varphi(t)$, $0\le t\le 1$, принадлежащей классу функций C_R, вычисляется по формуле

$$\sigma^2(F,\varphi)=\frac{\int_0^1\varphi^2(t)dt}{\left(\int_0^1\varphi'(t)f(F^{-1}(t))dt\right)^2}. \tag{10.11}$$

Определим отображения между классами функций C_M, C_L и C_R в виде

$$C_M \to C_L \,:\, J(t) = \frac{\psi'(F^{-1}(t))}{\int \psi'(x)dF(x)},\ 0 \le t \le 1; \tag{10.12}$$

$$C_L \to C_M \,:\, \psi(x) = \int_0^x J(F(y))dy,\ x \in R^1; \tag{10.13}$$

$$C_M \to C_R \,:\, \varphi(t) = \psi(F^{-1}(t)),\ 0 \le t \le 1; \tag{10.14}$$

$$C_R \to C_M \,:\, \psi(x) = \varphi(F(x)),\ x \in R^1; \tag{10.15}$$

$$C_L \to C_R \,:\, \varphi(t) = \int_{1/2}^{t} \frac{J(v)dv}{f(F^{-1}(v))},\ 0 \le t \le 1; \tag{10.16}$$

$$C_R \to C_L \,:\, J(t) = \frac{\varphi(t) f(F^{-1}(t))}{\int_0^1 \varphi'(t) f(F^{-1}(t))dt},\ 0 \le t \le 1. \tag{10.17}$$

Теорема 9.2.18. Отображения (10.12) – (10.17) между классами функций C_M, C_L и C_R обеспечивают для симметричных распределений равенство асимптотических дисперсий *M*-, *L*- и *R*-оценок.

Доказательство этой теоремы осуществляется непосредственной проверкой равенства дисперсий, вычисляемых по формулам (10.3), (10.7) и (10.11) при использовании отображений (10.12) – (10.17).

Пример 9.2.19. Рассмотрим супермодель вида (4.4), для которой в качестве идеального выбрано нормальное распределение. Такую супермодель называют гауссовской моделью с симметричным засорением и определяют в виде

$$\Im_\varepsilon(\Phi) = \{F : F(x) = (1-\varepsilon)\Phi(x) + \varepsilon H(x)\},\ \ 0 \le \varepsilon \le 1, \tag{10.20}$$

где $\Phi(x)$ – стандартная нормальная ф.р. и $H(x)$ – произвольное непрерывное симметричное распределение. Для этой супермодели Хьюбер [15] построил минимаксно-робастную M-оценку параметра положения θ, показав, что в семействе распределений $\Im_\varepsilon(\Phi)$

существует наименее благоприятное распределение F_0 и существует функция $\psi_0 \in C_M$ и они такие, что выполняется выражение

$$\sup_{F \in \Im_\varepsilon(\Phi)} \sigma^2(F, \psi_0) = \sigma^2(F_0, \psi_0) = \inf_{\psi \in C_M} \sigma^2(F_0, \psi). \tag{10.21}$$

Или, другими словами, для любой ф.р. из семейства $\Im_\varepsilon(\Phi)$ и любой функции ψ из класса функций C_M выполняются неравенства

$$\sigma^2(F, \psi_0) \le \sigma^2(F_0, \psi_0),\ \forall F \in \Im_\varepsilon(\Phi); \tag{10.22}$$

$$\sigma^2(F_0, \psi_0) \le \sigma^2(F_0, \psi),\ \forall \psi \in C_M, \tag{10.23}$$

где наименее благоприятное распределение F_0, для которого информация Фишера минимальна в классе распределений $\Im_\varepsilon(\Phi)$, имеет плотность

$$f_0(x) = \frac{1-\varepsilon}{\sqrt{2\pi}} \exp\{-\frac{1}{2}x^2\} \text{ для } |x| \le k$$

и
$$f_0(x) = \frac{1-\varepsilon}{\sqrt{2\pi}} \exp\{-k|x| + \frac{1}{2}k^2\} \text{ для } |x| > k. \tag{10.24}$$

Параметр k определяется при заданном значении ε согласно условию нормировки $\int f_0(x)dx = 1$ из выражения

$$\frac{\varepsilon}{1-\varepsilon} = \frac{2\phi(k)}{k} - 2\Phi(-k). \tag{10.25}$$

Функция ψ_0, определяющая минимаксно-робастную в супермодели $\Im_\varepsilon(\Phi)$ *M*-оценку параметра положения θ, имеет вид

$$\psi_0(x) = -\frac{f_0'(x)}{f_0(x)} = \begin{cases} k\,\mathrm{sign}(x), & |x| > k, \\ x, & |x| \le k \end{cases}$$

или
$$\psi_0(x) = \max\{-k, \min(k, x)\}. \tag{10.26}$$

Убедимся, что для наименее благоприятного распределения F_0 с приведенной плотностью f_0 вида (10.24) и для функции ψ_0 из (10.26) выполняется выражение (10.21). Для этого сначала убедимся в справедливости неравенства (10.23). Напомним, что функция влияния M-оценки, определяемой функцией $\psi \in C_M$, записывается в виде

$$IF(x;F,\psi) = \frac{\psi(x)}{\int \psi'(x)dF(x)}, \quad x \in R^1. \qquad \textbf{(10.27)}$$

Из этого выражения следует, что $\int IF'(x;F_0,\psi)dF_0(x) = 1$. Отсюда, после интегрирования по частям и применения неравенства Коши – Буняковского, получаем

$$1 \le \int IF^2(x;F_0,\psi)dF_0(x) \int \{-f_0'(x)/f_0(x)\}^2 dF_0(x).$$

Отметим, что в приведенном неравенстве достигается равенство, если функция влияния пропорциональна отношению f_0'/f_0, то есть записывается в виде $IF(x;F_0,\psi) = a\, f_0'(x)/f_0(x)$, но такой вид функции влияния соответствует M-оценке, которая определяется функцией $\psi_0(x)$. Таким образом, для приведенных F_0 и ψ_0 неравенство (10.23) выполняется, то есть $\sigma^2(F_0,\psi) \ge 1/I(f_0) = \sigma^2(F_0,\psi_0)$. Убедимся теперь в справедливости неравенства (10.22). В самом деле, наименее благоприятное распределение F_0 принадлежит семейству $\Im_\varepsilon(\Phi)$, поэтому оно может быть записано в виде $F_0(x) = (1-\varepsilon)\Phi(x) + \varepsilon H_0(x)$. С учетом этого замечания, из формулы (10.3) получаем

$$\sigma^2(F_0,\psi_0) = \frac{\int \psi_0^2(x)dF_0(x)}{\left(\int \psi_0'(x)dF_0(x)\right)^2} = \frac{(1-\varepsilon)\int \psi_0^2(x)d\Phi(x) + \varepsilon k^2}{\left((1-\varepsilon)\int \psi_0'(x)d\Phi(x)\right)^2}.$$

Далее,

$$\sigma^2(F,\psi_0) = \frac{(1-\varepsilon)\int \psi_0^2(x)d\Phi(x) + \varepsilon \int \psi_0^2(x)dH(x)}{\left((1-\varepsilon)\int \psi_0'(x)d\Phi(x) + \varepsilon \int \psi_0'(x)dH(x)\right)^2} \le$$

$$\le \frac{(1-\varepsilon)\int \psi_0^2(x)d\Phi(x) + \varepsilon k^2}{\left((1-\varepsilon)\int \psi_0'(x)d\Phi(x)\right)^2} = \sigma^2(F_0,\psi_0)$$

.

Отметим, что последнее неравенство в предыдущем выражении выполняется, так как $|\psi_0(x)| \le k$, $\forall x \in R^1$, и выполняется неравенство $\int \psi_0^2(x) dH(x) \le k^2$. Кроме того, функция ψ_0 неубывающая и, следовательно, $\psi_0'(x) \ge 0$. С учетом этого замечания выполняется неравенство

$$(1-\varepsilon)\int \psi_0'(x)d\Phi(x) + \varepsilon\int \psi_0'(x)dH(x) \ge (1-\varepsilon)\int \psi_0'(x)d\Phi(x).$$

Таким образом, в классе *M*-оценок, являющихся решением уравнения (10.1), для супермодели $\Im_\varepsilon(\Phi) = \{F : F(x) = = (1-\varepsilon)\Phi(x) + \varepsilon H(x)\}$, наименее благоприятное распределение F_0 с плотностью f_0 и функция ψ_0, определенные соответственно в (10.24) и (10.26), приводят к решению минимаксной задачи, то есть для них выполняется выражение (10.21).

Пример 10.28. Рассмотрим супермодель $\Im_\varepsilon(\Phi)$, определенную в (10.20). Используя отображение (10.14), построим минимаксно-робастную R-оценку в $\Im_\varepsilon(\Phi)$. Из (10.24) с учетом (10.25) имеем

$$F_0(-k) = \frac{1-\varepsilon}{\sqrt{2\pi}}\phi(k) = (\varepsilon/2) + (1-\varepsilon)\Phi(-k). \qquad \textbf{(10.29)}$$

Далее, так как $F_0(x) = (\varepsilon/2) + (1-\varepsilon)\Phi(x)$, $-k < x < k$, то обратная функция

$$F_0^{-1}(t) = \Phi^{-1}\left(\frac{t-(\varepsilon/2)}{1-\varepsilon}\right), \ \alpha_0 \le t \le 1-\alpha_0, \qquad \textbf{(10.30)}$$

где параметр α_0 вычисляется по формуле

$$\alpha_0 = F_0(-k) = (\varepsilon/2) + (1-\varepsilon)\Phi(-k).$$

Обозначив $\tilde{\alpha} = \Phi(-k)$, запишем α_0 в виде

$$\alpha_0 = (\varepsilon/2) + (1-\varepsilon)\tilde{\alpha}, \qquad \textbf{(10.31)}$$

где $\tilde{\alpha}$ при заданном значении ε, согласно (10.25), определяется из выражения

$$2\phi(\Phi^{-1}(\tilde{\alpha}))/\Phi^{-1}(1-\tilde{\alpha}) \ -2\tilde{\alpha} = \varepsilon/(1-\varepsilon). \qquad \textbf{(10.32)}$$

Используя (10.14), определим функцию меток $\varphi_0(t)$ для минимаксно-робастной R-оценки в супермодели $\Im_\varepsilon(\Phi)$ в виде $\varphi_0(t) = \psi_0(F_0^{-1}(t))$. Используя (10.26), с учетом (10.24) и (10.30), получаем

$$\varphi_0(t) = -\frac{f_0'(F_0^{-1}(t))}{f_0(F_0^{-1}(t))} = \Phi^{-1}\left(\frac{\alpha_0 - (\varepsilon/2)}{1-\varepsilon}\right),\ 0 \le t < \alpha_0,$$

$$\varphi_0(t) = \Phi^{-1}\left(\frac{t - (\varepsilon/2)}{1-\varepsilon}\right),\ \alpha_0 \le t \le 1-\alpha_0, \tag{10.33}$$

$$\varphi_0(t) = \Phi^{-1}\left(\frac{1-\alpha_0 - (\varepsilon/2)}{1-\varepsilon}\right),\ 1-\alpha_0 < t \le 1.$$

Таким образом, минимаксно-робастной R-оценкой в супермодели $\Im_\varepsilon(\Phi)$ является R_{α_0}-оценка, определяемая парой $\{\varphi_0, \alpha_0\}$, которые вычисляются по формулам в (10.33) и (10.31), и при этом выполняется выражение

$$\sup_{F \in \Im_\varepsilon(\Phi)} \sigma^2(F, \varphi_0) = \sigma^2(F_0, \varphi_0) = \inf_{\varphi \in C_R} \sigma^2(F_0, \varphi). \tag{10.34}$$

Пример 9.2.35. Построим минимаксно-робастную L-оценку в супермодели $\Im_\varepsilon(\Phi)$. Используя отображение (10.12), с учетом того, что $\psi_0'(F_0^{-1}(t)) = 1$ для $t \in [\alpha_0,\ 1-\alpha_0]$, и $\psi_0'(F_0^{-1}(t)) = 0$ для $t \notin [\alpha_0,\ 1-\alpha_0]$, а также, что $\int \psi_0'(F_0^{-1}(t))dt = 1 - 2\alpha_0$, получаем

$$J_0(t) = \begin{cases} 1/(1-2\alpha_0),\ \alpha_0 \le t \le 1-\alpha_0, \\ \quad 0,\ \ \text{в др. случаях}, \end{cases} \tag{10.36}$$

где параметр α_0 вычисляется по формуле (10.31). Таким образом, весовые коэффициенты L-оценки

$$a_{ni} = \int\limits_{(i-1)/n}^{i/n} J_0(t)dt = 1/(n - 2[\alpha_0 n])\ ,\ [\alpha_0 n] + 1 \le i \le n - [\alpha_0 n]. \tag{10.37}$$

Итак, в классе L-оценок минимаксно-робастной в супермодели $\Im_\varepsilon(\Phi)$ является α_0-урезанное среднее вида (8.42), для которого пропорция урезания α_0 исходной выборки вычисляется по формуле (10.31) и выполняется выражение

$$\sup_{F\in\Im_\varepsilon(\Phi)} \sigma^2(F,J_0)=\sigma^2(F_0,J_0)=\inf_{J\in C_L}\sigma^2(F_0,J). \qquad \textbf{(10.38)}$$

Замечание 10.39. Отметим, что приведенные результаты могут быть обобщены и для произвольной супермодели с симметричным засорением, для которой в качестве идеальной модели используется не нормальное распределение $\Phi(x)$, а некоторая заданная абсолютно непрерывная функция распределения $G(x)$ с конечной информацией Фишера $I(g)$. Эта более общая, по сравнению с (10.20), супермодель записывается в виде

$$\Im_\varepsilon(G)=\{F:F(x)=(1-\varepsilon)G(x)+\varepsilon H(x)\},\ 0\le\varepsilon\le 1, \qquad \textbf{(10.40)}$$

где $G(x)$ – заданная непрерывная, симметричная ф.р. с плотностью $g(x)$ и $H(x)$ – произвольное непрерывное симметричное распределение. Для данной супермодели было показано в работе [15], что наименее благоприятное распределение F_0, при котором количество информации Фишера минимально в классе распределений $\Im_\varepsilon(G)$, определяется плотностью f_0 вида

$$f_0(x)=\begin{cases}(1-\varepsilon)g(x_0)e^{b(x-x_0)}, & x<x_0,\\ (1-\varepsilon)g(x), & x_0\le x\le x_1,\\ (1-\varepsilon)g(x_1)e^{-b(x-x_1)}, & x>x_0,\end{cases} \qquad \textbf{(10.41)}$$

где x_0 и x_1 – конечные точки интервалов $|g'(x)/g(x)|\le b$, $b=-g'(x_0)/g(x_0)$, $b=g'(x_1)/g(x_1)$, причем

$$1/(1-\varepsilon)=\int_{x_0}^{x_1} g(x)dx+[g(x_0)+g(x_1)]/b. \qquad \textbf{(10.42)}$$

Используя данный результат, можно убедиться, как и в примере 10.28, что в классе R_α-оценок минимаксно-робастной оценкой в супермодели $\Im_\varepsilon(G)$ вида (10.40) является оценка, определяемая парой $\{\varphi_0,\alpha_0\}$, где $\alpha_0=(\varepsilon/2)+(1-\varepsilon)\tilde{\alpha}$ и параметр $\tilde{\alpha}$ в данном случае определяется из уравнения

$$2g^2(G^{-1}(\tilde{\alpha})) - 2\tilde{\alpha}g'(G^{-1}(\tilde{\alpha})) = g'(G^{-1}(\tilde{\alpha}))\varepsilon/(1-\varepsilon), \qquad \textbf{(10.43)}$$

а функция меток $\varphi_0(t)$, $0 \le t \le 1$, для R_α-оценки определяется выражением

$$\varphi_0(t) = \begin{cases} -b, & 0 \le t \le \alpha_0, \\ \dfrac{-g'\left(G^{-1}\left(\dfrac{t-(\varepsilon/2)}{1-\varepsilon}\right)\right)}{g\left(G^{-1}\left(\dfrac{t-(\varepsilon/2)}{1-\varepsilon}\right)\right)}, & \alpha_0 < t < 1-\alpha_0, \\ b, & 1-\alpha_0 \le t \le 1. \end{cases} \qquad \textbf{(10.44)}$$

Здесь $b = -g'(G^{-1}(\tilde{\alpha}))/g(G^{-1}(\tilde{\alpha}))$.

Отметим, что если в качестве идеальной модели G супермодели $\Im_\varepsilon(G)$ используется логистическое распределение, то минимаксно-робастной оценкой в супермодели $\Im_\varepsilon(G)$ является α_ε-урезанная оценка Ходжеса – Лемана вида (9.48), для которой $\alpha_\varepsilon = \sqrt{\varepsilon(2-\varepsilon)}/2$ (детали см. в работах [4, 9]).

Замечание 10.45. Пусть для заданной функции распределения $F(x) \in \Im_{S|0}$ с плотностью $f(x)$, $x \in R^1$, определена асимптотически эффективная M-оценка параметра положения θ, то есть её асимптотическая дисперсия равна обратной величине информации Фишера $I(f)$. Функцию ψ, определяющую эффективную M-оценку при ф.р. $F(x)$, будем помечать индексом F и записывать в виде

$$\psi_F(x) = -\frac{f'(x)}{f(x)}. \qquad \textbf{(10.46)}$$

Отметим, что асимптотически эффективные L- и R-оценки при ф.р. $F(x)$ определяются соответственно функциями J_F и φ_F вида

$$J_F(t) = \frac{\psi'_F(F^{-1}(t))}{\int \psi'_F(F^{-1}(t))dt} = \frac{\psi'_F(F^{-1}(t))}{I(f)}, \ 0 \le t \le 1, \qquad \textbf{(10.47)}$$

$$\varphi_F(t) = \psi_F(F^{-1}(t)), \ 0 \le t \le 1. \qquad \textbf{(10.48)}$$

Далее, обозначим через $G(x)$ некоторую абсолютно непрерывную симметричную функцию распределения с плотностью $g(x)$ и конечным количеством информации Фишера $I(g) < \infty$. Предположим далее, что асимптотически эффективные *M*-, *L*- и *R*-оценки вычисляются по выборке, порождённой распределением $G(x)$. Интересно выяснить в каком отношении будут находиться дисперсии эффективных оценок при ф.р. $G(x)$. Частичный ответ на этот вопрос дает следующая теорема Шольца [Sholz F.W. A comparison of efficient location estimators. Ann. Statist. 1974. V. 2. No. 6. P. 1323–1326].

Теорема 10.49. Пусть ф.р. $F(x) \in \Im_{S|0}$ и асимптотически эффективные *L*- и *R*-оценки параметра положения θ определены соответственно функциями J_F и φ_F вида (10.47) и (10.48). Тогда для любой абсолютно непрерывной ф.р. $G(x) \in \Im_{S|0}$ с плотностью $g(x)$ и конечным количеством информации Фишера $I(g) < \infty$ выполняется неравенство для асимптотических дисперсий вида $\sigma^2(G, \varphi_F) \le \sigma^2(G, J_F)$.

Доказательство этой теоремы также приводится в работах [4, 9]

Пример 10.50. Рассмотрим вариант, когда ф.р. *F* является нормальной, то есть $F = \Phi$. В этом случае в классе *L*-оценок асимптотически эффективной оценкой параметра положения θ является выборочное среднее $\bar{X}$, а в классе *R*-оценок таковой является *NS*-оценка с функцией меток $\varphi(t) = \Phi^{-1}(t)$, $0 \le t \le 1$. Пусть ф.р. $G(x) \in \Im_{S|0}$ абсолютно непрерывная с плотностью g и $I(g) < \infty$. Тогда выполняется неравенство $\sigma^2(G, NS) \le \sigma^2(G, \bar{X})$, причем равенство достигается только при $G = \Phi$.

11. ОЦЕНКИ ПАРАМЕТРОВ, ПОСТРОЕННЫЕ МЕТОДОМ МИНИМУМА РАССТОЯНИЙ

Метод минимума расстояний был предложен Вольфовитцем в 1957 году (см. работу [50]). Оценки параметров, построенные методом минимума расстояний, в литературе кратко называют *MD*-оценками. Обширная библиография работ по *MD*-оценкам составлена и опубликована Парром [51]. Отметим, что *MD*-оценки привлекают в последнее время повышенное внимание в связи с их свойствами робастности, на которые впервые обратил внимание Кнуселл в 1969 году и позднее Холм в дискуссии к работе Бикеля [23]. Отметим также, что *MD*-оценки, так же как и оценки максимального правдоподобия, являются состоятельными, асимптотически нормальными и эффективными оценками при достаточно общих условиях. В данном разделе изучаются асимптотические свойства *MD*-оценок и их обобщения в виде MD_α-оценок, которые вычисляются по «урезанным» выборкам.

MD-оценки

Предположим, что задана параметрическая статистическая модель $(\mathrm{X}, \Im_\theta)$. Здесь X – выборочное пространство и $\Im_\theta$ – заданное параметрическое множество допустимых распределений вероятностей, определяемое в виде $\Im_\theta = \{F : F_X(x,\theta), \theta \in \Theta\}$. Это предположение означает, что имеется выборка $X_1,...,X_n$ из распределения $F_X(x,\theta)$ с плотностью $f_X(x,\theta)$, $x \in R^1$, $\theta \in \Theta$, причем функциональный вид распределения задан с точностью до неизвестного параметра θ (скалярного либо векторного), который принадлежит заданному параметрическому множеству Θ. Требуется построить по выборке $X_1,...,X_n$ из распределения $F_X(x,\theta)$ оценку неизвестного параметра $\theta \in \Theta$. Метод минимума расстояний оценивания параметров состоит в том, что на множестве непрерывных функций распределений $\mathfrak{I}$, для пары распределений $F,G \in \mathfrak{I}$, задается метрика (или расстояние) $\rho(F,G)$. Оценка θ_n параметра θ, полученная методом выбранного расстояния $\rho(F,G)$, определяется из условия минимума этого расстояния между эмпирической функцией распределения $F_n(x)$, построенной по выборке $X_1,...,X_n$, и функцией распределения $F_\theta(x) = F_X(x,\theta)$, принятой в модели $(\mathrm{X},\mathfrak{I}_\theta)$. Таким образом, для выбранного расстояния $\rho(F,G)$, *MD*-оценки θ_n определяются в виде выражения

$$\theta_n = \arg\min_\theta \{\rho(F_n,F_\theta)\}. \qquad \textbf{(11.1)}$$

Отметим, что для построения *MD*-оценок могут быть использованы различные расстояния (см., например, работу [54]). Отметим также, что метод максимального правдоподобия основан на использовании расстояния вида

$$\rho(F_n,F_\theta) = -\int \ln f(x,\theta) dF_n(x).$$

В данном разделе мы рассмотрим *MD*-оценки, которые основаны на использовании взвешенного расстояния Крамера – Мизеса, определяемого при $G = F_n$, в виде

$$\rho_W(F_n,F_\theta) = \int [F_n(x) - F_\theta(x)]^2 W_\theta(x,F_\theta)\, dF_\theta(x), \qquad \textbf{(11.2)}$$

где $W_\theta(x,F_\theta)$ – заданная весовая функция, которая в общем случае может зависеть от ф.р. F_θ или плотности f_θ. Предполагая, что $\rho_W(F_n,F_\theta)$ – дифференцируемая по параметру θ функция, обозначим её производную через $\tilde{\lambda}_{F_n}(\theta) = \partial\rho_W(F_n,F_\theta)/\partial\theta$. С учетом этого обозначения, *MD*-оценка θ_n параметра $i = 2$, основанная на использовании взвешенного расстояния Крамера – Мизеса вида (11.2), является решением уравнения $\tilde{\lambda}_{F_n}(\theta) = 0$,

где $$\tilde{\lambda}_{F_n}(\theta) = -2\int [F_n(x) - F_\theta(x)] \frac{\partial F_\theta(x)}{\partial \theta} W_\theta(x) dF_\theta(x) +$$
$$+\int [F_n(x) - F_\theta(x)]^2 \frac{\partial}{\partial \theta}[W_\theta \cdot f_\theta(x)]dx.$$

Рассмотрим сначала *MD*-оценки параметра сдвига θ в одновыборочном варианте, то есть в этом случае $F_\theta(x) = F(x-\theta)$. Введем опорное семейство распределений в виде

$$\Im_0 = \{F : F_\theta(x) = F_0(x-\theta),\ \theta \in R^1\},$$

где F_0 – заданное опорное распределение с плотностью f_0. Перепишем (11.1) в виде

$$\rho_{F_n, F_0}(\theta, W) = \int [F_n(x) - F_0(x-\theta)]^2 W(x-\theta)dx. \qquad \textbf{(11.3)}$$

Отметим, что выбор весовой функции W в виде плотности опорного распределения, то есть в виде $W(x) = f_0(x)$, соответствует расстоянию Крамера – Мизеса, выбор весовой функции $W(x) = f_0(x)/F_0(x)(1-F_0(x))$ соответствует расстоянию Андерсона – Дарлинга (см., например, [4, 9]). Предполагая, что $\rho_{F_n,F_0}(\theta, W)$ из (11.3) дифференцируемая по параметру θ функция, обозначим её производную через $\lambda_{F_n}(\theta) = \partial\rho_{F_n,F_0}(\theta, W)/\partial\theta$. Уравнение $\lambda_{F_n}(\theta) = 0$ для нахождения *MD*-оценки записывается в виде

$$\frac{2}{n}\sum_{i=1}^{n}\left[\frac{2i-1}{2n} - F_0(X_{(i)} - \theta)\right] W(X_{(i)} - \theta) = 0, \qquad \textbf{(11.4)}$$

где $X_{(1)},\ldots,X_{(n)}$ упорядоченная статистика выборки $X_1,\ldots,X_n$.

Асимптотическая нормальность *MD*-оценок

Асимптотические свойства *MD*-оценок изучались различными авторами (см., например, [4, 9, 53–57]). В данной работе обсуждаются асимптотические свойства *MD*-оценок θ_n параметра сдвига θ, которые при заданной опорной ф.р. F_0, принятой в качестве исходной модели, и заданной весовой функции W являются решением уравнения (11.4). При этом различаются следующие два варианта оценивания параметра θ.

Вариант 1. Функция распределения F наблюдений $X_1,...,X_n$ *известна* и совпадает с опорной функцией распределения F_0, то есть $F = F_0$ (или $F \in \Im_0$).

Вариант 2. Функция распределения F наблюдений $X_1,...,X_n$ *неизвестна* и не обязательно совпадает с опорной функцией распределения F_0, то есть $F \neq F_0$ (или $F \notin \Im_0$).

Отметим, что MD-оценки θ_n параметра сдвига θ, которые являются решением уравнения (11.4), могут быть записаны в виде функционала от эмпирической функции распределения, то есть в виде $\theta_n = \theta(F_n)$, где функционал $\theta(F)$ определяется выражением $\min\limits_{\theta} \rho_{F,F_0}(\theta, W) = \rho_{F,F_0}(\theta(F), W)$ или, с принятым обозначением $T(F)$ для функционала, этот функционал $T(F) = \theta(F)$ задается неявно с помощью выражения вида

$$2\int [F(x+T(F)) - F_0(x)] f_0(x) W(x) dx - \\ -\int [F(x+T(F)) - F_0(x)]^2 W'(x) dx = 0. \tag{11.5}$$

Для изучения асимптотических свойств MD-оценок $\theta_n = \theta(F_n)$ параметра сдвига θ воспользуемся подходом Мизеса и рассмотрим разложение вида

$$\theta(F_n) = \theta(F) + V_{1n} + R_{1n}, \tag{11.6}$$

где V_{1n} – аппроксимационная статистика и $R_{1n} = \theta(F_n) - \theta(F) - V_{1n}$ – остаточный член разложения (11.6). Сначала конкретизируем аппроксимационную статистику V_{1n} и остаточный член R_{1n}. Для этого вычислим дифференциал Гато первого порядка $d_1 T(F; G-F)$ функционала $T(F)$, заданного выражением (11.5). Пусть $F_\lambda = F + \lambda(G-F)$, $0 \le \lambda \le 1$. Заменим в формуле (11.5) функцию распределения F на ф.р. F_λ, в результате получим выражение

$$2\int\{F(x+T(F_\lambda))+\lambda[G(x+T(F_\lambda))-F(x+T(F_\lambda))]-F_0(x)\}f_0(x)W(x)dx-$$

$$-\int\{F(x+T(F_\lambda))+\lambda[G(x+T(F_\lambda))-F(x+T(F_\lambda))]-F_0(x)\}^2 f_0(x)W'(x)dx=0.$$

Дифференцируя данное выражение по параметру λ, полагая $\lambda=0$ и учитывая, что

$$d_1T(F;G-F)=\partial T(F_\lambda)/\partial\lambda\,|_{\lambda=0} \text{ и } T(F_\lambda)\,|_{\lambda=0}=T(F)=\theta,$$

получим

$$d_1T(F;G-F)=$$

$$=\frac{\int[G(x+\theta)-F(x+\theta)]\{[F(x+\theta)-F_0(x)]W'(x)-f_0(x)W(x)\}dx}{\int f(x+\theta)f_0(x)W(x)dx-\int[F(x+\theta)-F_0(x)]f(x+\theta)W'(x)dx}=$$

$$=\frac{\int[G(x)-F(x)]\{[F(x)-F_0(x-\theta)]W'(x-\theta)-f_0(x-\theta)W(x-\theta)\}dx}{\int f(x)f_0(x-\theta)W(x-\theta)dx-\int[F(x)-F_0(x-\theta)]f(x)W'(x-\theta)dx}.$$

Из полученного выражения, в котором следует заменить ф.р. G на эмпирическую ф.р. F_n, получаем формулу для аппроксимационной статистики V_{1n} в виде

$$V_{1n}=d_1T(F;F_n-F)=n^{-1}\sum IF(X_i;F,F_0,W),$$

где $IF(u;F,F_0,W)=d_1T(F;\Delta_u-F)$ – функция влияния для *MD*-оценки $\theta_n=\theta(F_n)$ параметра сдвига θ, которая при заданной опорной ф.р. F_0 и заданной весовой функции W является решением уравнения (11.4). Отметим, что выражение для функции влияния также следует из предыдущей формулы путем замены ф.р. G на вырожденную в точке u функцию распределения Δ_u. Полученные формулы, а также разложение (11.6) служат основой для доказательства асимптотической нормальности *MD*-оценок, являющихся решением уравнения (11.4).

Отметим, что общие условия регулярности, накладывающие ограничения на поведение хвостов ф.р. F и весовой функции W, при которых выполняется выражение $\sqrt{n}R_{1n}\to^p 0$, $n\to\infty$, и при которых *MD*-оценки состоятельны и асимптотически нормальны, приводятся в [53–55]. Кроме того, рассматриваемые здесь *MD*-оценки входят в семейство MD_α-оценок, асимптотические свойства которых описаны в [57].

Для формулировки дальнейших результатов обозначим через $\Im_S$ семейство абсолютно непрерывных симметричных распределений. Выделим класс W_S весовых функций, для которых предполагаем, что они дифференцируемы, являются четными функциями, то есть $W(-x) = W(x)$, и $\int \{F(x)(1-F(x))\}^p W(x+c)dx < \infty$, $p > 0$, $c \in (-\infty, +\infty)$.

Теорема 11.7. Пусть $(F, F_0) \in \Im_S$ и $W \in W_S$. Тогда при выполнении неравенств

$$0 < \sigma^2(F; F_0, W) = \int IF^2(x; F, F_0, W)dF(x) < \infty$$

выполняется асимптотическое выражение вида

$$L\{\sqrt{n}[\theta(F_n) - \theta(F)] / \sigma(F; F_0, W)\} = N(0, 1),\ n \to \infty,$$

где асимптотическая дисперсия *MD*-оценки с опорной ф.р. F_0 и весовой функцией W при распределении F наблюдений $X_1, ..., X_n$ равна $D(F; F_0, W) = \sigma^2(F; F_0, W) / n$ и функция влияния Хампеля $IF(u; F, F_0, W) = -IF(-u; F, F_0, W)$ для *MD*-оценки вычисляется по формулам

$$IF(u; F, F_0, W) = A_{F,F_0}(u; W) / B_{F,F_0}(W),\ 0 \le u < \infty, \qquad \textbf{(11.8)}$$

$$A_{F,F_0}(u; W) = \int_0^u W(x)dF(x) - W(u)[F(u) - F_0(u)] =$$

$$= \int_0^u W(x)dF_0(x) - \int_0^u [F(x) - F_0(x)]W'(x)\,dx,$$

$$B_{F,F_0}(W) = \int_{-\infty}^{\infty} f_0(x)W(x)dF(x) - \int_{-\infty}^{\infty} [F(x) - F_0(x)]W'(x)dF(x). \qquad \textbf{(11.9)}$$

Доказательство этой теоремы может быть найдено в [53, 55].

Отметим, что для первого варианта оценивания параметра θ, когда $F \in \Im_0$, функция влияния $IF(u;F,W)$, $0 \le u < \infty$, определяется выражением

$$IF(u;F,W) = \frac{\int_{-\infty}^{+\infty} \{F(x) - I[u \le x]\} W(x) dF(x)}{\int_{-\infty}^{+\infty} f^2(x) W(x) dx} = \frac{\int_0^u W(x) dF(x)}{\int_{-\infty}^{+\infty} f(x) W(x) dF(x)} =$$

$$= J^{-1}(F,W) \int_0^u f(x) W(x) dx, 0 \le u < \infty,$$

$$IF(u;F,F_0,W) = -IF(-u;F,F_0,W) \qquad \textbf{(11.10)}$$

и асимптотическая дисперсия $\sqrt{n}\, MD$-оценки вычисляется по формуле

$$\sigma^2(F,W) = \frac{\int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} \{F(y) - I[u \le y]\} W(y) dF(y) \right)^2 dF(u)}{\left(\int_{-\infty}^{+\infty} f(x) W(x) dF(x) \right)^2} =$$

$$= \frac{\int_{-\infty}^{+\infty} \left(\int_0^x W(y) dF(y) \right)^2 dF(x)}{\left(\int_{-\infty}^{+\infty} f^2(x) W(x) dx \right)^2}. \qquad \textbf{(11.11)}$$

Эффективные *MD*-оценки

Для первого варианта оценивания параметра θ, когда функция распределения F наблюдений $X_1,...,X_n$ известна и совпадает с опорной симметричной функцией распределения F_0, в классе MD-оценок существует эффективная оценка параметра θ, асимптотическая дисперсия которой равна обратной величине информации Фишера $I(f_0)$ относительно параметра сдвига θ распределения $F_0(x-\theta)$ с плотностью f_0. Эта эффективная оценка определяется весовой функцией вида

$$W^*(x) = a\,\frac{d^2\{-\ln f_0(x)\}}{dx^2}\cdot\frac{1}{f_0(x)}. \qquad \textbf{(11.12)}$$

Данный результат отмечался ранее в работах [4, 9, 55]. В справедливости этого факта можно убедиться следующим образом. Обозначим $\psi(x) = -f'(x)/f(x)$, тогда $\psi'(x) = d^2\{-\ln f(x)\}/dx^2$ и выражение (11.12) перепишется с учетом того, что $F = F_0$, в виде $W(x) = a\psi'(x)/f(x)$. Подставив эту весовую функцию $W \in W_S$ в формулу (11.11), и учитывая, что для $F \in \Im_S$ выполняется равенство $\psi(0) = 0$, получаем

$$\sigma^2(F,W) = \frac{\int_{-\infty}^{+\infty}\left(\int_0^x W(y)dF(y)\right)^2 dF(x)}{\left(\int_{-\infty}^{+\infty} f^2(x)W(x)dx\right)^2} =$$

$$= \frac{a^2\int_{-\infty}^{+\infty}\psi^2(x)dF(x)}{a^2\left(\int_{-\infty}^{+\infty}\psi'(x)dF(x)\right)^2} = \frac{I(f)}{I^2(f)} = \frac{1}{I(f)}.$$

Пример 11.13. Отметим, что использование формулы (11.12) позволяет отыскать функцию распределения F_0, при которой MD-оценка Крамера – Мизеса с весовой функцией, равной опорной плотности, то есть $W^*(x) = f_0(x)$, является асимптотически эффективной оценкой параметра θ. В самом деле, использование равенства $W^*(x) = f_0(x)$ приводит к дифференциальному уравнению, $d^2\{-\ln f_0(x)\}/dx^2 = af_0^2(x)$, решением которого является плотность f_0 функции распределения F_0 гиперболического секанса. Эта плотность записывается в виде

$$f_0(x) = 2/\pi(e^x + e^{-x}) = (1/\pi)\sec h(x),\ x \in R^1,$$

где $\sec h(x) = 1/\mathrm{ch}(x)$, $\mathrm{ch} = (e^x + e^{-x})/2$. Отметим, что для этой плотности выполняется условие нормировки, в самом деле

$$\int_{-\infty}^{\infty} f_0(x)dx = \frac{1}{\pi}\int_{-\infty}^{\infty} \sec h(x)dx = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{dx}{\mathrm{ch}(x)} = \frac{2}{\pi}\mathrm{arctg}(e^x)\,|_{-\infty}^{+\infty} =$$
$$= \frac{2}{\pi}[\mathrm{arctg}(\infty) - \mathrm{arctg}(0)] = \frac{2}{\pi}[(\pi/2) - 0] = 1.$$

Далее, функция распределения гиперболического секанса определяется в виде

$$F_0(x) = \int_{-\infty}^{x} f_0(y)dy = \frac{1}{\pi}\int_{-\infty}^{x} \sec h(y)dy = \frac{2}{\pi}[\mathrm{arctg}(e^y)]\,|_{-\infty}^{x} = \frac{2}{\pi}\mathrm{arctg}(e^x),$$
$$x \in R^1.$$

Кроме того, справедливы следующие выражения:

$$F_0(x) = (2/\pi)\mathrm{arctg}(e^x),\ x \in R^1;\ F_0^{-1}(t) = \ln[\mathrm{tg}(\pi t/2)];$$
$$f_0(F_0^{-1}(t)) = (2/\pi)\sin(\pi t/2)\cos(\pi t/2),\ 0 \le t \le 1.$$

Непосредственной проверкой убеждаемся, что выбор весовой функции в виде, пропорциональном опорной плотности, то есть когда $W^*(x) = a\,f_0(x)$, приводит к эффективной *MD*-оценке для распределения гиперболического секанса с плотностью f_0. В самом деле, пусть $\psi_0(x) = -f_0'(x)/f_0(x)$. Тогда формула (11.12) запишется в виде $W(x) = a \cdot \psi_0'(x)/f_0(x)$. Учитывая теперь, что выполняются равенства

$$f_0'(x) = -(2/\pi)(e^x - e^{-x})/(e^x + e^{-x})^2,$$

$$\psi_0(x) = (e^x - e^{-x})/(e^x + e^{-x}),\ \psi_0'(x) = 4/(e^x + e^{-x})^2,$$

окончательно получаем требуемое равенство для весовой функции в виде

$$W(x) = a \cdot \psi_0'(x)/f_0(x) = a \cdot 2\pi/(e^x + e^{-x}) = a \cdot \pi^2 f_0(x).$$

Отметим, что информация Фишера для параметра сдвига θ плотности $f_0(x) = (1/\pi)\sec h(x)$ – гиперболический секанс, как и для распределения Коши, равна $I(f_0)=1/2$ и вычисляется с учетом того, что

$$f_0'\,(x)=(1/\pi)[-\sec h(x)\cdot th(x)],$$

$$d^2[-\ln f_0(x)]/dx^2=[-f_0'(x)/f_0(x)]'=[\mathrm{th}(x)]'=\sec h^2(x),$$

по формуле

$$I(f_0)=\int[-f_0'(x)/f_0(x)]^2\,dF_0(x)=$$

$$=-\int\{d^2[-\ln f_0(x)]/dx^2\}\,dF_0(x)=\frac{1}{\pi}\int\limits_{-\infty}^{\infty}\sec h^2(x)\sec h(x)dx=$$

$$=\frac{1}{\pi}\left[\frac{\mathrm{sh}(x)}{2\,\mathrm{ch}^2(x)}+\frac{1}{2}\mathrm{arctg}(\mathrm{sh}(x))\right]_{-\infty}^{+\infty}=\frac{1}{\pi}[0+(\pi/2)]=\frac{1}{2}.$$

Для весовой функции $W^*(x)=f_0(x)$ при $F=F_0$ функция влияния MD-оценки, согласно (11.10), определяется выражением

$$IF(u;F_0,W=f_0)=\frac{\int_{-\infty}^{+\infty}\{F_0(x)-I[u\le x]\}f_0^2(x)dx}{\int_{-\infty}^{+\infty}f_0^3(x)dx}=$$

$$=-2\cos(\pi F_0(u))=-2\cos[2\,\mathrm{arctg}(e^u)],\ \ u\in R^1.$$

Обратим внимание на неудачно используемые здесь обозначения. Чтобы не путать дальше опорное распределение F_0 с распределением гиперболического секанса, везде ниже для этого распределения будем обозначать функцию распределения F_5, а её плотность f_5. Отметим, что приведённую *ограниченную* функцию влияния (см. рис. 11.1) для *MD*-оценки с оптимальной весовой функцией $W_5^*=f_5$ при распределении гиперболический секанс (то есть при $F_0=F_5$) можно записать и в другом, эквивалентном варианте, с помощью выражения вида

$$IF(u;F_5,W_5^*=f_5)=\psi_0(u)/I(f_0)=2(e^u-e^{-u})/(e^u+e^{-u}),$$

$$u\in R^1.$$

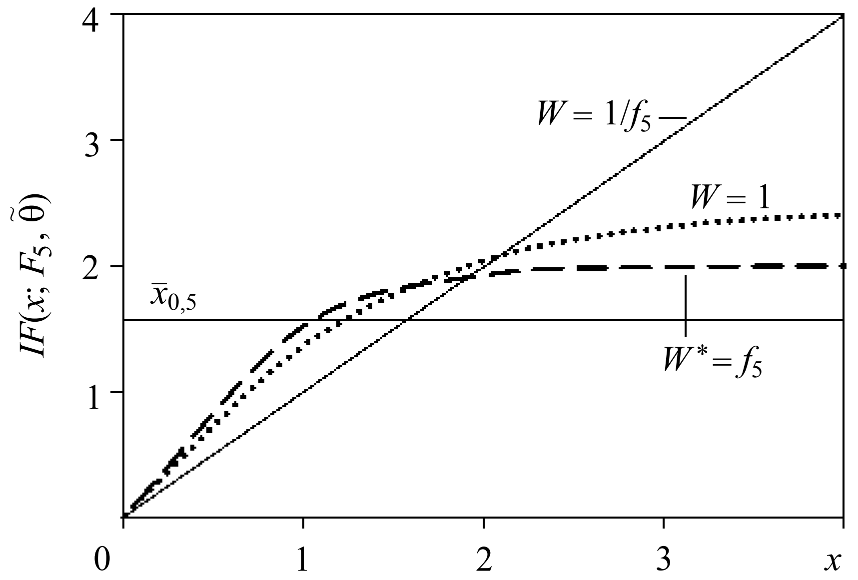

Рис. 11.1. Фунции влияния MD-оценок для $F_0 = F_5$

Далее, чувствительность к грубым ошибкам MD-оценки с оптимальной весовой функцией $W_5^* = f_5$ при распределении F_5 гиперболического секанса равна

$$\gamma^*(F_5, W = f_5) = \sup_x |IF(x; F_5, W = f_5)| = 2{,}00.$$

Асимптотическая дисперсия, в данном случае эффективной MD-оценки, вычисляется по формуле

$$\sigma^2(F_0, W = f_0) = \int_{-\infty}^{\infty} IF^2(x; F_0, W = f_0)\, dF_0(x) =$$

$$= \int_{-\infty}^{\infty} 4\cos^2(\pi F_0(u))\, dF_0(u) =$$

$$= \frac{8}{\pi} \int_{\pi/2}^{\pi} \cos^2(v)\, dv = \frac{8}{\pi}\left(\frac{v}{2} + \frac{\sin(2v)}{4}\right)\Bigg|_{\pi/2}^{\pi} = 2 = 1/I(f_0),$$

где $I(f_0)$ – информация Фишера относительно параметра сдвига для распределения гиперболического секанса с плотностью

$$f_0(x) = 2/\pi(e^x + e^{-x}) = (1/\pi)\sec h(x),$$

$$x \in R^1,$$

Напомним, что для $W \equiv 1$ при $F = F_0$ асимптотическая дисперсия *MD*-оценки совпадает с асимптотической дисперсией оценки Ходжеса – Лемана и для распределения гиперболического секанса (то есть $F_0 = F_5$) вычисляется по формуле

$$\sigma^2(F_0, W = 1) =$$

$$= \frac{1}{12\left(\int_0^1 f_0(F_0^{-1}(t))dt\right)^2} = \frac{1}{12\left((2/\pi)\int_0^1 \sin(\pi t/2)\cos(\pi t/2)dt\right)^2} =$$

$$= \frac{1}{12\left((4/\pi^2)\int_0^1 \sin(\pi t/2)d(\sin(\pi t/2))\right)^2} = \frac{1}{12(2/\pi^2)^2} = \frac{\pi^4}{48} \approx 2{,}029 .$$

Функция влияния *MD*-оценки с весовой функцией $W \equiv 1$ для распределения гиперболический секанс (то есть $F_0 = F_5$) определяется в виде

$$IF(x; F_0, W \equiv 1) = \frac{F_0(x) - (1/2)}{\int_0^1 f_0(F_0^{-1}(t))dt} = \frac{(2/\pi)\operatorname{arctg}(e^x) - (1/2)}{(2/\pi^2)} =$$

$$= \pi \operatorname{arctg}(e^x) - (\pi^2/4), \ \ x \in R^1,$$

и чувствительность к грубым ошибкам

$$\gamma^*(F_5, W \equiv 1) = \sup_x | IF(x; F_5, W \equiv 1) | = \pi^2/4 \approx 2{,}47 .$$

Приведённая функция влияния для MD-оценки с весовой функцией $W \equiv 1$ является *ограниченной функцией* и при распределении гиперболический секанс (то есть при $F_0 = F_5$) приведена на рис. 11.1.

Пример 11.14. Рассмотрим супермодель

$$\Im_S^* = \{F_{(1)}, F_{(2)}, F_{(3)}, F_{(4)}, F_{(5)}\}$$

в виде конечного семейства заданных распределений, где $F_{(1)} = \Phi$ – стандартное нормальное распределение с плотностью $f_{(1)} = \phi$, информация Фишера $I(f_{(1)}) = 1$; $F_{(2)}$ – логистическое, $I(f_{(2)}) = 1/3$; $F_{(3)}$ – Лапласа, $I(f_{(3)}) = 1$; $F_{(4)}$ – Коши, $I(f_{(2)}) = 1/2$; $F_{(5)}$ – гиперболический секанс, $I(f_{(5)}) = 1/2$. Оптимальные весовые функции вида (11.12) для распределений из $\Im_S^*$ приведены в табл. 11.1 и на рис. 11.2.

Таблица 11.1

Оптимальные весовые функции вида $W^*(x) = a \cdot \psi'(x)/f(x)$

$F_{(1)}$	$F_{(2)}$	$F_{(3)}$	$F_{(4)}$	$F_{(5)}$
$W_{(1)}^*(x) = 1/\phi(x)$	$W_{(2)}^*(x) \equiv 1$	$T_2(F) = F^{-1}(1/2)$	$0 \le u < \infty$	$W_{(5)}^*(x) =$ $= (2/\pi)(e^x + e^{-x})^{-1}$

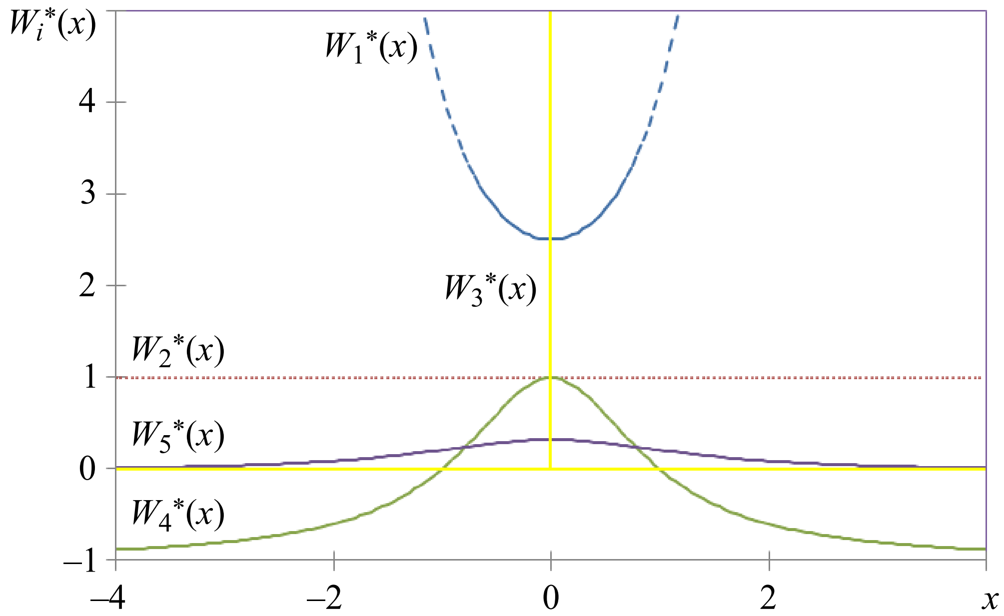

Рис. 11.2. Оптимальные весовые функции *MD*-оценок для $F \in \Im_S^*$

Отметим, что асимптотическая дисперсия *MD*-оценки с опорным распределением $F_0(x) = F(x)$ и весовой функцией $W(x) = 1/f(x)$ совпадает с асимптотической дисперсией выборочного среднего $\overline{X}$ и вычисляется по формуле

$$\sigma^2(F, W = 1/f) = \frac{\int_{-\infty}^{+\infty} \left(\int_0^x W(y)dF(y) \right)^2 dF(x)}{\left(\int_{-\infty}^{+\infty} f^2(x)W(x)dx \right)^2} =$$

$$=\frac{\int_{-\infty}^{+\infty}\left(\int_0^x (1/f(y))f(y)dy\right)^2 dF(x)}{\left(\int_{-\infty}^{+\infty} f^2(x)(1/f(x))dx\right)^2}=\int\limits_{-\infty}^{\infty} x^2 dF(x)=\sigma^2(F,\bar{X}).$$

Для весовой функции $W(x)=1/\phi(x)$, где $\phi(x)$ – стандартная нормальная плотность, *MD*-оценка является эффективной оценкой параметра сдвига θ нормального распределения, однако она, как и выборочное среднее $\bar{X}$, имеет *неограниченную* функцию влияния $IF(x;\Phi,W=1/\phi)=x$, $x\in R^1$, и её чувствительность к грубым ошибкам не является конечной, то есть $\gamma^*(\Phi,W=1/\phi)=\infty$. Отметим также, что выбор весовой функции $W(x)\equiv 1$ приводит к асимптотически эффективной *MD*-оценке при логистической ф.р. $F_{(2)}$ (её дисперсия в данном случае совпадает с дисперсией *HL*-оценки), при этом абсолютная эффективность *MD*-оценки с весовой функцией $W(x)=f_{(2)}(x)$ равна $АЭ(F_{(2)},W=f_{(2)})=[3{,}036(1/3)]^{-1}=0{,}988$. Напомним, что для логистического распределения $F_{(2)}$ с плотностью $f_{(2)}$, выполняется равенство $f_{(2)}=F_{(2)}(1-F_{(2)})$, поэтому, выбор весовой функции в виде $W(x)=f_0/F_0(1-F_0)$, соответствующий *MD*-оценке, основанной на использовании расстояния Андерсона – Дарлинга, также приводит к эффективной оценке при логистическом распределении. Для распределения Лапласа с плотностью $f_{(3)}(x)=(1/2)\exp(-|x|), x\in R^1$, функция $\psi(x)=-f'_{(3)}(x)/f_{(3)}(x)=$ $=\operatorname{sign}(x)$ и, следовательно, оптимальная весовая функция $W^*(x)=a\cdot\psi'(x)/f(x)$, определяемая формулой (11.12), при $a=1$ принимает вид

$$W^*_{(3)}(x)=\{\operatorname{sign}(x)\}'/f_{(3)}(x)=\delta(x-0)/f_{(3)}(x)=2e^{|x|}\delta(x-0).$$

Используя данное выражение для оптимальной весовой функции и формулу (11.11), можно убедится, что асимптотическая дисперсия *MD*-оценки будет совпадать с асимптотической дисперсией выборочной медианы $\bar{X}_{1/2}$, которая является асимптотически эффективной оценкой параметра сдвига θ для распределения Лапласа. В самом деле, из формулы (11.11) при весовой функции $W(x)=\delta(x-0)/f(x)$, получаем

$$\sigma^2(F,W)=\frac{\int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty}\{F(y)-I[u\le y]\}W(y)dF(y)\right)^2 dF(u)}{\left(\int_{-\infty}^{+\infty}f(x)W(x)dF(x)\right)^2}=$$

$$=\frac{\int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty}\{F(y)-I[u\le y]\}\delta(y-0)dy\right)^2 dF(u)}{\left(\int_{-\infty}^{+\infty}f(x)\delta(x-0)dx\right)^2}=$$

$$=\frac{\int_{-\infty}^{+\infty}\{F(0)-I[u\le 0]\}^2 dF(u)}{f^2(0)}=$$

$$=\frac{(1/4)-\int_{-\infty}^{+\infty}I[u\le 0]\}dF(u)+\int_{-\infty}^{+\infty}I^2[u\le 0]\}dF(u)}{f^2(0)}=$$

$$=\frac{1}{4f^2(0)}=\sigma^2(F,\bar{X}_{1/2}).$$

Отметим, что для распределения Коши оптимальная весовая функция $W^*_{(4)}(x)=a(1-x^2)/(1+x^2)$ отрицательна вне интервала $[-1,1]$. Этот факт можно пояснить следующим образом. Из формулы (11.10) следует, что весовая функция W выражается через производную функции влияния в виде

$$W(u)=J(F,W)\,IF^{/}(u;F,W)/f(u),\ 0\le u<\infty.$$

Таким образом, чтобы «уменьшить» влияние выбросов на MD-оценку, нужно, чтобы её функция влияния при больших значениях аргумента убывала и, следовательно, весовая функция должна быть *отрицательной*, что и наблюдается для оптимальной весовой функции $W^*_{(4)}(x)=a(1-x^2)/(1+x^2)$ при распределении Коши.

Пример 11.15. Рассмотрим семейство распределений Стьюдента $\Im_r\in\Im_S$, для которого плотность распределения $f_r(x)$ с r степенями свободы записывается в виде

$$f_r(x) = A(r)(1 + (x^2 / r))^{-(r+1)/2}, \; x \in R^1,$$
$$A(r) = \Gamma((r+1)/2) / \sqrt{r\pi}\,\Gamma(r/2).$$

Используя (11.11), можно убедиться, что оптимальные весовые функции для распределений этого семейства вычисляются по формуле

$$W_r^*(x) = a \cdot r^{-(r+1)/2}(r+1)A^{-1}(r)(r - x^2)(r + x^2)^{(r-3)/2}.$$

Отсюда, в частности при $r = 1$, получаем оптимальную весовую функцию для распределения Коши, $W_r^*(x)|_{r=1} = a \cdot 2\pi(1-x^2)/(1+x^2) = W_4^*(x)$. Случай $r \to \infty$ соответствует нормальному распределению. Учитывая, что при $r \to \infty$ выполняются выражения $A(r) \to 1/\sqrt{2\pi}$ и $(1+(x^2/r))^{-(r+1)/2} \to e^{-x^2/2}$, из общей формулы получаем

$$\lim_{r\to\infty} W_r^*(x) = a \cdot \sqrt{2\pi}\, \exp(x^2/2) = a \cdot 1/\phi(x) = W_{(1)}^*(x).$$

Свойства робастности *MD*-оценок

Для изучения свойств робастности *MD*-оценок рассмотрим два типа супермоделей, которые описывают отклонения от гауссовской модели наблюдений. Первая супермодель $\Im_S^*$ определяется в виде конечного семейства заданных распределений, то есть

$$\Im_S^* = \{F_{(1)}, F_{(2)}, F_{(3)}, F_{(4)}, F_{(5)}\}.$$

Вторую супермодель $\Im_{\varepsilon,\tau}(\Phi)$ называют гауссовской моделью с масштабным засорением и определяют в виде

$$\Im_{\varepsilon,\tau}(\Phi) = \{F : F_{\varepsilon,\tau}(x) = (1-\varepsilon)\Phi(x) + \varepsilon\Phi(x/\tau)\},\; W(x) = \phi(x),\; \tau \ge 1,$$

где $\Phi(x)$ – стандартная нормальная функция распределения с плотностью $\phi(x)$, ε – пропорция засорения выборки и τ – масштабный параметр засорения.

Пример 11.16. Первый вариант. Сначала рассмотрим свойства *MD*-оценок в рамках супермодели $\Im_S^*$ при различных вариантах задания опорной ф.р. F_0 и весовых функций *W*. Для первого варианта оценивания параметра θ, когда функция распределения *F* известна и равна опорной функции распределения F_0, то есть $F \in \Im_0$, функция влияния *MD*-оценки и её асимптотическая дисперсия вычисляются по формулам (11.10) и (11.11). Рассмотрим различные варианты задания весовой функции $W \in W_S$.

А) Пусть $W(x) \equiv 1$ и $F(x) = F_0(x)$. При этих условиях *MD*-оценки с весовой функцией $W(x) \equiv 1$ являются *B*-робастными, то есть они имеют *ограниченные* функции влияния, которые определятся в виде $IF(x;F,W \equiv 1) = \{2F(x)-1\}/2\int f^2(x)dx$. В гауссовском случае, при $F = \Phi$, функция влияния определяется выражением $IF(x;\Phi,W \equiv 1) = \sqrt{\pi}[2\Phi(x)-1]$. Чувствительность к грубым ошибкам $\gamma^*(F,T) = \sup_x |IF(x;F,T)|$ для *MD*-оценки с весовой функцией $W(x) \equiv 1$ равна $\gamma^*(\Phi,W \equiv 1) = \sqrt{\pi} \approx 1{,}77$.

Б) Пусть весовая функция совпадает с опорной плотностью, то есть $W(x) = f_0(x)$ и $F(x) = F_0(x)$. При этих предположениях из (11.11) следует, что асимптотическая дисперсия *MD*-оценки с опорным распределением $F_0(x) = F(x)$ и $W(x) = f(x)$ вычисляется по формуле

$$\sigma^2(F,W=f) = \frac{\int_{-\infty}^{+\infty}\left(\int_0^x f^2(y)dy\right)^2 dF(x)}{\left(\int_{-\infty}^{+\infty} f^3(x)dx\right)^2}.$$

Отметим, что при гауссовском распределении, то есть при $F(x) = \Phi(x)$, и весовой функции $W(x) = \phi(x) = (1/\sqrt{2\pi})\exp\{-x^2/2\}$ из (11.10) получаем *ограниченную* функцию влияния *MD*-оценки в виде

$$IF(x;\Phi,W=\phi) = (\sqrt{3\pi}/2)\tilde{\Phi}(x) = (\sqrt{3\pi}/2)[2\Phi(x\sqrt{2})-1],\ x \in R^1,$$

где $\tilde{\Phi}(x)$ – функция Лапласа, определяемая выражениями

$$\tilde{\Phi}(x) = (2/\sqrt{\pi})\int_0^x \exp\{-x^2\}dx,\ \tilde{\Phi}(x) = 2\Phi(x\sqrt{2}) - 1,\ x \ge 0,$$

$$\Phi(x) = (1/\sqrt{2\pi})\int_{-\infty}^x \exp\{-x^2/2\}dx.$$

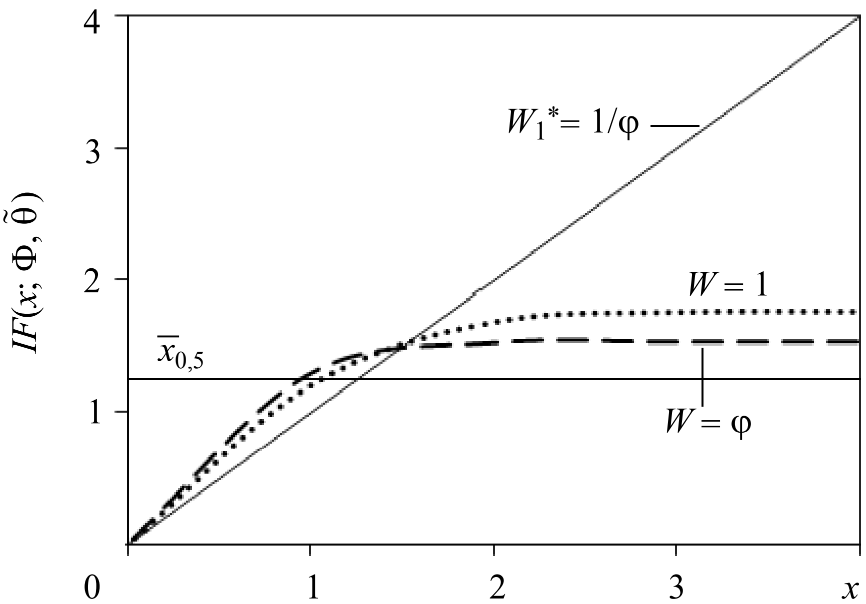

Рис. 11.3. Функции влияния *MD*-оценок для нормального распределения

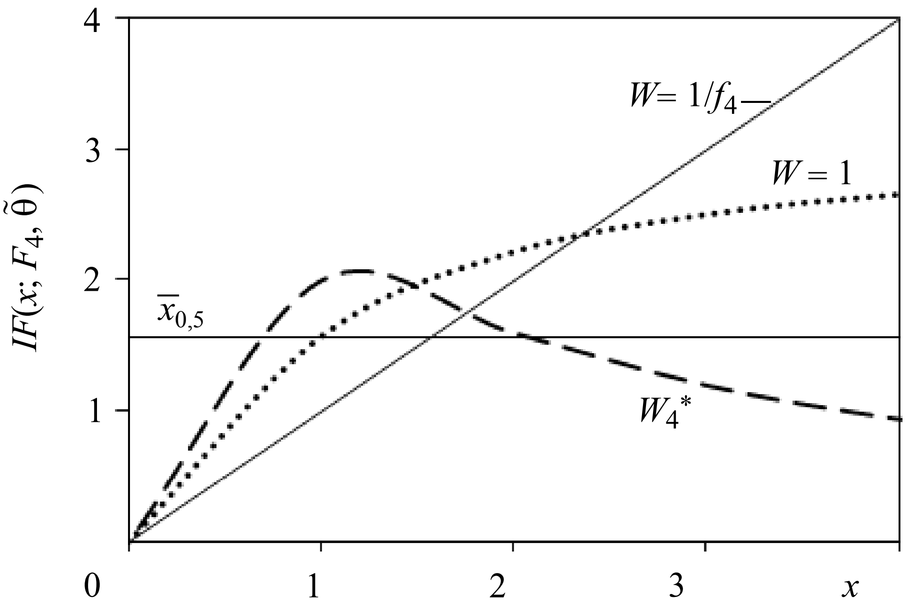

Рис. 11.4. Функции влияния *MD*-оценок для распределения Коши

Чувствительность к грубым ошибкам $\gamma^*(F,T)$ для *MD*-оценки с весовой функцией $W(x)=\phi(x)$ равна $\gamma^*(\Phi,W=\phi)=\sqrt{3\pi}/2=1{,}53$. В этом случае асимптотическая дисперсия $\sqrt{n}MD$-оценки равна

$$\sigma^2(\Phi,W=\phi)=2\int_0^\infty IF^2(x;\Phi,W=\phi)\,d\Phi(x)=\frac{3\pi}{2}\frac{1}{\sqrt{2\pi}}\int_0^\infty \tilde{\Phi}^2(x)e^{-x^2/2}dx=$$
$$=(3/2)\operatorname{arctg}(2/\sqrt{5})=1{,}095.$$

Значения асимптотических дисперсий *MD*-оценок для случаев (А) и (Б) вычислены для следующих распределений: $F_{(1)}$ – нормальное, $F_{(2)}$ – логистическое, $F_{(3)}$ – Лапласа, $F_{(4)}$ – Коши, $F_{(5)}$ – гиперболический секанс. Численные расчеты по полученным формулам при $F_0=F_{(i)}$, $i=1,\dots,5$ и при различных весовых функциях приведены в табл. 11.2.

Таблица 11.2

Асимптотические дисперсии $\sqrt{n}MD$-оценок для супермодели $\mathfrak{I}_S^*$ при $F_0=F_{(i)}$, $i=1,\dots,5$

Весовая функция	$F_{(1)}=\Phi$	$F_{(2)}$	$F_{(3)}$	$F_{(4)}$	$F_{(5)}$
$W\equiv 1$	1,047 (0,96)	3,000 **(1,00)**	1,333 (0,75)	3,287 (0,61)	2,029 (0,98)
$W_{(i)}(x)=f_{(i)}(x)$	1,095 (0,91)	3,036 (0,99)	1,200 (0,83)	2,573 (0,78)	2,000 **(1,00)**
$\tilde{W}_{(i)}(x)=f_{(i)}/F_{(i)}(1-F_{(i)})$	1,035 (0,97)	3,000 **(1,00)**	1,262 (0,79)	2,317 (0,86)	2,020 (0,99)
$W_{(i)}(x)=1/f_{(i)}(x)$	1,000 **(1,00)**	3,290 (0,91)	2,000 (0,50)	∞ (0,00)	2,467 (0,81)
$W^*_{(4)}(x)=(1-x^2)/(1+x^2)$	1,109 (0,90)	4,204 (0,71)	1,230 (0,81)	2,000 **(1,00)**	2,103 (0,95)

В данной таблице в круглых скобках приведены значения абсолютных эффективностей *MD*-оценок, которые вычислены по формуле $AЭ(F,W)=[\sigma^2(F,W)I(f)]^{-1}$. Отметим, что для распределений с «тяжелыми хвостами» (Коши и Лапласа) абсолютная эффективность *MD*-оценок в большей степени зависит от выбора весовой функции W. Весовая функция $W_{(1)}(x)=1/f_{(1)}(x)$ – оптимальная при нормальном распределении. Весовые функции $W\equiv 1$ и $W_{(2)}(x)=$ $=f_{(2)}/F_{(2)}(1-F_{(2)})$ – оптимальные для логистического распределения $F_{(2)}$. Весовая функция $W^*_{(4)}(x)=(1-x^2)/(1+x^2)$ – оптимальная для распределения Коши. Весовая функция $W_{(5)}(x)=f_{(5)}(x)$ – оптимальная для распределения $F_{(5)}$ – гиперболический секанс.

Пример 11.17. Второй вариант. Рассмотрим случай когда $F\neq F_0$ и супермодель в виде конечного семейства заданных распределений $\Im^*_S=\{F_{(1)},F_{(2)},F_{(3)},F_{(4)},F_{(5)}\}$, то есть $F\in\Im^*_S$. В этом случае асимптотическая дисперсия $\sqrt{n}MD$-оценок для весовой функции $W=1$ вычисляется по формуле

$$\sigma^2(F,F_0,W\equiv 1)=\frac{2\int_0^{\infty}[F_0(u)-(1/2)]^2\,dF(u)}{\left(\int f_0(x)f(x)dx\right)^2},F\in\Im^*_S. \qquad \textbf{(11.18)}$$

Численные значения асимптотических дисперсий $\sqrt{n}MD$-оценок для $F\in\Im^*_S$ и весовой функции $W\equiv 1$, вычисленные по формуле (11.18), приведены в табл. 11.3.

Таблица 11.3

Асимптотические дисперсии $\sqrt{n}MD$-оценок $\hat\theta_{(i)}=\hat\theta(F_0=F_{(i)},W\equiv 1)$, $i=1,\ldots,5$, для $F\in\Im^*_S$

$\hat\theta\setminus F$	$F_{(1)}$	$F_{(2)}$	$F_{(3)}$	$F_{(4)}$	$F_{(5)}$	$d(\hat\theta,\Im^*_S)$
$\hat\theta_{(1)}$	1,047 (0,96)	3,051 (0,98)	1,383 (0,72)	2,911 (0,69)	2,008 (0,99)	0,42
$\hat\theta_{(2)}$	1,016 (0,98)	3,000 (1,00)	1,524 (0,66)	3,679 (0,54)	2,069 (0,97)	0,57
$\hat\theta_{(3)}$	1,059 (0,94)	3,048 (0,98)	1,333 (0,75)	2,957 (0,68)	2,006 (0,99)	**0,41**
$\hat\theta_{(4)}$	1,046 (0,96)	3,025 (0,99)	1,385 (0,72)	3,290 (0,61)	2,017 (0,99)	0,48
$\hat\theta_{(5)}$	1,031 (0,97)	3,011 (0,99)	1,439 (0,70)	3,276 (0,61)	2,029 (0,98)	0,49

Отметим, что в табл. 11.3 в круглых скобках приведены абсолютные эффективности оценок, вычисленные по формуле $AЭ_F(F,\hat{\theta}) = \{\sigma^2(F, F_0, W \equiv 1) I(f)\}^{-1}$. В крайнем правом столбце таблицы приведены дефекты оценок в супермодели $\Im_S^*$, вычисленные по формулам (11.27).

Замечание 11.19. Отметим, что для сравнения совокупности оценок $\hat{\theta}_1,...,\hat{\theta}_k$ параметра сдвига θ при заданном симметричном распределении F используют понятие дефекта оценки (см., например, [4, 9]). Пусть имеется конечный набор асимптотически нормальных и несмещенных оценок $\hat{\theta}_1,...,\hat{\theta}_k$ параметра сдвига θ, построенных по выборке $X_1,...,X_n$ из распределения F, для которых выполняются выражения

$$L\left\{\frac{\sqrt{n}(\hat{\theta}_i - \theta)}{\sigma_F(\hat{\theta}_i)}\right\} = N(0,\ 1) \text{ при } n \to \infty,\ i = 1,...,k.$$

Дефект оценки $\hat{\theta}_i$, $i = 1,...,k$, среди сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$ параметра θ при заданном распределении F определяется в виде

$$DE_F(\hat{\theta}_i) = 1 - \min\{\sigma_F^2(\hat{\theta}_1),...,\sigma_F^2(\hat{\theta}_k)\} / \sigma_F^2(\hat{\theta}_i), i = 1,...,k. \quad \textbf{(11.20)}$$

Отметим, что если среди сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$ есть эффективная оценка $\hat{\theta}^*$ параметра θ при заданном распределении F, для которой $\sigma_F^2(\hat{\theta}^*) = 1/I(f)$, тогда $\min\{\sigma_F^2(\hat{\theta}_1),...,\sigma_F^2(\hat{\theta}_k)\} = 1/I(f)$ и, следовательно, в этом случае абсолютный дефект оценки $\hat{\theta}_i$ равен единице минус ее абсолютная эффективность, то есть

$$ADE_F(\hat{\theta}_i) = 1 - AЭ_F(\hat{\theta}_i), i = 1,...,k. \quad \textbf{(11.21)}$$

Замечание 11.22. При изучении свойств робастности сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$ параметра сдвига θ в рамках супермодели $\Im$, состоящей из конечного набора симметричных распределений $\Im = \{F_1,...,F_r\}$, изучают поведение дефективностей оценок на плоскости двух распределений. По оси абсцисс обычно откладывают дефективность для базовой (идеальной модели, обычно гауссовской), а по оси ординат откладывают дефективность для альтернативной модели, входящей в супермодель $\Im = \{F_1, ..., F_r\}$. При таком наглядном представлении дефективностей оценок на плоскости двух распределений предпочтение отдается той оценке, которая окажется ближе к началу координат. Для наглядности, абсолютные дефекты оценок в плоскости распределений «Гаусс – Лаплас» и «Гаусс – Коши» приведены на рис. 11.5 и 11.6.

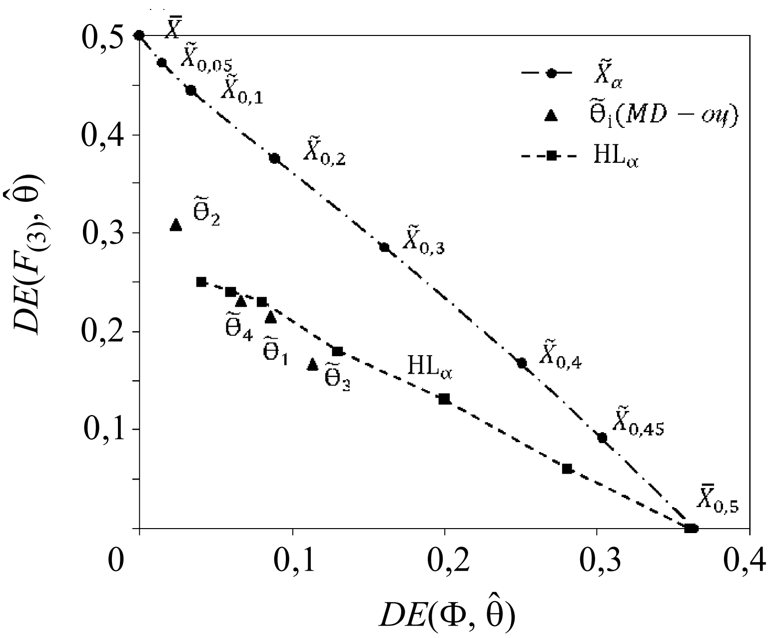

Рис. 11.5. Дефекты оценок в плоскости «Гаусс – Лаплас»

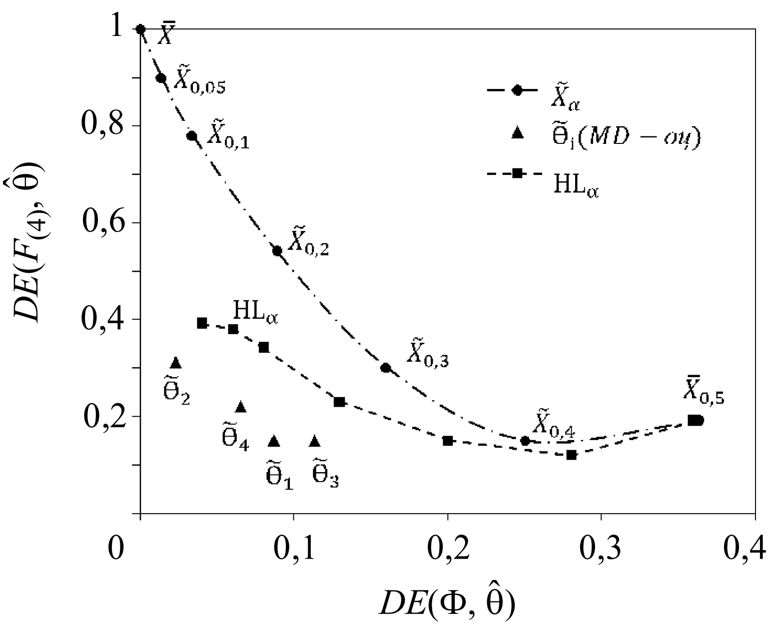

Рис. 11.6. Дефекты оценок в плоскости «Гаусс – Коши»

На этих рисунках наглядно видны преимущества *MD*-оценок $\hat{\theta}_{(i)} = \hat{\theta}(F_0 = F_{(i)}, W = f_{(i)})$, $i = 1,...,5$, для $F \in \Im_S^*$ (*они концентрируются ближе к началу координат*) перед семейством $\tilde{X}_\alpha$-винзоризованных средних и семейством HL_α-оценок Ходжеса – Лемана, $0 \le \alpha \le 1/2$.

Замечание 11.23. Если же мы хотим сделать вывод о предпочтительности оценки среди сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$ параметра θ в рамках всей рассматриваемой супермодели $\Im = \{F_1,...,F_r\}$, то можно использовать евклидову метрику, которая, с использованием введенных обозначений, запишется в виде

$$d(\hat{\theta}_i;\Im) = \left\{ \sum_{j=1}^{r} [DE_{F_j}(\hat{\theta}_i)]^2 \right\}^{1/2}, \tag{11.24}$$

или

$$Ad(\hat{\theta}_i;\Im) = \left\{ \sum_{j=1}^{r} [ADE_{F_j}(\hat{\theta}_i)]^2 \right\}^{1/2}, i = 1,..., k. \tag{11.25}$$

Предпочтение в рамках всей рассматриваемой супермодели $\Im = \{F_1,...,F_r\}$ отдается той оценке $\hat{\theta}_i$ среди сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$, для которой вычисленное значение $d(\hat{\theta}_i;\Im)$ минимальное, то есть

$$d(\hat{\theta}_i;\Im) = \min\{d(\hat{\theta}_1;\Im),...,d(\hat{\theta}_k;\Im)\}. \tag{11.26}$$

Для супермодели $\Im_S^* = \{F_{(1)}, F_{(2)}, F_{(3)}, F_{(4)}, F_{(5)}\}$ формула (11.24) запишется в виде

$$d(\tilde{\theta}_{(i)}, \Im_S^*) = \left(\sum_{j=1}^{5} [1 - \{\sigma^2(F_{(j)}, \tilde{\theta}_{(i)}) I(f_{(j)})\}^{-1}]^2 \right)^{1/2} =$$

$$= \left(\sum_{j=1}^{5} [1 - AЭ(F_{(j)}, \tilde{\theta}_{(i)})]^2 \right)^{1/2}, \; i = 1,..., 5. \tag{11.27}$$

Согласно критерию (11.26), среди сравниваемых оценок $\tilde{\theta}_{(1)},...,\tilde{\theta}_{(5)}$ в супермодели $\Im_S^*$ предпочтение следует отдать *MD*-

оценке $\tilde{\theta}_{(3)}$ с опорным распределением Лапласа, то есть $F_0 = F_{(3)}$, и весовой функцией $W \equiv 1$, так как для этих оценок $\hat{\theta}_{(i)} = \hat{\theta}(F_0 = F_{(i)}, W \equiv 1)$, $i = 1,...,5$, минимальное значение

$$d(\hat{\theta}_{(3)}, \Im_S^*) = \min\{d(\hat{\theta}_{(i)}, \Im_S^*),\ i = 1,...,5\} = 0{,}41$$

(см. правый столбец табл.11.3). Заметим для сравнения, что для оценки Ходжеса – Лемана $d(HL, \Im_S^*) = 0{,}47$; для $\tilde{X}_\alpha$-винзоризованного среднего $d(\tilde{X}_{0,45}, \Im_S^*) = 0{,}41$; для выборочной медианы $d(\bar{X}_{1/2}, \Im_S^*) = 0{,}51$; для выборочного среднего $d(\bar{X}, \Im_S^*) = 1{,}14$.

Пример 11.28. Второй вариант. Рассмотрим гауссовскую модель с масштабным засорением $\Im_{\varepsilon,\tau}(\Phi)$. В качестве опорного распределения выберем нормальное распределение, то есть $F_0 = \Phi$, а распределение F наблюдений $X_1,...,X_n$ характеризуется нормальным распределением с масштабным засорением, то есть $F \in \Im_{\varepsilon,\tau}(\Phi)$. Для принятых предположений асимптотическая дисперсия $\sqrt{n}\,MD$-оценки при $W = 1$ вычисляется по формуле

$$\sigma^2(F_{\varepsilon,\tau}, \Phi, W \equiv 1) = \frac{2\int_0^{+\infty}[\Phi(x) - (1/2)]^2[(1-\varepsilon)\phi(x) + (\varepsilon/\tau)\phi(x/\tau)]dx}{\left(\int_{-\infty}^{\infty}\phi(x)[(1-\varepsilon)\phi(x) + (\varepsilon/\tau)\phi(x/\tau)]dx\right)^2} =$$

$$\frac{[\pi(1-\varepsilon)/6] + [\varepsilon \operatorname{arctg}(\tau^2/\sqrt{2\tau^2+1})]}{\{[(1-\varepsilon)/\sqrt{2}] + (\varepsilon/\sqrt{\tau^2+1})\}^2}.$$

Для весовой функции $W(x) = f_0(x) = \phi(x)$ асимптотическая дисперсия $\sqrt{n}\,MD$-оценки вычисляется по формуле

$$\sigma^2(F_{\varepsilon,\tau}, \Phi, W = \phi) = \frac{2\int_0^{\infty}\left(\int_0^u \phi(x) f_{\varepsilon,\tau}(x)dx - \phi(u)[F_{\varepsilon,\tau}(u) - \Phi(u)]\right)^2 dF_{\varepsilon,\tau}(u)}{\left(\int_{-\infty}^{\infty}\phi^2(x)dF_{\varepsilon,\tau}(x) - \int_{-\infty}^{\infty}\phi'(x)[F_{\varepsilon,\tau}(x) - \Phi(x)]dF_{\varepsilon,\tau}(x)\right)^2} =$$

$$\frac{1}{4\pi^2 \cdot \tilde{B}^2(\varepsilon,\tau)}\sum_{i=1}^{20} A_i(\varepsilon,\tau),$$

где $\tilde{B}(\varepsilon,\tau)$ и $A_i(\varepsilon,\tau)$, $i=1,...,20$, заданные функции параметров ε и τ, которые приведены в [9]. Численные значения асимптотических дисперсий MD-оценок для $F \in \Im_{\varepsilon,\tau}(\Phi)$ при различных весовых функциях приведены в табл. 11.5.

Таблица 11.5

$W,\tau \setminus \varepsilon$	0,00	0,01	0,05	0,10	0,15	0,20	0,25	0,30
$\tau=3$ $W\equiv 1$, 5	1,047 (0,95)	1,071 (0,96)	1,171 (0,97)	1,307 (0,97)	1,458 (0,95)	1,625 (0,94)	1,811 (0,93)	2,019 (0,93)
	1,047 (0,95)	1,078 (0,95)	1,210 (0,93)	1,395 (0,90)	1,607 (0,86)	1,851 (0,83)	2,132 (0,80)	2,459 (0,78)
$\tau=3$ $W=\phi$, 5	1,095 (0,91)	1,117 (0,92)	1,209 (0,93)	1,333 (0,94)	1,470 (0,95)	1,620 (0,95)	1,786 (0,95)	1,972 (0,96)
	1,095 (0,91)	1,122 (0,92)	1,237 (0,91)	1,393 (0,90)	1,562 (0,89)	1,749 (0,88)	1,956 (0,87)	2,187 (0,87)

В этой таблице в скобках приведены абсолютные эффективности MD-оценок, вычисленные по формуле

$$AЭ(F_{\varepsilon,\tau},\hat{\theta}) = \{\sigma^2(F_{\varepsilon,\tau},W)I(f_{\varepsilon,\tau})\}^{-1},$$

где $I(f_{\varepsilon,\tau})$ – информация Фишера относительно параметра сдвига распределений из супермодели $\Im_{\varepsilon,\tau}(\Phi)$. На рис. 11.7 приведены абсолютные эффективности оценок для $F \in \Im_{\varepsilon,\tau}(\Phi)$ при $\tau=3$. На этом рисунке наглядно видно, что MD-оценки с опорной функцией $F_0=\Phi$ и весовой функцией $W(x)=\phi(x)$, так же как и с весовой функцией $W(x)=1$, обеспечивают высокую абсолютную эффективность при изменении пропорции засорения ε, $0\le\varepsilon\le 0{,}3$. При этом абсолютная эффективность для выборочного среднего $\bar{X}$ резко падает, а для выборочной медианы $\bar{X}_{1/2}$ медленно растет, оставаясь на низком уровне.

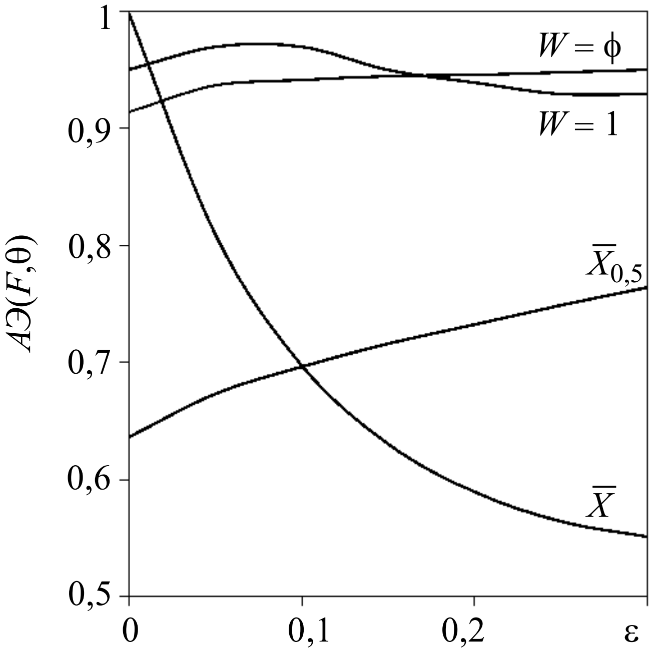

Рис.11.7 Абсолютные эффективности оценок для $F \in \Im_{\varepsilon,\tau}(\Phi)$, $\tau = 3$

Замечание 11.29. Отметим, что в работе Тьюки [11] был приведён подобный график, на котором вместо *MD*-оценок приводится поведение абсолютной эффективности α-урезанных средних. Этот график был наглядной и первой иллюстрацией свойств робастности семейства урезанных средних при отклонениях от гауссовской модели наблюдений в рамках супермодели $\Im_{\varepsilon,\tau}(\Phi)$.

Замечание 11.30. Пусть $X_1,...X_n$ – н.о.р. случайные величины с ф.р. $F(x-\theta)$ и $X_{(1)},...X_{(n)}$ – упорядоченная статистика. Напомним, что *MD*-оценка параметра θ определяется при выбранной нами опорной ф.р. F_0 и заданной весовой функции $W(x)$ в виде решения уравнения $\lambda_{F_n}(\theta) = 0$, где

$$\lambda_{F_n}(\theta) = \frac{2}{n}\sum_{i=1}^{n}\{[(2i-1)/2n] - F_0(X_{(i)}-\theta)\}W(X_{(i)}-\theta).$$

Для решения уравнения $\lambda_{F_n}(\theta) = 0$ рассмотрим дифференцируемую функцию параметра θ в виде

$$\lambda(\theta) = \sum_{i=1}^{n} [(2i-1)/2n \ - F_0(X_{(i)} - \theta)] W(X_{(i)} - \theta),$$

для которой, согласно теореме Лагранжа, выполняется выражение $\lambda(\theta) - \lambda(\theta_0) = (\theta - \theta_0)\lambda'(\theta_0 + \varepsilon(\theta - \theta_0))$, $0 \le \varepsilon \le 1$. Отсюда при $\varepsilon = 0$ уравнение $\lambda(\theta) = 0$ перепишется в виде

$$0 = \lambda(\theta_0) + (\theta - \theta_0)\lambda'(\theta_0)$$

или
$$\theta = \theta_0 - [\lambda(\theta_0)/\lambda'(\theta_0)].$$

Выбрав в качестве начального приближения некоторую робастную оценку $\hat{\theta}_0$ параметра θ (например, выборочную медиану, или оценку Ходжеса – Лемана), получаем формулу для вычисления *MD*-оценки параметра θ в виде

$$\hat{\theta} = \hat{\theta}_0 +$$

$$+ \frac{\sum_{i=1}^{n} [\{(2i-1)/2n\} - F_0(X_{(i)} - \hat{\theta}_0)] W(X_{(i)} - \hat{\theta}_0)}{\sum_{i=1}^{n} \{[(2i-1)/2n \ - F_0(X_{(i)} - \hat{\theta}_0)] W'(X_{(i)} - \hat{\theta}_0) - f_0(X_{(i)} - \hat{\theta}_0) W(X_{(i)} - \hat{\theta}_0)\}},$$

где $\hat{\theta}_0 = med(X_1, ..., X_n)$, F_0 – заданная опорная функция распределения с плотностью f_0 и $W(x)$ – заданная весовая функция. В частности, для $W \equiv 1$ получаем формулу для вычисления MD-оценки в виде

$$\hat{\theta} = \hat{\theta}_0 \ - \ \frac{\sum_{i=1}^{n} [\{(2i-1)/2n\} \ - F_0(X_{(i)} - \hat{\theta}_0)]}{\sum_{i=1}^{n} f_0(X_{(i)} - \hat{\theta}_0)}.$$

Замечание 11.31. Отметим, что *MD*-оценки имеют тесные связи с другими классами робастных оценок . В частности, в разделе 13 выясняются условия, при которых асимптотические дисперсии *MD*-оценок совпадают с асимптотическими дисперсиями, *M*-, *L*- и *R*-

оценок, которые обсуждались в разделах 7–9. Отметим также, что обсуждение свойств *MD*-оценок параметра θ, основанных на использовании взвешенного расстояния Крамера – Мизеса, для заданной *общей параметрической модели* $(\mathrm{X}, \Im_\theta)$ с функцией распределения $F_\theta(x) = F_X(x,\theta)$ приводится в [53]. В этой работе выписываются условия регулярности, при выполнении которых доказывается асимптотическая нормальность *MD*-оценок параметра θ на основе выражения

$$T(F_n) = T(F_\theta) + \frac{1}{n}\sum_{i=1}^{n} IF(X_i; F_\theta, W_\theta) + o_p(n^{-1/2}).$$

Отмечается, что асимптотически эффективная *MD*-оценка параметра θ для заданной параметрической модели $(\mathrm{X}, \Im_\theta)$ определяется весовой функцией

$$W_\theta^*(x) = a[-\partial^2 \ln f(x,\theta)/\partial\theta\,\partial x]/f_\theta(x),$$

а её функция влияния записывается в виде

$$IF(x; F_\theta, W_\theta^*) = \frac{\partial \ln f(x,\theta)/\partial\theta}{\int [\partial \ln f(x,\theta)/\partial\theta]^2\, dF_\theta(x)} = \frac{\partial \ln f(x,\theta)/\partial\theta}{I(f_\theta)},\ x \in R^1.$$

Заметим, что данная функция влияния *MD*-оценки совпадает с функцией влияния МП-оценки максимального правдоподобия. Подводя итог, отметим ещё раз, что *MD*-оценки параметра сдвига θ допускают обобщение для заданной общей параметрической модели $(\mathrm{X}, \Im_\theta)$, они обладает теми же оптимальными свойствами, что и МП-оценки максимального правдоподобия, но при этом *MD*-оценки параметра сдвига θ проявляют ещё дополнительно свойства робастности (см., например, рис. 11.7) в рамках различных супермоделей, в частности при отклонениях от принятой нормальной модели наблюдений. Построение *MD*-оценок параметров положения θ и масштаба σ в рамках параметрической модели $F_{\theta,\sigma}(x) = F((x-\theta)/\sigma)$ обсуждается в работе [81]. Дополнительные сведения относительно свойств *MD*-оценок могут быть найдены в обзорной работе [84].

12. *MD*-ОЦЕНКИ, ОСНОВАННЫЕ НА УРЕЗАННЫХ ВЫБОРКАХ (MD_α-ОЦЕНКИ)

В предыдущем разделе убедились, что свойства *MD*-оценок существенно зависят от выбора весовой функции $W(x)$. Для первого варианта оценивания параметра θ, когда $F = F_0$, оптимальная весовая функция, обеспечивающая эффективность *MD*-оценки, определяется выражением (11.12). В частности, для нормального распределения оптимальная весовая функция не является ограниченной функцией (см. рис. 11.1 в разделе 11) и равна $W(x) = 1/\phi(x)$. Отметим, что использование этой весовой функции приводит к *MD*-оценке, функция влияния которой также не является ограниченной функцией и, следовательно, такая оценка не является *B*-робастной (см. определение 4.15). Поэтому для обеспечения желательных свойств робастности *MD*-оценки следует отказаться от оптимального выбора весовой функции и выбирать её из других соображений. Рассмотрим MD_α-оценки с ограниченной весовой функцией вида

$$W_\alpha(x) = W(x)\, I[\,\xi_\alpha \le x \le \overline{\xi}_\alpha\,],\ W(x) \in W_S, \qquad \mathbf{(12.1)}$$

где α, $0 \le \alpha < 1/2$, – заданный параметр, который характеризует пропорцию урезания исходной выборки $X_1,...,X_n$ и $\xi_\alpha = F^{-1}(\alpha)$, $\overline{\xi}_\alpha = F^{-1}(1-\alpha)$. Отличие от рассматриваемых ранее Андерсоном и Дарлингом ограниченных весовых функций состоит в том, что границы весовой функции (12.1) связаны с ф.р. F наблюдений $X_1,...,X_n$. Далее, в отличие от рассмотренных выше *MD*-оценок, которые вычисляются на основе исходной выборки $X_1,...,X_n$, рассматриваемые MD_α-оценки, (предложенные в [57]), вычисляются по аналогии с α-урезанным средним и R_α-оценками на основе упорядоченной статистики $X_{(1)},...,X_{(n)}$, из которой предварительно удалены $[\alpha n]$ наименьших и $[\alpha n]$ наибольших порядковых статистик, n – объем выборки и α – заданный параметр, $0 \le \alpha < 1/2$.

Рассмотрим взвешенное расстояние Крамера – Мизеса $\rho_{F,F_0}(\theta, W_\alpha)$ с ограниченной весовой функцией $W_\alpha(x)$ вида (12.1), которое определяется выражением

$$\rho_{F,F_0}(\theta, W_\alpha) = \int_{-\infty}^{\infty} [F(x) - F_0(x-\theta)]^2 W_\alpha(x-\theta)dx =$$

$$= \int_{\theta+F^{-1}(\alpha)}^{\theta+F^{-1}(1-\alpha)} [F(x) - F_0(x-\theta)]^2 W(x-\theta)dx =$$

$$= \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} [F(x+\theta) - F_0(x)]^2 W(x)dx. \qquad \textbf{(12.2)}$$

Определим MD_α-оценки из условия минимума эмпирического взвешенного расстояния Крамера – Мизеса $\rho_{F_n,F_0}(\theta, W_\alpha)$, которое получено путем замены в $\rho_{F,F_0}(\theta, W_\alpha)$ функции распределения F на эмпирическую ф.р. F_n и записывается в виде

$$\rho_{F_n,F_0}(\theta, W_\alpha) = \int_{F_n^{-1}(\alpha)}^{F_n^{-1}(1-\alpha)} [F_n(x+\theta) - F_0(x)]^2 W(x)dx. \qquad \textbf{(12.3)}$$

Обозначим $\lambda_{F_n}(\theta, \alpha) = \partial\rho_{F_n,F_0}(\theta, W_\alpha)/\partial\theta$. С учетом этого обозначения, MD_α-оценки являются решением уравнения $\lambda_{F_n}(\theta, \alpha) = 0$, где

$$\lambda_{F_n}(\theta, \alpha) = \frac{2}{n} \sum_{i=[\alpha n]+1}^{n-[\alpha n]} \left[\frac{2i-1}{2n} - F_0(X_{(i)} - \theta) \right] W(X_{(i)} - \theta),$$

$$0 \le \alpha < 1/2. \qquad \textbf{(12.4)}$$

Отметим, что эти оценки могут быть представлены в виде функционала от эмпирической функции распределения, то есть в виде

$\theta_n = T(F_n)$, причем функционал $T(F)$ задается неявно с помощью выражения

$$\int\limits_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} [F(x+T(F)) - F_0(x)]^2 W(x)\, dF(x+T(F)) = 0\,. \qquad \textbf{(12.5)}$$

Асимптотические свойства MD_α-оценок исследуются по такой же схеме, что и для MD-оценок. Получим выражение функции влияния $IF(u; F, F_0, W_\alpha)$ для MD_α-оценки с весовой функцией $W_\alpha(x)$. Для удобства записи используем обозначения $\xi_\alpha = F^{-1}(\alpha)$, $\overline{\xi}_\alpha = F^{-1}(1-\alpha)$, и пусть $F_{\lambda,\,u}(x) = F(x) + \lambda[C(x-u) - F(x)]$, $x, u \in R^1$, $0 \le \lambda \le 1$, а F_λ^{-1} обозначает обратную функцию для ф.р. $F_{\lambda,u}(x)$. Заменим в формуле (12.5) ф.р. F на ф.р. $F_{\lambda,u}(x)$. В результате получим выражение вида

$$\int\limits_{F_\lambda^{-1}(\alpha)}^{F_\lambda^{-1}(1-\alpha)} \{F(x+T(F_\lambda)) + \lambda[C(x+T(F_\lambda)-u) - F(x+T(F_\lambda))] - F_0(x)\} W(x) \times$$

$$\times d\{F(x+T(F_\lambda)) + \lambda[C(x+T(F_\lambda)-u) - F(x+T(F_\lambda))]\} = 0\,. \qquad \textbf{(12.6)}$$

Дифференцируя данное выражение по параметру λ, полагая $\lambda = 0$ и учитывая, что $F_{\lambda,u}(x)|_{\lambda=0} = F(x)$, $T(F_{\lambda,u})|_{\lambda=0} = \theta(F)$, а также $IF(u) = [\partial T(F_{\lambda,\,u}) / \partial\lambda]_{\lambda=0}$, получим выражение

$$\int\limits_{\xi_\alpha}^{\overline{\xi}_\alpha} [f(x+\theta) IF(u) + C(x+\theta-u) - F(x+\theta)] W(x)\, dF(x+\theta) +$$

$$+\int\limits_{\xi_\alpha}^{\overline{\xi}_\alpha} [F(x+\theta) - F_0(x)] W(x)\, d[f(x+\theta) IF(u) + C(x+\theta-u) - F(x+\theta)] +$$

$$+[F(\overline{\xi}_\alpha + \theta) - F_0(x)] W(\overline{\xi}_\alpha) f(\overline{\xi}_\alpha + \theta) \frac{(1-\alpha) - C(\overline{\xi}_\alpha - u)}{f(\overline{\xi}_\alpha)} -$$

$$-[F(\xi_\alpha + \theta) - F_0(x)] W(\xi_\alpha) f(\xi_\alpha + \theta) \frac{\alpha - C(\xi_\alpha - u)}{f(\xi_\alpha)} = 0\,. \qquad \textbf{(12.7)}$$

Для принятых ранее предположений, то есть при $(F, F_0) \in \Im_{S|0}$ и $W \in W_S$, после преобразований в (12.7) получим окончательно выражение для функции влияния MD_α-оценки в виде

$$IF(u; F, F_0, W_\alpha) = A_{F,F_0}(u; W_\alpha) / B_{F,F_0}(W_\alpha),$$

$$0 \le u < \infty,\ 0 \le \alpha < 1/2, \qquad \textbf{(12.8)}$$

где функция $A_{F,F_0}(u; W_\alpha)$ в числителе (12.8) определяется выражением

$$A_{F,F_0}(u; W_\alpha) =$$

$$= \begin{cases} \int_0^u W(x)dF(x) - W(u)[F(u) - F_0(u)], & 0 \le u < \overline{\xi}_\alpha, \\ A_{F,F_0}(\overline{\xi}_\alpha; W_\alpha), & u \ge \overline{\xi}_\alpha, \end{cases} \qquad \textbf{(12.9)}$$

и знаменатель $B_{F,F_0}(W_\alpha) \ne 0$ в (12.8) вычисляется по формуле

$$B_{F,F_0}(W_\alpha) = \int_{\xi_\alpha}^{\overline{\xi}_\alpha} f_0(x)W(x)dF(x) - \int_{\xi_\alpha}^{\overline{\xi}_\alpha} [F(x) - F_0(x)]W'(x)dF(x) +$$

$$+2[F(\overline{\xi}_\alpha) - F_0(\overline{\xi}_\alpha)]W(\overline{\xi}_\alpha)f(\overline{\xi}_\alpha). \qquad \textbf{(12.10)}$$

Отметим, что функции влияния MD_α-оценок ограничены и, следовательно, эти оценки B-робастны и подвержены лишь ограниченному влиянию выбросов в выборке. Асимптотическая дисперсия $\sqrt{n}MD_\alpha$-оценок вычисляется как

$$\sigma^2(F, F_0, W_\alpha) = 2\int_0^\infty IF^2(u; F, F_0, W_\alpha)dF(u) =$$

$$= \frac{2}{B^2_{F,F_0}(W_\alpha)} \left[\int_0^{\overline{\xi}_\alpha} A^2_{F,F_0}(u, W_\alpha)dF(u) + \alpha A^2_{F,F_0}(\overline{\xi}_\alpha) \right]. \qquad \textbf{(12.11)}$$

Для первого варианта оценивания параметра θ, когда $F = F_0$, приведенные формулы (12.8) – (12.11) записываются в виде

$$IF(u; F, W_\alpha) = A_F(u; W_\alpha) / B_F(W_\alpha),\ 0 \le u < \infty,\ 0 \le \alpha < 1/2; \qquad \textbf{(12.12)}$$

$$A_F(u;W_\alpha)=\begin{cases}\int_0^u W(x)dF(x),\ 0\le u<\overline{\xi}_\alpha,\\ A_F(\overline{\xi}_\alpha;W),\quad u\ge\overline{\xi}_\alpha;\end{cases}$$

$$B_F(W_\alpha)=\int\limits_{\xi_\alpha}^{\overline{\xi}_\alpha} f(x)W(x)dF(x);\tag{12.13}$$

$$\sigma^2(F,W_\alpha)=\frac{2}{B_F^2(W_\alpha)}\left[\int\limits_0^{\overline{\xi}_\alpha} A_F^2(u,W_\alpha)dF(u)+\alpha\,A_F^2(\overline{\xi}_\alpha)\right].\tag{12.14}$$

Замечание 12.15. Приобретенные свойства робастности MD_α-оценки могут быть связаны с потерей эффективности оценки. Для сравнения MD- и MD_α-оценок соответственно с весовыми функциями $W(x)$ и $W_\alpha(x)$ рассмотрим асимптотическую относительную эффективность

$$АОЭ_F(W_\alpha:W)=\sigma^2(F,W)/\sigma^2(F,W_\alpha),$$

которая для первого варианта оценивания параметра θ, когда $F=F_0$, записывается в виде

$$АОЭ_F(W_\alpha:W)=\frac{\int_0^1\varphi^2(t)dt}{\int_0^1\varphi_\alpha^2(t)dt}\cdot\frac{\left(\int_\alpha^{1-\alpha}\varphi'(t)f(F^{-1}(t))dt\right)^2}{\left(\int_0^1\varphi'(t)f(F^{-1}(t))dt\right)^2},\tag{12.16}$$

где $\varphi(t)=\int\limits_0^{F^{-1}(t)} W(x)dF(x)$, $0\le t\le 1$, $\varphi_\alpha(t)=\int\limits_0^{F^{-1}(t)} W_\alpha(x)dF(x)$,

$$\varphi_\alpha(t)=\begin{cases}\varphi(t),\ 1/2\le t<1-\alpha,\\ \varphi(1-\alpha),\ 1-\alpha\le t\le 1.\end{cases}\tag{12.17}$$

Изучение асимптотической относительной эффективности $АОЭ_F(W_\alpha:W)$ для различных супермоделей с «упорядоченными» распределениями приводится в работах [58, 59].

13. СВЯЗИ МЕЖДУ *MD*-ОЦЕНКАМИ И *M*-, *L*- И *R*-ОЦЕНКАМИ ПАРАМЕТРА ПОЛОЖЕНИЯ

Обсудим связи между *MD*-оценками, построенными методом минимума взвешенного расстояния Крамера – Мизеса и *M*-, *L*- и *R*-оценками параметра положения θ. Обозначим через C_{MD} класс весовых функций W для *MD*-оценок, который содержит неотрицательные, дифференцируемые, четные функции, удовлетворяющие условию

$$\int \{F(x)(1-F(x))\}^p W(x+c)dx < \infty,\ p>0,\ c\in(-\infty,+\infty).$$

Определим отображения между классами функций C_{MD}, C_M, C_L и C_R в следующем виде:

$$C_{MD} \to C_M\ :\ \psi(x) = \int_0^x W(y)dF(y),\ x\in R^1; \qquad \textbf{(13.1)}$$

$$C_M \to C_{MD}\ :\ W(x) = \psi'(x)/f(x),\ x\in R^1; \qquad \textbf{(13.2)}$$

$$C_{MD} \to C_L\ :\ J(t) = W(F^{-1}(t))f(F^{-1}(t)),\ 0\le t\le 1; \qquad \textbf{(13.3)}$$

$$C_L \to C_{MD}\ :\ W(x) = J(F(x))/f(x),\ x\in R^1; \qquad \textbf{(13.4)}$$

$$C_{MD} \to C_R\ :\ W(x) = \varphi'(F(x)),\ x\in R^1; \qquad \textbf{(13.5)}$$

$$C_R \to C_{MD}\ :\ \varphi(t) = \int_0^{F^{-1}(t)} W(x)dF(x),\ 0\le t\le 1. \qquad \textbf{(13.6)}$$

Теорема 13.7. Пусть $X_1,...,X_n$ – последовательность н.о.р. случайных величин с функцией распределения $F(x-\theta)$. Предполагаем, что ф.р. F непрерывна и симметрична, то есть $F \in \Im_{S|0}$. Тогда, при выполнении отображений (13.1) – (13.6), асимптотическая дисперсия *MD*-оценок параметра θ для случая, когда ф.р. F совпадает с опорной ф.р. F_0, совпадает с асимптотическими дисперсиями *M*-, *L*- и *R*-оценок.

Доказательство этой теоремы осуществляется непосредственной проверкой путем использования формул (10.3), (10.7) и (10.11) при использовании отображений (13.1) – (13.6), с учетом того, что для $F \in \Im_{S|0}$, $W \in C_{MD}$, и для $F = F_0$ асимптотическая дисперсия *MD*-оценок вычисляется по формуле (11.11) (см. работы [9, 56, 59]).

Пример 13.8. Напомним (см. пример 11.13), что для $F \in \Im_{S|0}$, $W \in C_{MD}$ и при $F = F_0$, в классе *MD*-оценок существует асимптотически эффективная оценка параметра θ с весовой функцией $W(x)$ вида

$$W(x) = a\,\frac{d^2\{-\ln f_0(x)\}}{dx^2}\cdot\frac{1}{f_0(x)},$$

где f_0 – плотность опорного распределения F_0 (в данном случае она совпадает с плотностью наблюдений, то есть $f_0 = f$). Рассмотрим супермодель

$$\Im_\varepsilon(G) = \{F : F(x) = (1-\varepsilon)G(x) + \varepsilon H(x)\},$$

определенную в (10.40). В рамках данной супермодели *MD*-оценка является минимаксно-робастной при выборе весовой функции $W(x)$ в виде

$$W(x) = \frac{I[|x| \le a]}{(1-\varepsilon)g(x)}\frac{d^2\{-\ln g(x)\}}{dx^2}, \qquad \textbf{(13.9)}$$

где $g(x)$ – плотность, соответствующая ф.р. *G*, которая выбрана в качестве идеальной модели для супермодели $\Im_\varepsilon(G)$, и параметр *a* при заданном значении ε находится из уравнения

$$\varepsilon/2(1-\varepsilon) = G(a) - [g'(a) + g^2(a)]/g'(a). \qquad \textbf{(13.10)}$$

Замечание 13.11. В литературе приводятся примеры построения минимаксно-робастных оценок параметра θ для различных супер-моделей. Наиболее полно эти вопросы отражены в монографии Хьюбера [5] (см. также работы: *Collinc J.R.* Robustness comparisons of some classes of location parameter estimators // Ann. Inst. Statist. Math. 2000. V. 52. No. 2. P. 351–366; *Wiens D.P.* Robust weighted Cramer-von Mises estimators of location, with minimax variance in ε-contamination neighbourhoods // The Canadian Journal of Statistics. 1987. V. 15. No. 3. P. 269–278); *Collinc J.R. and Wiens D.P.* Minimax properties of M-, R- and L-estimators of location in Levy neighbourhoods // The Annals of Statistics. 1989. V. 17. No. 1. P. 327–336.

14. ОБОБЩЕННЫЕ *L*-ОЦЕНКИ

В данном разделе рассматривается класс статистик, предложенный Серфлингом [39], который включает *U*-статистики и обычные *L*-оценки в виде линейных комбинаций порядковых статистик. Отметим, что многие конкретные оценки, относящиеся к рассмотренным ранее разным классам *M*-, *L*- или *R*-оценок, могут быть объединены в одном классе обобщенных *L*-оценок (*GL*-оценок) и проанализированы с единых позиций при использовании теории *U*-статистик Хёфдинга и дифференциального подхода Мизеса.

Введем необходимые обозначения. Пусть $X_1,...,X_n$ – последовательность н.о.р. случайных величин с ф.р. F, и пусть $\theta(F)$ – допускающий оценку функционал степени m с ядром $h(X_1,...,X_m)$. Обозначим через $W_{(1)},...,W_{(n_m)}$ упорядоченные по возрастанию значения $h(X_{i_1},...,X_{i_m})$, общее число которых равно $n_m = n(n-1)\cdots(n-m+1)$ и которые порождены различными m-наборами индексов $(i_1,...,i_m)$ из множества $\{1,...,n\}$. Следуя работе [39], рассмотрим обобщенные *L*-оценки в виде линейных комбинаций упорядоченных значений $h(X_{i_1},...,X_{i_m})$, то есть в виде

$$\sum_{i=1}^{n_m} C_{ni} W_{(i)}, \qquad \textbf{(14.1)}$$

где C_{ni}, $1 \le i \le n_m$, – заданные константы (весовые коэффициенты). Отметим, что оценки вида (14.1) включают *U*-статистики с ядром $h(X_1,...,X_m)$ и весами $C_{ni} = 1/n_m$ для всех индексов *i*. Кроме того, они включают обычные *L*-оценки с ядром частного вида $h(X) = X$. Для удобства дальнейшей записи будем использовать наблюденную реализацию $x_1,...,x_n$ выборки $X_1,...,X_n$ и для заданного ядра

$h(x_1,...,x_m)$ определим эмпирическую функцию распределения значений $h(x_{i_1},...,x_{i_m})$ как

$$H_n(y) = \frac{1}{n_m}\sum_q I[h(x_{i_1},...,x_{i_m}) \le y],\ y \in R^1, \tag{14.2}$$

где $I[h(x_{i_1},...,x_{i_m}) \le y]$ обозначает единичную функцию, q – суммирование по всем n_m наборам индексов $(i_1,...,i_m)$ из множества $\{1,...,n\}$. Заметим, что $H_n(y)$ для каждого фиксированного значения аргумента y является U-статистикой с ядром $I[h(X_{i_1},...,X_{i_m}) \le y]$ для функции распределения $H_F(y)$, определенной в виде

$$H_F(y) = \int\cdots\int I[h(x_1,...,x_m) \le y]\prod_{i=1}^{m} dF(x_i). \tag{14.3}$$

Как и в случае обычных L-оценок, рассмотренных в разделе 8, для обобщенных L-оценок вида (14.1) весовые коэффициенты C_{ni}, $1 \le i \le n_m$, могут задаваться в различных вариантах. По аналогии с достаточно широким подклассом обычных L-оценок определим подкласс обобщенных L-оценок из (14.1) в виде

$$T(H_n) = \sum_{i=1}^{n_m}\left[\int_{(i-1)/n_m}^{i/n_m} J(t)dt\right] H_n^{-1}(i/n_m) + \sum_{j=1}^{d} a_j H_n^{-1}(p_j), \tag{14.4}$$

где $J(t)$, $0 \le t \le 1$, – заданная функция, определяющая веса C_{ni}, $1 \le i \le n_m$, a_j и p_j, $j = 1,...,d$, – заданные константы, причем $\sum a_j = 1$ и $0 < p_j < 1$ для всех $j = 1,...,d$. Отметим, что $T(H_n)$ является оценкой функционала $T(H_F)$, заданного как

$$T(H_F) = \int_0^1 J(t)H_F^{-1}(t)dt + \sum_{j=1}^{d} a_j H_F^{-1}(p_j). \tag{14.5}$$

Обсудим асимптотические свойства обобщенных L-оценок вида (14.4) по той же схеме, как это было сделано в разделе 8 для обычных L-оценок. Рассмотрим разложение вида

$$T(H_n) = T(H_F) + V_{1n} + R_{1n}, \tag{14.6}$$

которое позволяет описать ошибку оценивания, то есть разность $T(H_n)-T(H_F)$, путем конкретизации аппроксимационной статистики $V_{1n}=d_1T(H_F;H_n-H_F)$ и остаточного члена R_{1n}. Дифференциал Гато первого порядка функционала $T(H_F)$, по аналогии с формулами (8.12) и (8.45), запишется следующим образом:

$$d_1T(H_F;H_n-H_F)=-\int[H_n(y)-H_F(y)]J(H_F(y))dy+$$
$$+\sum_{j=1}^{d}a_j\frac{p_j-H_n(H_F^{-1}(y))}{h_F(H_F^{-1}(p_j))}, \tag{14.7}$$

где $h_F(y)$ – плотность и $H_F^{-1}(t)$ – квантильная функция для функции распределения $H_F(y)$, определенной в (14.5). Здесь не совсем удачные обозначения и следует различать плотность $h_F(y)$ и ядро $h(X_1,...,X_m)$. Из приведенного выражения следует важный вывод. Так как $H_n(y)$ для каждого фиксированного значения аргумента y является U-статистикой с ядром $I[h(X_{i_1},...,X_{i_m})\le y]$ для функции распределения $H_F(y)$, то аппроксимационная статистика $V_{1n}=d_1T(H_F;H_n-H_F)$ вида (14.7) также является U-статистикой с ядром

$$A(x_1,...,x_m)=-\int\{I[h(x_x,...,x_m)\le y]-H_F(y)\}J(H_F(y))dy+$$
$$+\sum_{j=1}^{d}a_j\frac{p_j-I[h(x_1,...,x_m)\le H_F^{-1}(p_j)]}{h_F(H_F^{-1}(p_j))} \tag{14.8}$$

для дифференциала Гато первого порядка $d_1T(H_F;G_F-H_F)$. Отметим, что в частном случае, который соответствует обычной L-оценке с ядром $h(X_1)=X_1$, из приведенного выражения следуют формулы для функций влияния вида (8.17) и (8.47). Таким образом, можно по аналогии заключить, что функция $A(x_1,...,x_m)$ характеризует влияние группы наблюдений $(X_{i_1},...,X_{i_m})$ на ошибку оценивания функционала $T(H_F)$ статистикой $T(H_n)$.

Замечание 14.9. Для оценки функционала $T(H_F)$ от функции распределения H_F, связанной с исходной функцией распределения F наблюдений $X_1,...,X_n$, естественно использовать статистику $T(H_n)$. Однако тот факт, что H_n является U-статистикой, приводит к сложностям , которых нет в обычных L-оценках вида $T(F_n)$. По этой причине представляет интерес рассмотрение функционала $T(H_F)$ в эквивалентной форме с помощью функционала $\tilde{T}(F)$, то есть в виде равенства

$$T(H_F) = \tilde{T}(F). \qquad \textbf{(14.10)}$$

Естественной оценкой функционала $\tilde{T}(F)$ является оценка $\tilde{T}(F_n)$ или эквивалентная оценка $T(H_{F_n})$, где

$$H_{F_n}(y) = \int \cdots \int I[h(x_1,...,x_m) \le y] \prod_{i=1}^{m} dF_n(x_i) =$$

$$= n^{-m} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} I[h(x_{i_1},...,x_{i_m}) \le y]. \qquad \textbf{(14.11)}$$

Отметим, что H_n и H_{F_n} – несколько различные, хотя и тесно связанные оценки (H_n является U-статистикой Хёфдинга, а H_{F_n} – V-статистикой Мизеса, эти оценки асимптотически эквивалентны. Таким образом, оценки $T(H_n)$ и $\tilde{T}(F_n) = T(H_{F_n})$ являются двумя различными оценками (но асимптотически эквивалентными) одного параметра, представленного с помощью функционалов двумя способами (см. равенство 14.10). Хотя H_n является более естественной оценкой, чем H_{F_n}, асимптотический анализ оценок $T(H_{F_n})$ может быть проведен непосредственно с помощью подхода Мизеса и использования понятия «функция влияния». Асимптотическая нормальность оценок $T(H_n)$ доказывается по аналогичной схеме, как это было проделано для обычных L-оценок в разделе 8, с использованием разложения вида (14.6) и теории U-статистик для доказательства асимптотической нормальности аппроксимационной статистики $V_{1n} = d_1 T(H_F; H_n - H_F)$ вида (14.7).

Прежде чем сформулировать строгие результаты, получим выражение для функции влияния оценки $\tilde{T}(F_n) = T(H_{F_n})$ функционала $\tilde{T}(F) = T(H_F)$, определенного в (14.5). Для этого предварительно вычислим дифференциал Гато первого порядка функционала

$$\tilde{T}_1(F) = H_F^{-1}(p),\ 0 < p < 1, \tag{14.12}$$

где $H_F^{-1}(p)$ – обратная функция для функции распределения $H_F(y)$, определенной в (14.3). Предположим, что ф.р. $H_F(y)$ имеет плотность $h_F(y) = dH_F(y)/dy$ и пусть $h_F(H_F^{-1}(p)) > 0$. Используя формулу (14.3), определим функционал $\tilde{T}_1(F)$ неявно с помощью выражения

$$p = \int \cdots \int I[h(x_1,...,x_m) \le \tilde{T}_1(F)] \prod_{i=1}^{m} dF(x_i). \tag{14.13}$$

Далее, действуя стандартным образом, получим

$$d_1\tilde{T}_1(F;G-F) \int \cdots \int \delta[H_F^{-1}(p) - h(x_1,...,x_m)] \prod_{i=1}^{m} dF(x_i) +$$

$$+ \int \cdots \int I[h(x_1,...,x_m) \le H_F^{-1}(p)]\ \{d[G(x_1) - F(x_1)] \prod_{i=2}^{m} dF(x_i) +$$

$$+ d[G(x_2) - F(x_2)] \prod_{i=1, i\ne 2}^{m} dF(x_i) + \cdots + d[G(x_m) - F(x_m)] \prod_{i=1}^{m-1} dF(x_i)\} = 0$$

После преобразований это выражение перепишется в виде

$$d_1\tilde{T}_1(F;G-F) h_F(H_F^{-1}(p)) + m \int \cdots \int I[h(x_1,...,x_m) \le H_F^{-1}(p)] \times$$

$$\times \prod_{i=1}^{m} dF(x_i) dG(x_m) - mp = 0.$$

Отсюда получаем выражение для дифференциала Гато первого порядка функционала $\tilde{T}(F) = T(H_F)$ в виде

$$d_1\tilde{T}_1(F;G-F) =$$

$$= \frac{m\{p - \int \cdots \int I[h(x_1,...,x_m) \le H_F^{-1}(p)] \prod_{i=1}^{m} dF(x_i) dG(x_m)\}}{h_F(H_F^{-1}(p))}. \tag{14.14}$$

Далее, функция влияния оценки $\tilde{T}_1(F_n)$ функционала $\tilde{T}_1(F)$

$$IF(x;F,\tilde{T}_1)=\frac{m\{p-\int\cdots\int I[h(x_1,...,x_{m-1},x)\le H_F^{-1}(p)]\prod_{i=1}^{m-1}dF(x_i)\}}{h_F(H_F^{-1}(p))},$$
$$x\in R^1. \qquad \mathbf{(14.15)}$$

Рассмотрим теперь функционал $\tilde{T}_2(F)$ вида

$$\tilde{T}_2(F)=\int_0^1 J(t)H_F^{-1}(t)dt. \qquad \mathbf{(14.16)}$$

Используя (14.14), получим

$$d_1\tilde{T}_2(F;G-F)=$$
$$=\int_0^1 J(t)\frac{m\{t-\int\cdots\int I[h(x_1,...,x_m)\le H_F^{-1}(p)]\prod_{i=1}^{m-1}dF(x_i)dG(x_m)\}}{h_F(H_F^{-1}(t))}dt. \qquad \mathbf{(14.17)}$$

Далее, после замены переменной $t=H_F(y)$ имеем

$$d_1\tilde{T}_1(F;G-F)=-m\int_{-\infty}^{\infty}J(H_F(y))\{\int\cdots\int I[h(x_1,...,x_m)\le y]\times$$
$$\times\prod_{i=1}^{m-1}dF(x_i)dG(x_m)-H_F(y)\}dy. \qquad \mathbf{(14.18)}$$

Отсюда следует, что функция влияния оценки $\tilde{T}_2(F_n)$ функционала $\tilde{T}_2(F)$

$$IF(x;F,\tilde{T}_2)=-m\int_{-\infty}^{\infty}J(H_F(y))\{\int\cdots\int I[h(x_1,...,x_{m-1},x)\le y]\times$$
$$\times\prod_{i=1}^{m-1}dF(x_i)-H_F(y)\}dy. \qquad \mathbf{(14.19)}$$

Окончательно, используя формулы (14.15) и (14.19), получаем выражение для функции влияния оценки $\tilde{T}(F_n)=T(H_{F_n})$ функционала

$\tilde{T}(F) = T(H_F)$, определенного в (14.5), в виде

$$IF(x;F,\tilde{T})=$$
$$=-m\int_{-\infty}^{\infty} J(H_F(y))\{\int\cdots\int I[h(x_1,...,x_{m-1},x)\le y]\prod_{i=1}^{m-1}dF(x_i)-H_F(y)\}dy+$$
$$+m\sum_{j=1}^{d}a_j\frac{\{p_j-\int\cdots\int I[h(x_1,...,x_{m-1},x)\le H_F^{-1}(p_j)]\prod_{i=1}^{m-1}dF(x_i)\}}{h_F(H_F^{-1}(p_j))},$$
$$x\in R^1. \qquad \textbf{(14.20)}$$

Замечание 14.21. Используя функцию $A(x_1,...,x_m)$ из (14.8), определим функцию $A_1(x)$ в виде

$$A_1(x) = M_F\{A(X_1,...,X_{m-1},x)\}. \qquad \textbf{(14.22)}$$

Сравнивая функцию $A_1(x)$ с (14.20), заключаем, что выполняется равенство

$$m\,A_1(x) = IF(x;F,\tilde{T}). \qquad \textbf{(14.23)}$$

Далее, учитывая, что аппроксимационная статистика V_{1n} в разложении (14.6), равная $V_{1n} = d_1T(H_F;H_n - H_F)$, определяется выражением (14.7) и является U-статистикой с ядром $A(x_1,...,x_m)$, определенным в (14.8), имеем

$$D_F\{\sqrt{n}\,V_{1n}\} = m^2D_F\{A_1(X)\} = D_F\{IF(X;F,\tilde{T})\} =$$
$$= \int IF^2(x;F,\tilde{T})dF(x). \qquad \textbf{(14.24)}$$

Таким образом, при выполнении условий регулярности, которые уточняются при доказательстве сходимости по вероятности к нулю нормированного остаточного члена R_{1n} в разложении (14.6), можно утверждать, что обобщенные L-оценки $T(H_n)$ функционала $T(H_F)$ асимптотически нормальны, то есть выполняется выражение

$$L\{\sqrt{n}[T(H_n) - T(H_F)]/\sigma(T,H_F)\} = N(0,1) \text{ при } n\to\infty,$$

где
$$\sigma^2(T,H_F) = \int IF^2(x;F,\tilde{T})dF(x) \qquad \textbf{(14.25)}$$

и функция влияния $IF(x;F,\tilde{T})$ определена в (14.20).

Для удобства дальнейших ссылок сформулируем теперь строгие результаты в виде теорем.

Теорема 14.26. Пусть функция распределения $H_F(y)$ имеет плотность $h_F(y)$, для которой выполняется неравенство $h_F(H_F^{-1}(p_j)) > 0$, $j = 1,...,d$. Пусть функция $J(t) = 0$ для $t \notin [\alpha,\beta]$, $0 < \alpha < \beta < 1$, а на интервале $[\alpha,\beta]$ эта функция ограничена и непрерывна. Предположим, что $0 < \sigma^2(T,H_F) < \infty$. Тогда выполняется асимптотическое выражение

$$L\{\sqrt{n}[T(H_n) - T(H_F)]/\sigma(T,H_F)\} = N(0,1) \text{ при } n \to \infty,$$

где
$$\sigma^2(T,H_F) = \int IF^2(x;F,\tilde{T})dF(x)$$

и функция влияния $IF(x;F,\tilde{T})$ определена в (14.20).

Доказательство асимптотической нормальности оценки $T(H_n)$ основано на разложении (14.6) с учетом того, что аппроксимационная статистика V_{1n} является асимптотически нормальной U-статистикой. Далее, согласно замечанию (14.21), выполняется вторая формула в (14.25). Условия, накладываемые на ф.р. $H_F(y)$, плотность $h_F(y)$ и функцию $J(t)$, $0 < t < 1$, позволяют доказать, что $\sqrt{n}R_{1n} \to^p 0$ при $n \to \infty$ точно по такой же схеме, как это было сделано для обычных L-оценок. Детали см. в работе [38].

Теорема 14.27. Пусть функция распределения $H_F(y)$ имеет плотность $h_F(y)$, для которой выполняется неравенство $h_F(H_F^{-1}(p_j)) > 0$, $j = 1,...,d$. Кроме того,

$$\int \{H_F(y)[1 - H_F(y)]\}^{1/2} dy < \infty. \quad \textbf{(14.28)}$$

Предположим, что функция $J(t)$ непрерывна на $[0,1]$ и $0 < \sigma^2(T,H_F) < \infty$. Тогда справедливо выражение (14.25).

Доказательство следует с помощью тех же аргументов, что и в предыдущей теореме. Справедливость выражения $\sqrt{n}R_{1n} \to^p 0$ при $n \to \infty$ проверяется так же, как в теореме (8.37). Детали можно найти в работе [39].

Замечание 14.29. Доказательство асимптотической нормальности оценок $\tilde{T}(F_n)$ проводится с помощью разложения вида

$$\tilde{T}(F_n) = \tilde{T}(F) + \frac{1}{n}\sum_{i=1}^{n} IF(X_i; F, \tilde{T}) + R_{1n},$$

где функция влияния $IF(x; F, \tilde{T})$ определена формулой (14.20). При выполнении условий регулярности для ф.р. F, функции $J(t)$ и ядра $h(x_1,...,x_m)$, при которых $n^{1/2}R_{1n} \to^p 0$, согласно центральной предельной теореме и теореме Слуцкого, выполняется выражение

$$L\{\sqrt{n}[\tilde{T}(F_n) - \tilde{T}(F)]/\sigma(F,\tilde{T})\} = N(0,1) \text{ при } n \to \infty,$$

где с учетом (14.24) выполняются равенства

$$\sigma^2(F,\tilde{T}) = \sigma^2(H_F, T) = \int IF^2(x; F, \tilde{T})dF(x).$$

Пример 14.30. Пусть $m = 2$ и функция $h(x_1, x_2) = (x_1 + x_2)/2$. Тогда функция распределения $H_F(y)$, определенная в (14.3), равна

$$H_F(y) = \int\int I[(x_1 + x_2)/2 \le y]dF(x_1)dF(x_2) = \int F(2y - x)dF(x), \ y \in R^1,$$

и, следовательно, плотность $h_F(y)$ ф.р. $H_F(y)$ равна

$$h_F(y) = 2\int f(2y - x)dF(x).$$

Эмпирическая функция распределения $H_n(y)$, определенная в (14.2), в данном случае записывается в виде

$$H_n(y) = \frac{1}{n(n-1)}\sum_{i<j} I[(X_i + X_j)/2 \le y].$$

Определим функционал $T(H_F)$, используя второе слагаемое в (14.5) при $d = 1$, $a_1 = 1$ и $p_1 = 1/2$, то есть в виде

$$T(H_F) = H_F^{-1}(1/2). \quad \textbf{(14.31)}$$

Обобщенной L-оценкой этого функционала в виде $T(H_n)$ является оценка Ходжеса – Лемана, определяемая выражением

$$T(H_n) = HL = med\{(X_i + X_j)/2, \ 1 \le i < j \le n\}. \quad \textbf{(14.32)}$$

Аппроксимационная статистика $V_{1n} = d_1 T(H_F; H_n - H_F)$ в разложении вида (14.6) определяется выражением (14.7), то есть вычисляется по формуле

$$V_{1n} = d_1 T(H_F; H_n - H_F) = \frac{(1/2) - H_n(H_F^{-1}(1/2))}{h_F(H_F^{-1}(1/2))} =$$

$$= \frac{1}{n(n-1)} \sum_{i<j} \frac{(1/2) - I[(X_i + X_j)/2 \le H_F^{-1}(1/2)]}{h_F(H_F^{-1}(1/2))}$$

и является U-статистикой с ядром

$$A(x_1, x_2) = \frac{(1/2) - I[(x_1 + x_2)/2 \le H_F^{-1}(1/2)]}{h_F(H_F^{-1}(1/2))}. \qquad \textbf{(14.33)}$$

Далее, при условии, что $h_F(H_F^{-1}(1/2)) > 0$, для остаточного члена $R_{1n} = T(H_n) - T(H_F) - V_{1n}$ выполняется выражение $n^{1/2} R_{1n} \to^p 0$. Таким образом, HL-оценка из (14.32), согласно теореме 14.26, асимптотически нормальна.

Рассмотрим теперь эквивалентный функционал $\tilde{T}(F)$, который в данном случае может быть определен неявно с помощью выражения

$$\int\int I[(x_1 + x_2)/2 \le \tilde{T}(F)] dF(x_1) dF(x_2) = \frac{1}{2}, \qquad \textbf{(14.34)}$$

или, что эквивалентно, с помощью уравнения вида

$$\int F(2\tilde{T}(F) - x) dF(x) = \frac{1}{2}. \qquad \textbf{(14.35)}$$

Оценка $\tilde{T}(F_n)$ функционала $\tilde{T}(F)$ находится путем решения уравнения

$$\int F_n(2\tilde{T}(F_n) - x) dF_n(x) = \frac{1}{2} \text{ или } \frac{1}{n} \sum_{i=1}^{n} F_n(2\tilde{T}(F_n) - X_i) = \frac{1}{2}.$$

Далее, так как

$$F_n(2\tilde{T}(F_n) - X_i) = \frac{1}{n} \sum_{j=1}^{n} I[X_j \le (2\tilde{T}(F_n) - X_i)],\ i = 1, \ldots, n,$$

то оценка $\tilde{T}(F_n)$ является решением уравнения

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} I[(X_i + X_j)/2 \ \le T(F_n)] = \frac{1}{2},$$

что окончательно определяет оценку $\tilde{T}(F_n)$ как выборочную медиану всех n^2 средних Уолша $(X_i + X_j)/2$, $1 \le i, j \le n$, то есть $\tilde{T}(F_n)$ является оценкой, асимптотически эквивалентной оценке Ходжеса – Лемана, и она записывается в виде

$$\tilde{T}(F_n) = H\tilde{L} = med\{(X_i + X_j)/2,\ 1 \le i, j \le n\}. \qquad \textbf{(14.36)}$$

Функция влияния оценки $\tilde{T}(F_n)$ вида (14.36) функционала $\tilde{T}(F)$, определенного выражением (14.34), согласно (14.20) при $m = 2$ и функции $h(x_1, x_2) = (x_1 + x_2)/2$, равна

$$IF(x; F, \tilde{T}) = \frac{(1/2) - F(2\tilde{T} - x)}{\int f(2\tilde{T} - x)dF(x)},\ x \in R^1. \qquad \textbf{(14.37)}$$

Заметим, что данное выражение совпадает, что естественно, с полученным ранее выражением для функции влияния оценки Ходжеса – Лемана (см. формулу (6.46)). Итак, оценка Ходжеса – Лемана, являясь R-оценкой, может быть получена и проанализирована и как обобщенная L-оценка.

Отметим также, что оценка Ходжеса – Лемана вида (14.32) может быть получена и как решение уравнения для соответствующей U-статистики. В самом деле, пусть U-статистика определяется ядром $h(X_1, X_2) = \mathrm{sign}[(X_1 + X_2)/2]$, тогда оценка Ходжеса – Лемана параметра сдвига θ в одновыборочном варианте является решением уравнения

$$\sum_{i<j} h(X_i - \theta; X_j - \theta) = 0. \qquad \textbf{(14.38)}$$

Асимптотически эквивалентная оценка в виде $\tilde{T}(F_n)$, которая от предыдущей отличается лишь тем, что допускается равенство индексов i и j, определяется функционалом $\tilde{T}(F_n)$, заданным неявно выражением вида

$$\iint \operatorname{sign}[(x+y)/2 - \tilde{T}(F)]dF(x)dF(y) = 0, \qquad \textbf{(14.39)}$$

которое, с учетом равенства $\operatorname{sign}(x) = 2C(x) - 1$, совпадает с (14.34).

Пример 14.40. Обобщенные L-оценки рассматривались в работах [4, 9, 61]. Пусть функция $h(x_1,...,x_m) = (x_1 + ... + x_m)/m$, $m \le n$. Для этой функции GL-оценка, названная в [61] обобщенной оценкой Ходжеса – Лемана, имеет вид

$$HL(m) = med\{(X_{i_1} + ... + X_{i_m})/m,\ 1 \le i_1 < ... < i_m \le n\}. \qquad \textbf{(14.41)}$$

В данном случае функция распределения $H_F(y)$, определенная в (14.3), равна

$$H_F(y) = \int \cdots \int I[x_1 + ... + x_{m-1}) \le my] \prod_{i=1}^{m} dF(x_i) =$$

$$= \int \cdots \int F[my - (x_1 + ... + x_{m-1})] \prod_{i=1}^{m-1} dF(x_i). \qquad \textbf{(14.42)}$$

Плотность $h_F(y)$ ф.р. $H_F(y)$

$$h_F(y) = m \int \cdots \int f[my - (x_1 + ... + x_{m-1})] \prod_{i=1}^{m-1} dF(x_i). \qquad \textbf{(14.43)}$$

Функционал, определяющий $HL(m)$-оценку задается в виде

$$T(H_F) = H_F^{-1}(1/2). \qquad \textbf{(14.44)}$$

Далее, эквивалентный функционал $\tilde{T}(F)$ в данном случае определяется выражением вида

$$\int \cdots \int \left\{ F\left(m\tilde{T}(F) - \sum_{i=1}^{m-1} x_i \right) - \frac{1}{2} \right\} \prod_{i=1}^{m-1} dF(x_i) = 0. \qquad \textbf{(14.45)}$$

Отметим, что функционал $\tilde{T}(F)$ является состоятельным по Фишеру для семейства симметричных распределений (см., например, [9]). Оценка $\tilde{T}(F_n)$ функционала $\tilde{T}(F)$ определяется путем решения уравнения

$$\frac{1}{n_m}\sum_{i_1=1}^{m}\cdots\sum_{i_m=1}^{m} I[(X_{i_1}+\ldots+X_{i_m})\le m\tilde{T}(F_n)]=\frac{1}{2}, \qquad \textbf{(14.46)}$$

что также определяет оценку $\tilde{T}(F_n)$ в виде асимптотически эквивалентном (14.41).

Функция влияния обобщенной оценки Ходжеса – Лемана $HL(m)$ вида (14.41), согласно (14.20), записывается в виде

$$IF(x;F,HL(m))=$$
$$=\frac{(1/2)-\int\cdots\int F[m\tilde{T}(F)-(x_1+\ldots+x_{m-2})-x]dF(x_1)\cdots dF(x_{m-2})}{\int\cdots\int f[m\tilde{T}(F)-(x_1+\ldots+x_m)]dF(x_1)\cdots dF(x_{m-1})},$$
$$x\in R^1. \qquad \textbf{(14.47)}$$

Для случая симметричных распределений, положив без потери общности $\tilde{T}(F)=0$, получим выражение для функции влияния в виде

$$IF(x;F,HL(m))=$$
$$=\frac{\int\cdots\int F[x-(x_1+\ldots+x_{m-2})]dF(x_1)\cdots dF(x_{m-2})-(1/2)}{\int\cdots\int f(x_1+\ldots+x_m)dF(x_1)\cdots dF(x_{m-1})},$$
$$x\in R^1, F\in\mathfrak{I}_S. \qquad \textbf{(14.48)}$$

Обозначим через $F^{(m)*}$ m-кратную свертку ф.р. F с собой, то есть $F^{(m)*}=F*F*\ldots*F$ – (m-раз), причем $F^{(1)*}(x)=F(x)$, $F^{(2)*}(x)=F*F=\int F(x-y)dF(y)$ и $F^{(m)*}(x)=F^{(m-1)*}*F$. Плотность ф.р. $F^{(m)*}(x)$ обозначим через $f^{(m)*}(x)$, то есть $f^{(m)*}$ является m-кратной сверткой плотности f для ф.р. F. Отметим, что с учетом введенных обозначений, выражение (14.45), определяющее задание функционала $\tilde{T}(F)$, перепишется в виде

$$\int F^{(m-1)*}(m\tilde{T}(F)-x)dF(x)=1/2 \text{ или } F^{(m)*}(m\tilde{T}(F))=1/2.$$

Далее, выражение (14.48) для функции влияния компактно запишется в виде

$$IF(x;F,HL(m)) = \frac{F^{(m-1)*}(x)-(1/2)}{\int f^{(m-1)*}(x)dF(x)}, x \in R^1, F \in \Im_S. \qquad \textbf{(14.49)}$$

Отсюда получаем формулу для асимптотической дисперсии $\sqrt{n}HL(m)$-оценок в виде

$$\sigma^2(F,HL(m)) = \int IF^2(x;F,HL(m))dF(x) =$$

$$= \frac{\int\limits_{-\infty}^{\infty}\left(\int\limits_{-\infty}^{x} f^{(m-1)*}(y)dy-(1/2)\right)^2 dF(x)}{\left(\int\limits_{-\infty}^{\infty} f^{(m-1)*}(x)dF(x)\right)^2}, \ F \in \Im_S. \qquad (14.50)$$

Отметим, что в классе обобщенных оценок Ходжеса – Лемана существует асимптотически эффективная оценка параметра сдвига θ в одновыборочной задаче для симметричных распределений, то есть, если ф.р. $F \in \Im_{S|\theta}$ и её плотность f является решением дифференциального уравнения

$$d^2\{-\ln f(x)\}/dx^2 = a\, f^{(m-1)*}(x),$$
$$a - \text{постоянная величина}, \qquad \textbf{(14.51)}$$

тогда $\sigma^2(F,HL(m)) = 1/I(f)$, где $I(f)$ – количество информации Фишера относительно параметра сдвига θ. Убедимся в этом. Пусть $\psi(x) = -f'(x)/f(x) = d\{-\ln f(x)\}/dx$, тогда (14.51) запишется в виде $(1/a)\psi'(x) = f^{(m-1)*}(x)$. Далее, пусть $a = 2\int_{-\infty}^{0}\psi'(x)dx < \infty$, тогда

$$\int_{-\infty}^{\infty} f^{(m-1)*}(x)dF(x) = (1/a)\int_{-\infty}^{\infty}\psi'(x)dF(x) = (1/a)I(f),$$

$$\int_{-\infty}^{x} f^{(m-1)*}(y)dy = (1/a)\{\int_{-\infty}^{0}\psi'(y)dy + \int_{0}^{x} d\psi(y)\} =$$
$$= (1/2) + (1/a)[\psi(x)-\psi(0)] = (1/2) + (1/a)\psi(x).$$

Учитывая, что $\int_{-\infty}^{\infty}\psi^2(x)dF(x)=I(f)$, окончательно получаем

$$\sigma^2(F,HL(m))=\frac{\int\limits_{-\infty}^{\infty}\left(\int\limits_{-\infty}^{x}f^{(m-1)*}(y)dy-(1/2)\right)^2 dF(x)}{\left(\int\limits_{-\infty}^{\infty}f^{(m-1)*}(x)dF(x)\right)^2}=$$

$$=\frac{(1/a)^2\int_{-\infty}^{\infty}\psi^2(x)dF(x)}{(1/a)^2 I^2(f)}=\frac{1}{I(f)}. \qquad \textbf{(14.52)}$$

Отметим, что при $m=2$, решением приведенного уравнения (14.51) является логистическая плотность

$$f(x)=\exp(-x)/\{1+\exp(-x)\}^2,\ x\in R^1,$$

при которой обычная оценка Ходжеса – Лемана (14.36) асимптотически эффективна для параметра сдвига θ (напомним,, что для логистического распределения $\sigma^2(F,HL)=1/I(f)=3$). При больших объемах выборки n и при $m\approx n$, $HL(m)$-оценки совпадают с выборочным средним, которое является эффективной оценкой при нормальном распределении. Численные значения, вычисленные по формуле (4.50) для различных функций распределений F и различных m, приведены в табл. 14.1.

Таблица 14.1

Асимптотическая дисперсия $HL(m)$-оценок для $F\in\mathfrak{I}_S^*$

Ф.р.	$m=1$	$m=2$	$m=3$	$m\approx n$
Гаусса	1,571	1.047	1,019	1,000
Логист.	4,000	3,000	-	3,290
Лапласа	1,000	1,333	1,514	2,000
Коши	2,467	3.290	4,621	∞

Приведем результаты расчетов числовых характеристик робастности $HL(m)$-оценок. Для нормального распределения, то есть при $F=\Phi$, из (14.49) получаем

$$IF(x;\Phi,HL(m)) = \frac{\Phi^{(m-1)*}(x)-(1/2)}{\int \phi^{(m-1)*}(x)d\Phi(x)} = \sqrt{m2\pi}\left[\Phi\left(\frac{x}{\sqrt{m-1}}\right)-\frac{1}{2}\right]=$$

$$=\sqrt{m\pi/2}\ \tilde{\Phi}(x/\sqrt{2(m-1)}), \qquad \textbf{(14.53)}$$

где $\tilde{\Phi}(x)=2\Phi(\sqrt{2}\,x)-1$. Асимптотическая дисперсия $\sqrt{n}HL(m)$-оценки вычисляется по формуле

$$\sigma^2(\Phi,HL(m)) = \int IF^2(x;\Phi,HL(m))d\Phi(x)=$$

$$=m\sqrt{\pi/2}\int_0^\infty \tilde{\Phi}^2(x/\sqrt{2(m-1)})e^{-x^2/2}dx = m\,\mathrm{arctg}(1/\sqrt{m^2-1}). \qquad \textbf{(14.54)}$$

Чувствительность $HL(m)$-оценки к грубым ошибкам равна

$$\gamma^*(\Phi,HL(m)) = \sqrt{m\pi/2}\,, \qquad \textbf{(14.55)}$$

и чувствительность $HL(m)$-оценки к локальным изменениям наблюдений вычисляется по формуле

$$\lambda^*(\Phi,HL(m)) = \sqrt{m/(m-1)}\,. \qquad \textbf{(14.56)}$$

Численные значения характеристик $HL(m)$-оценок при $F=\Phi$ приведены в табл. 14.2.

Таблица 14.2

Числовые характеристики робастности $HL(m)$-оценок при $F=\Phi$

$\backslash m$	1	2	3	4	5	10	$m \approx n \to \infty$
$\gamma^*(\Phi,HL(m))$	1,253	1,772	2,171	2,507	2,802	3,693	∞
$\lambda^*(\Phi,HL(m))$	∞	1,414	1,255	1,155	1,118	1,054	1,000
$\sigma^2(\Phi,HL(m))$	1,571	1,047	1,019	1,011	1,007	1,002	1,000

Отметим, что при гауссовском распределении чувствительность $HL(m)$-оценок к грубым ошибкам возрастает с увеличением m и достигает бесконечного значения (это является следствием неограниченной функции влияния при $m \approx n \to \infty$, см. рис. 14.1, на котором приведены функции влияния $HL(m)$-оценок при $F=\Phi$ и различных m).

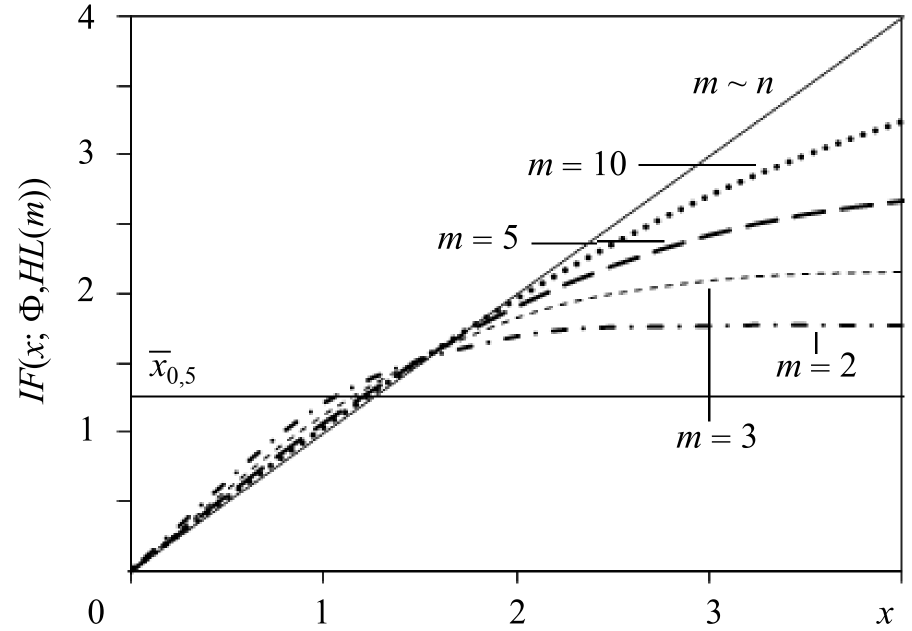

Рис. 14.1. Функции влияния $HL(m)$-оценок для $F = \Phi$

Далее, чувствительность $HL(m)$-оценок к локальным изменениям наблюдений и их асимптотическая дисперсия убывают до своих предельных значений, соответственно равных единице. Отметим также, что глобальная характеристика робастности оценок, называемая «пороговой точкой, или точкой срыва», для $HL(m)$-оценок вычисляется по формуле $\varepsilon^*(HL(m)) = 1 - 2^{-1/m}$ (см. раздел 9). В частности, при $m = 2$ получаем известный результат для точки срыва оценки Ходжеса – Лемана в виде $\varepsilon^*(HL) = 1 - 2^{-1/2} = 0{,}29$.

Пример 14.57. Пусть $m = 2$ и функция $h(x_1, x_2) = |x_1 - x_2|$. Тогда функция распределения $H_F(y)$, определенная в (14.3),

$$H_F(y) = \int\int I[|x_1 - x_2| \le y]dF(x_1)dF(x_2) =$$
$$= \int [F(x+y) - F(x-y)]dF(x), \quad y \in R^1. \qquad \textbf{(14.58)}$$

Плотность $h_F(y)$ ф.р. $H_F(y)$ равна

$$h_F(y) = \int [f(x+y) + f(x-y)]dF(x). \qquad \textbf{(14.59)}$$

Эмпирическая ф.р. $H_n(y)$ значений $|X_i - X_j|$, $1 \le i < j \le n$, определенная в (14.2), в данном случае записывается в виде

$$H_n(y) = \frac{1}{n(n-1)} \sum_{i<j} I[|X_i - X_j| \le y]. \qquad \textbf{(14.60)}$$

Определим функционал $T(H_F)$, используя второе слагаемое в (14.5) при $d = 1$, $a_1 = 1$ и $p_1 = 1/2$, то есть в виде

$$T(H_F) = H_F^{-1}(1/2), \qquad \textbf{(14.61)}$$

где H_F^{-1} – обратная функция для ф.р. $H_F(y)$ из (14.58). Обобщенная L-оценка этого функционала в виде $T(H_n)$, предложенная в [65] (см. также [4, 9]), является выборочной медианой абсолютных разностей, то есть

$$T(H_n) = med\{|X_i - X_j|,\ \ 1 \le i < j \le n\}. \qquad \textbf{(14.62)}$$

Согласно теореме 14.26, оценка $T(H_n)$ вида (14.62) асимптотически нормальна при выполнении неравенства $h(H_F^{-1}(1/2)) > 0$.

Рассмотрим теперь эквивалентный функционал $\tilde{T}(F)$, который в данном случае может быть определен неявно с помощью выражения

$$\int\int I[|x_1 - x_2| \le \tilde{T}(F)]dF(x_1)dF(x_2) = \frac{1}{2}, \qquad \textbf{(14.63)}$$

или, что эквивалентно, с помощью уравнения

$$\int [F(x + \tilde{T}(F)) - F(x - \tilde{T}(F)) - F(x)]dF(x) = 0. \qquad \textbf{(14.64)}$$

Стандартным способом убеждаемся, что дифференциал Гато первого порядка функционала $\tilde{T}(F)$, заданного выражением (14.64), вычисляется по формуле

$$d_1\tilde{T}(F; G - F) = \frac{(1/2) - \int [F(x + \tilde{T}) - F(x - \tilde{T})]dG(x)}{\int [f(x + \tilde{T}) + f(x - \tilde{T})]dF(x)}. \qquad \textbf{(14.65)}$$

Отсюда следует, что функция влияния оценки $\tilde{T}(F_n)$ функционала $\tilde{T}(F)$, заданного выражением (14.64), определяется в виде

$$IF(x;F,\tilde{T}) = \frac{1 + 2F(x-\tilde{T}) - 2F(x+\tilde{T})}{2\int [f(x+\tilde{T}) + f(x-\tilde{T})]\,dF(x)}, x \in R^1. \qquad \textbf{(14.66)}$$

Отметим, что для симметричных распределений функционал $\tilde{T}(F)$ определяется выражением

$$\int F(x+\tilde{T})dF(x) = 3/4, \qquad \textbf{(14.67)}$$

то есть соответствует квантилю уровня $(3/4)$ для ф.р. случайной величины $Y = |X_1 - X_2|$. Асимптотическая дисперсия $\sqrt{n}\,\tilde{T}(F_n)$-оценки вычисляется по формуле

$$\sigma^2(F,\tilde{T}) = \frac{\int [1 + 2F(x-\tilde{T}) - 2F(x+\tilde{T})]^2\,dF(x)}{4\left(\int [f(x+\tilde{T}) + f(x-\tilde{T})]dF(x)\right)^2}. \qquad \textbf{(14.68)}$$

Сравнение характеристик выборочной медианы абсолютных разностей вида (14.62) с другими оценками масштабного параметра приводится в [4, 9] (см. также раздел 16 и работы [82, 83]).

15. ОБОБЩЕННЫЕ *L*-ОЦЕНКИ, ОСНОВАННЫЕ НА УРЕЗАННЫХ ВЫБОРКАХ ($GL_{\alpha\beta}$-ОЦЕНКИ)

В данном разделе рассматривается класс $GL_{\alpha\beta}$-оценок, предложенных в [62]. В этот класс входят обычные L-оценки и обобщенные L-оценки. В частности, в этот класс входит оценка Ходжеса – Лемана и её α-урезанный вариант, а также обобщенные оценки Ходжеса – Лемана, медиана абсолютных отклонений и многие другие.

Пусть $X_1,...,X_n$ – последовательность н.о.р. случайных величин с ф.р. $F(x)$ и плотностью $f(x)$, $x \in R^1$, $X_{(1)},...,X_{(n)}$ – порядковые статистики и $X_{([\alpha n]+1)},...,X_{(n-[\beta n])}$ обозначает $\alpha\beta$-урезанную выборку, α и β – заданные пропорции урезания выборки, причем $0 \le \alpha, \beta \le 1/2$. Обозначим $n_{\alpha\beta} = n - [\alpha n] - [\beta n]$ и $n_{\alpha\beta m} = n_{\alpha\beta}(n_{\alpha\beta} - 1)\cdots(n_{\alpha\beta} - m + 1)$. Пусть задано «ядро» $h(x_1,...,x_m)$, $m < n$, которое является симметричной функцией своих аргументов. Множество m-наборов индексов $(i_i,...,i_m)$, удовлетворяющих условию $\{[\alpha n]+1 \le i_1 <,...,< i_m \le n - [\beta n]\}$, обозначим через $C_{\alpha\beta}$, то есть

$$C_{\alpha\beta} = \{ (i_1,...,i_m) : [\alpha n] + 1 \le i_1 <,...,< i_m \le n - [\beta n] \}. \qquad \textbf{(15.1)}$$

Определим функцию распределения значений $h(X_{(i_1)},...,X_{(i_m)})$, $(i_1,...,i_m) \in C_{\alpha\beta}$, в виде

$$H_{F,\alpha\beta}(y) = (1-\alpha-\beta)^{-m} \int_{\xi_\alpha}^{\xi_{1-\beta}} \cdots \int_{\xi_\alpha}^{\xi_{1-\beta}} I[h(x_1,...,x_m) \le y] \prod_{i=1}^{m} dF(x_i), \qquad \textbf{(15.2)}$$

где $I[A]$ – индикатор события A, $\xi_\alpha = F^{-1}(\alpha)$ и F^{-1} – обратная функция для ф.р. F. Далее плотность функции распределения $H_{F,\,\alpha\beta}(y)$ обозначим через $h_{F,\,\alpha\beta}(y)$. Для данного ядра $h(x_1,...,x_m)$ обозначим через $H_{n,\alpha\beta}$ эмпирическую функцию распределения значений $h(X_{(i_1)},...,X_{(i_m)})$, $(i_1,...,i_m) \in C_{\alpha\beta}$, и определим ее в виде

$$H_{n,\alpha\beta}(y) = (n_{\alpha\beta m})^{-1} \sum_{C_{\alpha\beta}} I[h(X_{(i_1)},...,X_{(i_m)}) \le y]. \tag{15.3}$$

Отметим, что для каждого фиксированного значения аргумента y эмпирическая функция распределения $H_{n,\alpha\beta}(y)$ может рассматриваться как $\alpha\beta$-урезанная U-статистика с ядром $I[h(x_1,...,x_m) \le y]$ для функции распределения $H_{F,\alpha\beta}(y)$. Рассмотрим класс оценок, которые будем кратко называть $GL_{\alpha\beta}$-оценками, в виде

$$GL_{\alpha\beta} = \sum_{j=1}^{d} a_j H_{n,\alpha\beta}^{-1}(p_j), \tag{15.4}$$

где $a_1,...,a_d$ – заданные константы, $0 < p_j < 1$, $j = 1,..., d$, и $H_{n,\alpha\beta}^{-1}$ обозначает квантильную функцию для эмпирической функции распределения $H_{n,\alpha\beta}$. Отметим, что оценки вида (15.4), являются обобщенными оценками в виде линейных комбинаций порядковых статистик, основанными на урезанных выборках. Эти оценки записаны в виде функционала $T(H_{n,\alpha\beta})$ от эмпирической функции распределения $H_{n,\alpha\beta}$, то есть $GL_{\alpha\beta} = T(H_{n,\alpha\beta})$, где функционал $T(\cdot)$ определен на множестве функций распределений $H_{F,\,\alpha\beta}$ и записывается в виде

$$T(H_{F,\,\alpha\beta}) = \sum_{j=1}^{d} a_j H_{F,\alpha\beta}^{-1}(p_j). \tag{15.5}$$

Класс оценок вида (15.4) является достаточно широким. В частности, при $d = 1$, $a_1 = 1$, $p_1 = p$ оценки (15.4) записываются в виде

$$T(H_{n,\alpha\beta}) = H^{-1}_{n,\alpha\beta}(p) = SQ_p\{h(X_{(i_1)},...,X_{(i_m)}) : (i_1,...,i_m) \in C_{\alpha\beta}\}, \quad \textbf{(15.6)}$$

где символ SQ_p обозначает выборочный квантиль уровня p, $0 < p < 1$. При $p = 1/2$ оценки (15.6) имеют вид

$$T(H_{n,\alpha\beta}) = med\{h(X_{(i_1)},...,X_{(i_m)}) : (i_1,...,i_m) \in C_{\alpha\beta}\}, \quad \textbf{(15.7)}$$

где символ «med», как обычно, обозначает выборочную медиану. В этот класс оценок также входят обобщенные оценки Ходжеса – Лемана, предложенные в [66, 67], которые соответствуют выбору ядра $h(x_1,...,x_m)$ в виде $h(x_1,...,x_m) = (x_1 + ... + x_m)/m$, $m < n$, и определяются следующим образом:

$$HL_{\alpha\beta}(m) = med\left\{ \left(X_{(i_1)} + ... + X_{(i_m)}\right)/m : (i_1,...,i_m) \in C_{\alpha\beta}\right\}. \quad \textbf{(15.8)}$$

Отметим также, что при $m = 2$, $p = 1/2$ и при выборе ядра $h(x_1,...,x_m)$ в виде $h(x_1,x_2) = |x_1 - x_2|$ оценки (15.7) записываются следующим образом:

$$T^*\left(H_{n,\alpha\beta}\right) = med\left\{ \left|X_{(i)} - X_{(j)}\right| : [\alpha n] + 1 \le i < j \le n - [\beta n] \right\}. \quad \textbf{(15.9)}$$

Эти оценки были предложены и изучены в [68]. Таким образом, представляет интерес изучение асимптотических распределений достаточно общего класса оценок вида (15.4).

Как и в предыдущем разделе, представление функционала $T(H_{F,\alpha\beta})$ с помощью эквивалентного функционала $\tilde{T}_{\alpha\beta}(F)$ от исходной функции распределения F позволяет получить выражения для функций влияния оценок (15.4) и их асимптотических дисперсий. Изучение асимптотических распределений $GL_{\alpha\beta}$-статистик основано на использовании разложения вида

$$T(H_{n,\alpha\beta}) = T(H_{F,\alpha\beta}) + d_1T(H_{F,\alpha\beta}; H_{n,\alpha\beta} - H_{F,\alpha\beta}) + R_{1n}, \quad \textbf{(15.10)}$$

где R_{1n} – остаточный член разложения, $d_1T(\cdot)$ – дифференциал Гато первого порядка функционала $T(H_{F,\alpha\beta})$ в точке $H_{F,\alpha\beta}$ по направлению эмпирической функции $H_{n,\alpha\beta}$. Рассмотрим сначала оценки вида (15.6), которые являются оценками функционала $T\left(H_{F,\alpha\beta}\right)=H^{-1}_{F,\alpha\beta}(p)$, $0<p<1$. В частности, при $d=1$, $a_1=1$, $p_1=p$ функционал $T(H_{F,\alpha\beta})$ может быть представлен эквивалентным функционалом $\tilde{T}_{\alpha\beta}(F)$ от исходной ф.р. F, который задается неявно с помощью выражения вида

$$\int\limits_{\xi_\alpha}^{\xi_{1-\beta}} \cdots \int\limits_{\xi_\alpha}^{\xi_{1-\beta}} \{I[h(x_1,...,x_m)\le\tilde{T}_{\alpha\beta}(F)]-(1-\alpha-\beta)^m p\}\prod_{i=1}^{m} dF(x_i)=0. \quad \textbf{(15.11)}$$

При замене функции распределения F на эмпирическую функцию распределения F_n получаем оценки $\tilde{T}_{\alpha\beta}(F_n)$, являющиеся V-статистиками Мизеса, асимптотически эквивалентные оценкам (15.6). Эквивалентное представление функционала $T(H_{F,\alpha\beta})$ с помощью функционала $\tilde{T}_{\alpha\beta}(F)$, то есть равенство $T(H_{F,\alpha\beta})=\tilde{T}_{\alpha\beta}(F)$, позволяет конкретизировать дифференциал Гато $d_1T(\cdot)$ и функцию влияния $IF\{x;F,H^{-1}_{n,\alpha\beta}(p)\}$ оценок (15.6). Прежде чем формулировать основной результат, получим общее выражение для дифференциала Гато первого порядка $d_1\tilde{T}_{\alpha\beta}(F;G-F)$ функционала $\tilde{T}_{\alpha\beta}(F)$ в точке F по направлению G, где $F,G\in\Im$ – множество абсолютно непрерывных распределений, и для простоты записи введем следующие обозначения:

$$q_F(x;\tilde{T}_{\alpha\beta})=$$
$$=I[\xi_\alpha\le x\le\xi_{1-\beta}]\int\limits_{\xi_\alpha}^{\xi_{1-\beta}} \cdots \int\limits_{\xi_\alpha}^{\xi_{1-\beta}} I[h(x_1,...,x_{m-1},x)\le\tilde{T}_{\alpha\beta}(F)]\prod_{i=1}^{m-1} dF(x_i) \quad \textbf{(15.12)}$$

и
$$A_{\alpha\beta}=q_F(\xi_\alpha;\tilde{T}_{\alpha\beta}),\ B_{\alpha\beta}=q_F(\xi_{1-\beta};\tilde{T}_{\alpha\beta}). \quad \textbf{(15.13)}$$

Для вычисления дифференциала Гато $d_1\tilde{T}_{\alpha\beta}(F;G-F)$ заменим ф.р. F в (15.11) на $F_\lambda=F+\lambda(G-F)$, $0<\lambda<1$. Дифференцируя полученное выражение по параметру λ и полагая λ равным нулю, получим следующее выражение:

$$d_1\tilde{T}_{\alpha\beta}(F;G-F)\int\limits_{\xi_\alpha}^{\xi_{1-\beta}}\cdots\int\limits_{\xi_\alpha}^{\xi_{1-\beta}}\delta[\tilde{T}_{\alpha\beta}-h(x_1,...,x_{m-1},x)\leq\tilde{T}_{\alpha\beta}(F)]\prod_{i=1}^{m-1}dF(x_i)+$$

$$+m\int\limits_{\xi_\alpha}^{\xi_{1-\beta}}\cdots\int\limits_{\xi_\alpha}^{\xi_{1-\beta}}I[h(x_1,...,x_{m-1},x_m)\leq\tilde{T}_{\alpha\beta}(F)]\prod_{i=1}^{m-1}dF(x_i)dG(x_m)-mp(1-\alpha-\beta)^m+$$

$$+mB_{\alpha\beta}(1-\beta-G(\xi_{1-\beta}))-mA_{\alpha\beta}(\alpha-G(\xi_\alpha))=0, \qquad \textbf{(15.14)}$$

где $\delta[\cdot]$ – дельта-функция Дирака. Из полученного выражения (15.14) с учетом того, что

$$IF\{x;F,H_{n,\alpha\beta}^{-1}(p)\}=d_1\tilde{T}_{\alpha\beta}(F;I[x\leq x_m]-F), \qquad \textbf{(15.15)}$$

получаем выражение для функций влияния оценок (15.6) в виде

$$IF\{x;F,H_{n,\alpha\beta}^{-1}(p)\}=$$

$$=\frac{\begin{gathered}-m\{q_F(x,\tilde{T}_{\alpha\beta})-p(1-\alpha-\beta)^m+B_{\alpha\beta}(1-\beta-I[x\leq\xi_{1-\beta}])-\\-A_{\alpha\beta}(\alpha-I[x\leq\xi_\alpha])\}\end{gathered}}{\int\limits_{\xi_\alpha}^{\xi_{1-\beta}}\cdots\int\delta[\tilde{S}_{\alpha\beta}-h(x_1,...,x_m)]\prod_{i=1}^{m}dF(x_i)}. \qquad \textbf{(15.16)}$$

Теорема 15.17. Предположим, что плотность $h_{F,\,\alpha\beta}(y)$ ф.р. $H_{F,\,\alpha\beta}(y)$ удовлетворяет условию $h_{F,\alpha\beta}\left(H_{F,\alpha\beta}^{-1}\left(p_j\right)\right)>0,\ 0<p_j<1$, $j=1,...,d$. Тогда случайные величины

$$\sqrt{n}\left(T\left(H_{n,\alpha\beta}\right)-T(H_{F,\alpha\beta})\right)/\sigma\left(F;T(H_{F,\alpha\beta})\right)$$

имеют асимптотически стандартное нормальное распределение, где

$$\sigma^2\left(F,T(H_{F,\alpha\beta})\right)=Var\left\{IF\left(X;F,T(H_{F,\alpha\beta})\right)\right\}=$$

$$=\int\limits_{-\infty}^{\infty}IF^2(x;F,T(H_{F,\alpha\beta}))dF(x), \qquad \textbf{(15.18)}$$

и функция влияния $IF\{x;F,T(H_{F,\alpha\beta})\}$ оценок $T(H_{n,\alpha\beta})$ записывается в виде

$$IF\{x;F,H_{n,\alpha\beta}^{-1}\}=$$

$$=\frac{\sum_{j=1}^{d} a_j m\{-q_F(x,\tilde{T}_{\alpha\beta})+p_j(1-\alpha-\beta)^m-B_{\alpha\beta}(1-\beta-I[x\le\xi_{1-\beta}])+ +A_{\alpha\beta}(\alpha-I[x\le\xi_\alpha])\}}{\int_{\xi_\alpha}^{\xi_{1-\beta}}\cdots\int\delta[\tilde{T}_{\alpha\beta}-h(x_1,...,x_m)]\prod_{i=1}^{m}dF(x_i)}. \quad \mathbf{(15.19)}$$

Доказательство теоремы следует с помощью тех же аргументов, что и в теореме (14.27) с использованием разложения (15.10) и полученного выражения (15.16). Справедливость выражения $\sqrt{n}R_{1n}\to^p 0$ при $n\to\infty$ проверяется, как в теореме 8.37. Детали можно найти в работах [4, 29, 38, 39].

Пример 8.5.20. Пусть функция $h(x_1,...,x_m)=(x_1+...+x_m)/m$, $m\le n$, и $\alpha=\beta$. Обобщенная L-оценка на урезанной выборке, предложенная в [67] и названная обобщенной α-урезанной оценкой Ходжеса – Лемана, имеет вид

$$HL_\alpha(m)=med\{(X_{(i_1)}+...+X_{(i_m)})/m, \\ [\alpha n]\le i_1<...<i_m\le n-[\alpha n]\}. \quad \mathbf{(15.21)}$$

В данном случае функционал $\tilde{T}_{\alpha\alpha}(F)$ определяется выражением

$$\int_{F^{-1}(\alpha)}^{F^{=1}(1-\alpha)}\cdots\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} F\left(m\tilde{T}_{\alpha\alpha}(F)-\sum_{i=1}^{m-1}x_i\right)\prod_{i=1}^{m-1}dF(x_i)=\frac{(1-2\alpha)^{m-1}}{2}; \quad \mathbf{(15.22)}$$

V-оценка Мизеса $\tilde{T}_{\alpha\alpha}(F_n)$ функционала $\tilde{T}_{\alpha\alpha}(F)$ асимптотически эквивалентна оценке (15.21) и записывается в виде

$$\tilde{T}_{\alpha\alpha}(\tilde{F})=H\tilde{L}_\alpha(m)=med\{(X_{i_1}+...+X_{i_m})/m, \\ [\alpha n]\le i_1,...,i_m\le n-[\alpha n]\}. \quad \mathbf{(15.23)}$$

Из формулы (15.16) следует, что функция влияния обобщенной α-урезанной оценки Ходжеса – Лемана определяется как

$$IF(x;F,HL_\alpha(m)) =$$

$$=\frac{\begin{gathered}[(1-2\alpha)^{m-1}/2]-V_F(x,\tilde{T}_{\alpha\alpha})-V_F(\xi_{1-\alpha},\tilde{T}_{\alpha\alpha})\{1-\alpha-I[x\le\xi_{1-\alpha}]\}+\\ +V_F(\xi_\alpha,\tilde{T}_{\alpha\alpha})\{\alpha-I[x\le\xi_\alpha]\}\end{gathered}}{\int_{\xi_\alpha}^{\xi_{1-\alpha}}\cdots\int_{\xi_\alpha}^{\xi_{1-\alpha}} f(m\tilde{T}_{\alpha\alpha}-(x_1+...+x_{m-2})-x)dF(x_1)\cdots dF(x_{m-2})}, \quad \mathbf{(15.24)}$$

где
$$V_F(x;\tilde{T}_{\alpha\alpha})=$$

$$=I[\xi_\alpha \le x\le\xi_{1-\alpha}]\int\limits_{\xi_\alpha}^{\xi_{1-\alpha}}\cdots\int\limits_{\xi_\alpha}^{\xi_{1-\alpha}} F(m\tilde{T}_{\alpha\alpha}-(x_1,...,x_{m-2})-x)\prod_{i=1}^{m-2}dF(x_i). \quad \mathbf{(15.25)}$$

Отметим, что из приведенной формулы (15.24) при соответствующих значениях параметров m и α следуют все полученные ранее выражения для функций влияния модифицированных вариантов оценок Ходжеса – Лемана.

16. *U*-СТАТИСТИКИ, ОСНОВАННЫЕ НА УРЕЗАННЫХ ВЫБОРКАХ ($U_{\alpha\beta}$-ОЦЕНКИ)

В класс U-статистик входят многие конкретные оценки, представляющие практический интерес. Обобщение класса U-статистик, описанное в работе [49], приводит к рассмотрению U-статистик, основанных на урезанных выборках, что позволяет изучать многие известные в теории робастности оценки с единых позиций и открывает широкие возможности для построения новых оценок.

Пусть $X_1,...,X_n$ – последовательность н.о.р. случайных величин с ф.р. $F(x)$ и плотностью $f(x)$, $x \in R^1$, $X_{(1)},...,X_{(n)}$ – порядковые статистики и $X_{([\alpha n]+1)},...,X_{(n-[\beta n])}$ обозначает $\alpha\beta$-урезанную выборку, α и β – заданные пропорции урезания выборки, причем $0 \le \alpha, \beta \le 1/2$. Обозначим $n_{\alpha\beta} = n - [\alpha n] - [\beta n]$. Пусть задано «ядро» $h(X_1,...,X_m)$, $m < n$, которое является симметричной функцией своих аргументов. Множество m-наборов индексов $(i_i,...,i_m)$, удовлетворяющих условию $\{\, [\alpha n]+1 \le i_1 <,...,< i_m \le n-[\beta n] \,\}$, обозначим через $C_{\alpha\beta}$, то есть

$$C_{\alpha\beta} = \{\, (i_1,...,i_m) : [\alpha n]+1 \le i_1 <,...,< i_m \le n-[\beta n] \,\}. \qquad \textbf{(16.1)}$$

Следуя работам [48, 49], рассмотрим U-статистики, основанные на $\alpha\beta$-урезанных выборках, которые определяются в виде

$$U_{n,\alpha\beta} = \binom{n_{\alpha\beta}}{m}^{-1} \sum_{C_{\alpha\beta}} h(X_{(i_1)},...,X_{(i_m)}). \qquad \textbf{(16.2)}$$

Рассмотрим для простоты случай $\alpha=\beta$ и переобозначим множество $C_{\alpha\beta}$ из (16.1) через

$$C_\alpha=\{(i_1,...,i_m):\ [\alpha n]+1\le i_1<,...,<i_m\le n-[\alpha n]\}.$$

Далее, пусть $n_\alpha=n-2[\alpha n]$ и

$$g_\alpha(x;F)=\frac{I[F^{-1}(\alpha)\le x\le F^{-1}(1-\alpha)]}{(1-2\alpha)^m}\times$$
$$\times\int\limits_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)}\cdots\int\limits_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)}h(x_1,...,x_{m-1},x)\prod_{i=1}^{m-1}dF(x_i).\qquad\textbf{(16.3)}$$

Пусть $\xi_\alpha=F^{-1}(\alpha)$, $\overline{\xi}_\alpha=F^{-1}(1-\alpha)$ и значения функции $g_\alpha(x;F)$ в точках ξ_α и $\overline{\xi}_\alpha$ обозначим соответственно в виде

$$A_\alpha=-g_\alpha(\xi_\alpha,F),\ B_\alpha=g_\alpha(\overline{\xi}_\alpha,F).\qquad\textbf{(16.4)}$$

Среднее значение и дисперсию $g_\alpha(X;F)$ соответственно обозначим в виде

$$U_\alpha(F)=M_F\{g_\alpha(X,F)\}=$$
$$=\frac{1}{(1-2\alpha)^m}\int\limits_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)}\cdots\int\limits_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)}h(x_1,...,x_m)\prod_{i=1}^{m}dF(x_i)\qquad\textbf{(16.5)}$$

и
$$\Delta_\alpha(F)=D_F\{g_\alpha(X,F)\}=\int\limits_{\xi_\alpha}^{\overline{\xi}_\alpha}g_\alpha^2(x,F)dF(x)-U_\alpha^2(F).\qquad\textbf{(16.6)}$$

Теорема 16.7. Предположим, что ф.р. F имеет плотность f, которая непрерывна и ограничена в точках ξ_α и $\overline{\xi}_\alpha$. Пусть функция $g_\alpha(x;F)$ непрерывна в точках ξ_α, $\overline{\xi}_\alpha$. Предположим, что для некоторых $a<\xi_\alpha$ и $b>\overline{\xi}_\alpha$ выполняется условие

$$\sup_{a\le x_1,...,x_m\le b}|h(x_1,...,x_m)|=M_0<\infty.\qquad\textbf{(16.8)}$$

Тогда при условии, что $\sigma^2(U_\alpha,F)>0$, статистика $U_{n,\alpha}$ из (16.2) имеет асимптотически нормальное распределение, то есть выполняется выражение

$$L\{\sqrt{n}[U_{n\alpha} - U_\alpha(F)] / \sigma(U_\alpha, F)\} = N(0,1) \text{ при } n \to \infty, \qquad \textbf{(16.9)}$$

где
$$\sigma^2(U_\alpha, F) = m^2\{\Delta_\alpha(F) + \alpha(1-\alpha)(A_\alpha^2 + B_\alpha^2) +$$
$$+2\alpha U_\alpha(F)(A_\alpha - B_\alpha) + 2\alpha^2 A_\alpha B_\alpha\}. \qquad \textbf{(16.10)}$$

Доказательство. При выполнении предположений для плотности f, функции $g_\alpha(x;F)$ и условия (16.8) для ядра $h(x_1,...,x_m)$ применимы леммы 2.1 – 2.4 из работы [49], согласно которым выполняется выражение

$$U_{n,\alpha} = U_\alpha(F) + n^{-1}\sum_{i=1}^{n} IF(X_i; F, U_\alpha) + o_p(n^{-1/2}), \qquad \textbf{(16.11)}$$

где $IF(x;F,U_\alpha)$ – функция влияния оценки $U_\alpha(F_n)$ функционала $U_\alpha(F)$, определенного формулой (16.5). Из выражения (16.11) следует, согласно теореме (6.20), асимптотическая нормальность $U_{n,\alpha}$-статистик. Справедливость формулы (16.10) проверяется непосредственно, путем вычисления функции влияния $IF(x;F,U_\alpha)$, с учетом того, что

$$\int IF(x;F,U_\alpha)dF(x) = 0$$

и
$$\sigma^2(U_\alpha, F) = \int IF^2(x;F,U_\alpha)dF(x). \qquad \textbf{(16.12)}$$

Для завершения доказательства осталось получить выражение для функции влияния $IF(x;F,U_\alpha)$ и воспользоваться формулами (16.12).

Получим выражение для функции влияния $IF(x;F,U_\alpha)$. Для этого введем ф.р. $F_{\lambda,x}(y)$ в виде $F_{\lambda,x}(y) = (1-\lambda)F(y) + \lambda C(y-x)$, $0 \le \lambda \le 1$, $-\infty < x, y < \infty$. Обозначим обратную функцию для ф.р. $F_{\lambda,x}(y)$ через $F_{\lambda,x}^{-1}(t)$, $0 \le t \le 1$. Используя формулу (6.16), получаем

$$IF(x;F,U_\alpha) = \frac{\partial U_\alpha(F_{\lambda,x})}{\partial\lambda}\Big|_{\lambda=0} = \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} \frac{\partial}{\partial\lambda}\left[g_\alpha(x_1, F_{\lambda,x})dF_{\lambda,x}(x_1)\right]_{\lambda=0} +$$

$$+g_\alpha(F_{\lambda,x}^{-1}(1-\alpha),F_{\lambda,x})f_{\lambda,x}(F_{\lambda,x}^{-1}(1-\alpha))|_{\lambda=0}\ \frac{(1-\alpha)+I[x\le F^{-1}(1-\alpha)]}{f(F^{-1}(1-\alpha))}-$$

$$-g_\alpha(F_{\lambda,x}^{-1}(\alpha),F_{\lambda,x})f_{\lambda,x}(F_{\lambda,x}^{-1}(\alpha))|_{\lambda=0}\ \frac{\alpha+I[x\le F^{-1}(\alpha)]}{f(F^{-1}(\alpha))}. \quad \textbf{(16.13)}$$

Вычисляя первое слагаемое и выполняя необходимые преобразования в последнем выражении, окончательно получаем

$$IF(x;F,U_\alpha)=m\{g_\alpha(x,F)-U_\alpha(F)+A_\alpha(\alpha-I[x\le\xi_\alpha])+$$
$$+B_\alpha(1-\alpha-I[x\le\overline{\xi}_\alpha])\}. \quad \textbf{(16.14)}$$

Используя полученное выражение (16.14) и вторую формулу в (16.12), получаем (16.10). Доказательство завершено.

Пример 16.15. Пусть задан параметр α, $0\le\alpha<1/2$, и ядро $h(x)=x$. Для такого ядра U-статистика, основанная на урезанной выборке, является α-урезанным средним, которое записывается в виде

$$U_{n\alpha}=\overline{X}_\alpha=\frac{1}{n-2[\alpha n]}\sum_{i=[\alpha n]+1}^{n-[\alpha n]}X_{(i)}\ ,\ 0\le\alpha<1/2. \quad \textbf{(16.16)}$$

В данном случае функция $g_\alpha(x,F)$ из (16.3) записывается в виде $g_\alpha(x,F)=x\,I[\xi_\alpha\le x\le\overline{\xi}_\alpha]/(1-2\alpha)$, а функционал $U_\alpha(F)$ из (16.5) определяется выражением

$$U_\alpha(F)=M_F\{g_\alpha(X,F)\}=\frac{1}{1-2\alpha}\int_{\xi_\alpha}^{\overline{\xi}_\alpha}xdF(x),$$

$$\xi_\alpha=F^{-1}(\alpha),\ \overline{\xi}_\alpha=F^{-1}(1-\alpha),\ 0\le\alpha<1/2.$$

Рассмотрим случай симметричных распределений, то есть $F\in\Im_{S|0}$, и предположим, что ф.р. F имеет непрерывную и ограниченную в точках ξ_α, $\overline{\xi}_\alpha$ плотность f. При выполнении этих предположений применима теорема (16.7), согласно которой $U_{n\alpha}$-оценки из (16.16) асимптотически нормальны, то есть выполняется выражение (16.9), и асимптотическая дисперсия $\sqrt{n}\,U_{n\alpha}$-оценок, вычисленная по формуле (16.10), с учетом того, что

$$\Delta_\alpha(F) = D_F\{g_\alpha(X,F)\} = \frac{1}{(1-2\alpha)^2}\int\limits_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x^2 dF(x),$$

$$A_\alpha = -F^{-1}(\alpha)/(1-2\alpha),\ B_\alpha = F^{-1}(1-\alpha)/(1-2\alpha),$$

$$F^{-1}(1-\alpha) = -F^{-1}(\alpha),$$

запишется в виде

$$\sigma^2(F,U_\alpha) = \frac{1}{(1-2\alpha)^2}\left\{\int\limits_{F^{-1}(\alpha)}^{F^{-}(1-\alpha)} x^2 dF(x) + 2\alpha[F^{-1}(\alpha)]^2\right\} =$$

$$= \frac{1}{(1-2\alpha)^2}\{\int\limits_{\alpha}^{1-\alpha}[F^{-1}(t)]^2 dt + 2\alpha[F^{-1}(1-\alpha)]^2\}.$$

Отметим, что приведенное выражение, полученное из общей формулы (16.10) для $U_{n\alpha}$-оценок, совпадает с полученной ранее формулой (8.61) для $\overline{X}_\alpha$-оценок.

Пример 16.17. Обсуждение средней разности Джини как меры разброса случайных величин и её связь с кривой Лоренца приводится в [20] (см. также [9]). Функционал $T(F)$, определяющий среднюю разность Джини, записывают в разных вариантах. Обычно, его записывают в виде

$$\Delta_F = T(F) = \int\int |x-y|\, dF(x)dF(y). \qquad \textbf{(16.18)}$$

Другая форма записи функционала $T(F)$ имеет вид

$$\Delta_F = T(F) = \int\limits_0^1 F^{-1}(t)J(t)dt,\ J(t) = 4t-2,\ 0 \le t \le 1. \qquad \textbf{(16.19)}$$

Выборочная оценка $\Delta_n = T(F_n)$ средней разности Джини Δ_F вида (16.18), построенная по выборке $X_1,...,X_n$ методом подстановки, записывается в виде

$$\Delta_n = T(F_n) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|X_i - X_j|. \qquad \textbf{(16.20)}$$

Если же при построении оценки не используются совпадающие индексы (i, j), то оценку записывают в асимптотически эквивалентном варианте вида

$$\tilde{\Delta}_n = \frac{1}{n(n-1)} \sum_{i \neq j=1}^{n} | X_i - X_j | . \qquad \textbf{(16.21)}$$

Отметим, что представление функционала $T(F)$ в виде (16.19), позволяет записать оценку в виде линейной комбинации порядковых статистик $X_{(1)},..., X_{(n)}$, то есть в виде

$$\hat{\Delta} = \sum_{i=1}^{n} a_{ni} X_{(i)} , \qquad \textbf{(16.22)}$$

где весовые коэффициенты a_{ni} вычисляются по формуле

$$a_{ni} = \int_{(i-1)/n}^{i/n} J(t)dt = \int_{(i-1)/n}^{i/n} (4t-2)dt = \frac{2i-1-n}{n^2} \qquad \textbf{(16.23)}$$

и $J(t) = 4t - 2$, $0 \le t \le 1$, – функция, определяющая L-оценку (см. раздел 8).

Асимптотические свойства выборочных оценок средней разности Джини приведены в [9]. Отметим, что оценка $\Delta_n = T(F_n)$ средней разности Джини $\Delta_F = T(F)$ асимптотически нормальна, то есть выполняется выражение

$$L\{\sqrt{n}(\Delta_n - \Delta_F)/\sigma_F(\Delta_n)\} = N(0,1) \text{ при } n \to \infty , \qquad \textbf{(16.24)}$$

где асимптотическая дисперсия $\sqrt{n}\Delta_n$-оценки вычисляется по формуле

$$\sigma_F^2(\Delta_n) = \int IF^2(x; F, \Delta_F) dF(x) =$$
$$= \int \varphi_F^2(x) dF(x) - \Big(\int \varphi_F(x) dF(x) \Big)^2 . \qquad \textbf{(16.25)}$$

Здесь функция $\varphi_F(x)$ имеет вид

$$\varphi_F(x) = 2x(2F(x)-1) - 4\mu_F(x), \ \mu_F(x) = \int_{-\infty}^{x} y dF(y) . \qquad \textbf{(16.26)}$$

Отметим, что функция влияния оценки $\Delta_n = T(F_n)$ средней разности Джини $\Delta_F = T(F)$ записывается в виде

$$IF(x;F,\Delta_F) = \varphi_F(x) - \int \varphi_F(x)dF(x) = \varphi_F(x) - 2(\Delta_F - \mu_F),$$
$$x \in R^1. \qquad \textbf{(16.27)}$$

Эта функция *не является ограниченной*, её график при нормальном распределении приведен на рис. 16.1. Для сравнения, на этом рисунке также приведены функции влияния других оценок масштабного параметра, свойства которых и их сравнение в рамках различных супермоделей приведены в [9, 82, 83]. Отметим, что оценка $\hat{S}_4$ медианы абсолютных разностей и интерквартильный размах $\hat{S}_3(0,25)$ имеют ограниченные функции влияния. Отметим также, что урезанные варианты стандартного отклонения $\hat{S}_1(\alpha)$, среднего абсолютных отклонений $\hat{S}_2(\alpha)$, и урезанный вариант средней разности Джини $\hat{\Delta}_\alpha$, $0 < \alpha < 1/2$, тоже имеют ограниченные функции влияния

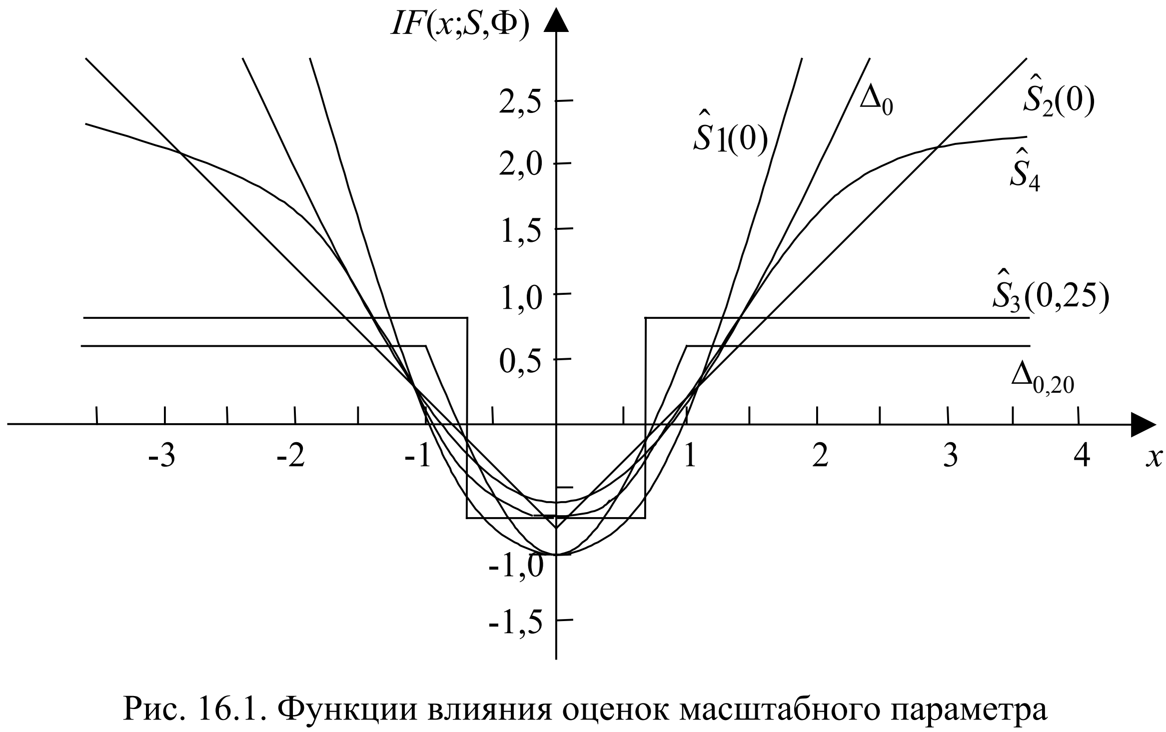

Рис. 16.1. Функции влияния оценок масштабного параметра

Обсудим подробнее урезанный вариант $\hat{\Delta}_\alpha$ средней разности Джини, которая была предложена в [63]. Эта оценка является *U*-статистикой, основанной на урезанной выборке, и имеет вид

$$U_{n,\alpha} = \{(n-2[\alpha n])(n-2[\alpha n]-1)\}^{-1} \sum_{C_\alpha} \left| X_{(i)} - X_{(j)} \right| . \qquad \textbf{(16.28)}$$

Функционал $U_\alpha(F)$ из (16.5), соответствующий этой оценке при $m=2$ и ядре $h(x_1, x_2) = | x_1 - x_2 |$, записывается в таком виде:

$$U_\alpha(F) = \frac{1}{(1-2\alpha)^2} \int\limits_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} \int |x-y| dF(x) dF(y). \qquad \textbf{(16.29)}$$

Для урезанной *U*-статистики данного примера асимптотически эквивалентная *V*-статистика Мизеса запишется в виде

$$\tilde{U}_{n,\alpha} = \{n-2[\alpha n]\}^{-2} \sum_{i,j=[\alpha n]+1}^{n-[\alpha n]} | X_{(i)} - X_{(j)} | . \qquad \textbf{(16.30)}$$

Далее, функция $g_\alpha(x;F)$из (16.3) для данного примера, при $m=2$ и $h(x_1, x_2) = | x_1 - x_2 |$, запишется в виде

$$g_\alpha(x;F) = \frac{I[F^{-1}(\alpha) \le x \le F^{-1}(1-\alpha)]}{(1-2\alpha)^2} \int\limits_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} | x-y | dF(x) dF(y) =$$

$$= \frac{I[F^{-1}(\alpha) \le x \le F^{-1}(1-\alpha)]}{(1-2\alpha)^2} \left\{ 2xF(x) - x + \int_{\xi_\alpha}^{\bar{\xi}_\alpha} y dF(y) - 2\int_{\xi_\alpha}^{x} y dF(y) \right\} =$$

$$= \frac{I[\xi_\alpha \le x \le \xi_{1-\alpha}]}{(1-2\alpha)^2} \{\varphi_\alpha(x,F)/2 + \mu_\alpha(F)\} , \qquad \textbf{(16.31)}$$

где $\varphi_\alpha(x,F), \mu_\alpha(F)$ вычисляются по формулам

$$\varphi_\alpha(x,F) = 4xF(x) - 2x - \int\limits_{F^{-1}(\alpha)}^{x} y dF(y),$$

$$\mu_\alpha(F) = \int\limits_{\alpha}^{1-\alpha} F^{-1}(t) dt . \qquad \textbf{(16.32)}$$

Величины A_α и B_α из (16.4) в данном случае равны

$$A_\alpha = \left[(1-2\alpha)\xi_\alpha - \mu_\alpha(F)\right]/(1-2\alpha)^2 \ ,$$
$$B_\alpha = \left[(1-2\alpha)\xi_{1-\alpha} - \mu_\alpha(F)\right]/(1-2\alpha)^2. \qquad \textbf{(16.33)}$$

Используя полученные формулы (16.20) – (16.23) и формулу (16.13), приведем выражение для функции влияния $IF(x;F,U_\alpha)$ оценки $U_{n\alpha}$ вида (16.18). Для простоты рассмотрим симметричные распределения, то есть $F \in \Im_{S|0}$, тогда $\mu_\alpha(F) = 0$ и $\xi_\alpha = -\overline{\xi}_\alpha$. С учетом этого, выражение (16.13) запишется в виде

$$IF(x;F,U_\alpha) = \frac{1}{(1-2\alpha)^2} \times$$
$$\times \begin{cases} [\varphi_\alpha(x,F) - 2(1-2\alpha)^2 U_\alpha(F) - 4\alpha(1-2\alpha)\overline{\xi}_\alpha], & |x| \le \overline{\xi}_\alpha \\ 2(1-2\alpha)^2[\overline{\xi}_\alpha - U_\alpha(F)], & |x| > \overline{\xi}_\alpha \end{cases}. \qquad \textbf{(16.34)}$$

Введем обозначения

$$J_1(F,\alpha) = \int_{\xi_\alpha}^{\overline{\xi}_\alpha} \varphi_\alpha(x,F)\,dF(x),$$
$$J_2(F,\alpha) = \int_{\xi_\alpha}^{\overline{\xi}_\alpha} \varphi_\alpha^2(x,F)\,dF(x). \qquad \textbf{(16.35)}$$

Используя их и учитывая, что

$$J_1(F,\alpha) = 2(1-2\alpha)^2 U_\alpha(F),$$

формулу (16.24) перепишем в виде

$$IF(x;F,U_\alpha) = \frac{1}{(1-2\alpha)^2} \times$$
$$\times \begin{cases} [\varphi_\alpha(x,F) - J_1(F,\alpha) - 4\alpha(1-2\alpha)\overline{\xi}_\alpha], & |x| \le \overline{\xi}_\alpha, \\ [2(1-2\alpha)^2\,\overline{\xi}_\alpha - J_1(F,\alpha)], & |x| > \overline{\xi}_\alpha. \end{cases} \qquad \textbf{(16.36)}$$

Подчеркнём ещё раз, что функции влияния $U_{n\alpha}$-оценок *ограничены* и при $\alpha = 0$ формулы (16.34) и (16.36) совпадают с выражением для функции влияния средней разности Джини, вычисляемой по исходной выборке, которая не является ограниченной функцией и определена в (16.27). С учетом введенных обозначений, формула (16.10) для вычисления асимптотической дисперсии $\sqrt{n}\,U_{n\alpha}$-оценок в данном случае принимает вид

$$\sigma^2(F,U_\alpha)=\frac{J_2(F,\alpha)-2J_1^2(F,\alpha)+8J_1(F,\alpha)\alpha(1-2\alpha)F^{-1}(\alpha)}{(1-2\alpha)^4}+$$
$$+\frac{8\alpha(1-2\alpha)^3[F^{-1}(\alpha)]^2}{(1-2\alpha)^4},\quad F\in\Im_{S|0}. \tag{16.37}$$

Численные расчеты этой характеристики для различных распределений приводятся в [4, 9, 63, 83] при сравнении $U_{n\alpha}$-оценок с другими оценками масштабного параметра. В частности, при $\alpha=0$ асимптотическая дисперсия оценки Δ_n средней разности Джини при нормальном распределении, то есть при $F=\Phi$, равна

$$\sigma_\Phi^2(\Delta_n)=\int\varphi_\Phi^2(x)dF(x)-\Big(\int\varphi_\Phi(x)dF(x)\Big)^2=$$
$$=\frac{4(\pi+6\sqrt{3})}{3\pi}-\frac{16}{\pi}=\frac{4(\pi+6\sqrt{3}-12)}{3\pi}\approx 0{,}651,$$

где

$$\varphi_\Phi(x)=2x(2\Phi(x)-1)-4\mu_\Phi(x)=2x\tilde{\Phi}(x/\sqrt{2})+\frac{4}{\sqrt{2\pi}}\exp\{-x^2/2\},$$
$$\tilde{\Phi}(x)=(2/\sqrt{\pi})\int_0^x\exp\{-x^2\}dx.$$

Асимптотическая стандартизованная дисперсия Δ_n-оценки при $F=\Phi$

$$\tilde{\sigma}^2(\Phi,\hat{\Delta})=4\cdot\frac{\int\varphi_\Phi^2(x)d\Phi(x)-\Big(\int\varphi_\Phi(x)d\Phi(x)\Big)^2}{\Big(\int\varphi_\Phi(x)d\Phi(x)\Big)^2}=$$
$$=4\frac{4(\pi+6\sqrt{3}-12)}{3\pi\cdot(16/\pi)}=\frac{\pi+6\sqrt{3}-12}{3}\approx 0{,}511.$$

Приведем в табл. 16.1 результаты сравнения различных оценок масштабного параметра, используя понятие дефекта оценки (3.20), в плоскости двух распределений «$F_{(1)}$ – Гаусс, $F_{(3)}$ – Лаплас».

Таблица 16.1

Дефекты оценок масштабного параметра для распределений $F_{(1)} = \Phi$ и $F_{(3)}$

	$\hat{S}_1(0)$	$\hat{S}_2(0)$	Δ_n	$\hat{S}_4$	$\hat{S}_3(0{,}10)$	$\hat{S}_3(0{,}25)$
$DE(F_{(1)},\hat{S})$	0,00	0,12	**0,02**	0,14	0,38	0,63
$DE(F_{(3)},\hat{S})$	0,20	0,00	**0,04**	0,21	0,35	0,52

Итак, среди сравниваемых оценок предпочтение следует отдать Δ_n-оценке средней разности Джини. Напомним, что для симметричных распределений функции влияния и асимптотические дисперсии $\hat{S}_3(0,25)$-оценки интерквартильного размаха и оценки $\hat{S}_3 = med\{|X_i - med(X)|,\ 1 \le i \le n\}$ совпадают. Отметим также, что оценка $\hat{S}_4 = med\{|X_i - X_j|,\ 1 \le i < j \le n\}$, в отличие от оценок $\hat{S}_1(0)$, $\hat{S}_2(0)$ и Δ_n, имеет ограниченную функцию влияния (см. рис. 16.1 и работы [4, 9]).

Пример 16.38. Рассмотрим супермодель с засорением

$$\Im_{\varepsilon,\tau}(\Phi) = \{F : F(x) = \Phi_{\varepsilon,\tau}(x)\},$$

для которой функция распределения $\Phi_{\varepsilon,\tau}(x)$ вида

$$\Phi_{\varepsilon,\tau}(x) = (1-\varepsilon)\Phi(x) + \varepsilon\Phi(x/\tau),\ 0 \le \varepsilon \le 1,\ \tau \ge 1,$$

имеет плотность $\phi_{\varepsilon,\tau}(x) = (1-\varepsilon)\phi(x) + (\varepsilon/\tau)\phi(x/\tau)$. Здесь $\phi(x)$ обозначает стандартную нормальную плотность $\phi(x) = (1/\sqrt{2\pi})\times$ $\times\exp(-x^2/2)$, $x \in R^1$. Отметим, что $\Phi_{\varepsilon,\tau}(x) \in \Im_{S|0}$ и, следовательно, приведенные выше формулы применимы. Численные значения асимптотической стандартизованной дисперсии $\tilde{\sigma}^2(F,\Delta_n)$ для $\sqrt{n}\Delta_n$-оценки средней разности Джини для супермодели $\Im_{\varepsilon,\tau}(\Phi)$, вычисленные в [9] по формуле (16.37) при $\alpha = 0$, приведены в табл. 16.2.

Таблица 16.2

Асимптотическая стандартизованная дисперсия $\tilde{\sigma}^2(F,\Delta_n)$ для $F \in \mathfrak{I}_{\varepsilon,\tau}(\Phi)$

$\tau \setminus \varepsilon$	0,00	0,001	0,005	0,01	0,05	0,10	0,20	0,30
$\tau=3$	0,511	0,523	0,566	0,618	0,928	1,137	1,255	1,204
$\tau=5$	0,511	0,558	0,735	0,933	1,887	2,256	2,159	1,831
$\tau=10$	0,511	0,753	1,601	2,443	4,901	4,781	3,542	2,592

Для сравнения оценки Δ_n средней разности Джини с оценкой $\hat{S}_1(0)$ стандартного отклонения в рамках супермодели $\mathfrak{I}_{\varepsilon,\tau}(\Phi)$ в табл. 16.3 приведены значения их асимптотических относительных эффективностей.

Таблица 16.3

Асимптотическая относительная эффективность $АОЭ_F(\Delta_n : \hat{S}_1(0))$ для $F \in \mathfrak{I}_{\varepsilon,\tau}(\Phi)$

$\tau \setminus \varepsilon$	0,00	0,001	0,005	0,01	0,05	0,10	0,20	0,30
$\tau=3$	0,978	1,046	1,274	1,468	1,679	1,612	1,304	1,140
$\tau=5$	0,978	1,634	3,011	3,518	2,512	1,712	1,183	1,010
$\tau=10$	0,978	8,740	10,53	7,730	2,114	1,269	0,907	0,825

Из этой таблицы следует, что оценка Δ_n средней разности Джини, проигрывая лишь 2 % в эффективности оценке $\hat{S}_1(0)$ стандартного отклонения при нормальном распределении становится её предпочтительнее уже при небольших отклонениях от нормального распределения в рамках супермодели $\mathfrak{I}_{\varepsilon,\tau}(\Phi)$. Отметим, что при увеличении параметра τ относительная эффективность $АОЭ_{\Phi_{\varepsilon,\tau}}(\Delta_n : \hat{S}_1(0))$ принимает неожиданно большие значения. По всей вероятности это является следствием того факта, что оценка $\hat{S}_1(0)$ стандартного отклонения имеет неограниченную (квадратично возрастающую) функцию влияния (см. рис. 16.1) и при увеличении τ и малых значениях ε её дисперсия резко возрастает (см. теорему 2 в работе [69]).

Замечание 16.39. Задачи оценивания масштабного параметра с помощью M-, L- и R-оценок, когда он выступает в качестве мешающего параметра при оценивании сдвига или в более общих регрессионных задачах, подробно рассмотрены Хьюбером [5]. Мы обсудим вариант, когда масштабный параметр функции распределения F используется в качестве меры, характеризующий степень разброса случайной величины X с ф.р. F. Рассмотрим такие меры, которые могут быть представлены в виде функционала $S(F)$, $F \in \Im$. Требования, которым должен удовлетворять функционал, описывающий разброс случайной величины X, сформулированы Бикелем и Леманом [69, 74].

Пусть X – случайная величина с функцией распределения F, которая абсолютно непрерывна и симметрична, то есть $F \in \Im_S$ и $F(x) = 1 - F(2\theta_x - x)$, $\forall x \in R^1$, где θ_x – точка симметрии. Определим разброс с.в. X относительно θ_x (масштабный параметр) в терминах расстояния X от θ_x, то есть с помощью величины $|X - \theta_x|$. Будем говорить, что с.в. X_1 имеет больший разброс относительно θ_{x_1}, чем с.в. X относительно θ_x, если $|X_1 - \theta_{x_1}|$ стохастически больше $|X - \theta_x|$. Различные меры масштабного параметра определим с помощью функционалов от ф.р. F, допускающих монотонность относительно стохастического возрастания распределений и удовлетворяющих условиям эквивариантности.

Определение 16.40. Функционал $S(F)$, $F \in \Im_S$, называют мерой разброса, или масштабным параметром ф.р. F, если он удовлетворяет следующим условиям [69]:

1. $S(aX + b) = |a| S(X)$ для всех a и b.
2. $S(F_1) \le S(F_2)$ для $F_1 <_{St} F_2$, **(16.41)**

где F_1 и F_2 – функции распределения вероятностей случайных величин $|X_1 - \theta_{x_1}|$ и $|X_2 - \theta_{x_2}|$.

Множество различных функционалов, характеризующих масштабный параметр, условно можно разделить на несколько групп (см. [4, 9]).

К первой группе относятся функционалы, построенные с помощью отклонений каждого члена генеральной совокупности от некоторого «центрального» (типичного значения) ф.р. F. Обычно в качестве такого значения используется параметр положения $T(F)$, либо в виде среднего $T_1(F)=\int x dF(x)$, либо в виде медианы $T_2(F)=F^{-1}(1/2)$. Обозначим ф.р. $|X-T(F)|$ через F_1, и ф.р. $|X_1-X_2|$, где X_1 и X_2 – независимые с ф.р. F, через F_2. Многих представителей первой группы можно описать при $i=1$ с помощью функционалов вида

$$\left\{\int_0^1 [F_i^{-1}(t)]^\gamma dV(t)\right\}^{1/\gamma}, \; i=1,2, \qquad \textbf{(16.42)}$$

где $V(t)$ – некоторая функция распределения на $[0,1]$ и $\gamma>0$. Например, если в (16.42) положить $i=1$, в качестве параметра положения выбрать среднее значение $T_1(F)$ и положить $V(t)=t$, $0\le t\le 1$, то при $\gamma=1$ получим среднее абсолютных отклонений. При $\gamma=2$ будем иметь стандартное отклонение. Если же положить $V(t)=t/(1-\alpha)$, $0\le t\le 1-\alpha$, $0\le\alpha<1/2$, то получим α-урезанные варианты указанных мер масштабного параметра (см. [4, 9]). Другая часть этой группы определяется функционалом $F_1^{-1}(1/2)$. Например, при использовании в качестве параметра положения $T_2(F)$, получаем широко используемую в теории робастного оценивания медиану абсолютных отклонений от медианы, то есть $MED|X-MED(X)|$.

Ко второй группе относятся функционалы, построенные с помощью отклонений между всеми членами генеральной совокупности. Некоторые представители этой группы также выражаются с помощью функционалов вида (16.42). Например, при $i=2$, $\gamma=1$ и $V(t)=t$, $0\le t\le 1$, из (16.42) получим среднюю разность Джини. При $\gamma=2$ получаем стандартное отклонение, умноженное на $\sqrt{2}$. К этой же группе относится и медиана абсолютных разностей

$MED\,|X_i - X_j|$, $1 \le i, j \le n$, определяемая с помощью функционала $S(F) = F_2^{-1}(1/2)$.

К третьей группе относятся функционалы. Построенные с помощью расстояний между точками, в которых ф.р. F имеет характерные особенности. К таким точкам могут относиться, например, квантили заданных уровней. Некоторых представителей этой группы можно описать с помощью функционалов вида

$$\left\{\int_0^1 |F^{-1}(t) - F^{-1}(1-t)|^\gamma \, dK(t)\right\}^{1/\gamma}, \qquad \textbf{(16.43)}$$

где $K(t)$ – некоторая функция распределения на $[0, 1]$ и $\gamma > 0$. В частности к этой группе относятся интер-α-квантильные размахи, определяемые в виде $[F^{-1}(1-\alpha) - F^{-1}(\alpha)]/2$ (см. [4, 9]). Отметим, что при $\alpha = 0{,}25$ получаем интерквартильный размах, который для симметричных распределений совпадает с медианой абсолютных отклонений от медианы, то есть определяется функционалом $F_1^{-1}(1/2)$ при использовании в качестве параметра положения функционала $T_2(F) = F^{-1}(1/2)$. Отметим, что выбор конкретного функционала для описания масштабного параметра может быть продиктован различными требованиями. Так, в работе [69] кроме естественных ограничений инвариантности относительно линейных преобразований накладывается требование непрерывности функционала относительно метрики, порождающей слабую сходимость. Выполнение этого требования приводит к оценкам функционалов, удовлетворяющим условиям качественной робастности. Окончательный выбор может осуществляться путем сравнения точностей, с которыми каждый функционал может быть оценен по наблюдениям в рамках заданной супермодели. В работах [4, 9] обсуждаются оценки масштабного параметра, которые получены методом подстановки, то есть записываются в виде функционала $S(F_n)$ от эмпирической ф.р. F_n.

Общая схема построения различных функционалов, описывающих масштабный параметр, сводится, по существу, к следующему.

Над исходной случайной величиной осуществляется некоторое преобразование вида $|X - T_1(F)|^\gamma$, $|X - T_2(F)|^\gamma$, $|X_1 - X_2|^\gamma$, $\gamma > 0$ и т.п. Затем к преобразованным случайным величинам применяется, либо операция усреднения, либо операция вычисления медианы, либо операция вычисления оценки Ходжеса – Лемана и т.п. Другими словами, функционал, описывающий масштабный параметр, определяется с помощью функционала, характеризующего параметр положения для преобразованных случайных величин. Например, «медианная операция», примененная к $|X - T_2(F)|^\gamma$ при $\gamma = 1$, приводит к медиане абсолютных отклонений от медианы; «медианная операция», примененная к $|X_1 - X_2|^\gamma$ при $\gamma = 1$, приводит к медиане абсолютных разностей [4, 9]; операция усреднения в этом случае приводит к средней разности Джини; «операция Ходжеса – Лемана» приводит к еще не изученным оценкам масштабного параметра, например такого вида:

$$med\{\,[|X_i - med(X)| + |X_j - med(X)|]/2,\ 1 \le i, j \le n\,\},$$

$$med\{\,[|X_i - X_j| + |X_k - X_l|]/2,\ 1 \le i, j, k, l \le n\,\}.$$

При этом также возможно использование обобщенных оценок Ходжеса – Лемана и их урезанных вариантов. Применение этой схемы отрывает большие возможности при построении новых мер масштабного параметра, при этом могут использоваться обширные результаты, полученные при изучении оценок параметра положения, включая их общие классы M-, L- и R-оценок (см. работу [5]). Сравнение оценок масштабного параметра проводится с помощью относительной эффективности, определенной через обратные отношения стандартизованных дисперсий, которые, в свою очередь, определяются в виде отношения асимптотической дисперсии оценки к квадрату оцениваемого функционала (см. работу [74]).

Отметим, что общая схема построения оценок масштабного параметра может быть распространена и на построение робастных оценок параметров для следующей модели наблюдений. Рассмотрим модель наблюдений $Y_1,...,Y_n$ в виде $Y_i = \varphi(x_i, \vec{\theta}) + \xi_i$, $i = 1,...,n$, где φ – заданная функция, $x_1,...,x_n$ – заданные значения независимой переменной x, $\vec{\theta} = (\theta_1, \ldots, \theta_r)$ – вектор неизвестных параметров, $\xi_1, \ldots, \xi_n$ – н.о.р. случайные величины с ф.р. F. Используя различные оценки масштабного параметра, построенные с помощью последовательности $\xi_1, \ldots, \xi_n$ – н.о.р. случайных величин, где $\xi_i = Y_i - \varphi(x_i, \vec{\theta})$, $i = 1, \ldots, n$, можно предложить следующие оценки векторного параметра $\vec{\theta} = (\theta_1, \ldots, \theta_r)$:

1. α-урезанные МНК-оценки вида

$$\vec{\theta}_n = \arg\min_{\vec{\theta}} \left(\sum_{i=[\alpha n]+1}^{n-[\alpha n]} \xi_{(i)}^2 \right) = \arg\min_{\vec{\theta}} \left(\sum_{i=[\alpha n]+1}^{n-[\alpha n]} [Y_{(i)} - \varphi(x_i, \vec{\theta})]^2 \right),$$
$$0 \le \alpha < 1/2;$$

2. α-урезанные МНМ-оценки вида

$$\vec{\theta}_n = \arg\min_{\vec{\theta}} \left(\sum_{i=[\alpha n]+1}^{n-[\alpha n]} |\, \xi_{(i)} \,| \right) = \arg\min_{\vec{\theta}} \left(\sum_{i=[\alpha n]+1}^{n-[\alpha n]} |\, Y_{(i)} - \varphi(x_i, \vec{\theta}) \,| \right),$$
$$0 \le \alpha < 1/2;$$

3. α-урезанные оценки, основанные на средней разности Джини, в виде

$$\vec{\theta}_n = \arg\min_{\vec{\theta}} \left(\sum_{i,j=[\alpha n]+1}^{n-[\alpha n]} |\, \xi_{(i)} - \xi_{(j)} \,| \right), \; 0 \le \alpha < 1/2;$$

4. Оценки «медианного» типа в виде

$$\vec{\theta}_n = \arg\min_{\vec{\theta}} \left\{ med\left(\,|\, \xi_1 \,|^p, \ldots, |\, \xi_n \,|^p \right), 1/2 \le p \le 2 \right\},$$

$$\vec{\theta}_n = \arg\min_{\vec{\theta}} \{ med\,(\, |\, \xi_i - med(\xi_1, \ldots, \xi_n)\,|), \, i = 1, \ldots, n \},$$

$$\vec{\theta}_n = \arg\min_{\vec{\theta}} \left\{ med\left(\,|\, \xi_i - \xi_j \,| \right), 1 \le i < j \le n \right\};$$

5. Оценки, построенные с помощью процедуры Ходжеса – Лемана, в виде

$$\vec{\theta}_n = \arg\min_{\vec{\theta}} \{ med\left([|\, \xi_{(i)} \,|^p + |\, \xi_{(j)} \,|^p]/2 \right),$$
$$[\alpha n] + 1 \le i < j \le n - [\alpha n], \; 1/2 \le p \le 2\},$$

$$\vec{\theta}_n = \arg\min_{\vec{\theta}} \{ med\Big([|\xi_{i_1} - med(\xi_1,...,\xi_n)| + ... + |\xi_{i_m} - med(\xi_1,...,\xi_n)|]/m \Big),$$
$$1 \le i_1 < ... < i_m \le n, m = 2,3... \}.$$

Эти оценки ещё не изучены и представляет интерес изучение их свойств в рамках гауссовской модели с масштабным засорением, то есть для $F \in \Im_{\varepsilon,\tau}(\Phi)$.

17. СРАВНЕНИЕ *R*- И R_α-ОЦЕНОК ПАРАМЕТРА ПОЛОЖЕНИЯ

При выполнении условий регулярности, накладываемых на ф.р. F наблюдений и функцию меток φ (см. раздел 9), R- и R_α-оценки имеют асимптотически нормальное распределение. Поэтому естественной характеристикой их качества являются асимптотические дисперсии $\sigma^2(F,\varphi)$ и $\sigma^2(F,\varphi,\alpha)$. Для сравнения различных оценок будем использовать асимптотическую относительную эффективность в виде

$$AOЭ_F(R_\alpha(\varphi_1):R_\alpha(\varphi_2))=\sigma^2(F,\varphi_2,\alpha)/\sigma^2(F,\varphi_1,\alpha). \qquad \textbf{(17.1)}$$

Нас будут интересовать асимптотические относительные эффективности для R- и R_α-оценок, которые обозначим через

$$AOЭ_F(R_\alpha(\varphi);R(\varphi)) \text{ и } AOЭ_F(R_\alpha(\varphi_1);R_\alpha(\varphi_2)). \qquad \textbf{(17.2)}$$

Эти характеристики могут быть вычислены при каждой конкретной ф.р. F, например для идеальной модели, либо для конкретного представителя рассматриваемой супермодели. Однако более общие результаты можно получить для супермоделей с «упорядоченными» распределениями. Уточним это понятие упорядоченности распределений.

Замечание 17.3. Различные типы «упорядоченности распределений по степени затянутости их хвостов» обсуждаются в работах [70, 71, 78, 79] (см. также разделы (1.4) и (1.5) в [4]). Степень затянутости хвостов распределений оказывает существенное влияние на качество статистических процедур. В литературе имеется достаточно широкий спектр определений, которые основаны либо на поведении квантильной функции $F^{-1}(t)$, либо на поведении плотности $f(F^{-1}(t))$, либо на производной $f'(F^{-1}(t))$, либо на второй производной $f''(F^{-1}(t))$. Некоторые из этих типов упорядоченности распределений нам понадобятся для сравнения R- и R_α-оценок. Приведем необходимые определения.

Определение 17.4. О случайных величинах X_1 и X_2 с функциями распределений F_1 и F_2 говорят, что с.в. X_2 стохастически больше, чем с.в. X_1 (при этом используют обозначение в виде $F_1 <_{St} F_2$), если для всех x выполняется неравенство $P(X_1 > x) \le P(X_2 > x)$.

Отметим, что $F_1 <_{St} F_2 \Rightarrow F_1(x) \ge F_2(x)$ для всех x и $F_1^{-1}(t) \le F_2^{-1}(t)$, $0 \le t \le 1$. Кроме того, если $S(x)$ неубывающая функция и $F_1 <_{St} F_2$, то выполняется неравенство

$$\int S(x)dF_1(x) \le \int S(x)dF_2(x). \qquad \textbf{(17.5)}$$

Попутно также отметим, что если $r(x)$ и $S(x)$ – неубывающие функции и $F(x)$ – функция распределения вероятностей, то выполняется неравенство

$$\int S(x)r(x)dF(x) \ge \int S(x)dF(x) \cdot \int r(x)dF(x). \qquad \textbf{(17.6)}$$

Определение 17.7 (Ван Цвет [79]). Пусть $F, G \in \Im_{S|\theta}$. Говорят, что хвосты ф.р. F легче хвостов ф.р. G (или G имеет хвосты тяжелее, чем F, и это записывают в виде $F <_S G$), если функция $G^{-1}(F(x))$ выпуклая для $x \ge \theta$, где θ – точка симметрии для F и G.

Отметим, что если $F <_S G$, то также говорят, что распределения F и G являются S-упорядоченными. Отметим также, что при изучении S-упорядоченности можно ограничится лишь классом распределений $\Im_{S|0}$, то есть можно без потери общности положить точку симметрии $\theta = 0$ (см., например, [79]). Для проверки выражения $F <_S G$ часто бывает удобным использовать следующий критерий выпуклости функции $G^{-1}(F(x))$. Если $F, G \in \Im_{S|0}$, то $F <_S G$ тогда и только тогда, когда существует строго возрастающая функция $S(t)$, $1/2 < t \le 1$, такая, что выполняется равенство

$(G^{-1}(t))' = S(t)(F^{-1}(t))'$, которое также может быть переписано в виде $f(F^{-1}(t)) = S(t)g(G^{-1}(t))$, $1/2 < t \le 1$.

Пример 17.8. Рассмотрим супермодель $\Im_S^* = \{U, F_{(1)}, F_{(2)}, F_{(3)}, F_{(4)}\}$, которая содержит следующий конечный набор стандартных распределений:

равномерное в (–1, 1) с ф.р. $U(x) = \begin{cases} 0, \ x < -1, \\ (x+1)/2, \ -1 \le x < 1, \\ 1, \ x \ge 1; \end{cases}$

нормальное с ф.р. $F_{(1)}(x) = \Phi(x) = \dfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{1}{2}x^2} dx$;

логистическое с ф.р. $F_{(2)}(x) = 1/(1 + e^{-x})$, $x \in R^1$;

Лапласа с ф.р. $F_{(3)}(x) = \begin{cases} (1/2)e^x, \ x < 0, \\ 1 - (1/2)e^{-x}, \ x \ge 0; \end{cases}$

Коши с ф.р. $F_{(4)}(x) = \dfrac{1}{2} + \dfrac{1}{\pi}\operatorname{arctg}(x)$, $x \in R^1$.

Можно убедиться (см. [45, с.123]), что данная супермодель $\Im_S^*$ содержит S-упорядоченные распределения, причем выполняется выражение вида

$$U <_S F_{(1)} <_S F_{(2)} <_S F_{(3)} <_S F_{(4)}. \qquad \textbf{(17.9)}$$

Кроме того, если $\Im(SU)$ – семейство симметричных распределений с сильно одновершинными плотностями, то $\forall F \in \Im(SU)$ выполняется выражение $U <_S F <_S F_{(3)}$, то есть распределение Лапласа является S-наибольшим в семействе $\Im(SU)$, а равномерное распределение U является S-наименьшим в семействе $\Im(SU)$. Доказательство этого факта имеется в [80]. Заметим, что $F_{(4)} \notin \Im(SU)$.

Определение 17.10 (Гаек [78]). Пусть $F, G \in \Im_S$. Говорят, что ф.р. G имеет хвосты тяжелее, чем F, и это записывают в виде $F <_h G$, если выполняется равенство $F^{-1}(t) = a(t)G^{-1}(t)$, где $a(t)$, $1/2 \le t < 1$, строго возрастающая функция.

Напомним, что в теории ранговых критериев и R-оценок, построенных с использованием ранговых статистик, важную роль играет φ-функция Гаека, которая для альтернатив сдвига определяется в виде

$$\varphi_F(u) = -f'(F^{-1}(u))/f(F^{-1}(u)),\ 1/2 < u \le 1. \qquad \mathbf{(17.11)}$$

Отметим, что если плотность $f(x)$ сильно одновершинна, то соответствующая φ-функция неубывающая. Справедливо также и обратное утверждение. Гаек [78] установил связь между «тяжестью хвостов» распределений и поведением φ-функций (см. теорему 17.13) и при этом использовал следующее понятие.

Определение 17.12 (Гаек [78]). Пусть $\varphi_1(u)$ и $\varphi_2(u)$ – две кососимметричные неубывающие функции на (0, 1). Говорят, что функция $\varphi_1(u)$ возрастает более быстро, чем $\varphi_2(u)$ ($\varphi_1(u)$ increases more rapidly than $\varphi_2(u)$) и кратко это записывают в виде $\varphi_1(u)\,\mathrm{imr}\,\varphi_2(u)$, если существует строго возрастающая функция $a(u)$, такая, что выполняется равенство $\varphi_1(u) = a(u)\varphi_2(u)$, $1/2 < u \le 1$.

Теорема 17.13 (Гаек [78]). Пусть $F, G \in \Im(SU)$ – семейство симметричных распределений с сильно одновершинными плотностями. Пусть далее, согласно (17.11), соответствующие φ-функции записываются в виде

$$\varphi_F(u) = -f'(F^{-1}(u))/f(F^{-1}(u)),$$
$$\varphi_G(u) = -g'(G^{-1}(u))/g(G^{-1}(u)),\ 0 \le u \le 1.$$

Тогда, если $\varphi_F(u)\,\mathrm{imr}\,\varphi_G(u)$, то выполняется выражение $F <_h G$.

В работе [70] также вводится понятие затянутости хвостов распределений не через исходные функции распределений, а через соответствующие им φ-функции.

Определение 17.14 (Гаствирт [70]). Пусть $F, G \in \Im_S$. Говорят, что ф.р. G имеет хвосты тяжелее, чем F, и это записывают в виде $F <_g G$, если выполняется выражение $\varphi_F(u)\,\mathrm{imr}\,\varphi_G(u)$.

Определение 17.15. Функция φ_F характеризует (представляет) середину распределения F меньше, чем функция φ_G это делает для распределения G (φ_F emphasizes the middle less than φ_G), кратко это записывают в виде $\varphi_F(u)\,\mathrm{eml}\,\varphi_G(u)$, если существует строго возрастающая функция $a(u)$, такая, что выполняется равенство $\varphi'_F(u) = a(u)\varphi'_G(u)$, $1/2 < u \le 1$.

Определение 17.16 (Гаствирт [70]). Пусть $F, G \in \Im_S$. Говорят, что ф.р. G имеет хвосты тяжелее, чем F, и это записывают в виде $F <_k G$, если выполняется выражение $\varphi_F(u)\,\mathrm{eml}\,\varphi_G(u)$.

Отметим, что в эквивалентном виде это определение формулируется следующим образом. Если $\varphi_G^{-1}(\varphi_F(u))$ – выпуклая функция, $1/2 < u \le 1$, то выполняется выражение $F <_k G$.

Замечание 17.17. Все приведенные выше определения связаны с таким общим свойством распределений, как степень затянутости (или «тяжести их хвостов»). Как отмечается в [73, с. 316], термин «тяжелые хвосты » не следует воспринимать слишком буквально. В частности, он не обязательно отражает ту скорость, с которой плотность $f(x) \to 0$ при $x \to \infty$. Для подтверждения этого Э. Леман рассмотрел пример с распределением Коши и с плотностью $f(x) = 1/\pi(1+x^2)$, $-\infty < x < \infty$. Пусть это распределение усечено на уровнях $\pm A$. Усеченное распределение Коши определяется плотностью $f(x)/P(|X| \le A)$ при $|x| \le A$ и $f(x) = 0$ – в других случаях. Это усеченное распределение вообще не имеет хвостов. При этом отмечается, что асимптотическая относительная эффективность выборочной медианы $\bar{X}_{1/2}$ по отношению к выборочному среднему $\bar{X}$ ведет себя следующим образом: $АОЭ_F(\bar{X}_{1/2}:\bar{X}) = 4f^2(0)\sigma^2 \to \infty$ при $A \to \infty$, хотя для каждого конечного A, сколь большим бы оно не было, усеченная плотность вообще не имеет никаких хвостов. Отметим, что приведенные выше определения тесно связаны между собой. В частности, в [71] отмечается справедливость следующих импликаций:

$$F <_k G \Rightarrow F <_g G \Rightarrow F <_S G \Rightarrow F <_h G. \qquad \textbf{(17.18)}$$

В общем случае эти импликации не могут быть обратными. Гаствирт [70] приводит пример, когда S-упорядоченные распределения не являются g-упорядоченными. Примеры семейств распределений, упорядоченных по степени затянутости хвостов в различных вариантах, приводятся в [4].

Для удобства сравнения R- и R_α-оценок введем следующие обозначения:

$$C(\varphi,\alpha)=\frac{1}{B(\varphi,\alpha)}\int_0^1\varphi^2(t)dt; \tag{17.19}$$

$$B(\varphi,\alpha)=2\alpha\varphi^2(\alpha)+\int_\alpha^{1-\alpha}\varphi^2(t)dt; \tag{17.20}$$

$$A(F,\varphi,\alpha)=\left(\int_\alpha^{1-\alpha}\varphi'(t)f(F^{-1}(t))dt\right)^{-1}. \tag{17.21}$$

Теорема 17.22. Пусть ф.р. F и H симметричны, то есть $F,H\in\Im_S$. Определим φ-функции в виде

$$\varphi_F(u)=-f'(F^{-1}(u))/f(F^{-1}(u)),$$
$$\varphi_H(u)=-h'(H^{-1}(u))/h(H^{-1}(u)).$$

Тогда, если $\varphi_F(u)\,\mathrm{eml}\,\varphi_G(u)$ и $F<_k H$, то при любом α, $0\le\alpha<1/2$, выполняется неравенство

$$АОЭ_F(R_\alpha(\varphi_H):R_\alpha(\varphi_F))\le АОЭ_H(R_\alpha(\varphi_H):R_\alpha(\varphi_F)). \tag{17.23}$$

Кроме того, для $\forall Q\in\Im_S$ выполняется неравенство

$$АОЭ_Q(R_\alpha(\varphi_H):R_\alpha(\varphi_F))\le\frac{B(\varphi_F,\alpha)}{B(\varphi_H,\alpha)}\left(\frac{\varphi'_H(1/2)}{\varphi'_F(1/2)}\right)^2,\ 0\le\alpha<1/2. \tag{17.24}$$

Далее, для любого α, $0\le\alpha<1/2$, выполняются неравенства

$$АОЭ_F(R_\alpha(\varphi_H):R(\varphi_F))\le АОЭ_H(R_\alpha(\varphi_H):R(\varphi_F)); \tag{17.25}$$

$$АОЭ_Q(R_\alpha(\varphi_H):R(\varphi_F))\le C(\varphi_F,\alpha)\frac{B(\varphi_F,\alpha)}{B(\varphi_H,\alpha)}\left(\frac{\varphi'_H(1/2)}{\varphi'_F(1/2)}\right)^2. \tag{17.26}$$

Доказательство. Для доказательства неравенства (17.23) необходимо показать, что при $F <_k H$ выполняется неравенство

$$\frac{A(F,\varphi_F,\alpha)}{A(F,\varphi_H,\alpha)} \le \frac{A(H,\varphi_F,\alpha)}{A(H,\varphi_H,\alpha)}. \qquad \textbf{(17.27)}$$

Для этого достаточно убедиться в справедливости следующего неравенства:

$$\frac{\int_{1/2}^{1-\alpha} \varphi'_H(t) f(F^{-1}(t))dt}{\int_{1/2}^{1-\alpha} \varphi'_F(t) f(F^{-1}(t))dt} \le \frac{\int_{1/2}^{1-\alpha} \varphi'_H(t) h(H^{-1}(t))dt}{\int_{1/2}^{1-\alpha} \varphi'_F(t) h(H^{-1}(t))dt}. \qquad \textbf{(17.28)}$$

Итак, пусть $\varphi_F(u)\,\mathrm{eml}\,\varphi_H(u)$ и $F <_k H$. Согласно определениям 17.15 и 17.16, отношение $\varphi'_F(t)/\varphi'_H(t) = a(t)$ является строго возрастающей функцией при $1/2 < t < 1$. Кроме того, $F <_k H \Rightarrow F <_S H$ и, согласно 17.8), отношение $f(F^{-1}(t))/h(H^{-1}(t)) = S(t)$ является строго возрастающей функцией $1/2 < t < 1$. Далее, определим функцию распределения $G(t)$, $t \in [0, 1]$, в виде

$$G(t) = 0,\ 0 \le t < 1/2,$$

$$G(t) = \frac{\int_{1/2}^{t} \varphi'_H(t) h(H^{-1}(t))dt}{\int_{1/2}^{1-\alpha} \varphi'_H(t) h(H^{-1}(t))dt},\ 1/2 \le t \le 1-\alpha, \qquad \textbf{(17.29)}$$

$$G(t) = 1,\ 1-\alpha < t \le 1.$$

Используя теперь неравенство (17.6) для возрастающих функций $a(t)$, $S(t)$ и функции распределения $G(t)$, $t \in [0, 1]$, получаем неравенство

$$\int_{1/2}^{1-\alpha} \frac{\varphi'_F(t)}{\varphi'_H(t)} \frac{f(F^{-1}(t))}{h(H^{-1}(t))} dG(t) \ge \int_{1/2}^{1-\alpha} \frac{\varphi'_F(t)}{\varphi'_H(t)} dG(t) \int_{1/2}^{1-\alpha} \frac{f(F^{-1}(t))}{h(H^{-1}(t))} dG(t).$$

Учитывая, что дифференциал функции $G(t)$ из (17.28) равен

$$dG(t) = \frac{\varphi'_H(t)h(H^{-1}(t))dt}{\int_{1/2}^{1-\alpha} \varphi'_H(t)h(H^{-1}(t))dt}, \;\; 1/2 \le t \le 1-\alpha,$$

из предыдущего неравенства следуют (17.28), (17.27) и (17.23).

Для доказательства неравенства (17.24) отметим, что $\varphi'_F(t)/\varphi'_H = a(t)$ является строго возрастающей функцией, $1/2<t<1$, и $\varphi'_F(t)q(Q^{-1}(t)) > 0$. С учетом данного замечания, получаем неравенство

$$\int_{1/2}^{1-\alpha} \varphi'_F(t)q(Q^{-1}(t))dt \ge a(1/2) \int_{1/2}^{1-\alpha} \varphi'_H(t)q(Q^{-1}(t)dt).$$

Далее, используя (17.1), с учетом предыдущего неравенства, получаем (17.24). Справедливость (17.25) и (17.26) проверяется аналогично. Доказательство завершено.

Рассмотрим теперь более слабый тип S-упорядоченности распределений (см. определение 17.7 и выражение (17.18)).

Теорема 17.30. Пусть $F, H \in \Im_S$ и p. Тогда, если выполняется выражение $F <_S H$, то при любом α, $0 \le \alpha < 1/2$, выполняется неравенство

$$АОЭ_F(R_\alpha(\varphi) : R(\varphi)) \le АОЭ_H(R_\alpha(\varphi) : R(\varphi)) \qquad \textbf{(17.31)}$$

и для $\forall Q \in \Im_S$ выполняется неравенство

$$АОЭ_Q(R_\alpha(\varphi) : R(\varphi)) \le C(\varphi, \alpha), \qquad \textbf{(17.32)}$$

где $C(\varphi, \alpha)$ определено в (17.19).

Доказательство. Для доказательства (17.31) достаточно показать, что выполняется неравенство

$$A(F, \varphi, 0) / A(F, \varphi, \alpha) \le A(H, \varphi, 0) / A(H, \varphi, \alpha),$$

или

$$\frac{\int_{1/2}^{1-\alpha} \varphi'(t) f(F^{-1}(t))dt}{\int_{1/2}^{1} \varphi'(t) f(F^{-1}(t))dt} \le \frac{\int_{1/2}^{1-\alpha} \varphi'(t)h(H^{-1}(t))dt}{\int_{1/2}^{1} \varphi'(t)h(H^{-1}(t))dt}. \qquad \textbf{(17.33)}$$

Определим функции распределения $G(t)$ и $G_\alpha(t)$, $0 \le t \le 1$, в виде

$$G(t) = 0,\ 0 \le t < 1/2,$$

$$G(t) = \frac{\int_{1/2}^{t} \varphi'(t) h(H^{-1}(t)) dt}{\int_{1/2}^{1} \varphi'(t) h(H^{-1}(t)) dt},\ 1/2 \le t \le 1, \qquad \textbf{(17.34)}$$

и

$$G_\alpha(t) = 0,\ 0 \le t < 1/2;$$

$$G_\alpha(t) = \frac{\int_{1/2}^{t} \varphi'(t) h(H^{-1}(t)) dt}{\int_{1/2}^{1-\alpha} \varphi'(t) h(H^{-1}(t)) dt},\ 1/2 \le t \le 1, \qquad \textbf{(17.35)}$$

$$G_\alpha(t) = 1,\ 1-\alpha < t \le 1.$$

Отметим, что $G_\alpha(t) <_{St} G(t)$ (см. определение 17.4). Далее, согласно условию $F <_S H$ и, следовательно, $f(F^{-1}(t))/h(H^{-1}(t)) = S(t)$ является строго возрастающей функцией, $1/2 < t < 1$. Воспользуемся теперь неравенством (17.5), которое перепишем в виде

$$\int_{1/2}^{1} \frac{f(F^{-1}(t))}{h(H^{-1}(t))} dG_\alpha(t) \le \int_{1/2}^{1} \frac{f(F^{-1}(t))}{h(H^{-1}(t))} dG(t). \qquad \textbf{(17.36)}$$

Учитывая, что дифференциалы функций $G(t)$ и $G_\alpha(t)$ из (17.34) и (17.35) соответственно равны

$$dG(t) = \frac{\varphi'(t) h(H^{-1}(t)) dt}{\int_{1/2}^{1} \varphi'(t) h(H^{-1}(t)) dt},\quad 1/2 \le t \le 1,$$

$$dG_\alpha(t) = \frac{\varphi'(t) h(H^{-1}(t)) dt}{\int_{1/2}^{1-\alpha} \varphi'(t) h(H^{-1}(t)) dt},\quad 1/2 \le t \le 1-\alpha,$$

и

$$dG_\alpha(t) = 0,\quad 1-\alpha < t \le 1,$$

из (17.36) получаем неравенство

$$\frac{\int_{1/2}^{1-\alpha} \varphi'(t) f(F^{-1}(t)) dt}{\int_{1/2}^{1-\alpha} \varphi'(t) h(H^{-1}(t)) dt} \le \frac{\int_{1/2}^{1} \varphi'(t) f(F^{-1}(t)) dt}{\int_{1/2}^{1} \varphi'(t) h(H^{-1}(t)) dt},$$

из которого следует неравенство (17.33). Далее, неравенство (17.32) следует из того факта, что при $1/2 \le t \le 1$ функция $\varphi'(t) f(F^{-1}(t))$ неотрицательна. Доказательство завершено.

***Следствие* 17.37.** Пусть $\varphi \in C_R$ и $\Im(SU)$ – класс симметричных унимодальных распределений. Тогда при любом α, $0 < \alpha < 1/2$, асимптотическая относительная эффективность $R_\alpha(\varphi)$-оценки по отношению к $R(\varphi)$-оценке удовлетворяет выражению

$$\inf_{F \in \Im(SU)} АОЭ_F(R_\alpha(\varphi) : R(\varphi)) = C(\varphi,\alpha)\frac{\varphi^2(\alpha)}{\varphi^2(1)}, \qquad \textbf{(17.38)}$$

где $C(\varphi,\alpha)$ определено в (17.19). Этот результат следует из неравенства (17.31) с учетом того, что в семействе распределений $\Im(SU)$ «S-наименьшим» является равномерное распределение (см. пример 17.8).

Пример 17.39. Рассмотрим асимптотическую относительную эффективность урезанной оценки Ходжеса – Лемана (HL_α-оценка) и R_α-оценки с нормальными метками (HS_α-оценка). Из выражения (17.24) получаем неравенство

$$АОЭ_F(HL_\alpha : NS_\alpha) \le \frac{6(1-2\alpha)+2\Phi^{-1}(\alpha)\phi(\Phi^{-1}(\alpha))+2\alpha[\Phi^{-1}(\alpha)]^2)}{\pi(1+4\alpha)(1-2\alpha)^2},$$
$$\forall F \in \Im_S, \qquad \textbf{(17.40)}$$

Численные значения правой части приведены в табл. 17.1

Таблица 17.1

Численные значения правой части (17.40)

α	0,00	0,05	0,10	0,15	0,20	0,25	0,30	0,40	0,50
(17.40)	1,91	1,54	1,45	1,31	1,21	1,14	1,09	1,03	1,00

Асимптотическая относительная эффективность урезанных вариантов оценки Ходжеса – Лемана и R-оценки с нормальными метками для $F \in \Im_S$ и $0 \le \alpha < 1/2$ записывается в виде

$$АОЭ_F(HL_\alpha : NS_\alpha) = \frac{12\left(\int_{1/2}^{1-\alpha} f(F^{-1}(u)du\right)^2}{\left(\int_{1/2}^{1-\alpha}[f(F^{-1}(u))/\phi(\Phi^{-1}(u))]du\right)^2}\times$$

$$\times\frac{\{1-2\alpha+2\Phi^{-1}(\alpha)\phi(\Phi^{-1}(\alpha))+2\alpha[\Phi^{-1}(\alpha)]^2\}}{(1+4\alpha)(1-2\alpha)^2}.$$

Далее, при выполнении неравенств $1/2<u<v<1$ имеем $\Phi^{-1}(u)<\Phi^{-1}(v)$ и $[1/\phi(\Phi^{-1}(u))]\le[1/\phi(\Phi^{-1}(v))]$. Следовательно, выполняется неравенство

$$\int\limits_{1/2}^{1-\alpha}\frac{f(F^{-1}(u))}{\phi(\Phi^{-1}(u))}du\ge\frac{1}{\phi(\Phi^{-1}(1/2))}\int\limits_{1/2}^{1-\alpha}f(F^{-1}(u))du.$$

Используя данное неравенство и учитывая, что $\phi(\Phi^{-1}(1/2))=$ $=1/\sqrt{2\pi}$, получаем неравенство (17.40). Отметим, что при $\alpha=0$, $АОЭ_F(HL:NS)\le 6/\pi\approx 1{,}91$, $\forall F\in\Im_S$.

Пример 17.41. Приведем результаты сравнения урезанной оценки Ходжеса – Лемана (HL_α-оценки) с HL-оценкой. Асимптотическая относительная эффективность этих оценок вычисляется по формуле

$$АОЭ_F(HL_\alpha:HL)=\frac{\sigma_F^2(HL)}{\sigma_F^2(HL_\alpha)}=$$

$$=\frac{1}{(1+4\alpha)(1-2\alpha)^2}\frac{\left(\int_\alpha^{1-\alpha}f(F^{-1}(t))dt\right)^2}{\left(\int_0^1 f(F^1(t))dt\right)^2},\quad 0\le\alpha<1/2. \qquad \textbf{(17.42)}$$

Рассмотрим, как и в примере 17.8, семейство S-упорядоченных распределений $\Im_S^*=\{U,F_{(1)},F_{(2)},F_{(3)},F_{(4)}\}$, тогда для асимптотической относительной эффективности $АОЭ_F(HL_\alpha:HL)$ можно указать нижнюю и верхнюю границы в виде

$$\frac{1}{1+4\alpha}\le АОЭ_F(HL_\alpha:HL)\le\frac{1}{(1+4\alpha)(1-2\alpha)^2},\ F\in\Im_S. \qquad \textbf{(17.43)}$$

Отметим, что нижняя граница является точной, она достигается при равномерном распределении. Численные значения $АОЭ_F(HL_\alpha:HL)$ для различных S-упорядоченных распределений приведены в табл. 17.2 (см. также пример 17.8).

Таблица 17.2

Численные значения $АОЭ_F(HL_\alpha : HL)$, $F \in \mathfrak{I}_S^*$ и границы из (17.43)

α	0,1	0,2	0,3	0,4	0,5
Н.гр. (17.43)	0,714	0,556	0,455	0,385	0,333
$F_{(1)}$	0,965	0,906	0,834	0,753	0,666
$F_{(2)}$	0,995	0,968	0,917	0,842	0,750
$F_{(3)}$	1,029	1,089	1,164	1,245	1,333
$F_{(4)}$	1,089	1,258	1,403	1,441	1,334
В.гр. (17.43)	1,116	1,543	2,841	9,615	∞

Из данных приведенной таблицы следует, что относительные достоинства урезанной оценки Ходжеса – Лемана по сравнению с HL-оценкой возрастают при переходе к «S-большим» распределениям. Сравнение этих оценок для семейства распределений Стьюдента и в рамках гауссовской модели с масштабным засорением приводится в [9].

Отметим также, что для нормального распределения абсолютная эффективность HL_α-оценок падает с 0,955 до 0,637 при изменении параметра α, $0 \le \alpha \le 0,5$. Далее, при $\alpha = 0$ оценка Ходжеса – Лемана является асимптотически эффективной оценкой параметра сдвига θ для логистического распределения. Для распределения Коши абсолютная эффективность меняется от 0,6 до 0,80, при изменении параметра α, $0 \le \alpha \le 0,5$, достигая максимальное значение 0,89 при $\alpha = 0,375$. При $\alpha \to 1/2$ асимптотическая дисперсия HL_α-оценки совпадает с асимптотической дисперсией выборочной медианы $\overline{X}_{1/2}$, которая является асимптотически эффективной оценкой параметра сдвига θ для распределения Лапласа.

18. АДАПТИВНЫЕ ОЦЕНКИ ПАРАМЕТРА ПОЛОЖЕНИЯ

Изучение эффективности многих статистических процедур при изменении распределения вероятности наблюдений в некотором заданном классе (в рамках заданной супермодели) показывает, что эффективность часто зависит монотонно от некоторых общих свойств распределений. В частности, к таким общим свойствам относится «затянутость хвостов» распределений или «тяжесть хвостов» распределений. В литературе (см., например, [4, 45, 70, 71]) описаны различные подходы для упорядочивания распределений в заданном классе по степени тяжести хвостов. Следуя работе [60], рассмотрим количественную меру «тяжести хвостов» распределения $F(x)$, $x \in R^1$, в виде функционала $Q(F;\nu,\mu)$:

$$Q(F;\nu,\mu) = \frac{(1/\nu)\{\int\limits_{1-\nu}^{1} F^{-1}(t)dt - \int\limits_{0}^{\nu} F^{-1}(t)dt\}}{(1/\mu)\{\int\limits_{1-\mu}^{1} F^{-1}(t)dt - \int\limits_{0}^{\mu} F^{-1}(t)dt\}},$$

$$0 < \nu < \mu \le 0{,}5. \qquad \textbf{(18.1)}$$

Значения этого функционала для различных супермоделей вычислены в [4, 9]. Изучение асимптотических свойств α-урезанных средних ($\overline{X}_\alpha$-оценок), α-винзоризованных средних ($\tilde{X}_\alpha$-оценок) и HL_α-оценок Ходжеса – Лемана показало, что качество этих оценок существенно зависит от выбора параметра α, $0 \le \alpha < 1/2$. Например, если мы заинтересованы в уменьшении асимптотической дисперсии оценок, то для распределений «близких по затянутости хвостов» к нормальному, величину α следует выбирать близкой к нулю, для распределений с «тяжелыми хвостами» (например, Лапласа, Коши) параметр α следует выбирать близким к $1/2$. Таким образом, выбор параметра α можно связать с поведением функционала $Q(F;\nu,\mu)$, который характеризует степень «тяжести хвостов» распределений при их изменении в заданной супермодели. Однако на практике функция распределения F наблюдений $X_1,...,X_n$ обычно неизвестна, поэтому естественно использовать вместо функционала $Q(F;\nu,\mu)$ его выборочную оценку, построенную по исходной выборке $X_1,...,X_n$. Выборочная оценка $Q(F_n)$ функционала $Q(F;\nu,\mu)$, построенная по выборке $X_1,...,X_n$ методом подстановки, записывается в виде

$$Q(F_n;\nu,\mu)=\frac{\dfrac{m}{k}\left(\sum\limits_{i=n-k+1}^{n}X_{(i)}-\sum\limits_{i=1}^{k}X_{(i)}\right)}{\left(\sum\limits_{i=n-m+1}^{n}X_{(i)}-\sum\limits_{i=1}^{m}X_{(i)}\right)},$$

$$k=[\nu n],\ m=[\mu n], \qquad \textbf{(18.2)}$$

где $0<\nu<\mu\le 0{,}5$ и $X_{(1)},...,X_{(n)}$ – порядковые статистики выборки $X_1,...,X_n$. Следуя работе [60], везде ниже полагаем $\nu=0{,}2$ и $\mu=0{,}5$. Отметим, что $Q(F_n)\xrightarrow{p}Q(F;\nu,\mu)$ при $n\to\infty$. Кроме того, результаты моделирования (см., например, [9]) показали, что уже при объемах выборки $n\ge 20$ статистика $Q(F_n)$ может быть использована для определения типов распределений, различающихся их степенью «тяжести хвостов». Итак, сопоставляя поведение абсолютных эффективностей оценок в зависимости от параметра α и поведение функционала $Q_F(\nu;\mu)$ в рамках заданной супермодели, можно предложить процедуру выбора параметра α на основе информации, содержащейся в исходной выборке $X_1,...,X_n$, точнее на основе выборочной оценки $Q(F_n)$ функционала $Q(F;\nu,\mu)$, построенной по исходной выборке $X_1,...,X_n$ из распределения F. Проиллюстрируем на примерах такой адаптивный выбор параметра $\hat{\alpha}(X_1,...,X_n)$ для различных оценок в рамках различных супермоделей.

Пример 18.3. Рассмотрим супермодель с засорением

$$\Im_{\varepsilon,\tau}(\Phi) = \{F : F(x) = \Phi_{\varepsilon,\tau}(x)\},$$

где $\Phi_{\varepsilon,\tau}(x) = (1-\varepsilon)\Phi(x) + \varepsilon\Phi(x/\tau)$. Предполагаем, что пропорция засорения ε может изменяться и удовлетворяет неравенствам $0 \le \varepsilon \le 0{,}3$, а масштабный параметр τ известен и $\tau = 3$. Определим параметр α для $\overline{X}_\alpha$-оценки в виде

$$\hat{\alpha}(X_1,...,X_n) = \begin{cases} 0{,}00\,, 1{,}58 < Q(F_n) \le 1{,}75, \\ 0{,}05, 1{,}75 < Q(F_n) \le 1{,}80, \\ 0{,}10, 1{,}80 < Q(F_n) \le 1{,}87, \\ 0{,}20, 1{,}87 < Q(F_n) \le 2{,}00. \end{cases} \quad \mathbf{(18.4)}$$

При таком выборе параметра $\hat{\alpha}(X_1,...,X_n)$ абсолютная эффективность адаптивной $\overline{X}_{\hat{\alpha}}$-оценки не опускается ниже уровня 0,95 при изменении пропорция засорения ε, $0 \le \varepsilon \le 0{,}3$, то есть в рамках заданной супермодели для абсолютной эффективности $\overline{X}_{\hat{\alpha}}$-оценки выполняются неравенства $0{,}95 \le АЭ(\Phi_{\varepsilon,\tau}, \overline{X}_{\hat{\alpha}}) \le 1$, при $\tau = 3$, $0 \le \varepsilon \le 0{,}3$, $n \ge 40$. Если же не адаптировать выбор параметра α, а использовать фиксированное значение, например $\alpha = 0{,}05$, то абсолютная эффективность $\overline{X}_\alpha$-оценки в рамках данной супермодели может опуститься до уровня 0,74, а при $\alpha = 0{,}00$ – до уровня 0,55.

Пример 18.5. Рассмотрим супермодель с засорением

$$\Im_{\varepsilon,\tau}(\Phi) = \{F : F(x) = \Phi_{\varepsilon,\tau}(x)\}.$$

Предполагаем, что пропорция засорения ε может изменяться в большем диапазоне, то есть удовлетворяет неравенствам $0 \le \varepsilon \le 0{,}5$, а масштабный параметр τ, как и прежде, известен и равен $\tau = 3$. Определим параметр α для винзоризованной $\tilde{X}_\alpha$-оценки в виде

$$\hat{\alpha}(X_1,...,X_n) = \begin{cases} 0{,}00\,, 1{,}58 < Q(F_n) \le 1{,}76, \\ 0{,}10, 1{,}76 < Q(F_n) \le 1{,}84, \\ 0{,}20, 1{,}84 < Q(F_n) \le 1{,}88, \\ 0{,}30, 1{,}88 < Q(F_n) \le 2{,}00. \end{cases} \quad \mathbf{(18.6)}$$

При таком выборе параметра $\hat{\alpha}(X_1,...,X_n)$ абсолютная эффективность адаптивной винзоризованной $\tilde{X}_{\hat{\alpha}}$-оценки не опускается ниже уровня 0,92 при изменении пропорция засорения ε, $0 \le \varepsilon \le 0{,}5$, то есть в рамках заданной супермодели для абсолютной эффективности адаптивной винзоризованной $\tilde{X}_{\hat{\alpha}}$-оценки выполняются неравенства $0{,}92 \le АЭ(\Phi_{\varepsilon,\tau}, \tilde{X}_{\hat{\alpha}}) \le 1$, при $\tau = 3$, $0 \le \varepsilon \le 0{,}5$, $n \ge 40$. Если же не адаптировать выбор параметра α, а использовать фиксированное значение, например $\alpha = 0{,}10$, то абсолютная эффективность $\tilde{X}_{\alpha}$-оценки в рамках данной супермодели может опуститься до уровня 0,57.

Пример 18.7. Рассмотрим супермодель с засорением

$$\Im_{\varepsilon,\tau}(\Phi) = \{F : F(x) = \Phi_{\varepsilon,\tau}(x)\}.$$

Предполагаем, что пропорция засорения ε может изменяться и удовлетворяет неравенствам $0 \le \varepsilon \le 0{,}5$, а масштабный параметр τ, как и прежде, известен и равен $\tau = 3$. Определим параметр α для HL_{α}-оценки в виде

$$\hat{\alpha}(X_1,...,X_n) = \begin{cases} 0{,}00, & 1{,}58 < Q(F_n) \le 1{,}84, \\ 0{,}10, & 1{,}84 < Q(F_n) \le 1{,}90, \\ 0{,}20, & 1{,}90 < Q(F_n) \le 2{,}00, \\ 0{,}25, & 2{,}00 < Q(F_n). \end{cases} \tag{18.8}$$

При таком выборе параметра $\hat{\alpha}(X_1,...,X_n)$ абсолютная эффективность адаптивной $HL_{\hat{\alpha}}$-оценки не опускается ниже уровня 0,95 при изменении пропорция засорения ε, $0 \le \varepsilon \le 0{,}5$, то есть в рамках заданной супермодели для абсолютной эффективности $HL_{\hat{\alpha}}$-оценки выполняются неравенства $0{,}95 \le АЭ(\Phi_{\varepsilon,\tau}, \overline{X}_{\hat{\alpha}}) \le 1$, при $\tau = 3$, $0 \le \varepsilon \le 0{,}3$, $n \ge 40$. Если же не адаптировать выбор параметра α, а использовать фиксированное значение, например $\alpha = 0{,}00$, то абсолютная эффективность HL-оценки Ходжеса – Лемана в рамках данной супермодели может опуститься до уровня 0,86.

Замечание 18.9. В рассмотренных примерах при адаптивном выборе параметра α использовались дискретные значения этого параметра, и параметр $\hat{\alpha}(X_1,...,X_n)$ вычислялся с помощью ступенчатой, разрывной функции от $Q(F_n)$ вида (18.4). Отметим, что эту разрывную функцию можно аппроксимировать непрерывной функцией, например линейной функцией, и вычислять параметр $\hat{\alpha}(X_1,...,X_n)$ по формуле

$$\hat{\alpha}(X_1,...,X_n) = \begin{cases} \alpha_1, \; Q(F_n) \le Q_1 \\ \dfrac{\alpha_2 - \alpha_1}{Q_2 - Q_1}\{Q(F_n) - Q_1\} + \alpha_1, \; Q_1 < Q(F_n) < Q_2, \\ \alpha_2, \; Q(F_n) \ge Q_2 \end{cases} \quad \textbf{(18.10)}$$

где параметры α_1, α_2, Q_1 и Q_2 задаются в соответствии с рассматриваемым типом супермодели и выборочная оценка $Q(F_n)$ функционала $Q(F;\nu;\mu)$ определена в (18.2).

Пример 18.11. Рассмотрим супермодель в виде конечного набора $\Im_S^* = \{F_{(1)}, F_{(2)}, F_{(3)}, F_{(4)}\}$ стандартных симметричных распределений: $F_{(1)}$ – ф.р. нормальная, $F_{(2)}$ – ф.р. логистическая, $F_{(3)}$ – ф.р. Лапласа, $F_{(4)}$ – ф.р. Коши. Отметим, что эти распределения s-упорядочены (см. [9]), то есть выполняется выражение $F_{(1)} <_S F_{(2)} <_S F_{(3)} <_S F_{(4)}$. В рамках данной супермодели сравним асимптотические дисперсии выборочной медианы $\overline{X}_{1/2}$, оценки Ходжеса – Лемана (HL-оценки) и адаптивной $HL_{\hat{\alpha}}$-оценки, для которой параметр $\hat{\alpha}(X_1,...,X_n)$ вычисляется по формуле (18.10) при следующих значениях параметров: $\alpha_1 = 0$, $\alpha_2 = 0,4$ и $Q_1 = 1,75$, $Q_{21} = 1,95$. Результаты сравнения приведены в табл. 18.1 в виде отношения дисперсии оценки к минимальной среди них дисперсии при заданном распределении $F \in \Im_S^*$.

Из приведенной таблицы наглядно видны в рамках рассматриваемой супермодели $\Im_S^*$ преимущества адаптивной $HL_{\hat{\alpha}}$-оценки перед выборочной медианой $\overline{X}_{1/2}$ и HL-оценкой Ходжеса – Лемана. Отметим, что приведенные результаты являются асимптотическими. Однако, как показывают результаты моделирования при конечных объемах выборки (см. табл. 18.2), можно ожидать, что это преимущество сохраниться и при умеренных объемах выборки (например, при $n \geq 40$).

Таблица 18.1

Отношения дисперсий оценок $\hat{\theta}$ для распределений $F \in \mathfrak{I}_S^*$

$\hat{\theta} \backslash F$	$F_{(1)}$	$F_{(2)}$	$F_{(3)}$	$F_{(4)}$
$\bar{X}_{1/2}$	1,50	1,33	1,00	1,08
HL	1,00	1,00	1,33	1,43
$HL_{\hat{\alpha}}$	1,00 ($\hat{\alpha} = 0,00$)	1,02 ($\hat{\alpha} = 0,10$)	1,08 ($\hat{\alpha} = 0,34$)	1,00 ($\hat{\alpha} = 0,40$)

Пример 18.12. При изучении свойств робастности оценок используют различные числовые характеристики. Для асимптотически нормальных оценок наибольший интерес представляет их дисперсия. Однако, с точки зрения влияния выбросов в выборке на оценку, важными являются и такие характеристики, как чувствительность к грубым ошибкам, чувствительность к группировке и округлению наблюдений, предел устойчивости (или точка срыва). Каждая из этих характеристик может быть положена в основу адаптации с целью улучшения качества оценки. Например, свойства *MD*-оценок существенно зависят от выбора весовой функции *W*. В связи с этим, адаптивный выбор весовой функции *W* позволяет обеспечивать требуемые качества *MD*-оценки для заданной супермодели. Рассмотрим в качестве примера супермодель с засорением $\mathfrak{I}_{\varepsilon,\tau}(\Phi) = \{F : F(x) = \Phi_{\varepsilon,\tau}(x)\}$. Предполагаем, что пропорция засорения ε может изменяться и удовлетворяет неравенствам $0 \leq \varepsilon \leq 0,3$, а масштабный параметр τ известен и $\tau = 3$. Для *MD*-оценки выберем опорное распределение $F_0 = \Phi$ и определим адаптивную весовую функцию в виде

$$\hat{W}(x;\, X_1,\dots,X_n) = \begin{cases} 1/\phi(x), & 1,71 < Q(F_n) \leq 1,76, \\ 1, & 1,76 < Q(F_n) \leq 1,86, \\ \phi(x), & 1,86 < Q(F_n) \leq 1,91. \end{cases} \qquad \mathbf{(18.13)}$$

При таком выборе весовой функции $\hat{W}(x;\, X_1,...,X_n)$, абсолютная эффективность адаптивной MD-оценки не опускается ниже уровня 0,95 при изменении пропорция засорения ε, $0 \le \varepsilon \le 0{,}3$, то есть в рамках заданной супермодели для абсолютной эффективности выполняются неравенства $0{,}95 \le AЭ(\Phi_{\varepsilon,\tau}, \hat{W}) \le 1$, при $\tau = 3$, $0 \le \varepsilon \le 0{,}3$, $n \ge 40$. Если же выбрать весовую функцию Андерсона – Дарлинга в виде $\tilde{W}(x, \phi) = \phi(x)/\Phi(x)(1-\Phi(x))$, то абсолютная эффективность MD-оценки с такой весовой функцией в рамках данной супермодели может опуститься до уровня 0,47.

Пример 18.14. Рассмотрим супермодель

$$\Im_S^* = \{F_{(1)}, F_{(2)}, F_{(3)}, F_{(4)}\}$$

в виде конечного набора стандартных распределений. В рамках данной супермодели приведем асимптотические и выборочные дисперсии семейства HL_α-оценок и адаптивной $HL_{\hat{\alpha}}$-оценки, для которой параметр $\hat{\alpha}(X_1,...,X_n)$ вычисляется по формуле (18.10) при следующих значениях параметров: $\alpha_1 = 0$, $\alpha_2 = 0{,}5$, $Q_1 = 1{,}75$, $Q_2 = 2{,}50$. В табл. 18.2 приведены асимптотические дисперсии (строки $n = \infty$) для $\sqrt{n}HL_\alpha$-оценок, вычисленные по формуле (9.51), и выборочные дисперсии (строки $n = 20$) для HL_α-оценок, полученные методом статистических испытаний при числе испытаний $M = 10\,000$ и $n = 20$. Выборочные дисперсии вычислялись по формуле

$$S_n^2(\hat{\theta}) = (n/M)\sum_{i=1}^{M}(\hat{\theta}_i - \overline{\theta})^2,\ \overline{\theta} = (1/M)\sum_{i=1}^{M}\hat{\theta}_i,$$

где $\hat{\theta}$ обозначает одну из рассматриваемых оценок, построенных по исходной выборке $X_1,...,X_n$ при $n = 20$.

В строках $(20/\infty)$ этой таблицы приведены отношения выборочных дисперсий HL_α-оценок при $n = 20$ к асимптотическим дисперсиям. В правом столбце таблицы приведены данные для адаптивной $HL_{\hat{\alpha}}$-оценки Ходжеса – Лемана, для которой параметр $\hat{\alpha}(X_1,...,X_n)$ определен выражением (18.10) при значении параметров: $\alpha_1 = 0$, $\alpha_2 = 0{,}5$, $Q_1 = 1{,}75$, $Q_2 = 2{,}50$.

Таблица 18.2

Асимптотические и выборочные дисперсии HL_α-оценок для супермодели $\mathfrak{I}_S^* = \{F_1, F_2, F_3, F_4\}$

Ф	n	HL	$HL_{0,05}$	$HL_{0,10}$	$HL_{0,20}$	$HL_{0,30}$	$HL_{0,40}$	$HL_{0,50}$	$HL_{\hat{\alpha}}$
$F_{(1)}$ – Гаусс	$n=20$	1,04	1,07	1,09	1,15	1,25	1,39	1,43	1,04
	$n=\infty$	1,047	1,060	1,085	1,156	1,256	1,390	1,571	1,047
	$20/\infty$	0,99	1,01	1,00	0,99	1,00	1,00	0,91	0,99
$F_{(2)}$ – логист.	$n=20$	3,06	3,09	3,11	3,24	3,50	3,87	3,96	3,14
	$n=\infty$	3,000	3,002	3,016	3,099	3,273	3,561	4,000	3,000
	$20/\infty$	1,02	1,03	1,03	1,05	1,07	1,09	0,99	1,05
$F_{(3)}$ – Лаплас	$n=20$	1,40	1,37	1,33	1,26	1,23	1,23	1,25	1,39
	$n=\infty$	1,333	1,322	1,296	1,224	1,146	1,070	1,000	1,306
	$20/\infty$	1,05	1,04	1,03	1,03	1,07	1,15	1,25	1,06
$F_{(4)}$ – Коши	$n=20$	3,93	3,66	3,36	2,81	2,47	2,43	2,45	2,66
	$n=\infty$	3,290	3,208	3,025	2,616	2,345	2,283	2,467	2,467
	$20/\infty$	1,19	1,14	1,11	1,07	1,05	1,06	0,99	1,08
$d(HL_\alpha, \mathfrak{I}_S^*)$	$n=20$	0,40	0,35	0,29	0,18	0,21	0,33	0,36	**0,15**

Из данных табл. 18.2 следует, что приведенные выше асимптотические результаты являются вполне приемлемой аппроксимацией дисперсий $\sqrt{n}HL_\alpha$-оценок при конечных объемах выборки $n \geq 20$, за исключением некоторых значений при распределениях Лапласа и Коши. Для этих распределений с «тяжелыми хвостами», как показали результаты моделирования, качество асимптотики существенно улучшается при объемах выборки $n \geq 40$. Чтобы проиллюстрировать преимущество предложенных адаптивных $HL_{\hat{\alpha}}$-оценок, применим критерий сравнения оценок, основанный на понятии дефекта оценки. Напомним, что дефект оценки $\hat{\theta}_i$, $i = 1,...,k$, среди сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$ параметра θ при заданном распределении F определяется в виде

$$DE(F, \hat{\theta}_i) = 1 - \min\{\sigma^2(F, \hat{\theta}_1),...,\sigma^2(F, \hat{\theta}_k)\} / \sigma^2(F, \hat{\theta}_i),$$
$$i = 1,...,k. \qquad \textbf{(18.15)}$$

Если же мы хотим сделать вывод о предпочтительности оценки среди сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$ параметра θ в рамках всей рассматриваемой супермодели $\Im = \{F_1,...,F_r\}$, то вычислим евклидову метрику, которая, с использованием введенных обозначений, запишется в виде

$$d(\hat{\theta}_i;\Im) = \left\{\sum_{j=1}^{r}[DE(F_j,\hat{\theta}_i)]^2\right\}^{1/2}, \; i = 1,...,k. \qquad \mathbf{(18.16)}$$

Предпочтение в рамках всей рассматриваемой супермодели $\Im = \{F_1,...,F_r\}$ отдается той оценке $\hat{\theta}_i$ среди сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$, для которой вычисленное значение евклидовой метрики $d(\hat{\theta}_i;\Im)$ *минимальное*, то есть $d(\hat{\theta}_i;\Im) = \min\{d(\hat{\theta}_1;\Im),...,d(\hat{\theta}_k;\Im)\}$. Вычисленное значение критерия (18.15) для семейства HL_α-оценок, с включением адаптивной $HL_{\hat{\alpha}}$-оценки, приведены в нижней строке табл. 18.2. Отметим, что в этой таблице значения евклидовой метрики были вычислены по эмпирическим данным (при $n = 20$) с использованием формул (18.15) и (18.16).

Итак, в рамках супермодели $\Im_S^* = \{F_1, F_2, F_3, F_4\}$, согласно критерию (18.16), среди рассматриваемых оценок предпочтение следует отдать адаптивной $HL_{\hat{\alpha}}$-оценке, так как значение $d(HL_{\hat{\alpha}}, \Im_S^*) = 0{,}15$ является минимальным в нижней строке таблицы (18.2).

Пример 18.17. Рассмотрим супермодель с засорением

$$\Im_{\varepsilon,\tau}(\Phi) = \{F : F(x) = \Phi_{\varepsilon,\tau}(x)\}.$$

Предполагаем, что пропорция засорения ε может изменяться, то есть удовлетворяет неравенствам $0 \le \varepsilon \le 0{,}4$, а масштабный параметр τ, как и прежде, известен и равен $\tau = 3$.

Определим параметр $\hat{\alpha}(X_1,...,X_n)$ для адаптивной $HL_{\hat{\alpha}}$-оценки по формуле (18.10) и примем следующие значения параметров: $\alpha_1 = 0$, $\alpha_2 = 0{,}5$, $Q_1 = 1{,}75$, $Q_2 = 2{,}00$. В табл. 18.3 приведены асимптотические дисперсии (строки $n = \infty$) HL_α-оценок, вычисленные по формуле (9.52), и выборочные дисперсии (строки $n = 20$) HL_α-

оценок, полученные методом статистических испытаний при числе испытаний $M = 10000$ и объёме выборки $n = 20$. Для сравнения HL_α-оценок с адаптивной $HL_{\hat\alpha}$-оценкой в рамках супермодели $\Im_{\varepsilon,\tau}(\Phi)$, как и в примере 18.14, воспользуемся критерием, основанным на евклидовой метрике $d(HL_\alpha, \Im_{\varepsilon,\tau})$, которая была вычислена по эмпирическим данным (при $n = 20$) табл. 18.3 с использованием формул (18.15) и (18.16).

Таблица 18.3

Асимптотические и выборочные дисперсии HL_α-оценок для $F \in \Im_{\varepsilon,\tau}(\Phi)$ при $\tau = 3$

при	n	HL	$HL_{0,05}$	$HL_{0,10}$	$HL_{0,20}$	$HL_{0,30}$	$HL_{0,40}$	$HL_{0,50}$	$HL_{\hat\alpha}$
$\varepsilon = 0,00$	$n = 20$	1,04	1,06	1,10	1,19	1,30	1,42	1,46	1,05
	$n = \infty$	1,047	1,060	1,085	1,156	1,256	1,390	1,571	1,047
	$20/\infty$	0,99	1,00	1,01	1,03	1,04	1,02	0,93	1,00
$\varepsilon = 0,05$	$n = 20$	1,17	1,18	1,20	1,29	1,40	1,55	1,58	1,17
	$n = \infty$	1,171	1,172	1,189	1,252	1,351	1,490	1,681	1,171
	$20/\infty$	1,00	1,01	1,01	1,03	1,04	1,04	0,94	1,00
$\varepsilon = 0,10$	$n = 20$	1,38	1,38	1,40	1,46	1,55	1,67	1,70	1,32
	$n = \infty$	1,311	1,302	1,308	1,360	1,457	1,600	1,803	1,303
	$20/\infty$	1,05	1,06	1,07	1,07	1,06	1,04	0,94	1,01
$\varepsilon = 0,20$	$n = 20$	1,80	1,77	1,75	1,75	1,85	1,94	1,97	1,65
	$n = \infty$	1,651	1,628	1,605	1,622	1,709	1,861	2,091	1,617
	$20/\infty$	1,09	1,09	1,09	1,08	1,08	1,04	0,94	1,02
$\varepsilon = 0,30$	$n = 20$	2,14	2,08	2,00	1,97	2,05	2,21	2,24	1,95
	$n = \infty$	2,090	2,062	2,004	1,966	2,032	2,191	2,454	1,993
	$20/\infty$	1,02	1,01	1,00	1,00	1,01	1,01	0,91	0,98
$\varepsilon = 0,40$	$n = 20$	2,63	2,57	2,49	2,44	2,53	2,79	2,87	2,50
	$n = \infty$	2,655	2,627	2,543	2,425	2,455	2,616	2,921	2,483
	$20/\infty$	0,99	0,98	0,98	1,01	1,03	1,07	0,98	1,01
$d(HL_\alpha, \Im_{\varepsilon,\tau})$	$n = 20$	0,15	0,12	0,11	0,19	0,32	0,48	0,52	**0,03**

Итак, в рамках супермодели $\Im_{\varepsilon,\tau}(\Phi)$, согласно критерию (18.16), среди рассматриваемых оценок предпочтение следует отдать адаптивной $HL_{\hat\alpha}$-оценке, так как значение $d(HL_{\hat\alpha}, \Im_{\varepsilon,\tau}) = 0,03$ является минимальным в нижней строке табл. 18.3. На рис. 18.1 приведены зависимости абсолютных эффективностей оценок от параметра засорения ε при $\tau = 3$. Из рисунка наглядно видно, что адаптивная $HL_{\hat{\alpha}}$-оценка обладает преимуществом для $0 \le \varepsilon < 0,25$ перед HL_{α}-оценками с фиксированной пропорцией урезания α исходной выборки $X_1,...,X_n$.

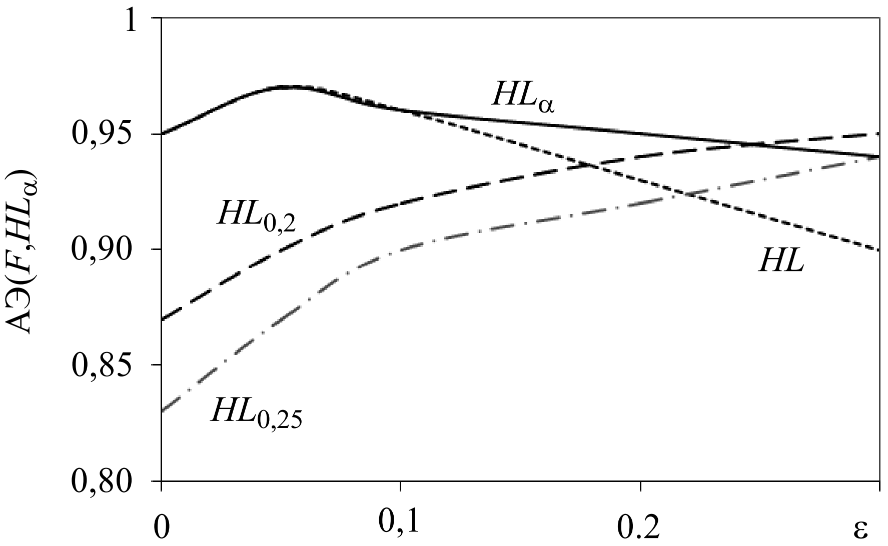

Рис. 18.1. Абсолютные эффективности HL_{α}-оценок в рамках супермодели $\Im_{\varepsilon,\tau}(\Phi)$, $\tau = 3$

Замечание 18.18. Отметим, что в литературе используют различные подходы к построению адаптивных оценок. Например, в работах Джекеля [25,26] для построения адаптивной $\overline{X}_{\alpha}$-оценки используется выборочная оценка асимптотической дисперсии $\sqrt{n}\overline{X}_{\alpha}$-оценок, которая строится по исходной выборке $X_1,...,X_n$ и записывается в виде

$$S^2(\alpha)=\frac{1}{(1-2\alpha)^2}\left\{\frac{1}{n}\sum_{i=k+1}^{n-k}(X_{(i)}-\overline{X}_{\alpha})^2+\alpha(X_{(k+1)}-\overline{X}_{\alpha})^2+\alpha(X_{(n-k)}-\overline{X}_{\alpha})^2\right\},$$

$$k=[\alpha n].$$

Далее, параметр $\hat{\alpha}(X_1,...,X_n)$ адаптивной $\overline{X}_{\hat{\alpha}}$-оценки определяется выражением $\hat{\alpha}(X_1,...,X_n)=\arg\min\limits_{\alpha} S^2(\alpha)$. Асимптотические свойства такой адаптивной $\overline{X}_{\hat{\alpha}}$-оценки исследованы в работе Джекеля [26], а её свойства при конечных объемах выборки (при $n = 20$) иллюстрируются в табл. 4.6 книги Лемана [73] , в которой также приводятся (на стр. 327) ссылки на работы других авторов. Некоторые модификации адаптивных оценок Джекеля приводятся в [24]. Дополнительные сведения по адаптивным оценкам могут быть найдены в работах [60, 72]. Заключая данный раздел, отметим, что адаптивный подход является практически важным инструментом, который на основе изучения свойств оценок для различных супермоделей позволяет обеспечивать требуемое качество оценок в рамках заданной супермодели.

ПРИЛОЖЕНИЯ

П.1. Метод проекций

Предположим, что $X_1,...,X_n$ – независимые, но необязательно одинаково распределенные случайные величины с функциями распределения $F_{X_i}(x)$, $i=1,...,n$, $x \in R^1$. Пусть $S^*(\vec{X}) = S^*(X_1,...,X_n)$ – некоторая статистика с конечным математическим ожиданием, для которой требуется конкретизировать асимптотическое распределение с помощью метода проекций. Обозначим Гильбертово пространство статистик через $\aleph = \{S^*(\vec{X}) : M\{S^*(\vec{X})\} < \infty\}$, для него определено скалярное произведение в виде $M\{S^*T^*\}$, S^* и $T^* \in \aleph$. Выделим линейное подпространство $\hbar$ статистик, которые записываются в виде суммы некоторых заданных функций $\psi_i(X_i)$, $i=1,...,n$, с конечными математическими ожиданиями, то есть

$$\hbar = \{\tilde{S}^*(\vec{X}) = \sum \psi_i(X_i) : M\{\psi_i(X_i)\} < \infty, i=1,..., n\}. \quad \textbf{(П.1.1)}$$

Проекция статистики $S^* \in \aleph$ на подпространство $\hbar$ линейных статистик определяется следующим образом.

Определение П.1.2. Проекция статистики $S^*(\vec{X}) \in \aleph$ на подпространство $\hbar$ линейных статистик вида $\tilde{S}^*(\vec{X}) = \sum \psi_i(X_i)$ определяется выражением

$$\tilde{S}^*_p(\vec{X}) = \sum_{i=1}^{n} M\{S^*(\vec{X}) \mid X_i\} - (n-1)M\{S^*(\vec{X})\},$$

причем

$$M(\tilde{S}^*_p(\vec{X})) = M\{S^*(\vec{X})\}, M\{S^*(\vec{X}) - \tilde{S}^*_p(\vec{X})\}^2 = D\{S^*(\vec{X})\} - D\{\tilde{S}^*_p(\vec{X})\}.$$

Напомним, что статистика $S^*(\vec{X})$ имеет конечное математическое ожидание, поэтому обычно ее центрируют, то есть переходят к рассмотрению статистики $S(\vec{X}) = S^*(\vec{X}) - M\{S^*(\vec{X})\}$, для которой математическое ожидание равно нулю. В этом случае теорема о проекции статистики $S(\vec{X})$ на подпространство $\hbar$ линейных статистик вида $\tilde{S}^*(\vec{X}) = \sum \psi_i(X_i)$ формулируется следующим образом.

Теорема П.1.3. Проекция $S_p(\vec{X})$ статистики $S(\vec{X}) \in \aleph$, для которой $M\{S(\vec{X})\} = 0$, на подпространство $\hbar$ линейных статистик вида

$$\tilde{S}(\vec{X}) = \sum_{i=1}^{n} \psi_i(X_i) \qquad \textbf{(П.1.4)}$$

определяется такими функциями $\psi_i^*(x)$, $i = 1,...,n$, которые минимизируют $M\{S(\vec{X}) - \tilde{S}(\vec{X})\}^2$ и записываются в виде

$$\psi_i^*(x) = M\{S(\vec{X}) \,|\, X_i = x\},\ i = 1,...,n, \qquad \textbf{(П.1.5)}$$

то есть проекция $S_p(\vec{X})$ определяется выражением

$$S_p(\vec{X}) = \sum_{i=1}^{n} \psi_i^*(X_i) = \sum_{i=1}^{n} M_{X_i}\{S(\vec{X}) \,|\, X_i\}, \qquad \textbf{(П.1.6)}$$

причем
$$M(S_p(\vec{X})) = M\{S(\vec{X})\} = 0 \qquad \textbf{(П.1.7)}$$

и
$$M\{S(\vec{X}) - S_p(\vec{X})\}^2 = D\{S(\vec{X})\} - D\{S_p(\vec{X})\}. \qquad \textbf{(П.1.8)}$$

В частном случае, если $X_1,...,X_n$ – независимые и одинаково распределенные случайные величины, то есть $F_{X_i}(x) = F_X(x)$, $\forall i = 1,...,n$, то функции $\psi_i^*(x) = M\{S(\vec{X}) \,|\, X_i = x\}$ не зависят от индекса i и проекция $S_p(\vec{X})$ записывается в виде суммы независимых и одинаково распределенных случайных величин $\psi^*(X_1),...,\psi^*(X_n)$, то есть в виде

$$S_p(\vec{X}) = \sum_{i=1}^{n} \psi^*(X_i) = \sum_{i=1}^{n} M\{S(\vec{X}) \,|\, X_i\}. \qquad \textbf{(П.1.9)}$$

Доказательство. Приведем доказательство для случая, когда $X_1,...,X_n$ независимые и одинаково распределенные случайные величины с ф.р. $F_X(x)$, $x \in R^1$. Напомним, что исходная статистика $S(\vec{X})$ имеет нулевое математическое ожидание. Учитывая, что $X_1,...,X_n$ – н.о.р. случайные величины с ф.р. $F_X(x)$, согласно формуле полного математического ожидания, получаем

$$M\{S_p(\vec{X})\} = M\{\sum_{i=1}^{n} \psi^*(X_i)\} = \sum_{i=1}^{n} M[M\{S(\vec{X}) \,|\, X_i\}] = \sum_{i=1}^{n} M\{S(\vec{X})\} = 0. \qquad \textbf{(П.1.10)}$$

Убедимся теперь в том, что выбор функции $\psi^*(x)$ в виде

$$\psi^*(x) = M\{S(\vec{X}) \,|\, X_i = x\}$$

минимизирует выражение $M\{S(\vec{X})-\tilde{S}(\vec{X})\}^2$. Распишем это выражение в виде

$$M\{S(\vec{X})-\tilde{S}(\vec{X})\}^2 = M\{[S(\vec{X})-S_p(\vec{X})]+[S_p(\vec{X})-\tilde{S}(\vec{X})]\}^2 =$$
$$= M\{S(\vec{X})-S_p(\vec{X})\}^2 + M\{S_p(\vec{X})-\tilde{S}(\vec{X})\}^2 +$$
$$+2M\{S(\vec{X})-S_p(\vec{X})\}\{S_p(\vec{X})-\tilde{S}(X)\} \qquad \textbf{(П.1.11)}$$

и покажем, что последнее слагаемое в этом выражении равно нулю. В самом деле, используя (П.1.4) и (П.1.8), а также свойства математического ожидания, получаем

$$M[\{S_p(\vec{X})-\tilde{S}(X)\}\{S(\vec{X})-S_p(\vec{X})\}] = M[\sum_{i=1}^{n}[\psi^*(X_i)-\psi(X_i)\}\{S(\vec{X})-S_p(X)\}] =$$
$$= \sum_{i=1}^{n} M[\psi^*(X_i)-\psi(X_i)\}\{S(\vec{X})-S_p(X)\}].$$

Рассмотрим теперь отдельно каждое слагаемое последней суммы, то есть выражение

$$M[\{S(\vec{X})-S_p(\vec{X})\}\{\psi^*(X_i)-\psi(X_i)\}].$$

Используя формулу для полного математического ожидания, это выражение перепишем в виде

$$M[\{S(\vec{X})-S_p(\vec{X})\}\{\psi^*(X_i)-\psi(X_i)\}] =$$
$$= M_{X_i}[M[\{S(\vec{X})-S_p(\vec{X})\}\{\psi^*(X_i)-\psi(X_i)\}] \mid X_i] =$$
$$= M_{X_i}[\{\psi^*(X_i)-\psi(X_i)\}M[\{S(\vec{X})-S_p(\vec{X})\} \mid X_i]].$$

Убедимся теперь, что $M[\{S(\vec{X})-S_p(\vec{X})\} \mid X_i] = 0$. В самом деле, используя (П.1.5) и (П.1.9), а также выражение

$$M\{S_p(\vec{X}) \mid X_i = x\} = \sum_{i=1}^{n} M[\psi^*(X_i) \mid X_i = x] = \psi^*(x) + (n-1)M\{\psi^*(X_j)\},$$

с учетом того, что $M\{\psi^*(X_j)\} = M[M[\{S(\vec{X})\} \mid X_j]] = M\{\vec{S}(X)\} = 0$, для любого j, удовлетворяющего неравенствам $1 \le j \le n$ и $j \ne i$, получаем

$$M[\{S(\vec{X})-S_p(\vec{X})\} \mid X_i = x] = M\{S(\vec{X}) \mid X_i = x\} - \psi^*(x) -$$
$$-(n-1)M\{\psi(X_j)\} = \psi^*(x) - \psi^*(x) = 0.$$

Таким образом, убедились, что $M[\{S_p(\vec{X})-\tilde{S}(X)\}\{S(\vec{X})-S_p(\vec{X})\}] = 0$ и, следовательно, выражение (П.1.11) запишется в виде

$$M\{S(\vec{X})-\tilde{S}(\vec{X})\}^2 = M\{S(\vec{X})-S_p(\vec{X})\}^2 + M\{S_p(\vec{X})-\tilde{S}(\vec{X})\}^2. \qquad \textbf{(П.1.12)}$$

Отсюда следует выполнение неравенства

$$M\{S(\vec{X})-S_p(\vec{X})\}^2 \le M\{S(\vec{X})-\tilde{S}(\vec{X})\}. \qquad \textbf{(П.1.13)}$$

Отметим, что второе слагаемое в (П.1.12) обращается в ноль при $\tilde{S}(\vec{X})=S_p(\vec{X})$. Таким образом, выражение $M\{S(\vec{X})-\tilde{S}(\vec{X})\}^2$ достигает минимального значения при $\tilde{S}(\vec{X})=S_p(X)$. Доказательство завершено.

Итак, после того как для интересующей нас статистики $S(\vec{X})$ найдена проекция $S_p(\vec{X})$ вида (П.1.9), ее асимптотическая нормальность может быть установлена с помощью центральной предельной теоремы. Далее, можно доказать, что статистики $\sqrt{n}\,S(\vec{X})$ и $\sqrt{n}\,S_p(\vec{X})$ имеют одинаковые предельные распределения, опираясь на следующую теорему.

Теорема П.1.14. Пусть статистика W_n имеет асимптотически нормальное распределение, то есть $L\{W_n\}=N\{0;D(W_n)\}$, и пусть выполняется выражение $M\{U_n-W_n\}^2 \to 0$ при $n\to\infty$. Тогда статистика U_n имеет то же предельное распределение, что и W_n, то есть $U_n \xrightarrow{d} W_n$ при $n\to\infty$.

Доказательство. Согласно неравенству Чебышева, при выполнении условия $M\{U_n-W_n\}^2 \to 0$, $n\to\infty$, выполняется следующее выражение:

$$\forall\varepsilon>0\ \ P\{|U_n-W_n|\ge\varepsilon\}\le\frac{M(U_n-W_n)^2}{\varepsilon^2}\to 0 \text{ при } n\to\infty. \qquad \textbf{(П.1.15)}$$

Отсюда следует, что $(U_n-W_n)\xrightarrow{p}0$ и, согласно теореме Слуцкого, выполняется выражение

$$U_n \xrightarrow{d} W_n \text{ при } n\to\infty. \qquad \textbf{(П.1.16)}$$

Доказательство завершено.

П.2. Асимптотическая эффективность Питмена

Всякий T_n-критерий проверки статистических гипотез H_0, H_1, основанный на статистике $T_n(\vec{X})$, вычисляемой по выборке $X_1,...,X_n$ из распределения $F(x)$, характеризуется двумя вероятностями ошибок первого и второго рода. Вероятность ошибки первого рода используется для определения критической области тестовой статистики T_n путем задания уровня значимости α-критерия. Вероятность ошибки второго рода связана с понятием функции мощности критерия, которое отражает поведение вероятности правильного решения в пользу верной альтернативы и которое зависит от многих параметров. От таких, как уровень значимости α, объем выборки n и параметр θ, который характеризует различие между проверяемыми гипотезами H_0, H_1. При сравнении двух состоятельных критериев с одинаковыми уровнями значимости α, с помощью поведения их функций мощности при конечных объемах выборок, возникают сложности, в связи с большим числом параметров, характеризующим поведение функций мощности критериев. Кроме того, асимптотическое сравнение критериев также невозможно, так как для состоятельных критериев их функции мощности при *фиксированных* альтернативах вырождаются в единицу при неограниченном увеличении объема выборки. Для сравнения критериев предложены различные подходы (см., например, [29]). Наиболее часто для этой цели используют асимптотическую относительную эффективность Питмена (см. [20, 28, 29, 45]), которая вычисляется для последовательности контигуальных альтернатив, сходящихся к нулевой гипотезе при неограниченном увеличении объема выборки. На качественном уровне асимптотическая относительная эффективность двух критериев, обозначаемая через $ARE_F(T_n : S_n) = Э_F^2(T_n) / Э_F^2(S_n)$ (см. теорему П.2.4), характеризует отношение объемов выборок n_1 и n_2, при которых T_n-критерий и S_n-критерий проверки заданных гипотез H_0, H_1 с одинаковыми уровнями значимости обеспечивают одинаковую асимптотическую мощность против одной и той же последовательности альтернатив, сходящихся к нулевой гипотезе. Следуя [45], приведем необходимые определения.

Определение П.2.1. Пусть имеется два критерия размера α для проверки гипотез $H_0 : \theta = 0, H_1 : \theta > 0$ по выборке $X_1, ..., X_n$ из распределения $F(x - \theta)$, основанные на статистиках $V_n^{(i)}$, $i = 1, 2$. Для заданных θ и W, $\alpha < W < 1$, обозначим через $n^{(i)}$, $i = 1, 2$, объемы выборок, при которых $V_n^{(i)}$-критерии обеспечивают заданную мощность W, то есть $P\{V_n^{(i)} \geq k_{n,1-\alpha}^{(i)} \mid H_1\} = W$. Тогда эффективность $V_n^{(1)}$-критерия относительно $V_n^{(2)}$-критерия определяется в виде отношения объемов выборок $n^{(2)} / n^{(1)}$.

Например, если $n^{(2)} / n^{(1)} = 0{,}5$, то для достижения той же мощности $V_n^{(1)}$-критерию потребуется число наблюдений в два раза большее, чем $V_n^{(2)}$-критерию.

Определение П.2.2. Пусть имеется два критерия для проверки гипотез $H_0 : \theta = 0, H_1 : \theta > 0$ по выборке $X_1, ..., X_n$ из распределения $F(x - \theta)$, основанные на статистиках $V_n^{(i)}$, $i = 1, 2$, причем они имеют одинаковые асимптотические уровни значимости α, то есть

$$P\{V_n^{(i)} \geq k_{n,1-\alpha}^{(i)} \mid H_0\} \to \alpha \text{ при } n \to \infty,\ i = 1, 2 .$$

Далее, пусть для фиксированного W, $\alpha < W < 1$, имеется последовательность альтернатив $H_{1,n_j}:\{\theta_j\}$, с соответствующей ей последовательностью объемов выборки $\{n_j\}$, такая, что $\theta_j \to 0$, причем

$$P\{V_{n_j}^{(i)} \ge k_{n,1-\alpha}^{(i)} \mid H_{1,n_j}\} \to W \text{ при } n_j \to \infty,\ i = 1, 2\ .$$

Асимптотическая относительная эффективность Питмена $V^{(1)}$-критерия по отношению к $V^{(2)}$-критерию в англоязычной литературе обозначается через $ARE_F(V^{(1)}:V^{(2)})$ – аббревиатура от «Asymptotic Relative Efficiency», и определяется в виде предела отношения $n_j^{(2)}/n_j^{(1)}$, при $n_j \to \infty$ и при условии, что этот предел существует и не зависит от $\{\theta_j\}$, α, W, то есть

$$ARE_F(V^{(1)}:V^{(2)}) = \lim_{n_j \to \infty} \frac{n_j^{(2)}}{n_j^{(1)}}. \qquad \textbf{(П.2.3)}$$

Таким образом, асимптотическая относительная эффективность Питмена $V^{(1)}$-критерия по отношению к $V^{(2)}$-критерию, также часто обозначаемая в виде $АОЭ_F(V^{(1)}:V^{(2)})$, на качественном уровне характеризует отношения объемов выборки, которые необходимы при одинаковых уровнях значимости сравниваемых критериев для достижения ими одинаковой мощности при альтернативах, сходящихся к нулевой гипотезе при неограниченном увеличении объемов выборок.

Теорема П.2.4. Пусть имеется два критерия для проверки гипотез $H_0:\theta = 0, H_1:\theta > 0$ по выборке $X_1,...,X_n$ из распределения $F(x-\theta)$, основанные на статистиках $V_n^{(i)}$, $i = 1,\ 2$, причем они имеют одинаковые асимптотические уровни значимости α, то есть

$$P\{V_n^{(i)} \ge k_{n,1-\alpha}^{(i)} \mid H_0\} \to \alpha \text{ при } n \to \infty,\ i = 1, 2\ .$$

Далее, пусть для фиксированного W, $\alpha < W < 1$, имеется последовательность альтернатив $H_{1,n_j}:\{\theta_j\}$ с соответствующей ей последовательностью объемов выборки $\{n_j\}$, такая, что $\theta_j \to 0$, причем

$$P\{V_{n_j}^{(i)} \ge k_{n,1-\alpha}^{(i)} \mid H_{1,n_j}\} \to W \text{ при } n_j \to \infty,\ i = 1, 2\ .$$

Предполагается, что статистики $V_n^{(i)}$, $i = 1,\ 2$ удовлетворяют следующим условиям:

(A1) $V_n^{(i)}$-критерий состоятельный, $i = 1,\ 2$.

(A2) Существуют такие последовательности $\{\mu_n(\theta)\}$ и $\{\sigma_n(\theta)\}$, что

$$L\{[V_n - \mu_n(\theta)]/\sigma_n(\theta)\} = N(0;1),\ \forall\theta \text{ в окрестности } \theta = 0,$$

где $\mu_n(\theta) = M(V_n \mid H_1)$ и $\sigma_n(\theta) = \sqrt{D(V_n \mid H_1)}$.

(A3) Существует $ARE_F(V_n^{(1)} : V_n^{(2)}) = Э_F^2(V_n^{(1)}) / Э_F^2(V_n^{(2)})$.

(A4) Для последовательности $\{\theta_n\}$, такой, что $\theta_n \to 0$ при $n \to \infty$,

$$\sigma_n(\theta_n)/\sigma_n(0) \to 1 \text{ и } \mu_n'(\theta_n)/\mu_n'(0) \to 1 \text{ при } n \to \infty.$$

(A5) Эффективность Питмена $V_n^{(i)}$-критерия, $i = 1, 2$, определяется в виде

$$Э_F(V_n^{(i)}) = \lim_{n\to\infty} \frac{d}{d\theta} M\{V_n^{(i)} \mid H_1\}|_{\theta=0} \Big/ \sqrt{n D(V_n^{(i)} \mid H_0)},\ i = 1, 2\ . \qquad \textbf{(П.2.5)}$$

Тогда асимптотическая относительная эффективность Питмена для $V_n^{(1)}$-критерия по отношению к $V_n^{(2)}$-критерию определяется в виде

$$ARE_F(V_n^{(1)} : V_n^{(2)}) = \lim_{n_j\to\infty} \{n_j^{(2)} / n_j^{(1)}\} = Э_F^2(V_n^{(1)}) / Э_F^2(V_n^{(2)}). \qquad \textbf{(П.2.6)}$$

Теорема П.2.7. Пусть критерий размера α для проверки гипотез $H_0 : \theta = 0, H_1 : \theta > 0$ по выборке $X_1, ..., X_n$ из распределения $F(x - \theta)$, основан на статистике V_n, которая удовлетворяет условиям регулярности (A1) – (A5) теоремы П.2.4. Пусть критерий определяется критической областью размера α вида $X_{1\alpha}^{+} = \{(\vec{x}) : v_n \geq k_{n,1-\alpha}\}$, для которой асимптотическое критическое значение $k_{n,1-\alpha}$ определяется выражением $\lim\limits_{n\to\infty} P\{V_n \geq k_{n,1-\alpha} \mid H_0\} = \alpha$ и записывается в виде

$$k_{n,1-\alpha} = M(V_n \mid H_0) + \lambda_{1-\alpha}\sqrt{D(V_n \mid H_0)},\ \lambda_{1-\alpha} = \Phi^{-1}(1-\alpha). \qquad \textbf{(П.2.8)}$$

Тогда для последовательности альтернатив $H_{1,n} : \theta_n = \theta / n^{1/2}$ для фиксированного $\theta > 0$ асимптотическая функция мощности V_n-критерия вычисляется по формуле

$$\lim_{n\to\infty} P\{V_n \geq k_{n,1-\alpha} \mid H_{1,n}\} = 1 - \Phi(\lambda_{1-\alpha} - \theta \cdot Э_F(V_n)), \qquad \textbf{(П.2.9)}$$

где $Э_F(V_n)$ – эффективность Питмена V_n-критерия, определяемая выражением (П.2.5).

Доказательство этих теорем может быть найдено в [20, с. 351; 45, с. 81; 28].

П.3. Процедура Ходжеса – Лемана

Эта процедура является формализованным изложением свойства двойственности задач проверки статистических гипотез и задач построения интервальных оценок параметров (см., например, [20, 33, 45]).

Пусть $\vec{X} = (X_1,...,X_n)$ – выборка из распределения $F(x-\theta)$. Проверяются гипотезы $H_0:\theta = 0, H_1:\theta \neq 0$ с помощью критерия, основанного на статистике $V(\vec{X})$. Определим $V(\vec{X};\theta)$ путем замены X_i на $X_i - \theta$, $i = 1,...,n$, то есть $V(\vec{X};\theta) = V(X_1-\theta,...,X_n-\theta)$. Предполагается, что $V(\vec{X};\theta)$ удовлетворяет следующим условиям:

(А) Гипотеза $H_0:\theta = 0$ отвергается при больших значениях статистики $V(\vec{X})$.

(Б) $V(\vec{x};\theta)$ – невозрастающая функция от θ при заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X} = (X_1,...,X_n)$.

(В) При нулевой гипотезе статистика $V(\vec{X};\theta)$ свободна от распределения и ее распределение симметрично относительно μ_0.

Определим значения θ^* и θ^{**} следующим образом:

$$\theta^* = \sup\{\theta : V(\vec{x};\theta) > \mu_0\},\ \theta^{**} = \inf\{\theta : V(\vec{x};\theta) < \mu_0\}.$$

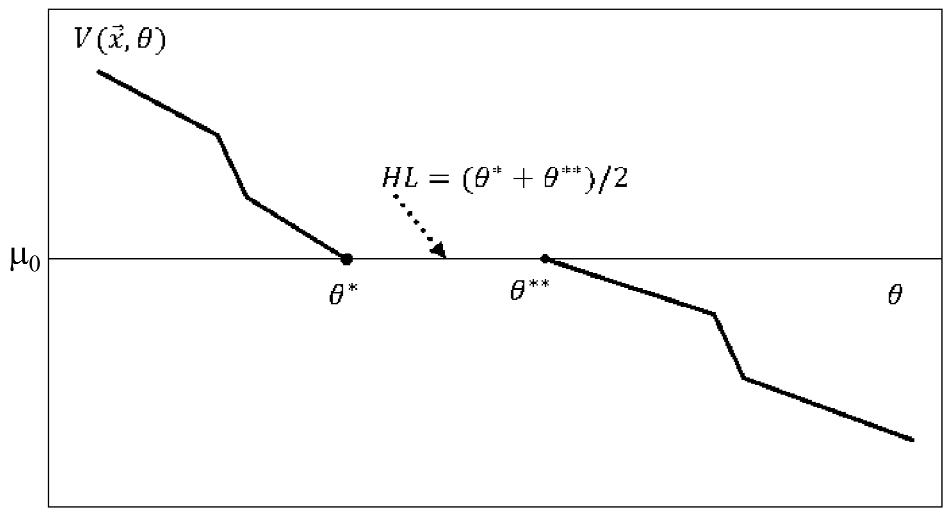

Рис. П.3.1. Определение оценки Ходжеса – Лемана $\hat{\theta} = HL$ параметра θ

Оценка $\hat{\theta} = HL$ параметра θ (см. рис. П.3.1) называется оценкой Ходжеса – Лемана и определяется в виде среднего арифметического значений θ^* и θ^{**}, то есть $\hat{\theta} = HL$ определяется в виде

$$HL = (\theta^* + \theta^{**})/2. \quad \textbf{(П.3.1)}$$

Отметим, что если $V(\vec{x};\theta)$ – непрерывная и строго монотонная функция, то оценка Ходжеса – Лемана находится из выражения $V(\vec{X};\hat{\theta}) = \mu_0$ и может рассматриваться как оценка моды, или оценка точки, соответствующей максимуму плотности распределения вероятности тестовой статистики $V(X_1,...,X_n)$ критерия проверки гипотез $H_0:\theta = 0, H_1:\theta \neq 0$. При этом точка максимальной вероятности распределения тестовой статистики при нулевой гипотезе называется модальной точкой.

Далее, определим значения $\hat{\theta}_L$ и $\hat{\theta}_U$ следующим образом:

$$\hat{\theta}_L = \inf\{\theta : V(\theta) < C_1\},\ \hat{\theta}_U = \sup\{\theta : V(\theta) > C_2\},$$

где числа C_1 и C_2 находятся из равенств

$$P\{V(\vec{X};\theta) \geq C_1 \mid H_0\} = P\{V(\vec{X};\theta) \leq C_2 \mid H_0\} = \alpha/2.$$

Тогда $(1-\alpha)$-доверительный интервал параметра θ, основанный на статистике $V(\vec{X})$, запишется в виде $[\hat{\theta}_L, \hat{\theta}_U]$, для которого выполняется выражение

$$P\{\hat{\theta}_L \leq \theta \leq \hat{\theta}_U\} = 1-\alpha. \quad \textbf{(П.3.2)}$$

Если C_1 и C_2 – целые числа, то $\hat{\theta}_L$ и $\hat{\theta}_U$ определяются как наименьшее и наибольшее решения уравнений $V(\hat{\theta}_L) = C_1 - 1$ и $V(\hat{\theta}_U) = C_2 + 1$.

Замечание П.3.3. Отметим, что если статистика $V(\vec{X})$ эквивариантна относительно сдвига, то есть для нее выполняется равенство $V(x_1+a,...,x_n+a)=$ $=V(x_1,...,x_n)+a$, то оценки $\hat{\theta}$, $\hat{\theta}_L$, $\hat{\theta}_U$ также эквивариантны, то есть при сдвиге исходных наблюдений к построенным оценкам по исходным данным надо лишь добавить постоянную величину сдвига. *Это свойство эквивариантности оценок относительно сдвига позволяет положить без потери общности* $\theta = 0$ *при изучении свойств распределений оценок.*

Применение процедуры Ходжеса – Лемана с использованием статистик критерия знаков, одновыборочного критерия знаковых рангов Уилкоксона, двухвыборочного рангового критерия Уилкоксона, приводится в [8].

Замечание П.3.4. Процедура Ходжеса – Лемана с использованием двухвыборочных критериев формулируется следующим образом. Имеем две выборки $\vec{X} = (X_1,...,X_m)$ из распределения $F(x)$ и $\vec{Y} = (Y_1,...,Y_n)$ из распределения $F(x-\Delta)$. Пусть $T(\vec{X},\vec{Y}) = T(X_1,...,X_m;Y_1,...,Y_n)$ – статистика критерия проверки гипотезы $H_0:\Delta = 0$ против альтернативы $H_1:\Delta > 0$, и эта статистика удовлетворяет следующим условиям:

(А) Гипотеза $H_0:\Delta = 0$ отвергается при больших значениях статистики $T(\vec{X},\vec{Y})$.

(Б) $T(x_1,...,x_m;y_1-h,...,y_n-h)$ – невозрастающая функция от h при заданных реализациях $\vec{x} = (x_1,...,x_m)$ и $\vec{y} = (y_1,...,y_n)$ выборок $\vec{X} = (X_1,...,X_m)$ и $\vec{Y} = (Y_1,...,Y_n)$.

(В) При нулевой гипотезе статистика $T(\vec{X},\vec{Y})$ свободна от распределения и ее распределение симметрично относительно ξ.

Определим значения Δ^* и Δ^{**} следующим образом:

$$\Delta^* = \sup\{\Delta : T(X_1,...,X_m;Y_1-\Delta,...,Y_n-\Delta) > \xi\},$$
$$\Delta^{**} = \inf\{\Delta : T(X_1,...,X_m;Y_1-\Delta,...,Y_n-\Delta) < \xi\}. \quad \textbf{(П.3.5)}$$

Оценка Ходжеса – Лемана $\hat{\Delta}$ параметра сдвига Δ в двухвыборочной задаче с альтернативой сдвига определяется в виде среднего арифметического значений Δ^* и Δ^{**}, то есть в виде

$$\hat{\Delta} = (\Delta^* + \Delta^{**})/2. \quad \textbf{(П.3.6)}$$

Отметим, что если $T(x_1,...,x_m;y_1-h,...,y_n-h)$ – непрерывная и строго монотонная функция от h, то оценка Ходжеса – Лемана находится из выражения $T(X_1,...,X_m;Y_1-\hat{\Delta},...,Y_n-\hat{\Delta}) = 0$. Отметим также, что оценка Ходжеса – Лемана $\hat{\Delta}$ параметра сдвига Δ в двухвыборочной задаче, которая записывается в виде $\hat{\Delta} = med\{Y_j - X_i,\ i=1,...,m,\ j=1,...,n\}$ и построена с использованием критерия Манна – Уитни, обсуждается в [8].

ЛИТЕРАТУРА

1. *Тарасенко Ф.П.* Непараметрическая статистика. – Томск: Изд-во Том. ун-та, 1976. – 292 с.
2. *Лонер Р.Л.*, *Уилкинсон Г.Н.* Устойчивые статистические методы оценки данных: пер. с англ. под ред. Н.Г. Волкова. – М.: Машиностроение, 1984. – 229 с.
3. *Box G.E.P.* Non-normality and tests on variances // Biometrika. – 1953. – V. 40. – P. 318–335.
4. *Шуленин В.П.* Введение в робастную статистику. – Томск: Изд-во Том. ун-та, 1993. – 227 с.
5. *Хьюбер П.* Робастность в статистке. – М.: Мир, 1984. – 304 с.
6. *Хампель Ф.*, *Рончетти Э.*, *Рауссей П.*, *Штаэль В.* Робастность в статистике. Подход на основе функций влияния. – М.: Мир, 1989. – 512 с.
7. *Шуленин В.П.* Математическая статистика. Ч.1. Параметрическая статистика: учебник. – Томск: Изд-во НТЛ, 2012. – 540 с.
8. *Шуленин В.П.* Математическая статистика. Ч.2. Непараметрическая статистика: учебник. – Томск: Изд-во НТЛ, 2012. – 388 с.
9. *Шуленин В.П.* Математическая статистика. Ч.3. Робастная статистика: учебник. – Томск: Изд-во НТЛ, 2012.– 520 с.
10. *Stigler S.M.* Simon Newcomb, Percy Daniel and history of robust estimations // J. Amer. Statist. Assoc. – 1973. – V. 68. – P. 872–879.
11. *Tukey J.W.* A survey of sampling from contaminated distributions // Contributions to Prob. Statist. / Ingram Olkin, ed. – Stanford Univ. Press, 1960. – P. 448–485.
12. *Tukey J.W.* Bias and confidence in not-quite large samples (Abstract) // Ann. Math. Statist. – 1958. – V. 29. – P. 614.
13. *Tukey J.W.* Data Analysis, Computation and Mathematics // Quarterly of Applied Mathematics, V. XXX, April 1972. – No. I, Special Issue: Symposium on the Future of Applied Mathematics. – P. 51–65. (См. также: Современные проблемы математики: сб.: пер. с англ. // Знание. Сер.: Математика, кибернетика. – 1977. – № 12. – С. 41–64).
14. *Tukey J.W.* Exploratory Data Analysis. – Reading, Mass.: Addison Wesley, 1977.
15. *Huber P.J.* Robust estimation of location parameter // Ann. Math. Statist. – 1964. – V. 35. – No. 1. – P. 73–101.

16. *Huber P.J.* Robust statistics: a review // Ann. Math. Statist. – 1972. – V. 43. – P. 1041–1067.
17. *Anscomb F.Z.* Topics in the investigation of linear relation fitted by the method of least squares // J. Roy. Stat. Soc., ser. B. – 1967. – V. 29. – P. 1–52.
18. *Крамер Г.* Математические методы статистики. – М.: Мир, 1974. – 648 с.
19. *Hampel F.R.* The influence curve and its role in robust estimation // J. Amer. Statist. Assoc. – 1974. – V. 69. – No. 346. – P. 383–393.
20. *Кендэлл М.*, *Стьюарт А.* Статистические выводы и связи. – М.: Наука, 1973. – 899 с.
21. *Hampel F.R.* Contribution to the Theory of Robust Estimation: Ph. D. diss. – Berkeley: Univ. California, 1968. – 103 p.
22. *Hampel F.R.* A general qualitative definition of robustness // Ann. Math. Statist. – 1971. – V. 42. – P. 1887–1896.
23. *Bickel P.J.* Another look at robustness: a review of reviews and some new development // Scand. J. Statist. Theory and Appl. – 1976. – V. 3. – P. 145–168.
24. *Andrews D.F.*, *Bickel P.J.*, *Hampel F.R.*, *et al.* Robust Estimation of Location: Survey and Advances. – Princeton, N.Y.: Princeton Univ. Press, 1972. – 375 p.
25. *Jaeckel L.A.* Robust estimation of location: symmetry and asymmetric contamination // Ann. Math. Statist. – 1971. – V. 42. – No. 3. – P. 1020–1034.
26. *Jaeckel L.A.* Some flexible estimates of location // Ann. Math. Statist. – 1971. – V. 42. – P. 1540–1552.
27. *Hodges J.L.* Efficiency in normal samples and tolerance of extreme values for some estimates of location // Proc. 5th Berkeley Symp. Math. Statist. and Prob. 1965 – 1966. – V. I. – Berkeley – Los Angeles, 1967. – P. 163–186.
28. *Noether G.E.* On a theorem of Pitman // Ann. Math. Statist. – 1955. – V. 26. – P. 64–68.
29. *Serfling R.J.* Approximation Theorems of Mathematical Statistics. – N.Y.: Wiley, 1980.–371 p.
30. *Reeds J.A.* On the definition of von Mises functionals: Ph. D. Thesis. Dept. of Statistics. – Harvard, 1976. – 178 p.
31. *Von Mises R.* On the asymptotic distributions of differentiable statistical functions // Ann. Math. Statist. – 1947. – V. 18. – P. 309–348.
32. *Ferenholz L.T.* Von Mises calculus for statistical functionals // Lecture Notes in Statistics. – 1983. – V. 19. – No. 8. – P. 1–124.
33. *Randles R.H. and Wolfe D.A.* Introduction to the Theory of Nonparametric Statistics. – N.Y.: Wiley, 1979.
34. *Gibbons J.D.* Nonparametric Statistical Inference. – N.Y.: McGraw-Hill, 1971.
35. *Королюк В.С.*, *Бобровских Ю.В.* Теория *U*-статистик. – Киев: Наукова думка, 1989.– 383 с.
36. *Ghosh J.K.* A new proof of the Bahadur representation of quantiles and application // Ann. Math. Statist. – 1971. – V. 42. – No. 6. – P. 1957–1961.

37. *Boos D.D. and Serfling R.J.* A note on differentials and CLT and LIL for statistical functions, with application to *M*-estimates // Ann. Statist. – 1980. – V. 8. – P. 618–624.
38. *Boos D.D.* A differential for *L*-statistics // Ann. Statist. – 1979. – V. 7. – P. 955–959.
39. *Serfling R.J.* Generalized L-M-R-statistics // Ann. Statist. – 1984. – V. 12. – P. 76–86.
40. *Бикел П.*, *Доксум К.* Математическая статистика. – М.: Финансы и статистика, 1983. – Вып. 1. – 280 с.; Вып. 2. – 197 с.
41. *Chernoff H.*, *Gastwirth J.L.*, *and Johns M.V.* Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation // Ann. Math. Statist. – 1976. – V. 38. – P. 52–72.
42. *Jureckova J. M-L-R*-estimators. Handbook of Statistics. V. 4 / eds. P.R. Krishnaiah, P.K. Sen. – Elsevier Science Publishers, 1984. – P. 463–485.
43. *Stigler S.M.* Linear function of order statistics with smooth weight functions // Ann. Statist. – 1974. – V. 2. – P. 676–693.
44. *Hodges J.L. and Lehmann E.L.* Estimation of location based on rank tests // Ann. Math. Statist. – 1963. – V. 34. – P. 598–611.
45. *Хеттсманспергер Т.* Статистические выводы, основанные на рангах. – М.: Финансы и статистика, 1987. – 334 с.
46. *Гаек Я.*, *Шидак З.* Теория ранговых критериев. – М.: Наука, 1971. – 376 с.
47. *Azencott R.* Robustass des R-estimateurs // Asterisque. – 1977. – V. 43–44. – P. 189–202.
48. *Janssen P.*, *Serfling R.*, *and Veraverbeke M.* Asymptotic normality for a general class of statistical functions and applications to measures of spread // Ann. Statist. – 1984. – V. 12. – No. 4. – P. 1369–1379.
49. *Janssen P.*, *Serfling R.*, *and Veraverbeke M.* Asymptotic normality of U-statistics based on trimmed samples // J. Statist. Planning and Inference. – 1987. – V. 16. – P. 63–74.
50. *Wolfowitz J.* The minimum distance method // Ann. Math. Statist. – 1957. – V. 28. – P. 75–88.
51. *Parr W.C.* Minimum distance estimation: a bibliography // Comm. Statist. – 1981. – V. A10. – P. 1205–1224.
52. *Parr W.C.* A note on Hajek protections an the influence curve // Statistics and Probability Letters. – 1983. – V. 1. – P. 177–179.
53. *Parr W.C. and De Wet.* On minimum weighted Cramer-von Mists statistical estimation // Comm. Statist. – 1981. – V. A10(12). – P. 1149–1166.
54. *Parr W.S. and Schucany W.R.* Minimum distance and robust estimation // J. Amer. Statist. Assoc. – 1980. – V. 75. – No. 371. – P. 616–624.
55. *Boos D.D.* Minimum distance estimators for location and goodness of fit // J. Amer. Statist. Assoc. – 1981. – V. 76. – No. 375. – P. 663–670.

56. *Shulenin V.P. and Tarasenko F.P.* Connection of *MD*-estimates with classes of robust estimates of location parameter // 12-th PRAGUE Conf. Inf. Theory, Stat. Decision Functions, Random Processes. – Prague, 1994. – P. 220–223.
57. *Шуленин В.П.* Асимптотические свойства и робастность *MD*-оценок // Теория вероятностей и её применение. – 1992. – Т. 37. – Вып. 4. – С. 816–818.
58. *Shulenin V.P.* On estimation of parameters by the minimum distance method. Reliability: Theory and Applications. Gnedenko-Forum, San Diego, 2013, Vol. 8, No 2, p. 24–38.
59. *Шуленин В.П.* Границы эффективности оценок, построенных методом минимума расстояния Крамера – Мизеса // Изв. вузов. Физика. – 1995. – Т. 38. – № 9. – С. 84–89.
60. *Hogg R.V.* Adaptive robust procedures: A partial review and some suggestions for future applications and theory // J. Amer. Statist. Assoc. – 1974. – V. 69. – P. 909–923.
61. *Тарасенко Ф.П.*, *Шуленин В.П.* Об одном классе оценок параметра положения // VI Международный симпозиум по теории информации: тез. докл. Ч. I. – Москва – Ташкент, 1984. – С. 171–173.
62. *Shulenin V.P.* Asymptotic properties of the trimmed *GL*- and *U*-statistics // 6th Prague Symposium on Asymptotic Statistics. Prague, 1998, August 23 – 28. PRAGUE STOCHATICS'98: Abstracts. – P. 84.
63. *Шуленин В.П.* Исследование устойчивости и асимптотических свойств урезанной средней разности Джини // Тр. IV Международной конференции по теории вероятности и математической статистике. – Вильнюс, 1985. – С. 330–332.
64. *Шуленин В.П.* Асимптотические свойства *GL*- и *U*-статистик // Вестник Томского государственного университета. Приложение. – 2004. – № 9 (11). – С. 184–190.
65. *Шуленин В.П.* Исследование свойств оценки медианы абсолютных разностей // V совещание-семинар по непараметрическим и робастным методам статистики в кибернетике. – Томск: Изд-во Том. ун-та, 1987. – Ч. II. – С. 460–467.
66. *Шуленин В.П.* Асимптотические свойства обобщенных оценок Ходжеса – Лемана // IX Всесоюзная конференция по теории кодирования и передачи информации. – Одесса, 1988. – Ч. II. – С. 111–113.
67. *Шуленин В.П.* Асимптотические свойства и робастность обобщенных *L*-оценок // Тр. V Международной конференции по теории вероятности и математической статистике. – Вильнюс, 1989. – Т. 4. – С. 377–378.
68. *Шуленин В.П.* Асимптотические свойства обобщенных *L*-оценок, вычисляемых по урезанным выборкам // Непараметрические и робастные статистические методы в кибернетике и информатике. – Томск: Изд-во Том. ун-та, 1990. – С. 564–570.

69. *Bickel P.J. and Lehmann E.L.* Descriptive statistics for nonparametric models. III. Dispersion // Ann. Statist. – 1976. – V. 4. – No. 6. – P. 1139–1158.
70. *Gastwirth J.L.* On Robust Rank Tests. Nonparametric Techniques in Stat. Inference. – Cambrige: Univ. Press., 1970. – P. 89–109.
71. *Hettmansperger T.R. and Keenan M.A.* Tail weight statistical inference and families of distribution a brief survey // Mod. Course Statist. Distrib. Sci. Work. – Dordrecht – Boston, 1975. – P. 161–172.
72. *Шуленин В.П.* Свойства адаптивных оценок Ходжеса – Лемана в асимптотике и при конечных объемах выборки. Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2010. – № 2(11). – С. 96–112.
73. *Леман Э.* Теория точечного оценивания. – М.: Наука, 1991. – 448 с.
74. *Bickel P.J. and Lehmann E.L.* Measures of location and scale // Proc. Prague Symp. Asymptotic Statist. 1973. V. 1. – Prague: Prague Charles Univ., 1974. – P. 25–36.
75. *Shulenin V.P. and Deeva T.A.* The numerical characteristics of robustness of the class of the Hodges-Lehmann estimators // Proc. of the Third Russian – Korean International Symposium on Science and Technology. KORUS’99. – V. 2. – P. 510 – 514.
76. *Tarasenko F.P., Shulenin V.P.* Rank as Proxy for the Observation in Statistical Procedures. Applied Methods of Statistical Analysis. Nonparametric Approach – AMSA’2015, Novosibirsk, Russia, 14–19 September, 2015: Proceedings of the International Workshop. – Novosibirsk: NSTU publisher 2015, p. 12–17.
77. *Шуленин В.П., Серых А.П.* Робастные и непараметрические алгоритмы обработки данных физических экспериментов // Изв. вузов. Физика. – 1993. – Т. 36. – № 10. – С. 128–136.
78. *Hajek J.* Course in Nonparametric Statistics. – Academic Press, 1970. – 184 p.
79. *Van Zwet W.R.* Convex transformations of Random Variables // Math. Centrum. Amsterdam. – 1964.
80. *Loh Wei-Yin* . Bounds on ARE’s for restricted classes of distribution defined via tail-orderings // Ann. Statist. – 1984. – V.12. – No. 2. – P. 685 – 701.
81. *Hettmansperger T.P., I. Hueter and J.Husler.* Minimum distance estimators // J. Statist. Plann. Inference. – 1993. – 76. – P. 663 – 670.
82. *Шуленин В.П.* Асимптотические свойства робастных оценок масштабного параметра // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2016. – № 2(35). – С. 62–70.
83. *Шуленин В.П.* Робастные альтернативы стандартному отклонению при обработке данных физических экспериментов // Изв. вузов. Физика. – 2016. – Т. 59 – № 6. – С. 62–69.
84. *Beran R.* Minimum distance procedure // Krishnaiah P.R. and Sen P.K., eds. Handbook of Statistics. – Amsterdam: Elsevier Science, 1984. – V. 4. – P. 741–754.

СПИСОК ОСНОВНЫХ ОБОЗНАЧЕНИЙ

- $P\{A\}$ – вероятность события A
- $P\{A \mid B\}$ – условная вероятность события $V_n^{(i)}$ при условии B
- $F_X(x)$ – функция распределения (ф.р.) случайной величины (с.в.) X
- $F^{-1}(t) = \inf\{x : F_X(x) \geq t\}$, $0 \leq t \leq 1$ – квантильная функция случайной величины (с.в.) X
- $x_p = F^{-1}(p)$ – квантиль уровня p, $0 < p < 1$
- $f_X(x)$ – плотность распределения вероятностей (п.р.в.) случайной величины X
- $M(X)$ – математическое ожидание случайной величины X
- $D(X)$ – дисперсия случайной величины X
- $M(X \mid Y = y)$ – математическое ожидание случайной величины X при условии, что случайная величина Y приняла значение y
- $\vec{X} = (X_1, ..., X_n)$ – случайный вектор, компоненты которого являются независимыми и одинаково распределенными (н.о.р.) случайными величинами ($X_1, ..., X_n$ – выборка объема n)
- $F_{\vec{X}}(x_1, ..., x_n)$ – функция распределения случайного вектора $\vec{X} = (X_1, ..., X_n)$
- $f_{\vec{X}}(x_1, ..., x_n)$ – плотность распределения вероятностей случайного вектора $\vec{X} = (X_1, ..., X_n)$
- $\vec{x} = (x_1, ..., x_n)$ – наблюденная реализация наблюдаемого случайного вектора $\vec{X} = (X_1, ..., X_n)$
- $X_{(1)}, ..., X_{(n)}$ – порядковые статистики выборки $X_1, ..., X_n$
- $x_{(1)}, ..., x_{(n)}$ – вариационный ряд наблюденной реализации $\vec{x} = (x_1, ..., x_n)$ случайного вектора $\vec{X} = (X_1, ..., X_n)$
- $\vec{R} = (R_1, ..., R_n)$ – ранговый вектор выборки $X_1, ..., X_n$
- $\vec{r} = (r_1, ..., r_n)$ – наблюденная реализация рангового вектора $\vec{R} = (R_1, ..., R_n)$
- $T(F)$, $F \in \mathfrak{I}$ – функционал, заданный на множестве $\mathfrak{I}$ допустимых в эксперименте распределений

- $T_n = T_n(X_1,...,X_n) = T(F_n)$ – выборочная оценка функционала $T(F)$
- $F_n(x)$ – эмпирическая функция распределения (э.ф.р.), построенная по выборке $X_1,...,X_n$
- $(\mathrm{X}, \Im)$ – статистическая модель эксперимента, где X-выборочное пространство, а $\Im$-множество допустимых функций распределения вероятностей в условиях данного опыта, результатом которого является *наблюденная* реализация $\vec{x} = (x_1,...,x_n)$ *наблюдаемого* случайного вектора $\vec{X} = (X_1,...,X_n)$
- $\Im_\theta = \{F_X : F_X(x\,;\theta)\,, \theta \in \Theta\}$ – параметрическое семейство функций с конечным числом неизвестных параметров θ, которые принадлежат заданному параметрическому множеству Θ
- $d_L(F,G) = \inf\{\,\varepsilon : G(x-\varepsilon) - \varepsilon \le F(x) \le G(x+\varepsilon) + \varepsilon,\ \forall x \in R^1\},\ \ 0 \le \varepsilon \le 1$ – метрики Леви
- $\Im_{L,\varepsilon}(G) = \{F \in \Im : d_L(F,G) \le \varepsilon\}, 0 \le \varepsilon \le 1$ – супермодель Леви
- $d_K(F,G) = \sup_x |\,F(x) - G(x)\,|$ – метрика Колмогорова
- $\Im_{K,\varepsilon}(G) = \{F \in \Im : d_K(F,G) \le \varepsilon\}, 0 \le \varepsilon \le 1$ – супермодель Колмогорова
- $\Im_{\varepsilon,\tau}(\Phi) = \{F : F(x) = \Phi_{\varepsilon,\tau}(x)\}$ – супермодель с засорением (Тьюки), для которой функция распределения $\Phi_{\varepsilon,\tau}(x)$ определяется в виде $\Phi_{\varepsilon,\tau}(x) = (1-\varepsilon)\Phi(x) + \varepsilon\Phi(x/\tau)$, $0 \le \varepsilon \le 1$, $\tau \ge 1$
- $\Im_\varepsilon(F_0) = \{F : F(x) = (1-\varepsilon)F_0(x) + \varepsilon H(x)\}$, $0 \le \varepsilon \le 1$ – супермодель с засорением (Хьюбера), где F_0 – заданная функция распределения (идеальная модель) и $H(x)$ – произвольная непрерывная функция распределения вероятности
- $\Im_S^* = \{F_{(1)}, F_{(2)}, F_{(3)}, F_{(4)}, F_{(5)}\}$ – супермодель в виде конечного набора стандартных симметричных распределений: $F_{(1)}$ – ф.р. нормальная, $F_{(2)}$ – ф.р. логистическая, $F_{(3)}$ – ф.р. Лапласа, $F_{(4)}$ – ф.р. Коши, $F_{(5)}$ – ф.р. гиперболический секанс
- $\Im_{S|\theta} = \{F : F(x) = 1 - F(2\theta - x), \forall x \in R^1\}$ – семейство распределений симметричных относительно точки θ
- $\hat{x}_p = F_n^{-1}(p)$ – выборочный квантиль уровня p, $0 < p < 1$
- $\bar{X}_{1/2} = med\{X_1,...,X_n\}$ – выборочная медиана
- $\bar{X}_\alpha$ – выборочная оценка α-урезанных средних, $0 \le \alpha < 1/2$
- $\tilde{X}_\alpha$ – выборочная оценка α-винзоризованных средних, $0 \le \alpha < 1/2$
- $T_n = HL$ – выборочная оценка Ходжеса – Лемана

- HL_α – выборочная оценка α-урезанных оценок Ходжеса – Лемана, $0 \le \alpha < 1/2$
- $\hat{\Delta}_\alpha$ – выборочная оценка средней разности Джини, $0 \le \alpha < 1/2$
- $d_k T(F; G-F)$ – дифференциал Гато *k*-го порядка функционала $T(F)$ в «точке» F по направлению функции распределения G
- $SC_n(x)$ – кривая чувствительности (sensitivity curve) Тьюки
- $IF(x; F, T)$ – функция влияния (Influence Function) Хампеля оценки T_n функционала $T(F)$
- $IF(x; F, T) = d_1 T(F; \Delta_x - F)$, Δ_x – вырожденная в токе x функция распределения
- $\gamma^*(F,T) = \sup\limits_x |IF(x; F, T)|$ – чувствительность оценки T_n функционала $T(F)$ к грубым ошибкам
- $\lambda^*(F,T) = \sup\limits_{x \ne y} |IF(x; F, T) - IF(x; F, T)| / |x-y|$ – чувствительность оценки T_n к локальным изменениям наблюдений
- $\rho^*(F,T) = \inf\{ r > 0 : IF(x; F, T) = 0,\ |x| > r \}$ – точка отбраковки для оценки T_n
- $\varepsilon^*(T,F) = \sup\{\varepsilon : \sup\limits_{F: d_L(F, F_0) < \varepsilon} |T(F) - T(F_0)| < \infty\}$ – предел устойчивости (пороговая точка, «точка срыва») оценки T_n функционала $T(F)$, $d_L(F, F_0)$ – метрика Леви, F_0 – идеальная модель
- $CVF(x; F, T) = \lim\limits_{\lambda \to 0+} \dfrac{V\{T, [(1-\lambda)F + \lambda\Delta_x]\} - V\{T, F\}}{\lambda}$, $0 \le \lambda \le 1$ – функция изменения дисперсии $V(T,F)$ оценки $T_n = T(F_n)$ функционала $T(F)$, Δ_x – вырожденная в токе x функция распределения
- $\omega^*(T,F) = \sup\limits_x CVF(x; F, T) / V(T, F)$ – чувствительность к изменению дисперсии $V(T,F)$ оценки $T_n = T(F_n)$
- $T(F_n) \to^p T(F)$ при $n \to \infty$ – оценка $T(F_n)$ сходится по вероятности к функционалу $T(F)$ при $n \to \infty$
- $L\{\sqrt{n}[T_n - T(F)] / \sigma_F(T_n)\} = N(0, 1)$ при $n \to \infty$ – закон распределения с.в. T_n является асимптотически стандартным нормальным, где $\sigma_F^2(T_n)$ – асимптотическая дисперсия $\sqrt{n}\, T_n$-оценки функционала $T(F)$
- $L(X) = N(a, \sigma^2)$ – закон распределения вероятностей с. в. X нормальный с параметрами a и σ^2

- $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{x} e^{-\frac{1}{2}x^2}dx$ – стандартная нормальная функция распределения вероятностей
- $\Phi\{(x-a)/\sigma\}$ – функция распределения нормальной с. в. X с параметрами a и σ^2
- $\Phi^{-1}(t)$ – обратная функция для стандартной нормальной функции распределения $\Phi(x)$
- $\lambda_\alpha=\Phi^{-1}(\alpha)$ – квантиль заданного уровня α стандартного нормального распределения
- $\phi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$ – стандартная нормальная плотность распределения вероятностей
- $А Э_F(\hat{\theta})=АОЭ_F(\hat{\theta}:\hat{\theta}_{эф})=\frac{\sigma^2(\hat{\theta}_{эф},F)}{\sigma^2(\hat{\theta},F)}=\{\sigma^2(\hat{\theta},F)i_1(\theta)\}^{-1}$ – асимптотическая абсолютная эффективность оценки $\hat{\theta}$ параметра θ, где $\hat{\theta}_{эф}$ – эффективная оценка (или МГД-оценка, оценка с минимальной граничной дисперсией) параметра θ, дисперсия которой равна обратной величине количества информации Фишера $i_1(\theta)$ в параметрическом семействе распределений $\Im_\theta=\{F_X:F_X(x\,;\theta)\,,\theta\in\Theta\}$
- $I(F)$– информация Фишера относительно параметра сдвига θ ф.р. $F(x-\theta)$
- $DE_F(\hat{\theta}_i)=1-\min\{\sigma_F^2(\hat{\theta}_1),...,\sigma_F^2(\hat{\theta}_k)\}/\sigma_F^2(\hat{\theta}_i)$ – дефект оценки $\hat{\theta}_i$, $i=1,...,k$, среди сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$ параметра θ при заданном распределении F наблюдений $X_1,...,X_n$
- $ADE_F(\hat{\theta}_i)=1-АЭ_F(\hat{\theta}_i)$ – абсолютный дефект оценки $\hat{\theta}_i$, $i=1,...,k$ среди сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$
- $ARE_F(V_n^{(1)}:V_n^{(2)})=Э_F^2(V_n^{(1)})/Э_F^2(V_n^{(2)})$ – асимптотическая относительная эффективность Питмена (см. П.2) для $V_n^{(1)}$-критерия по отношению к $V_n^{(2)}$-критерию, где эффективность Питмена
$$Э_F(V_n^{(i)})=\lim_{n\to\infty}\frac{d}{d\theta}M\{V_n^{(i)}\mid H_1\}|_{\theta=0}/\sqrt{n\,D(V_n^{(i)}\mid H_0)}\,,\ i=1,2\ .$$

ОГЛАВЛЕНИЕ

Научное издание

Шуленин Валерий Петрович

РОБАСТНЫЕ МЕТОДЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Редактор *Т.С. Портнова*

Дизайн, верстка *Д.В. Фортеса*

ООО «Издательство научно-технической литературы»
634050, г. Томск, пл. Новособорная, 1, тел. (3822) 533-335

Изд. лиц. ИД № 04000 от 12.02.2001. Подписано к печати 20.07.2016.
Формат 60 × 84 $^{1}/_{16}$. Бумага офсетная. Печать офсетная. Гарнитура «Таймс».
Усл. п. л. 15,11. Уч.-изд. л. 16,93. Тираж 300 экз. Заказ № 23.

Отпечатано в типографии «М-Принт», г. Томск,
пер. Добролюбова, 10, ст. 3, тел. (3822) 258-279